TU Delft

Delft University of Technology

A hybrid approach to structural modeling of individualized HRTFs

Miccini, Riccardo; Spagnol, Simone

DOI
10.1109/VRW52623.2021.00022

Publication date
2021

Document Version
Accepted author manuscript

Published in
Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VRW 2021

Citation (APA)
Miccini, R., & Spagnol, S. (2021). A hybrid approach to structural modeling of individualized HRTFs. In *Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VRW 2021* (pp. 80-85). Article 9419096 (Proceedings - 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops, VRW 2021). IEEE. https://doi.org/10.1109/VRW52623.2021.00022

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

A hybrid approach to structural modeling of individualized HRTFs

Riccardo Miccini*
Aalborg University, Denmark

Simone Spagnol†
Aalborg University, Denmark
Delft University of Technology, Netherlands

ABSTRACT

We present a hybrid approach to individualized head-related transfer function (HRTF) modeling which requires only 3 anthropometric measurements and an image of the pinna. A prediction algorithm based on variational autoencoders synthesizes a pinna-related response from the image, which is used to filter a measured head-and-torso response. The interaural time difference is then manipulated to match that of the HUTUBS dataset subject minimizing the predicted localization error. The results are evaluated using spectral distortion and an auditory localization model. While the latter is inconclusive regarding the efficacy of the structural model, the former metric shows promising results with encoding HRTFs.

Index Terms: Hardware—Digital signal processing; Computing methodologies—Neural networks; Applied computing—Sound and music computing

1 INTRODUCTION

Providing users with a personalized head-related transfer function (HRTF) set is paramount for an immersive VR experience, free from localization errors and inside-the-head sound perception. However, direct acoustic measurement of the user's HRTF requires specialized apparatuses and is often strenuous and expensive. HRTF individualization technologies have been developed to tackle this problem, employing a variety of approaches such as HRTF selection, adaptation, or synthesis. Some of these are based on anthropometric measurements of the head, torso, and pinnae, while others involve calibration procedures where the user must provide feedback.

HRTFs mathematically encode the impact of a user's morphology on an incoming sound as a function of its spatial location. This is due to the acoustic phenomena occurring on the interface between different media — in this case, the air and the human body — such as reflection, diffraction, and diffusion, causing the human body to act as a filter. Most notably, the parts of the human body interacting with the incoming wavefront and therefore known to contribute to the HRTF are the shoulders, the torso, the head, and most prominently the pinnae of the listener [1].

Perceptual cues affecting localization along the vertical direction have been researched since the 70s. In particular, Hebrank and Wright [11] established that spectral cues for vertical localization exist between 4 kHz and 16 kHz, and that only sounds occupying this frequency range can be reliably localized along the median plane. These cues take the form of spectral peaks and notches. According to Shaw [21], the pinna resonant modes are thought to cause the most prominent peaks in the HRTF. While the center frequency of peaks is relatively insensitive to changes in elevation of the sound source [12], pinna notches, especially the lowest-frequency one, are generally seen to increase with the elevation angle, providing a salient elevation cue [11]; conversely, notches exhibit little variation with changes in azimuth [16].

*e-mail: rmicci18@student.aau.dk
†e-mail: s.spagnol@tudelft.nl

The individual contributions of head, torso, and pinna anatomy can be isolated and investigated, in order to replicate the spectral effects of the underlying physical phenomena. A *structural model* is a system whereby such spectral effects are independently modeled according to anthropometric data and combined to create a personalized HRTF.

This paper extends the *mixed structural modeling* approach [8] by the use of deep learning (DL) models and, starting from a previous work from the authors [18], offers the following contributions:

- a deep-learning-based solution for synthesizing pinna-related responses (PRTFs) from user pictures;
- a hybrid approach for combining such PRTFs with the best-matching interaural time difference from a dataset and a generic head-and-torso response through a customized structural HRTF model;
- an evaluation of the performances of the resulting HRTFs in vertical localization using an objective metric and an auditory localization model.

2 RELATED WORK

Structural modeling of HRTFs finds its origin in the work of Brown and Duda [4], where the physical sources of sound diffraction, delay, and reflection are simulated in the time domain. In more recent times, frequency-domain structural models of the pinna have also been proposed. In the work of Spagnol et al. [24], the 2D reflection paths derived from three pinna edges are used to predict pinna notches, whereas Mokhtari et al. [20] estimate spectral peaks from individual anthropometric parameters.

Several individualization methods involving DL technologies have also been proposed. Luo et al. [17] trained a stacked denoising autoencoder to encode and reconstruct HRTFs from multiple subjects. The resulting latent representation is then manipulated using feedback from the user to optimize their localization performances. The solution devised by Yamamoto and Igarashi [27] similarly relies on user feedback to adjust the latent vector from which an HRTF is synthesized. However, their model features a richer input data representation comprising neighboring HRTFs, parallel frequency and time representations of the HRTF, and a sophisticated variational autoencoder architecture. Lee and Kim [14] use two separate neural networks: a densely-connected one processing the anthropometric measurements of head and torso, and a convolutional one taking an image of the user's pinna as input. The outputs of these networks are then fed into a DNN which is trained to predict a given HRTF.

Several recent studies focused on autoencoding HRTF data. Most notably, T. Y. Chen et al. [5] train a dense autoencoder to reconstruct the magnitude response of an HRTF from its latent representation and its azimuth coordinate. The encoder part of the network projects each HRTF onto a latent space of reduced dimensionality. A separate DNN is then trained to predict these latent vectors using anthropometric data as input. Lastly, in a very recent paper, W. Chen et al. [6] train a convolutional denoising autoencoder on 2D frequency-elevation input features derived from listener-specific directional components of HRTFs, with the purpose of optimizing HRTF dataset storage.

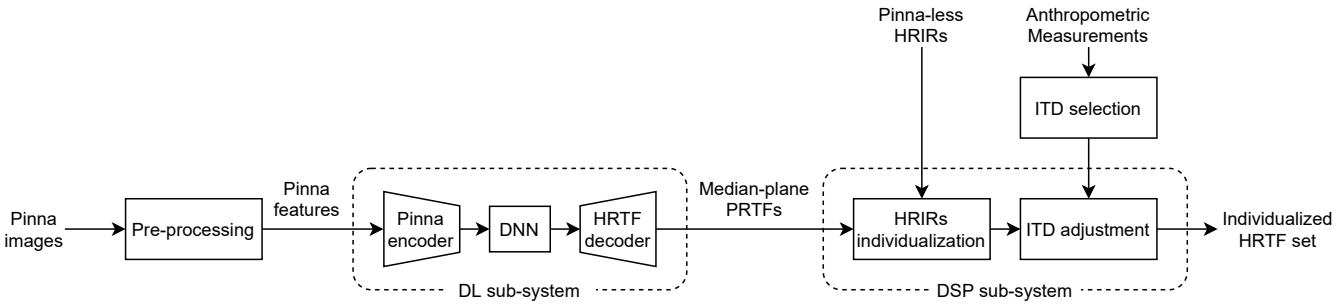

Figure 1: HRTF individualization pipeline, with its inputs, outputs, and constituting elements.

3 The hybrid structural HRTF model

The structural model presented in this paper is based on a pipeline composed of three cascaded DL models, as well as conventional DSP blocks. It has been aptly defined as *hybrid* in that it combines synthesized, selected, and measured components. An overview of the processing pipeline and its constituting elements can be seen in Fig. 1. In particular, the architecture comprises:

- a DL sub-system capable of synthesizing PRTFs from an image of the pinna or analogous 2D features, such as the pinna edges;
- a DSP sub-system implementing a structural model where an HRTF set comprising only of shoulders and head reflection effects is filtered using the PRTF described above and processed to match the interaural time difference (ITD) of a fitting subject from an HRTF database.

The neural networks employed in the first step are trained separately on their respective datasets and then combined into a prediction script capable of generating an individualized PRTF. Once trained, the prediction script can be used independently and only requires the pre-trained model weights to work. A separate MATLAB script script takes care of performing the last steps of the pipeline and generate an individualized HRTF set in the SOFA format[1], which is compatible with most binaural rendering engines.

3.1 Deep learning sub-system

The DL sub-system, implemented using the PyTorch library, comprises the following three building blocks:

- a *variational autoencoder* (VAE) whose encoder is used for deriving a compact representation, called z_{ear}, of 2D pinna features such as pictures, depth maps, or edge maps;
- a *conditional variational autoencoder* (CVAE) whose decoder is analogously used to synthesize a PRTF from a compact representation, called z_{hrtf};
- a *deep neural network* (DNN) capable of predicting the compressed representation z_{hrtf} from z_{ear}.

Fig. 2 provides an overview of the algorithm, as used during training and evaluation (in green). The theoretical background and architecture details of each of these DL models are covered in the following subsections.

[1] www.sofaconventions.org

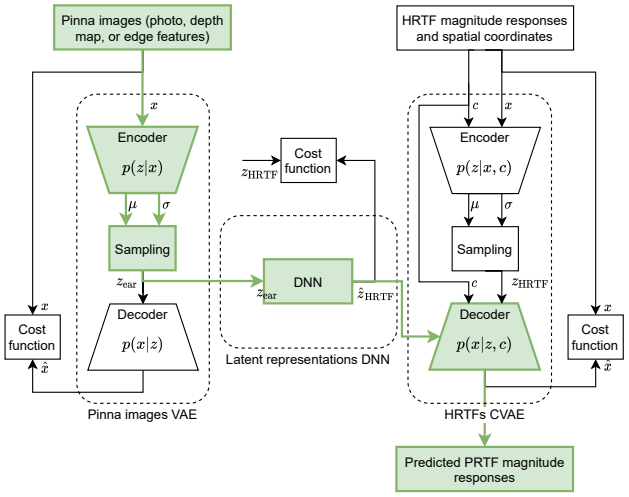

Figure 2: DL sub-system for PRTF synthesis; data flow during evaluation is highlighted in green.

3.1.1 Encoding of pinna images

This first part focuses on deriving useful features from pinna images, which could later be used as predictors for the HRTFs. Specifically, these features correspond to the compressed encoding, known as *latent representation*, of an artificial neural network. We employed a variant of autoencoders called *variational autoencoders* (VAE). VAEs are probabilistic models mapping an input sample to a probability distribution. Furthermore, the latent space distribution is constrained into an isotropic multivariate Gaussian with the help of a Kullback-Leibler divergence term in the loss function. These properties prevent the input data points to be mapped to limited sparse areas of the latent space and make VAEs particularly suitable as generative models; however, in this case, only the encoder part is used. This encoder network approximates the probability $p(z|x)$ — that is, the distribution of the latent variable z given the input data x. This distribution is parameterized by its mean value μ and the logarithm of its variance σ, which are the two outputs of the encoder. During training, a value of $z \sim \mathcal{N}(\mu, \sigma)$ is sampled from the distribution and fed into the decoder, which attempts to reconstruct the original input. The model is shown in Fig. 2 (left side).

A VAE architecture based on *inception modules* was designed and implemented. Inception modules are a particular configuration of convolutional layers found in the work of Szegedy et al. [26]. Within one such module, several convolutional layers with differently sized kernels are applied in parallel, and their respective output features are concatenated or summed. The rationale behind this design choice is to let the network discover the most relevant feature for the task

at hand, which may be expressed by kernels of different sizes at different points of the computational graph.

Each module is composed of four convolutional stacks with kernels of shape 1×1, 3×3, 5×5, and 7×7 respectively, preceded by a single 1×1 convolutional layer for dimensionality reduction. Each stack is composed of 2D convolution, batch normalization, and ReLU activation. Furthermore, two extra paths with max pooling layers of shape 3×3 and 5×5 are provided. The input of each parallel path is padded to obtain equal output shapes, and their respective outputs are summed together. The modules are arranged into four stages and separated by downsampling or upsampling layers for the encoder and decoder respectively. Finally, dense layers with no activations are featured before and after the network bottleneck. The configurable hyperparameters are the size of the latent space and the number of inception modules per stage.

Three datasets were tested, consisting of grayscale images, depth maps, and edge maps. These have been generated from the 3D head meshes of 55 HUTUBS [3] subjects, out of the 58 available ones, thereby discarding repeated measurements and non-human subjects. A script was developed to load the mesh, place it in a 3D scene together with a camera pointing on either side of the head, align the entrance of the ear canal with the cameras, and render it into either a depth map or a grayscale image. In the latter case, the scene was illuminated using a directional lamp, simulating an infinitely far point light source, located at the camera position and pointing towards the head. Due to the limited amount of available data, augmentations were introduced in the form of slight variations in camera yaw and pitch, using the ear canal entrance as the pivot point so as to keep it at the center of the image. Finally, the third dataset of edge maps has been created by feeding the depth map dataset into a Canny edge detection algorithm.

Each dataset was parametrically generated with an image size of 256×256 pixels and variations in pitch and yaw from $-15°$ to $15°$ with a step of $5°$, for a total of 49 different orientations. This accounted for 5390 images of left and right pinnae. Of these, 392 images belonging to 4 subjects were set aside for testing, while a randomly selected subset corresponding to 10 % of the remaining ones was used for validation during training. Furthermore, the grayscale renderings were merged with the following datasets of ear pictures: the AMI Ear Database [9], the AWE Dataset [7], and the IITD Ear Dataset [13]. Finally, to further extend the size of the datasets, three types of noises were introduced: gaussian noise, salt and pepper noise, and speckle noise, to be applied during training. Each type of noise had a 25 % probability of being applied (including a 25 % of no noise), thus the datasets were effectively augmented by a factor of 4.

3.1.2 Encoding of HRTFs

The second part of this project consisted in autoencoding the HRTF magnitude responses. As mentioned earlier, novel output data — in this case, a customized magnitude response — can be synthesized by sampling the latent space of a trained VAE. HRTFs, however, depend not only on the individual characteristics of the users expressed by their latent parameters, but also on the spatial coordinate of interest. Therefore, the distribution $p(x|z)$ appnroximated by the decoder must be conditioned by said spatial coordinates. This is achieved using a class of ANNs called *conditional variational autoencoders* (CVAEs). These are analogous to VAEs but trained in a supervised fashion: the encoder learns a probability distribution $p(z|x,c)$ where c is a vector of data labels, while the decoder learns the distribution of the input data based on the latent variables z and the labels c, i.e. $p(x|z,c)$. This process is shown in Fig. 2 (right side).

Since previous attempts at leveraging the spatial-frequential hierarchy of HRTF magnitudes using 1D convolutional layers proved ineffective [18], a fully-connected architecture was employed. The CVAE network implemented here comprises a customizable number of dense layers with ReLU activation. The final encoder layers, in charge of predicting the mean and log-variance of z, uses dense layers without an activation function, so as not to clamp the data into an arbitrary range.

The input used by the models is composed of pairs of HRTF magnitude responses and data labels, generated using the following pipeline. Individual HRIRs are extracted from a SOFA file, along with their respective spatial coordinates consisting of azimuth and elevation angles. Subsequently, they are converted to the frequency domain using a Fast Fourier Transform algorithm. The number of frequency bins n_{fft} is a customizable hyperparameter of the system, and yields a complex-valued spectrum comprising $\lfloor n_{\text{fft}} \rfloor + 1$ real-valued frequency bins. The resulting magnitude response is then converted to logarithmic units and clipped to a dynamic range of 120 dB. The test set comprised the same four subjects kept aside in the previous subsection, while the validation set used during training was composed of a random 20 % subset of all other HRTFs.

The CVAE model was trained on two variations of the HUTUBS dataset: one containing HRTFs across the entire spatial grid, and one based on median-plane data only. We hypothesized that training with a larger dataset would improve the generalization capabilities of the model, thus avoiding overfitting. Furthermore, to verify whether a continuous elevation label would benefit from the smoothness of the decoded median plane HRTFs, another set of trainings was performed on the same two datasets, but with the HRTFs labeled according to an interaural-polar coordinate system.

3.1.3 Prediction of encoded representations

This section tackles the translation of the 2D pinna features-encoded representation z_{ear} into the HRTF-encoded representation z_{hrtf}. This can be formulated as a supervised learning task where z_{ear} constitutes the input data and z_{hrtf} is the target. The prediction of the HRTF latent representation is believed to be possible because, in order to faithfully reconstruct its input, the pinna images VAE must encode information pertaining to the individual morphology of the pinnae within its latent dimensions. According to widely established existing literature, the anthropometric parameters of the pinnae are known to affect their frequency response, although their exact impact is an open topic of inquiry. While manually extracted anthropometric data such as those available in the HUTUBS dataset are arbitrarily chosen and prone to systematic measurement biases, in this work it is hypothesized that automatically derived ones may offer additional insight.

An artificial neural network was therefore designed and implemented to perform this task, in the form of a deep multilayer perceptron. The model is therefore composed of an input layer taking the z_{ear} vector plus a spatial coordinate vector, a number of hidden fully-connected layers with leaky ReLU activation, and optional dropout layers for regularization. In the previous subsection, the spatial coordinates of a given HRTF were used to condition the CVAE encoder and decoder, to be able to sample points in the latent space belonging to the given spatial orientation. However, this external conditioning is not sufficient to ensure the complete disentanglement between the latent representation of a given HRTF and its spatial coordinates, hence their inclusion in the DNN input.

The pairs of corresponding z_{ear} and z_{hrtf} training data points are extracted from the pinna VAE and HRTF CVAE respectively, using their encoder sub-networks. The resulting latent vectors are then stored along with labels comprising metadata such as the subject ID, the pinna under consideration (left or right), and spatial coordinates.

For this model, several datasets were tested, differing in the content of the target HRTF latent data and input features. The former included the z_{hrtf} vectors extracted from the CVAE trained across the entire spatial grid, those extracted from the same model but filtered by median plane only, and median plane z_{hrtf} extracted from the CVAE trained on the median plane only. Regarding the input

features used as predictors, the choice included the latent representations extracted from the pinna images, depth maps, and edge maps.

Beside the training scheme mentioned earlier, two other alternative approaches were tested. One involved the latent vectors from HRTFs labeled using the aforementioned interaural-polar system for spatial coordinates instead of the conventional vertical-polar one. The last approach consisted in using a subset of the principal components of the HRTF latent vectors as prediction target, calculated by principal component analysis (PCA). The rationale behind this was to simplify the regression task by reducing the number of target variables and projecting it onto a more interpretable set of basis functions. The part of the DL sub-system associated to this model is shown in Fig. 2 (center).

3.2 HRTF set generation

The final step consists in using the PRTF synthesized through the previous steps to generate an entire HRTF set. It is important to notice that, although the CVAE model is trained on HRTF magnitude responses and amongst its latent representation there may be dimensions responsible for imparting the effect of head, shoulders, and torso, the DNN input consists only of spatial coordinates and pinna latent features, which are not trustworthy predictors of the head, shoulders, and torso impact. Thus, the output of the DL sub-system is thought to represent the individualized PRTF.

While the system described earlier can generate PRTFs over a spatial grid of different azimuths and elevations, pinna notches and other spectral features are known to be relatively stable across the horizontal direction [16], so only the responses along the median plane are extracted and used. This is in line with the fact that pinna images alone do not contain the necessary predictors for deriving binaural cues, which are known to be related to the anthropometry of the head. For these reasons, spectral features and binaural cues that are not caused by the pinnae must be accounted for separately.

The structural model presented herein identifies and integrates two external contributions: the effect of head, torso, and shoulders, and the interaural time difference. The former is derived from one of the VIKING dataset [25] subjects consisting of a KEMAR mannequin with its original pinnae removed and the slots filled with a silicone baffle. The spectral features introduced in this way are a coarse approximation of the effect of shoulders and torso. However, these are thought to provide localization cues at frequencies below 3 kHz, which are only crucial when localizing narrow-band sounds under that threshold [1]. The pinna contribution is applied to the pinna-less HRIRs as a minimum-phase IIR filter matching the magnitude response of the PRTFs and derived using the Yule-Walker method with a sufficiently large order.

The ITD, on the other hand, is extracted from a HUTUBS subject and applied to the HRIRs of the generated set. The relevant HUTUBS subject is chosen using the HRTF selection algorithm proposed by Spagnol [23], where three anthropometric parameters — corresponding to head width, head depth, and shoulder circumference — are used as features of a linear regression model predicting a horizontal localization error metric. The metric is computed for all HUTUBS subjects, and the one minimizing the error is selected. Since the spatial grid employed by the VIKING dataset differs from the HUTUBS one, an interpolation algorithm to convert the former to the latter was implemented accordingly. The choice of adopting the spatial grid of the HUTUBS dataset simplifies the comparison between its subjects and an individualized HRTF.

4 Results

This section presents and discusses the results obtained from the trainings as well as from using the individualized HRTF on a virtually simulated listener. In particular, the CVAE and DNN models

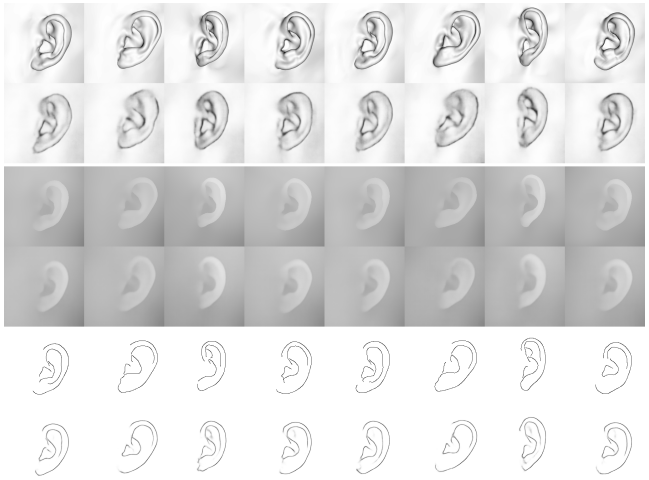

Figure 3: Comparison of original and reconstructed 2D pinna features (grayscale images, depth maps, and edge maps) for unseen subjects.

were evaluated in terms of *spectral distortion* (SD), which is often used in the relevant literature [24] and is calculated as:

$$\mathrm{SD}_{\mathrm{dB}}(H,\hat{H}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(20\log_{10}\frac{|H(f_i)|}{|\hat{H}(f_i)|}\right)^2} \qquad (1)$$

where H is the original HRTF, $\hat{H}$ is the predicted one, N is the number of frequency bins, and f_i is a given frequency bin.

Conversely, the individualized HRTFs are evaluated using the vertical localization model by Baumgartner et al. [2]. The model comprises an auditory processing and a spatial mapping block, which can estimate the probability of user responses across sagittal planes. From these probability vectors, we derive two psychoacoustic performance parameters: the *quadrant error rate* (QE) and the *polar root-mean-square error* (PE) [22]. The former represents the fraction of responses within $\pm 90°$ from the target angle, also called local responses, while the latter corresponds to the RMSE of the local responses. All the results presented in the following subsections are collected from the data of four unseen HUTUBS subjects.

4.1 Deep learning sub-system

Reconstructed 2D features along with their respective input are shown in Fig. 3. It is worthwhile to note how extending one of the training sets with real-world photographs caused poorer reconstruction performances than with the other, more homogeneous datasets. Nevertheless, both the model trained on the depth maps and the one trained on edge maps show a satisfactory level of coherence. However, the reconstruction performances are in part dependent on the decoder sub-network, which is not needed in the final pipeline. One of the main challenges that arise from using variational autoencoders is the interpretability of their latent space representation. Indeed, a given latent variable may account for several observable or perceptual factors at once, or provide no discernable contribution to the encoding.

Table 1 shows the mean SD of all four trained models, calculated from the median plane HRTFs of the unseen test subjects over a frequency range of 500 Hz to 16 kHz. These values represent the theoretical maximum performance achievable by the entire system. It is possible to notice how, while there is a slight increase in performance from vertical-polar to interaural-polar systems, the amount of training data accounts for most of the difference in spectral distortion. Interestingly, when looking at the distribution of the SD

Table 1: CVAE model performances for each training strategy.

CVAE training	SD [dB]
Full grid	1.91
Median plane	2.93
Full grid (interaural)	1.84
Median plane (interaural)	2.73

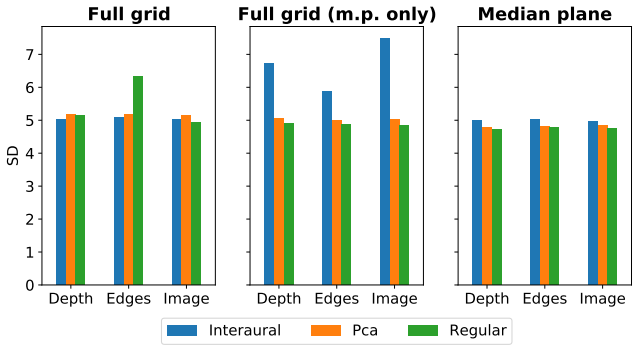

Figure 4: Comparison between training strategies (colors), for a range of input features (x-axis) and datasets (subplots).

over the range of elevations for all models in Table 1, a clear pattern emerges whereby HRTFs at low elevations ($< -60°$) present the largest errors. This may be explained by a steep change in response in the HUTUBS HRTFs, known to be caused by torso-induced damping [3].

A comparison of the aforementioned models is shown in Fig. 4. We observe similar performances across the entire range of training schemes, with a slight advantage for the regular procedure using vertical-polar coordinates. Furthermore, the SD seems to be unaffected by the type of input feature used, while the best results are generally achieved by training on the median plane dataset and using the CVAE decoder trained on the median plane only. Two alternative hypotheses for this phenomenon are advanced: in the worst case, the z_{ear} are not particularly useful for the task at hand and the DNN minimizes its prediction error by assigning low weights and arbitrary biases to the feature inputs, thereby coalescing all input data points into a specific region of the z_{hrtf} manifold. Alternatively, the features extracted by the pinna images VAE from each of the datasets are similar and thus they have equivalent prediction power.

4.2 Generated HRTFs

This section evaluates individualized HRTFs generated using the hybrid structural model, according to anthropometric data and 2D pinna features of the unseen HUTUBS subjects. The evaluation is limited to the most relevant combinations of models. Specifically, models using z_{ear} vectors derived from the pinna depth maps were excluded, as well as combinations using PCA decomposition or interaural-polar coordinates. The performances are compared against a baseline generic HRTF set collected from the FABIAN head-and-torso simulator [15] included in HUTUBS, and summarized in Fig. 5.

Most of the models appear to reduce QE by up to 5 %. However, PE similarly increases by up $\sim 5°$ on average. Most notably, the standard deviation of the metrics for the individualized HRTFs is significantly lower than the baseline's, indicating a more consistent behavior across subjects. Fig. 6 exemplifies these results, showing how the predicted responses are focused around the listener's zenith. It is thought that the shallow spectral features synthesized by the DL sub-system are interpreted by the localization model as belonging to elevations above the head, where spectral notches gradually soften.

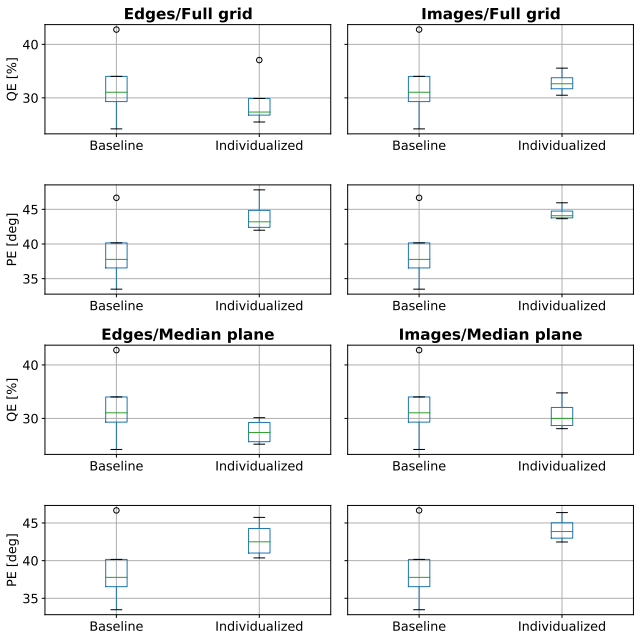

Figure 5: Psychoacoustic metrics (rows) computed using the localization model, for each of the chosen model combinations (columns).

A recent unpublished work, aimed at evaluating different HRTF profiles within a gamified interactive 3D environment, employed the pipeline described here to generate HRTF sets from pinna edge maps of the users. Based on the gathered data, the individualized HRTF improved the vertical localization error by 1° and the horizontal error by 2.5° on average compared to a generic HRTF, which proves promising despite the lackluster results shown by the psychoacoustic metrics.

5 Conclusion and Future Work

The implemented systems fulfills the initial requirement of generating customized HRTF sets from easily obtainable user data. However, the metrics considered in the evaluation are inconclusive regarding the efficacy of this structural model, and more investigation is needed. Nevertheless, it is believed that a thorough optimization of the building blocks of the pipeline may provide substantial performance gains. Inspecting each building block of the DL sub-system yielded the following findings:

- It is unsure whether the latent variables extracted by the pinna VAE are good predictors of HRTF spectral characteristics, compared to conventional anthropometric measurements such as those available in HRTF datasets. More user data would be needed, although training on larger and more heterogenous datasets seems to degrade performances.
- The conditional VAE can effectively encode an entire HRTF dataset and behaves faithfully on unseen subjects, although the reconstructed notches lack depth.
- The deep neural network used for mapping the two latent representations is prone to overfitting due to the lack of training data where there is a correspondence between 2D pinna features and HRTFs. Furthermore, it may benefit from employing additional predictors.

If considered separately, the CVAE may find applications in compressing HRTFs or interpolating HRTF sets over a finer spatial

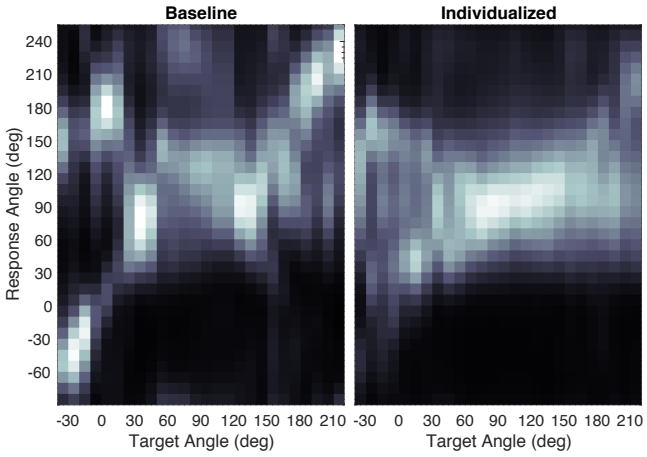

Figure 6: Probability distribution (color) of the predicted responses (y-axis) against the target angle (x-axis).

grid. Its low number of trainable parameters makes it suitable for embedded applications such as hearing aids and wireless earbuds. Furthermore, it could constitute the foundation of an HRTF individualization approach based on perceptual feedback from the user.

The lack of large-scale HRTF datasets containing rich anthropometric data such as head scans or pinna images is one of the main hindrances to the application of DL techniques. Recent efforts have been made with regards to generating arbitrarily sized synthetic PRTF sets [10], which may improve the generalization abilities of both the CVAE and the DNN models by providing more data to learn from.

Finally, the head-and-torso effect provided by the VIKING pinnaless subject could be adapted to reflect the anatomical differences between users, by means of frequency scaling parametrized according to a set of anthropometric measurements [19].

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 NordForsk's Nordic University Hubs programme under grant agreement No. 86892.

References

[1] V. R. Algazi, C. Avendano, and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122, Feb. 2001.

[2] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2):791–802, Aug. 2014.

[3] 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. *Journal of the Audio Engineering Society*, 67(9):705–718, Sept. 2019.

[4] C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5):476–488, Sept. 1998.

[5] 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)*, pp. 271–275. IEEE, Brighton, United Kingdom, May 2019.

[6] W. Chen, R. Hu, X. Wang, and D. Li. HRTF Representation with Convolutional Auto-encoder. In Y. M. Ro, W.-H. Cheng, J. Kim, W.-T. Chu, P. Cui, J.-W. Choi, M.-C. Hu, and W. De Neve, eds., *MultiMedia Modeling*, vol. 11961, pp. 605–616. Springer International Publishing, Cham, 2020.

[7] Ž. Emeršič, V. Štruc, and P. Peer. Ear Recognition: More Than a Survey. *arXiv:1611.06203 [cs]*, Feb. 2019.

[8] M. Geronazzo, S. Spagnol, and F. Avanzini. Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pp. 1–8. IEEE, Fira, Santorini, Greece, July 2013.

[9] E. Gonzalez, L. Alvarez, and L. Mazorra. AMI Ear Database. http://ctim.ulpgc.es/research_works/ami_ear_database/#whole.

[10] C. Guezenoc and R. Séguier. A wide dataset of ear shapes and pinna-related transfer functions generated by random ear drawings. *The Journal of the Acoustical Society of America*, 147(6):4087–4096, June 2020.

[11] J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *The Journal of the Acoustical Society of America*, 56(6):1829–1834, Dec. 1974.

[12] Y. Kahana and P. A. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of Sound and Vibration*, 300(3):552–579, Mar. 2007.

[13] A. Kumar and C. Wu. Automated human identification using ear imaging. *Pattern Recognition*, 45(3):956–968, Mar. 2012.

[14] G. Lee and H. Kim. Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear. *Applied Sciences*, 8(11):2180, Nov. 2018.

[15] A. Lindau, T. Hohn, and S. Weinzierl. Binaural Resynthesis for Comparative Studies of Acoustical Environments. In *Audio Engineering Society Convention 122*. Audio Engineering Society, May 2007.

[16] E. A. Lopez-Poveda and R. Meddis. A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, 100(5):3248–3259, Nov. 1996.

[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*, pp. 1–4. IEEE, New Paltz, NY, USA, Oct. 2013.

[18] R. Miccini and S. Spagnol. HRTF Individualization using Deep Learning. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 390–395, Mar. 2020.

[19] J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan. Psychophysical customization of directional transfer functions for virtual sound localization. *The Journal of the Acoustical Society of America*, 108(6):3088–3091, Dec. 2000.

[20] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Vertical normal modes of human ears: Individual variation and frequency estimation from pinna anthropometry. *The Journal of the Acoustical Society of America*, 140(2):814–831, Aug. 2016.

[21] E. A. G. Shaw. The acoustics of the external ear. *Acoustical Factors Affecting Hearing Aid Performance*, 1980.

[22] S. Spagnol. Auditory model based subsetting of head-related transfer function datasets. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2020)*, pp. 391–395. Barcelona, Spain, May 2020.

[23] S. Spagnol. HRTF Selection by Anthropometric Regression for Improving Horizontal Localization Accuracy. *IEEE Signal Processing Letters*, 27:590–594, 2020.

[24] S. Spagnol, M. Geronazzo, and F. Avanzini. On the Relation Between Pinna Reflection Patterns and Head-Related Transfer Function Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(3):508–519, Mar. 2013.

[25] S. Spagnol, R. Miccini, and R. Unnthorsson. The Viking HRTF dataset v2, Oct. 2020. doi: 10.5281/zenodo.4160401

[26] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, June 2015.

[27] K. Yamamoto and T. Igarashi. Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder. *ACM Transactions on Graphics*, 36(6):1–13, Nov. 2017.