HRTF individualization using deep learning

Miccini, Riccardo; Spagnol, Simone

Published in:
Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Workshops (VRW 2020)

DOI (link to publication from Publisher):
10.1109/VRW50115.2020.00084

Creative Commons License
Other

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

HRTF Individualization using Deep Learning

Riccardo Miccini* Simone Spagnol†

Aalborg University
Copenhagen, Denmark

Abstract

The research presented in this paper focuses on Head-Related Transfer Function (HRTF) individualization using deep learning techniques. HRTF individualization is paramount for accurate binaural rendering, which is used in XR technologies, tools for the visually impaired, and many other applications. The rising availability of public HRTF data currently allows experimentation with different input data formats and various computational models. Accordingly, three research directions are investigated here: (1) extraction of predictors from user data; (2) unsupervised learning of HRTFs based on autoencoder networks; and (3) synthesis of HRTFs from anthropometric data using deep multilayer perceptrons and principal component analysis. While none of the aforementioned investigations has shown outstanding results to date, the knowledge acquired throughout the development and troubleshooting phases highlights areas of improvement which are expected to pave the way to more accurate models for HRTF individualization.

Index Terms: Hardware—Communication hardware, interfaces and storage—Signal processing systems—Digital signal processing; Applied computing—Arts and humanities—Sound and music computing—

1 Introduction

Virtual/augmented/mixed reality (XR) research has made substantial progress over the last decades, and XR environments created using binaural sound rendering technologies find applications in a wide array of areas, ranging from travel aids for the visually impaired to entertainment systems [3, 25].

Binaural sound rendering techniques are based on the application of a particular filter called *Head-Related Transfer Function* (HRTF), which colors a sound according to its location in the virtual environment. However, HRTFs derived from generic anthropometries such as dummy heads often result in localization errors and limited spatial perception [18]. In fact, while generic HRTFs may acceptably approximate the interaural cues used to perceive the horizontal direction of a sound source, the monaural cues needed to discern its vertical direction are highly dependent on the anthropometric characteristics of the individual ear [1].

In order to provide the most realistic and immersive experience possible, it is necessary for users to have their custom set of HRTFs measured, which can prove quite impractical due to the need for dedicated facilities and the overall invasiveness of the procedure. Over the past decades, several strategies have been devised in order to avoid the burden of conducting strenuous acoustical measurements with human subjects. In a recent review, Guezenoc and Seguier [8] divide such alternative approaches into *numerical simulation*, *anthropometrics*-based, and *perceptual feedback*-based customization.

The first method consists in simulating the propagation of acoustic waves around the subject, using 3D scans; the most common simulation schemes include the Fast Multipole Accelerated Boundary Element Method (FM-BEM) [9] and the Finite Difference Time Domain (FDTD) method [26]. With the help of databases of publicly available HRTFs and machine learning techniques, anthropometric measurements can be used to choose, adapt, or estimate a subject's HRTF set. For instance, in 2010 Zeng et al. [30] implemented a hybrid model based on Principal Component Analysis (PCA) and multiple linear regression, which used anthropometric parameters to select the most suitable HRTF set for a given user. Similarly, user feedback on perceptual tests can be used to inform regression models for tasks such as those listed above.

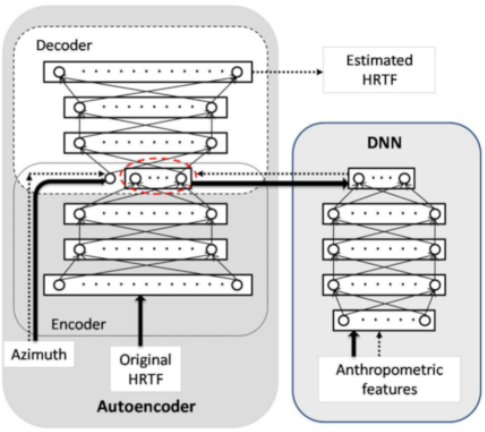

Figure 1: Architecture proposed by Chen et al., with HRTF autoencoding and latent representation estimation from anthropometry with a DNN. Solid arrows indicate the training process, dotted arrows indicate usage, and the red dashed oval highlights the latent space. Figure reproduced from [5].

In more recent times, there has been an interest in solving these tasks using deep learning techniques [7]. In 2017, Yao et al. [29] used anthropometric measurements to select the most suitable HRTF sets from a larger database. In their work, a dataset of user anthropometry and fitness scores for each available HRTF is compiled — by means of conducting perceptual tests with users — and neural networks for each HRTF are trained to predict their suitability. Again in 2017, Yamamoto and Igarashi [28] trained a variational autoencoder (VAE) on HRTF data, and devised a perceptual calibration procedure to fine-tune the latent variable used as input by the generative part of the model. In 2018, Lee and Kim [15] developed a double-branched neural network that processes anthropometric data with a multilayer perceptron (MLP) and edge-detected pictures of the ear with convolutional layers, combining the outputs of the two into a third network to estimate HRTF sets. Finally, in 2019, Chen et al. [5] trained an autoencoder to reconstruct HRTFs along the horizontal plane, and subsequently used the resulting latent representations as targets for a MLP which feeds on anthropometric data and azimuth angle, allowing users to synthesize new HRTFs using the MLP and decoder — this is shown in Fig. 1.

*e-mail: rmicci18@student.aau.dk
†e-mail: ssp@create.aau.dk

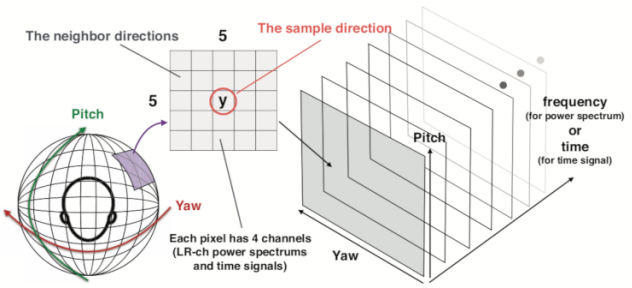

Figure 2: Data structure, called *HRTF patch*, used as input to the autoencoder by Yamamoto and Igarashi. Figure reproduced from [28].

Since there is not a clear consensus on what is the most effective strategy for deep learning based HRTF individualization, this paper investigates further methods — in particular using newly developed deep learning algorithms and alternative data representations — and expands on the topic by documenting the experiments conducted as part of the research. Section 2 details the computational techniques used in selected works from the literature as well as in the original research carried out herein, with a particular focus on deep learning methods. In Section 3, the applications and outcomes of the aforementioned techniques are discussed, with the purpose of assessing their effectiveness. Finally, closing remarks as well as pointers for future research are stated in Section 4.

2 Methods

This section presents some of the most relevant computational methods found in the relevant literature on HRTF individualization. The aspects covered in the following subsections include the encoding of generated HRTFs, the extraction and choice of viable predictors, and the deep neural network (DNN) architectures adopted.

2.1 HRTF representation

A single HRTF is defined as the the far-field frequency response of a given ear, measured from a point in the free field to a point in the ear canal [6]. An HRTF set is composed of the HRTFs of both left and right ears, measured at a fixed distance from the center of the head, and across several elevations and azimuths. According to Kulkarni et al. [14], HRTFs specified as minimum-phase FIR filters have been empirically proved to be perceptually acceptable. Thus, HRTFs can be stripped of the ITD information and stored as real-valued log-magnitude responses.

While this is the preferred way of storing, exchanging, and using HRTF sets, neural networks have different requirements that call for ad-hoc formats. Notably, Yamamoto and Igarashi [28] use different representations for the input and output of their autoencoder. The input data format, which is dubbed *HRTF patch*, consists of a 4-dimensional tensor of shape $(5 \times 5 \times 128 \times 4)$. The first two dimensions describe the HRTF under investigation and its neighbors along the elevation and azimuth directions, for a total of 25 HRTFs in each given patch. The remaining ones describe the content of each HRTF in the patch: the last dimension, also called *channel*, encodes frequency power spectrum or time-domain signal for either left or right ear, where 128 is their length. This data representation provides a substantial amount of contextual information, which can be learnt by 3D convolutional layers. The structure of the HRTF patch can be seen in Fig. 2. The output of the autoencoder does not contain any neighbor HRTF, but instead of encoding the frequency power spectrum or time-domain information as a continuous signal, it uses a quantized format where each sample can have one of 256 possible discrete values that are then mapped to another dimension using one-hot encoding. The continuous signals can be reconstructed by taking the index of the value with highest magnitude and passing it to a μ-law algorithm. This strategy — which can be found in certain WaveNet implementations [27] — makes sure to retain some of the high-frequency details of the continuous signals, which are often lost when reconstructing data with autoencoders.

In this paper we consider mappings where HRTFs sharing the same azimuth or elevation are combined in a 2-dimensional image-like representation with either elevation or azimuth along one axis and frequency along the other; the color of each pixel represents the log-magnitude of the spectrum. The structure expressed by adjacent HRTFs could therefore be learnt using 2D convolutional layers. The downside of combining data in this way is the reduction of available data points to use for training.

Another possible format is a compact representation of individual HRTFs consisting of their first N principal components. While it has been observed that as little as 5 components are enough to explain approximately 90% of the variance in the original HRTF magnitude functions [13], the loadings of the PCA must be learned, thus becoming an essential part of the representation.

2.2 User data extraction

A fundamental aspect of HRTF individualization is the kind of data used to personalize the frequency response. Most often, acquiring data about a subject is faster and less strenuous than collecting an entire HRTF set, as well as having looser requirements in terms of external conditions and tools. The kind of data that can be collected comprises anthropometric measurements, 3D models, and perceptual feedback.

The CIPIC dataset [2] released in 2001 sets a convention for anthropometric data collection and reporting, which has been adopted by later datasets too [4]. Its format specifies 17 anthropometric parameters for the head and torso, and 10 for each pinna. It has the disadvantage of having loosely defined measurement points, which translate into systematic biases which makes merging different datasets particularly prone to errors. Moreover, anthropometric features are only unique to each given subject and as such, may not have enough predictive power to be used for the regression of several HRTFs per subject. This shortcoming can be partially addressed by introducing elevation-dependent anthropometric measurements as predictors [11, 23], based on the length of segments spanning from the ear canal entrance to each of the three contours outlined by the helix and concha, and oriented according to a given elevation angle.

Another source of useful predictors for regression tasks can be found in 3-dimensional representations of the subject. Recent HRTF datasets include digital scans of subjects' heads and/or pinnae [4, 24] stored as 3D models, which can be used for feature extraction. In particular, 2-dimensional projections such as digital renderings or depth maps can be conveniently processed in neural networks using convolutional and pooling layers. As will be mentioned in the following section, convolutional autoencoders can extract salient features from images of the pinna, which can then be used as predictors.

Finally, perceptual feedback consists in letting a user evaluate and rate the performance of a given HRTF set, and is most commonly adopted in HRTF selection or adaptation tasks. Nevertheless, Yamamoto and Igarashi use perceptual feedback to navigate the latent space in order to synthesize suitable HRTFs for a given user [28]. It is worth noting how, while validating models based on anthropometric data is quite trivial, models that use perceptual data require a user study or the implementation of a virtual agent.

2.3 Autoencoder

Most conventional neural networks are used to predict a target y from an input x, a task known as *supervised learning*. On the other hand, autoencoders learn a compressed representation z of the input data x called *latent representation*, which is then used to generate a reconstructed version $\hat{x}$. Thus, the purpose of autoencoders is

to extract useful features from the input data in an unsupervised manner [7]. One such example can be seen back in Fig. 1, where an autoencoder is used to derive a compact HRTF representation which can then be used as the target of a prediction task [5].

An autoencoder usually consists of a feed-forward neural network, in turn composed of two subnets: an encoder network $f()$ and a decoder network $g()$ such that $g(f(x)) = g(z) = \hat{x}$. Training an autoencoder usually involves iteratively updating the weights and biases of the two networks through backpropagation, in order to minimize a cost function representing the mean squared error (MSE) between x and $\hat{x}$.

Over time, several variants of autoencoder have been developed. Each variant extends the conventional autoencoder architecture by promoting different properties of the latent space, thereby catering to different tasks such as denoising, classification, or — as is the case here — generative applications. Two common autoencoder-based generative models are described below.

2.3.1 Variational autoencoder

A VAE is a probabilistic model where the encoder maps the probability distribution of a certain latent representation given a data point, and the decoder outputs the probability distribution of the data, given a point in the latent space. It is often desirable to model the latent space prior distribution as an isotropic multivariate Gaussian; in order to enforce this, the Kullback-Leibler divergence between the aforementioned prior and the encoder is introduced.

This probabilistic framework proves useful when synthesizing HRTFs, because it can learn causal factors of variations in the data [12]. However, there exists no way of generating a data point with specific characteristics, such as the HRTF at a given azimuth and elevation angles. While points in the latent space are likely to generate plausible new data, one can only sample randomly. The class of autoencoders described below aims at addressing this shortcoming.

2.3.2 Conditional variational autoencoder

Conditional variational autoencoders (CVAEs) are an extension of VAEs, where an input data label c modulates the prior distribution of the latent variables that generate the output [21]. Thus, the encoding process is conditioned by c instead of the data content alone. Furthermore, the decoder too is conditioned by the label. The influence of c is incorporated into the VAE structure by means of concatenating its value to the input data x before feeding it into the encoder, as well as to the latent variables z before feeding them into the decoder. Yamamoto and Igarashi [28] use a customized deep CVAE where labels consisting of a subject ID and a spatial orientation, both provided as one-hot encoded vectors, are used to condition each layer of the encoder and decoder.

3 Experiments and Results

The research conducted as part of this work can be grouped into three main threads: (1) extraction of user data to be used as predictor; (2) unsupervised learning of HRTF data; and (3) synthesis of HRTFs from anthropometric data using deep multilayer perceptrons and principal component analysis. The following subsections elaborate on the aforementioned topics.

The recent release of the HUTUBS HRTF database [4], comprising 93 different human subjects, prompted its adoption for the experiments described herein. This dataset is remarkably extensive compared to previously released ones since it features both acoustically measured and numerically simulated HRTFs, as well as anthropometric measurements and 3D models of the head — these latter available for 55 subjects only.

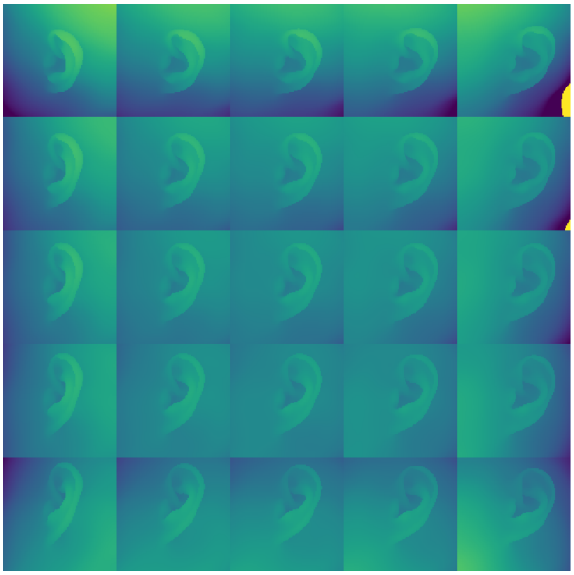

Figure 3: Pinna depth maps for a given subject, over a grid of different azimuth and elevation tilts.

3.1 Autoencoding ear images

The shortcomings of anthropometric measurements mentioned in Section 2.2, together with their limited predictive power highlighted in previous works [17, 23], prompted the exploration of alternative features to be used in HRTF prediction tasks. In the literature, features from pinna images have been extracted using convolutional neural networks for the purpose of biometrics-based identification [20]. Thus, it was thought to employ convolutional layers in a VAE in order to derive salient features from its compact representation in an unsupervised manner.

Digital renderings of the z-buffer (also known as *depth maps*) of the 3D head models in the HUTUBS dataset [4] have been extracted, using the `pyrender` and `trimesh` packages for Python, and converted into 8-bit grayscale images. For each of the 55 unique head meshes, the point of view has been placed on either side of the head, so as to show each pinna separately.

In order to increase the amount of images used for training, several data augmentation techniques have been adopted. Firstly, variations of the point of view have been introduced, by tilting the camera along both elevation and azimuth. Secondly, slight vertical and horizontal offsets have also been applied. Lastly, each of the images thus generated has been duplicated and processed with sparse, discrete noise. A tool for selecting subsets based on the augmentation parameters has been developed, allowing the size of the dataset used for training to be anywhere from a few to well over a million pictures. Figure 3 shows the azimuth and elevation variations for a given subject.

The model used in these experiments is a VAE. The architecture is similar to the one described in Section 2.3.1, except it uses 2D convolutional layers. Within the encoder part, several dimensionality reduction techniques have been tested, such as *max pooling* layers or *strides* in the convolutional kernels, with no discernible difference in performance. Similarly, the decoder part has been originally implemented using transpose convolution with strides, which caused noticeable artifacts in the output images. In order to address this, a combination of *upsampling* layers and regular convolution has instead been used [19]. The hyperparameters of the architecture included the number of stacked layers, the number of convolutional filters for each layer, the number of latent dimensions, and batch normalization.

Figure 4: True and reconstructed pinna depth maps, using a convolutional VAE and data from only the frontal point of view.

Multiple experiments have been conducted, using different combinations of input data and model hyperparameters. The main criterion for evaluating the model effectiveness was the quality of the reconstruction. Indeed, images that have been faithfully reconstructed are indicative of a meaningful compact representation that can effectively encode the physical characteristics of the pinnae; conversely, ear pictures that fail to be distinguishable are not satisfactory. Furthermore, it was expected that latent variables show some degree of correlation with the anthropometric measurements related to the pinna, since they also describe factors of variance across pinnae.

The experiments have generally shown how, despite the elevation- and azimuth-dependent data augmentations exposing or hiding different parts of the pinnae thereby affecting their appearance, most of the variance in the data occurs in the surrounding area of the head, which is of no interest. This results in latent variables encoding mostly features related to the gradient in the background areas, while pinnae appear blurry and indistinguishable. Using data rendered from the same point of view and only augmenting the data using noise seems to alleviate the problem, as seen in the reconstruction in Fig. 4. However, the decrease in available training data negatively affects generalization, causing artefacts when sampling the latent space. Furthermore, the compact representation of the data points exhibits little to no correlation with the anthropometric measurements.

3.2 Autoencoding HRTFs

Synthesizing HRTFs from a set of parameters is a fundamental aspect of the individualization task. Since it is not yet fully understood what the optimal parameters are, it might be interesting to let a neural network derive its own by means of autoencoding the HRTFs. These parameters can then be either the target of another prediction task [5] or adjusted through perceptual feedback from the user [28]. Therefore, several VAE/CVAE models have been developed and trained, using different data layouts and architectures.

The architecture of the networks employed in this series of experiments is heavily dependent on the data formats, which are described in Section 2.1. In particular, given different dimensionalities, the following formats have been tested:

- 3D: HRTF patch processed using 3D convolutional layers, where only the middle HRTF is decoded;
- 2D: group of HRTFs ($elevation \times frequency$, $azimuth \times frequency$) or HRTF bins ($elevation \times azimuth$), processed using 2D convolutional layers;
- 1D: single HRTF, processed using either dense or 1D convolutional layers.

In the first case, the reconstruction target of the autoencoder is similar to the one-hot encoded output representation used by Yamamoto and Igarashi [28]. In the last case, a layer topology reminiscent of *ResNet* [10] has also been tested. This variant allows for deeper networks that do not suffer from vanishing gradient during training. For each scenario, a different number of convolutional filters, convolutional layers, and latent dimensions has been tried.

Just as for the pinna images case, the main goals here were a satisfactory reconstruction of the input and a meaningful latent space mapping. While the former can be assessed using quantitative metrics such as the *spectral distortion* (SD) between true and reconstructed HRTF, the latter is a more elusive property which can be inferred from the correlation with known HRTF predictors such as pinna anthropometric data, or azimuth and elevation angles.

The experiments conducted so far show mediocre reconstruction performances and little correlation with the aforementioned anthropometric variables, highlighting the need for more sophisticated models or more effective data representations. Indeed, reconstructed 2D and 1D representations appear blurry and lacking sharpness on the distinctive notches and peaks of the HRTFs. For the 1D case, which can be observed in Fig. 5, the average SD on the test set is 5.2 dB. Similarly, the one-hot output encoding used in the 3D representation experiments results in a mostly erratic behavior, most likely because without the Gaussian distribution constructed along the quantized levels dimension [28], the reconstruction task proves too difficult. The sharpest reconstructions are related to the 2D elevation-frequency representation using convolutional layers, which may suggest how contextual data can aid the training process; however, the reduced amount of data points negatively impacts the generalization capabilities of the model.

3.3 PCA-based HRTF prediction

This last set of experiments is based on the notion that autoencoders perform a similar task as PCA, while also learning non-linear feature spaces [16]. Accordingly we found that, with as little as 20 principal components, it is possible to reconstruct HRTFs with an average SD of 1.7 dB — see Fig. 6 (left). Moreover, some of these principal components show a high degree of correlation with elevation and azimuth angles. The models developed and evaluated here are therefore aimed at predicting these principal components from user data, and are inspired by the work of Chen et al. [5], who focused on predicting HRTFs over the horizontal plane only.

The type of data used by the models as predictors were either anthropometric measurements or the pinna depth maps used in Section 3.1. Since the aforementioned features do not change depending on the HRTF direction, the elevation and azimuth angles were also introduced as predictors, and HRTFs across the entire range of both elevation and azimuth were used. In order to limit the amount of features introduced by the depth maps, PCA has been performed on their pixels, and the first few principal components were used as input variables. Due to the nature of the input and output data, only architectures with dense layers have been tested. The main hyperparameters were: number of principal components for HRTF representation, number of principal components for depth map representation, and number and size of hidden layers.

The setup used for these experiments implemented a potential complete HRTF individualization procedure, where user data is fed as input, a DNN composed of fully-connected layers derives principal components for the HRTF, and the PCA loadings learned from a training set are used to derive a new HRTF. Thus, the entire system has been embedded in a 10-fold validation routine where, at each iteration, a training set comprising $\frac{9}{10}$ of the data was used to fit the neural network and calculate the PCA loadings, whereas a smaller test set with the remaining $\frac{1}{10}$ was used for validation. The metric observed throughout the process was the SD between true and reconstructed HRTF, calculated between 3 and 16 kHz.

The result of using only anthropometric measurements as predictors and a full HRTF range is an average SD of 4.5 dB and 4.7 dB for training and test sets respectively. While this is indeed promising, upon close inspection most generated HRTFs look similar, and while the general trend of the spectrum is correct, the sharp spectral features are not clearly distinguishable. When trying to derive principal components from pinna depth maps, a large number of

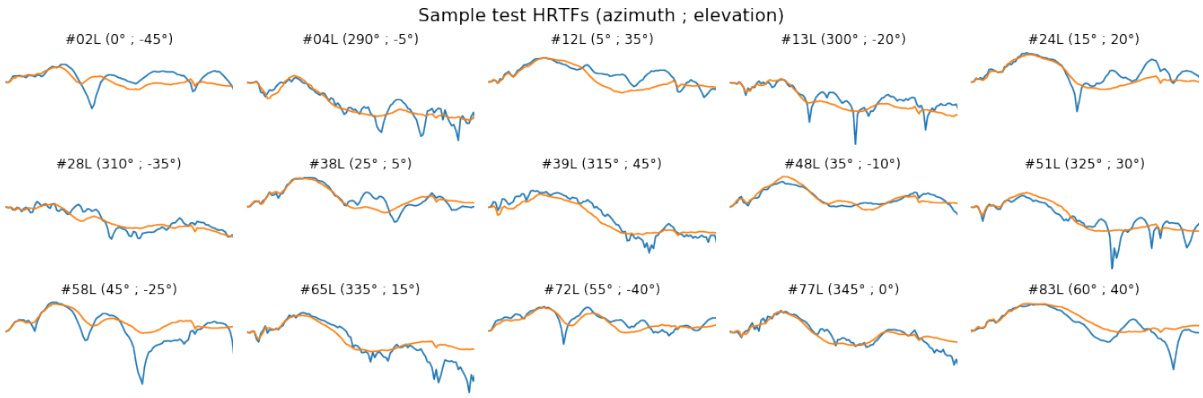

Figure 5: True (blue) and reconstructed (orange) HRTFs from different test subjects at different azimuth and elevation angles, using a convolutional VAE with residual layers.

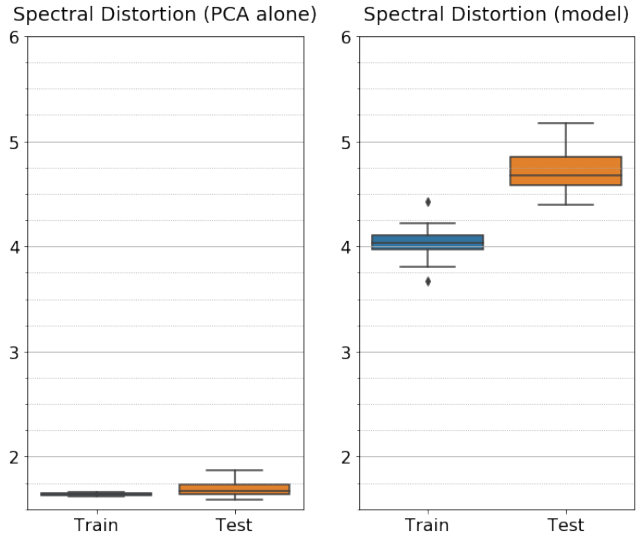

Figure 6: Box-and-whiskers plots for HRTF reconstruction using 20 principal components (left) and HRTF synthesis from anthropometric data and 40 pinna depth map principal components using a DNN (right), across 10 validation folds.

principal components proves necessary for a satisfactory reconstruction, which is always worse in the test set. Thus, features obtained from the principal components fail at generalizing the characteristics of the pinna, and result in poorer performances with more severe overfitting. A model combining both predictors has also been tested, peaking at 4.1 dB and 4.7 dB SD for training and test sets respectively as shown in Fig. 6 (right).

4 Conclusions

This paper presented some of the most promising advances in HRTF individualization, introduced the deep learning techniques associated with them, and detailed the results of further experiments based on the underlying knowledge base. Our contribution focused in particular on (1) extraction of meaningful predictors from user data; (2) unsupervised learning of HRTF data; and (3) synthesis of HRTFs from anthropometry.

While none of the above three threads of investigation has shown outstanding results, the knowledge acquired throughout the development and troubleshooting phases highlighted areas of improvement which are expected to pave the way to more accurate models for HRTF prediction from user data. No major difference with network size or hyperparameter tuning has been observed, although several setups proved more effective than others. In particular, providing a larger amount of data during the training process positively affects generalization, and so do deeper networks with batch normalization layers. It is also worth noting how SD alone is not a solid measure of perceptual fitness, and user tests in a XR environment might be necessary to reliably assess the performances of generated HRTF sets.

There exist several possible improvements for each of the three directions. Autoencoding pinna images could benefit from more sophisticated models such as those using *ResNet* or *Inception* convolutional layers. Autoencoding HRTFs in multiple dimensions may prove more useful when performed with a CVAE, where elevation and azimuth directions are fed either as two scalars or as one-hot encoded vectors. Moreover, in order to improve the sharpness of the most salient HRTF spectral features, the probabilistic one-hot encoding output representation introduced by Yamamoto and Igarashi [28] could be adopted. Finally, all deep models presented above would certainly benefit from having access to larger datasets with more subjects. This may be addressed by merging multiple datasets [22], which would require a normalization step to ensure that biases — such as those caused by different measurement setups — are not introduced into the learning process.

The code for the experiments, along with additional results, can be found on GitHub[1] in the form of Jupyter notebooks for Python.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 797850, and from Nord-Forsk's Nordic University Hubs programme under grant agreement No. 86892.

References

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon. A survey on ear biometrics. *ACM Trans. Embedded Computing Systems*, 9(4):39:1–39:33, March 2010.

[2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, pp. 1–4. New Paltz, New York, USA, October 2001.

[3] J. Blauert, ed. *The Technology of Binaural Listening*. Springer, New York, NY, USA, 2013.

[1] `https://github.com/miccio-dk/hrtf_individualize_dl`

[4] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl. A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses. *J. Audio Eng. Soc.*, 67(9):705–718, Sept. 2019.

[5] T.-Y. Chen, T.-H. Kuo, and T.-S. Chi. Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features. In *Proc. 44th IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2019)*, pp. 271–275. Brighton, UK, May 2019.

[6] C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space. *J. Audio Eng. Soc.*, 49(4):231–249, April 2001.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[8] C. Guezenoc and R. Seguier. HRTF individualization: A survey. In *Proc. 145th Conv. Audio Eng. Soc.* New York, NY, USA, Oct. 2018.

[9] 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. *J. Acoust. Soc. Am.*, 127(1):370–386, January 2010.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. Las Vegas, NV, USA, June 2016.

[11] K. Iida, H. Shimazaki, and M. Oota. Generation of the amplitude spectra of the individual head-related transfer functions in the upper median plane based on the anthropometry of the listener's pinnae. *Appl. Acoust.*, 155:280–285, Dec. 2019.

[12] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, Nov. 2019.

[13] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647, March 1992.

[14] A. Kulkarni, S. K. Isabelle, and H. S. Colburn. On the minimum-phase approximation of head-related transfer functions. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.* New Paltz, NY, USA, Oct. 1995.

[15] G. W. Lee and H. K. Kim. Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear. *Appl. Sci.*, 8(11), Nov. 2018.

[16] Y. Luo, D. N. Zotkin, and R. Duraiswami. Virtual autoencoder based recommendation system for individualizing head-related transfer functions. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.* New Paltz, NY, USA, Oct. 2013.

[17] R. Miccini and S. Spagnol. Estimation of pinna notch frequency from anthropometry: An improved linear model based on principal component analysis and feature selection. In *Proc. 1st Nordic Sound and Music Computing Conf. (Nordic SMC 2019)*, pp. 5–8. Stockholm, Sweden, Nov. 2019.

[18] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.*, 44(6):451–469, June 1996.

[19] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003

[20] H. Sinha, R. Manekar, Y. Sinha, and P. K. Ajmera. Convolutional neural network-based human identification using outer ear images. In J. Bansal, K. Das, A. Nagar, K. Deep, and A. Ojha, eds., *Soft Computing for Problem Solving*, vol. 817 of *Advances in Intelligent Systems and Computing*, pp. 707–719. Springer, Singapore, 2018.

[21] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., *Advances in Neural Information Processing Systems 28*, pp. 3483–3491. Curran Associates, Inc., 2015.

[22] S. Spagnol. Auditory model based subsetting of head-related transfer function datasets. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2020)*. Barcelona, Spain, May 2020.

[23] S. Spagnol and F. Avanzini. Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model. In *Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15)*, pp. 231–236. Trondheim, Norway, December 2015.

[24] S. Spagnol, K. B. Purkhús, S. K. Björnsson, and R. Unnthórsson. The Viking HRTF dataset. In *Proc. 16th Int. Conf. Sound and Music Computing (SMC 2019)*, pp. 55–60. Malaga, Spain, May 2019.

[25] S. Spagnol, G. Wersényi, M. Bujacz, O. Balan, M. Herrera Martínez, A. Moldoveanu, and R. Unnthórsson. Current use and future perspectives of spatial audio technologies in electronic travel aids. *Wireless Comm. Mob. Comput.*, 2018:17 pp., March 2018.

[26] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida. Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *J. Acoust. Soc. Am.*, 132(6):3832–3841, December 2012.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv*, Sept. 2016.

[28] K. Yamamoto and T. Igarashi. Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder. *ACM Trans. Graphics*, 36(6):212:1–12, Nov. 2017.

[29] S.-H. Yao, T. Collins, and C. Liang. Head-related transfer function selection using neural networks. *Arch. Acoust.*, 42(3):365–373, Sept. 2017.

[30] X.-Y. Zeng, S.-G. Wang, and L.-P. Gao. A hybrid algorithm for selecting head-related transfer function based on similarity of anthropometric structures. *J. Sound Vibr.*, 329(19):4093–4106, Sept. 2010.