HRTF Individualization using Deep Learning

Riccardo Miccini
Aalborg University
rmicci18@student.aau.dk

***Abstract*—This is the abstract. There is no abstract yet.**

***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 functions (HRTFs), 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 succesfully approximate the interaural time difference (ITD) and interaural level difference (ILD) cues which are used to percieve 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.

II. State of the art

Over the past decades, several strategies have been devised, in order to avoid the burden of conducting strenous acoustical measurements with human subjects. In a recent review, Guezenoc [2] divides such alternative approaces 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 combining deep learning to some of the aforementioned approaches. In 2017, Yao [6] uses anthropometric measurements to select the most suitable HRTF sets from a larger database. In 2018, Lee [7] feeds anthropometric data into a multilayer perceptron and egde-detected pictures of the ear into a convolutional network, which are then combined in a third networt to estimate HRTF sets. 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.

III. Problem statement

The research focuses on investigating HRTF individualization techniques using data collected from human subjects and deep learning approaches, with an emphasis on HRTF estimation/synthesizing approaches, such as [8].

IV. Milestone plan

With regards to the work of Yamamoto [8], I will set to:

- Fully understand the methods employed
- Implement VAE with relevant novel features
- Replicate training of VAE with HRTF data
- Inspect resulting latent space for coherent mapping of features
- Replicate parameter-tuning mechanism, or devise novel one (e.g. based on anthropometric measurements)
- Validate on secondary dataset (HUTUBS, VIKING HRTF, etc
- Evaluate alternative input representations, such as Real Spherical Harmonic [9]

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] 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.