HRTF Individualization using Deep Learning

Riccardo Miccini
Aalborg University
rmicci18@student.aau.dk

***Abstract*—The research presented in this paper focuses on investigating HRTF individualization techniques using deep learning approaches, with an emphasis on HRTF estimation and synthesizing. (also mention FABIAN dataset, autoencoders, depthmaps etc, see slides)**

***Index Terms*—Spatial Audio; HRTF; Deep Learning**

I. Introduction

Virtual reality (VR) and augmented reality (AR) research has made substantial progress over the last decades, and virtual environments created using binaural sound rendering technologies find applications in a wide array of areas, such as aids for the visually impaired, tools for audio professionals, and VR-based entertainment systems.

These 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 standard anthropometric pinnae, such as those in dummy heads, often results in localization errors and wrong spatial perception [1]. In fact, while generic HRTFs may successfully approximate the interaural time difference (ITD) and interaural level difference (ILD) cues which are 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 each ear.

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. Recently, attempts have been performed at synthesizing or customizing HRTFs using various data from users such as anthropometric measurements, 3D scans, or perceptual feedback.

This paper investigates methods for generating individualized HRTFs, in particular using newly-developed deep learning algorithms, and further expands on the topic by documenting the replication attempts and experiments conducted as part of the research. In the next section, current relevant contributions to the field are introduced. The Methods section details the computational techniques used in selected works from the literature as well as in the research carried out as part of the seminar, with particular focus on deep learning methods. In Results, the applications and outcomes of the aforementioned techniques for replications and other experiments are discussed, with the purpose of assessing their effectiveness. Finally, closing remarks as well as pointers for future research are stated in the Conclusions section.

II. State of the art

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 [2] divides such alternative approaches into *numerical simulation*, *anthropometrics*-based, and *perceptual feedback*-based.

The former method consists in simulating the propagation of acoustic waves around the subject, using 3D scans; the most common simulation schemes include Fast-Multipole-accelerated Boundary Element Method (FM-BEM) [3] and Finite Difference Time Domain (FDTD) [4] for frequency and time domain respectively.

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. In 2010, Zeng [5] implements an hybrid model based on principal component analysis (PCA) and multiple linear regression, which uses anthropometric parameters to select the most suitable HRTF set for the given user. Similarly, user feedback on perceptual tests can be used to inform regression models for tasks such as those listed above.

In more recent times, there has been an interest in solving the aforementioned tasks using deep learning techniques. In 2017, Yao [6] uses anthropometric measurements to select the most suitable HRTF sets from a larger database. In 2018, Lee [7] develops a double-branched neural network that processes anthropometric data with a multi-layer perceptron and edge-detected pictures of the ear with convolutional layers, combining the outputs of the two into a third network to estimate HRTF sets. Again in 2017, Yamamoto [8] trains a variational autoencoder on HRTF data, and devises a perceptual calibration procedure to fine-tune the latent variable used as input by the generative part of the model. Finally, in 2019, Chen et al. [9] train an autoencoder to reconstruct HRTFs along the horizontal plane, and subsequently uses the resulting latent representations as targets for a multilayer perceptron which feeds on anthropometric data and azimuth angle, allowing users to synthesize new HRTFs using the MLP and decoder.

III. 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 predictors, and the deep neural network architectures adopted.

A. HRTF representation

single HRTF, 2d repr, 3d repr, patch [8], principal components

B. Anthropometric data extraction

FABIAN dataset [fabian_hutubs_2019], depthmaps, CIPIC [algazi_cipic_2001] anthropometric parameters

C. Autoencoder

Most conventional neural networks are employed to predict a target y from an input x in a supervised manner. 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 $\hat{x}$. The purpose of autoencoders is to learn useful features from the input data in an unsupervised manner [10].

Autoencoders usually consists of a feed-forward neural network composed of two sub-nets: 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 the autoencoder have been developed. Each variant extends on the conventional autoencoder architecture by promoting different properties of the latent space, thereby catering to different tasks such as denoising, classification, or — as it is this case here — generative applications. Two common generative models based on the autoencoder are described below.

1) Variational autoencoder (VAE): Variational autoencoders are a class of generative models, extending the classic autoencoder. A VAE is a probabilistic model where the encoder $q_\theta(z|x)$ maps the probability distribution a certain latent representation given a data point, and the decoder $p_\phi(x|z)$ outputs the probability distribution of the data, given a point in the latent space. In VAEs, it is often desireable to model the latent space as an isotropic multivariate Gaussian distribution. This constraint is enforced by introducing a measure of distance between the aforementioned prior distribution $p_\theta(z) \sim \mathcal{N}(0,1)$ and the encoder distribution, called Kullback-Leibler divergence. This probabilistic framework proves useful when synthesizing HRTFs, because it can learn causal factors of variations in the data [11]. However, there exists no way of generating a specific HRTF — i.e. one for a given combination of azimuth and elevation angles; while the points in latent space are likely to generate plausible new data, one can only obtain random instances of said data. The class of autoencoders described below aims at addressing this shortcoming.

2) Conditional variational autoencoder (CVAE):

D. Architectures and models

IV. Results

Introduce the three main themes.

A. Autoencoding ear images

B. Autoencoding HRTFs

C. Autoencoding principal components

1) Predicting principal components:

V. Conclusions

[12]

Bibliography

[1] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, “Binaural Technique: Do We Need Individual Recordings?” *J. Audio Eng. Soc*, vol. 44, no. 6, pp. 451–469, 1996.

[2] C. Guezenoc and R. Seguier, “HRTF Individualization: A Survey,” in *Audio Engineering Society Convention 145*, 2018.

[3] N. A. Gumerov, R. Duraiswami, and D. N. Zotkin, “Fast Multipole Accelerated Boundary Elements for Numerical Computation of the Head Related Transfer Function,” in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07*, 2007, vol. 1, pp. I–165–I–168.

[4] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, “Comparison of Simulated and Measured HRTFs: FDTD Simulation Using MRI Head Data,” in *Audio Engineering Society Convention 123*, 2007.

[5] 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,” *Journal of Sound and Vibration*, vol. 329, no. 19, pp. 4093–4106, Sep. 2010.

[6] S.-N. Yao, T. Collins, and C. Liang, “Head-Related Transfer Function Selection Using Neural Networks,” *Archives of Acoustics*, vol. 42, no. 3, pp. 365–373, Sep. 2017.

[7] G. Lee and H. Kim, “Personalized HRTF Modeling Based on Deep Neural Network Using Anthropometric Measurements and Images of the Ear,” *Applied Sciences*, vol. 8, no. 11, p. 2180, Nov. 2018.

[8] K. Yamamoto and T. Igarashi, “Fully perceptual-based 3D spatial sound individualization with an adaptive

variational autoencoder," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–13, Nov. 2017.

[9] 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)*, 2019, pp. 271–275.

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

[11] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[12] G. D. Romigh, D. S. Brungart, R. M. Stern, and B. D. Simpson, "Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 921–930, Aug. 2015.