HRTF Estimation in the Wild

Vivek Jayaram
vjayaram@cs.washington.edu
University of Washington
Seattle, WA, USA

Ira Kemelmacher-Shlizerman
kemelmi@cs.washington.edu
University of Washington
Seattle, WA, USA

Steven M. Seitz
seitz@cs.washington.edu
University of Washington
Seattle, WA, USA

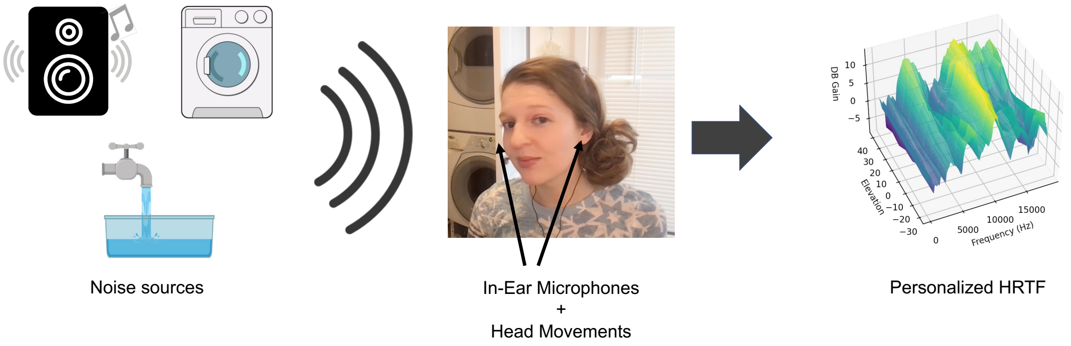


Figure 1: Our method uses binaural recordings of everyday noises along with head tracking information to create a personalized HRTF for the listener.

ABSTRACT

Head Related Transfer Functions (HRTFs) play a crucial role in creating immersive spatial audio experiences. However, HRTFs differ significantly from person to person, and traditional methods for estimating personalized HRTFs are expensive, time-consuming, and require specialized equipment. We imagine a world where your personalized HRTF can be determined by capturing data through earbuds in everyday environments. In this paper, we propose a novel approach for deriving personalized HRTFs that only relies on in-the-wild binaural recordings and head tracking data. By analyzing how sounds change as the user rotates their head through different environments with different noise sources, we can accurately estimate their personalized HRTF. Our results show that our predicted HRTFs closely match ground-truth HRTFs measured in an anechoic chamber. Furthermore, listening studies demonstrate that our personalized HRTFs significantly improve sound localization and reduce front-back confusion in virtual environments. Our approach offers an efficient and accessible method for deriving personalized HRTFs and has the potential to greatly improve spatial audio experiences.

UIST '23, October 29–November 01, 2023, San Francisco, CA, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0132-0/23/10.
https://doi.org/10.1145/3586183.3606782

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **Sound-based input / output**.

KEYWORDS

Spatial Audio, Head-Related Transfer Function, Virtual Reality, sound localization

ACM Reference Format:
Vivek Jayaram, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. 2023. HRTF Estimation in the Wild. In *The 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23), October 29–November 01, 2023, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3586183.3606782

1 INTRO

Spatial audio is an important aspect of many audio applications, including virtual and augmented reality, gaming, music, and audio for film and television. The fundamental challenge of spatial audio is to create the perception that sound is coming from any location in space, even though the sound is played back through headphones. Humans are remarkably good at perceiving the location of incoming sounds in the real world, with as little as 3.5° error even in noisy environments [24]. This ability is achieved through the Head Related Transfer Function (HRTF), which is the direction-dependent filtering of sound by the head, ears, and torso. By using a listener's HRTF to render virtual sounds, it is possible to create an immersive audio experience that simulates sound coming from any position in 3D space. The HRTF is comprised of two components: interaural time differences (ITD) and interaural level differences (ILD). While both components are important for accurate spatial localization, this paper focuses on the frequency dependent ILDs, also called spectral features which describe the different frequencies arriving at each ear. These are more easily obtainable from in-the-wild recordings and have been shown to be more important for HRTF personalization compared to ITDs [45].

A key problem is that HRTFs vary significantly from person to person, and using a personalized HRTF is necessary to create high fidelity spatial audio. This is because using someone else's HRTF or a generic HRTF will lead to localization errors and an unpleasant listening experience [27, 45]. Despite its importance, accurately measuring an individual HRTF is a difficult task. This is due to the fact that HRTF is a complex, dynamic phenomenon that is affected by a variety of factors, including an individual's ear shape, head size, and general anatomy. Traditional methods require the listener to sit in an anechoic chamber while sine-sweeps are played from all possible angles. Other methods involve taking complex 3D scans of the head and ears along with anatomical measurements. This problem of personalized spatial audio has also received increasing attention from companies, such as Apple, Sony, and Logitech, which have recently developed methods to create personalized HRTFs through head scans, imaging, and user feedback [2, 22, 39]. Despite these advances, achieving high-fidelity spatial audio remains an ongoing challenge and an active area of research and development.

We are particularly motivated by the rapid proliferation of earbuds systems, with 100 million AirPods sold in 2020 alone [30]. These systems typically contain a microphone in each ear as well as a head tracking IMU, making them ideal for capturing personalized HRTFs. As more and more people use earbuds, we envision a future where collecting data for personalized HRTFs is as simple as wearing earbuds and moving around in different environments. By analyzing the changes in sound arriving at the listener's ears over time, we can infer their personalized HRTF and use it across a wide range of spatial audio applications. This approach has the potential to be more efficient and less burdensome than traditional methods that require 3D scans or anthropometric measurements.

As a step towards that, in this paper we present a method for measuring individualized HRTFs that leverages environmental sounds recorded by the listener in everyday settings. Our approach is designed for scenarios where there is a single stationary noise source, and we demonstrate its effectiveness using a wide range of noise sources such as music, home appliances, and outdoor sounds. By analyzing the recorded sounds, we can extract features that are specific to the listener's HRTF and use them to construct a personalized HRTF. Our method utilizes machine learning along with synthetic and real training data in order to predict the frequency-dependent filtering of a subject from natural recordings.

Because binaural microphone data and head tracking information is not available through the public APIs of current earbud systems, we built our own physical system to resemble the data available from these earbuds. There have already been some commercial headphones that enable binaural recordings for developers [7, 37] so we expect to see this data becoming more accessible to developers over time.

To validate our approach, we conducted user studies with real listeners and show three key experimental results. First, our predicted HRTFs closely match the ground truth HRTF recorded in an anechoic chamber. Second, our HRTFs significantly improve the sound localization accuracy of users in a virtual auditory display when compared to a generic HRTF. Third, our HRTFs greatly reduce front-back confusion when used to render sounds. Overall, our proposed method of measuring individualized HRTFs in-the-wild has the potential to offer a more efficient and less burdensome alternative to traditional methods, and we hope that it will inspire further research and development in this field.

2 RELATED WORKS

Traditional methods of measuring an individual's HRTF involve dense acoustic measurement in an anechoic chamber [1, 29, 40, 42, 44]. The listener is positioned in the center of the chamber and provided with in-ear microphones, while a series of loudspeakers are arranged in a spherical array to cover all possible azimuth and elevation angles. A reference signal, such as an exponential sine sweep, is played one at a time from each speaker location, and the resulting sound wave that reaches the ear is compared to the reference signal to determine the HRTF. In some cases, the speakers are placed on an arc and rotated around the subject to reduce the number of required speakers. While these approaches are accurate and provide high fidelity HRTFs, they are time-consuming and resource-intensive and require the listener to come to a specialized lab for measurement.

In order to speed up this process, other methods have been proposed that only use a single loud speaker [21, 32, 33]. In these works, a reference signal is played from a stationary loudspeaker while the subject rotates their head through different directions under the measurement of an IMU or other head tracking device. This has the possibility to greatly simplify the HRTF measurement process, and our method builds on this idea of measuring the listener's motion relative to the noise source.

Acoustic simulations on 3D scans of individuals represent another broad category of HRTF estimation methods. For example, the algorithms described in [16, 17, 25, 26, 53] use 3D mesh data with boundary-element methods to simulate the diffraction of sound waves through the head and ear. It has also been shown that HRTFs can be calculated directly from a point cloud of the head, which is slightly easier to obtain than a 3D mesh [41]. Although not published, the method released by Apple [2] uses the depth sensor to create a 3D scan of the head. 3D scan based methods are more accessible than traditional acoustic methods, but still suffer from several drawbacks. For one, they rely on an accurate 3D mesh which can be hard to obtain, either requiring a depth sensor or many images. In addition, 3D scans and imaging raise more privacy concerns for users compared to audio only methods.

It is also possible to estimate the HRTF directly from anthropometric measurements, given the availability of large HRTF datasets with associated head and ear measurements [1, 44]. The works in [54, 55] show positive results when selecting the HRTF with the closest anthropometric measurements to a new user. Other works, [8, 9, 15, 51], use regression methods or deep learning to predict HRTF features from these anthropometric measurements, including works like [28, 51, 52] that use images of the ears along with anthropometric measurements. Anthropometric methods face the same challenges as 3D scan-based methods, as obtaining accurate measurements is difficult and can be time-consuming.

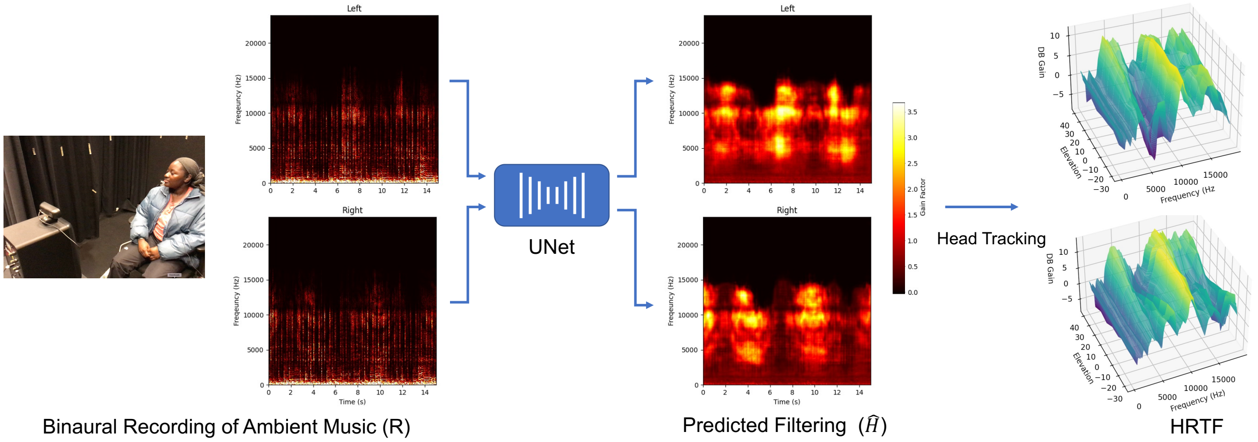


Figure 2: An overview of our method. We use binaural recordings of in-the-wild sounds to predict the filtering from the HRTF at each time step. We then use the head tracking data to map this predicted filtering to the user's location dependent HRTF.

Recently, methods have been proposed to measure HRTFs acoustically in less controlled environments. The method in [10] proposed measuring the HRTF from everyday recordings, but uses a third microphone in the room as a way to record the clean reference signal. Another method [50] allows the user to play sine sweeps from their smartphone, but requires capturing this signal from at least 60 unique locations around the head. Similarly the method in [49] asks the user to play predefined sounds from their smartphone while they move the phone around their head. Finally, the method in [48] allows users to answer pairwise comparison questions in a listening study to determine the best HRTF.

In contrast to these methods, our method has a few key advantages. First, we don't require an anechoic chamber or specialized speakers. Second, we don't require any 3D scans or imaging of the head. Finally, the recordings are collected passively from sound sources in the environment. In our method, the user is in an everyday environment and we capture their natural head movements after an initial calibration step. This is meant to be less cumbersome than existing methods that involve answering questions, moving a smartphone around, or take detailed head measurements and scans.

3 METHOD

Suppose that a listener is wearing earbuds which contain a microphone in each ear as well as a head tracking IMU. The listener may be in the presence of a sound source s, and the microphones at each ear will pick up the binaural recording given by r_l and r_r for left and right respectively. We also use the 3DoF head rotation: $\boldsymbol{\theta}_h(t)$ which describes the head rotation at any given moment in time t This data is available from recent airpod devices [3], or could be captured through the webcam. Our goal is to learn the HRTF solely from r_l, r_r, and $\boldsymbol{\theta}_h(t)$. We use the uppercase notation S, R_l, and R_r to refer to the time-frequency representation of the audio, and the lowercase to refer to the waveform representation. Similarly we use H to denote the filtering function imposed by the listener during a recording, and the time domain version, the head related impulse response (HRIR) is written as h. In this work, we limit the method to scenarios with a single stationary sound source, and its position relative to the head is written as $\theta_s(t)$, and $\varphi_s(t)$.

Under the simplest assumptions, the captured audio is a convolution between the HRIR and the original source. For example, the recorded audio r can be written as

$$r = h * s \tag{1}$$

In the frequency domain, this is

$$R = H \cdot S \tag{2}$$

or equivalently

$$H = R/S \tag{3}$$

Furthermore, the recording may include multi-path signals and other ambient noise, denoted as ϵ, which add ambiguity to the scenario. Breaking this down by left and right separately we get

$$H_l = (R_l - \epsilon_l)/S \tag{4}$$

$$H_r = (R_r - \epsilon_r)/S \tag{5}$$

As we can see, this is a highly underdetermined problem, since we do not have access to the original sound source S or multipath contributions ϵ, only the captured recordings R_l and R_r. Therefore, at any given moment, we would not know whether a given frequency was modified by the listener's HRTF or by the emitting sound source.

Our goal is then to predict H_l and H_r from R_l and R_r without access to the actual ground truth source S. We can then use H_l and H_r along with the head tracking information to create a listener specific HRTF, which is a function of the source location and frequency: $HRTF(\theta_s, \varphi_s, f)$

3.1 Deep Network

To solve this problem, we can use the fact that most sound sources have a repeated or predictable frequency distribution over time which can be learned. Furthermore, the recording from both ears together provide clues towards the relative filtering at each ear. We frame this problem as a supervised learning problem and use a deep network for this prediction task. Deep networks can learn the underlying structure of sounds such as speech and various noise sources, solving one of the ambiguities. Secondly, these networks can use the temporal information of the source along with the data captured at both ears to predict which frequencies are being modified by the HRTF instead of by the sound source or multipaths.

Our network is a modification of the Unet Convolutional Neural Network [35] with an initial convolutional block comprising 32 features, and composed of of 4 downsampling and 4 upsampling convolutional blocks. R_l and R_r are produced using the magnitude of a short-time Fourier transform of the captured audio. They are concatenated channel wise, and feed through the network to produce 2 channels of output of the same dimension. The output represents the predicted level change at a certain frequency due to the HRTF as a scalar factor. We found that learning the filtering function as multiplicative gain was easier than dB due to the fact that cutting out a frequency would require learning a dB gain of $-\infty$. Training details are described in Section 4.

3.2 Source Localization and Head Tracking

Head tracking through an IMU or camera can provide the 3DoF rotation angle of the head. However, because the sound source may not be located directly in front of the user, it is also necessary to know the location of the signal relative to the user. In our system, we require an initial localization input from the user. They are asked to point their head directly towards the sound source (or directly away for sounds coming from behind). They then press a button which allows the system to record the initial location of the sound source, $\theta_s(0)$ and $\varphi_s(0)$. During the rest of the recording process, the rotation matrix of the head orientation can be applied to the initial source location to give the relative position of the sound source at that time, $\theta_s(t)$ and $\varphi_s(t)$.

It may also be possible to infer the initial source location using localization algorithms, but we leave that as future work as the manual localization by the user is quick and very accurate.

3.3 HRTF Estimation from Aggregated Results

By aggregating predictions across many recordings with different sound sources and head rotations, we can obtain a more accurate and full representation of the listener's HRTF. Let F be the number of frequency bins in the spectrogram representation. For each binaural recording $R \in \mathbb{R}^{2\times T\times F}$, we use a deep network to predict the filtering function $\hat{H} \in \mathbb{R}^{2\times T\times F} = \text{UNet}(R)$.

To build a model of the listener's HRTF, we first initialize an empty HRTF for all source locations and frequencies. Then, we use the UNet to predict how the listener's HRTF filtered the sound source for each recording. If the entirety of a recording contains minimal energy at a given frequency, we assume that this frequency was absent from the source signal and do not use it. Finally, we use the known relative location of the source over time to create a HRTF prediction for each location-frequency bin. Our method does not explicitly solve for directions with no data, but in such scenarios, we could use HRTF extrapolation/interpolation methods which have shown good results when we only have a sparse HRTF [6, 18]

Algorithm 1 Create HRTF, HRIRs from Binaural Recordings

```
for $\forall\theta, \forall\varphi, \forall f$ do
    $\widehat{HRTF}(\theta, \phi, f) \leftarrow []$                ▷ Initialize empty HRTF
end for
for $R \in$ Recordings do
    $\hat{H} \leftarrow \text{UNet}(R)$                            ▷ Network inference
    for $t \in 0..T, f \in 0..F$ do
        if $R(f)$.mean() $> \epsilon$ then
            $\widehat{HRTF}(\theta_s(t), \phi_s(t), f)$.append($\hat{H}(t, f)$)
        end if
    end for
end for
for $\forall\theta, \forall\varphi, \forall f$ do                ▷ Use phase from generic HRTF
    $|\widehat{HRTF}(\theta, \phi, f)| \leftarrow \widehat{HRTF}(\theta, \phi, f)$.mean()
    $\angle\widehat{HRTF}(\theta, \phi, f) \leftarrow \angle HRTF_{\texttt{generic}}(\theta, \phi, f)$
end for
$HRIR(\theta, \varphi) = $ iFFT($HRTF(\theta, \varphi)$)
```

Across time steps, the predicted filtering function for the same location-frequency bin may vary due to the changes in the underlying sound source, reverb, or other effects not modeled such as doppler effects. Because of this, we average the predicted HRTF values at each location-frequency bin to obtain the listener's HRTF magnitude at each location and frequency. One of the advantages of our method is that over time, we can collect more and more information about the HRTF and use that to produce a better estimate. We explored both the mean and median and found that the mean worked better.

To obtain the head-related impulse response (HRIR) for use in spatial audio applications, we also need the phase information which describes the interaural time differences (ITDs). We use ITDs from a generic HRTF and apply inverse fast Fourier transform (IFFT) to obtain the HRIR. Although some previous works in similar domains [34, 43] predict the phase as well as the magnitude of the impulse response, we found that phase was much harder to predict in a reverberant environment due to multipath effects. At many frequencies, the captured phase was completely different from the actual ITD phase due to multipath interference. Our user studies also showed that generic ITDs still produced a strong ability to localize sounds. The full algorithm is described in Algorithm 1.

3.4 System Implementation

Our method is general and designed to work with any device that supports binaural recordings and head tracking. This could include earbuds, VR headsets, or smart glasses. However, with the exception of certain headsets paired with certain phones [4, 37], these devices do not currently expose the required functionality to third-party developers. We therefore built our own physical system with commercially available hardware.

For the binaural recordings, we used the Sound Professionals SP-TFB-2 in-ear Binaural Microphones [31]. These wired headphones are capable of capturing frequencies up to 20kHz. It's noteworthy that our microphones, unlike those used in numerous previous studies such as [1, 41, 42], are positioned at the entrance of the ear canal rather than fully blocking it. Our research demonstrates that it is feasible to generate an accurate HRTF even without perfect microphone placement. Extending this methodology to commercial earbuds would require re-training with data captured using those specific headphones to learn their unique transfer function.

For head tracking, we used the face pose detector provided by the Google MediaPipe Library [13]. This algorithm uses a forward facing webcam to detect the 3DoF head position, and is based on BlazeFace [5] and AttentionMesh [14]. The head tracking runs in less than 10ms on a Macbook pro, and we use a HRTF with bin size $\theta = 5^\circ$ and $\varphi = 5^\circ$. This means that as long as the user is not rotating their head faster than $\sim 300^\circ/s$, the head tracking will assign the sound to the correct HRTF bin.

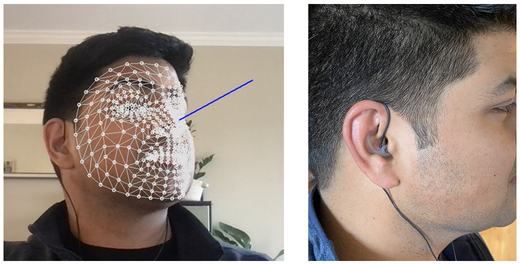


Figure 3: Left: Our head tracking implementation uses the webcam to determine the 3DoF head rotation during recording. The normal vector is drawn in blue to help visualize the direction the head is pointing. Right: An image of the binaural microphone used in our implementation. The microphone sits near the ear canal.

4 DATA AND TRAINING

To train our network, we adopt an approach that combines synthetic and real data. We begin with large amounts of synthetically rendered data, which enables us to learn from a wide range of noise types and simulated environments, including multi-path scenarios. However, such data does not capture all nuances of real-world audio and fails to generalize completely to actual recordings. To address this, we incorporate real data, which is more time-consuming to collect but provides more effective training for the network. By leveraging both sources of data, we are able to benefit from the strengths of each approach. This mix of synthetic and real training data has been explored in previous works as well [7, 19, 38]. Below, we describe the two data sources in more detail.

4.1 Synthetic Training Data

We first train the network on synthetically rendered spatial data. For HRTFs, we use the RIEC dataset [23], which contains 109 HRTFs measured in an anechoic chamber for different individuals. This was split into a training set of 64 and a test set of 45 HRTFs. To render sounds, we use the Steam Audio C++ API which allows realistic sound rendering for moving sources with custom HRTFs and multi-path environments.

The noise sources come from the WHAM! dataset [46] and AudioSet dataset [12]. These datasets contain a wide variety of noise sources such as music, speech, appliances, and machinery. Sound sources without sufficient frequency ranges (requiring a minimal energy up to at least 5khz) and without sufficient regularity (e.g. impact only sounds) were filtered out. Some example of sound categories that were removed included chewing, clapping, snapping, and whistling. Some of the most effective noise sources included water, kitchen appliances, pop music, and machinery. For both datasets, a 80/20 train/test split was maintained. None of the audio samples used for training the network were used during any of the synthetic or real evaluations

Each generated recording was 3s long and created by placing the sound source at a random azimuth and random elevation 1.5m away from the listener while the listener moved their head in a random direction at a random speed. Multipaths were simulated by create walls at distances between 2 and 10 meters from the listener with RT60 values from 0.4 to 0.9 seconds. For each recording we also obtained the ground truth filtering at the source locations to use as a training label, $\tilde{H}$. The train set contains 10,000 generated examples, and the test set contains 1,000 examples.

4.2 Real Training Data and Anechoic HRTFs

Although large amounts of synthetic data can be easily collected, a network solely trained on synthetic data does not perform well in real-world scenarios. To address this issue, we augment the training data with in-the-wild recordings that more closely resemble the acoustic environments and sounds that would be encountered by users during inference. The main challenge is that the ground-truth filtering function is required as a training supervised label for the model. In order to generate these labels, we measure the ground-truth HRTFs of the subjects in an anechoic chamber. These HRTFs can also be used as a baseline to evaluate the in-the-wild inferred HRTFs.

Our anechoic HRTF measurement procedure is most similar to the method described in [33]. Subjects are seated in an anechoic chamber while a single loudspeaker emits a reference signal. They are instructed to move their head slowly to cover a broad range of azimuth and elevation angles. Unlike [33], we place the speaker at 3 different elevation angles when capturing the ground-truth HRTF. This better captures the filtering effects of the torso at different sound elevations which are not captured by simply rotating the head up and down with respect to a single speaker location. Furthermore, we use a broadband Gaussian noise signal instead of a sine sweep, as we only care about the frequency-dependent level differences and not the ITDs. This allows us to capture the filtering across all frequencies at each time step. The speaker used is the KRK Classic 5 Studio Monitor which contains 2 drivers. To account for an imperfect speaker response, a reference signal $\tilde{S}$ is first recorded. The ground truth filtering function is then obtained by dividing the recording R by $\tilde{S}$. This also has the effect of cancelling out any frequency response imposed by the microphones as both R and $\tilde{S}$ contain the same microphone response. A full discussion of speaker and microphone compensation is provided in [20].

After collecting the anechoic HRTFs, we generated real training data for the neural network by having 2 subjects listen to 1 hour of noise sources, from the training partition of our audio datasets, played back through the loudspeaker in regular environments. The

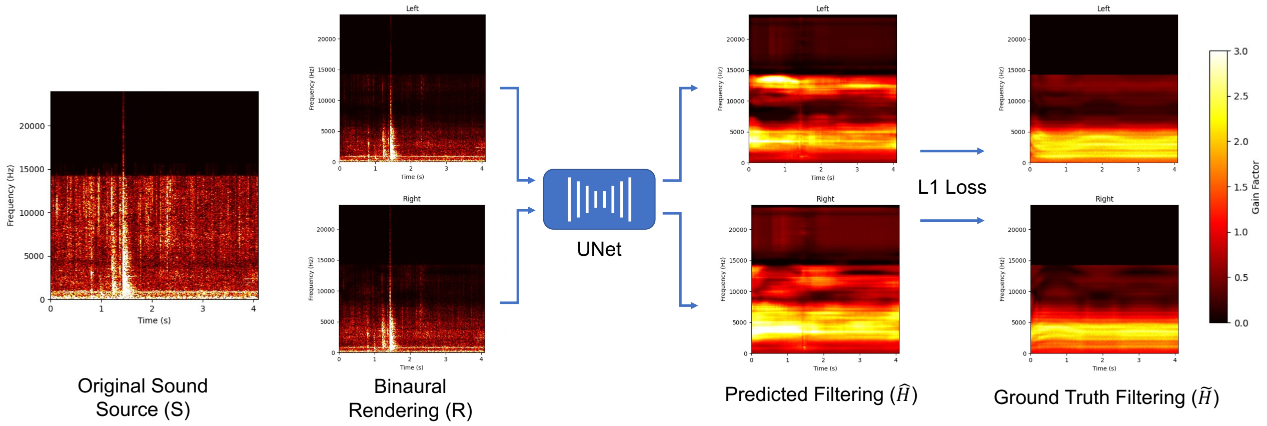

Left
Right
Frequency (Hz)
Time (s)
UNet
L1 Loss
Gain Factor
Original Sound Source (S)
Binaural Rendering (R)
Predicted Filtering ($\hat{H}$)
Ground Truth Filtering ($\tilde{H}$)

Figure 4: The process for training the network. We use create binaural renderings of a sound source with simulated multi-path environments. We then use the ground truth filtering of the HRTF to train the network with an L1 loss between the predicted filtering, $\hat{H}$ an the ground truth filtering $\tilde{H}$. The real training data is used in an identical way except R is a binaural recording, not a binaural rendering, and we don't have access to the original sound source S.

speaker location was known, and the training label $\tilde{H}$ could be generated from the anechoic HRTFs.

4.3 Training Details

All recordings were captured at 48kHz sample rate. Each training example contained 3s of binaural audio, and mini-batch size 32 was used. The STFT was conducted with a window size of 2048. Training occured on a Nvidia Titan Xp GPU and took approximately 10 hours for 100 epochs of training. Data augmentation techniques included random left-right flip, random volume changes, and the addition of random noise. Samples from the real and synthetic dataset were randomly sampled with equal probability

5 RESULTS

We evaluate the effectiveness of our method through a user study, and present three key results to show the strength of the method. First, we show that our predicted HRTFs closely match the ground-truth HRTFs. Second, we demonstrate that our HRTFs improve localization by listeners in a virtual environment. Finally, we show that our HRTFs significantly reduce front-back confusion with rendered sounds.

5.1 User Study and In-the-Wild HRTF

8 individuals with regular hearing abilities (4 male, 4 female, mean age 28) participated in the user study. First, we measured their ground-truth HRTF in an anechoic chamber as described in Section 4.2. Next, we used our in-the-wild method to measure their HRTF in a regular environment. The subjects were in a normal sized reverberant room, that was not particularly quiet. The background noise in the room was measured to be around 50dB due to electric hum and other noises. Next, a variety of noise types were played from a loud speaker in the room. This included music, running water, kitchen appliances, and other sounds from the test partition of the WHAM! and AudioSet datasets. The speaker was placed at 3 elevations and a variety of azimuth angles relative to the listener at distances that varied from $1 - 3\mathrm{m}$. The listener was instructed to rotate their head through a normal range of angles as they listened to the audio sounds. In total, roughly 15 minutes of audio were captured per user across all the locations.

Method	LSD (dB)
Random RIEC Subject	8.23
Generic HRTF	7.32
Zandi et. al [50]	4.5
Ours	**4.38**
Hu et. al [15]	3.5

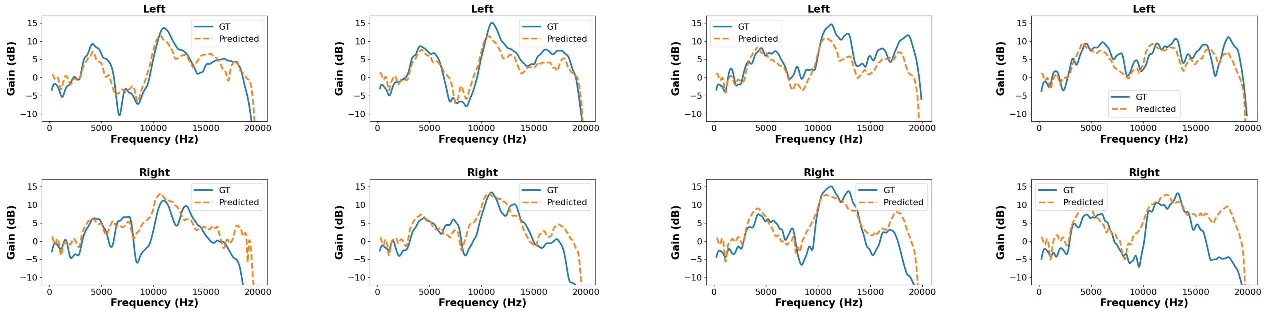
Table 1: Log-spectral distortion between ground-truth HRTF and the output HRTF for several methods. We note that the method in [15] requires additional physical measurements and the method in [50] requires significantly more active input from the user.

5.2 Comparison with Ground-Truth HRTF

The first metric we use to evaluate the correctness of our HRTFs is the agreement with the ground-truth HRTF. A visual comparison between the two is shown in Figure 5 which plots the results for a given subject at four consecutive elevations and $\theta = 0°$. To evaluate the similarity quantitatively, at every azimuth and elevation, we compute the Log-Spectral Distortion (LSD) in dB which is given by

$$LSD(\hat{H}, \tilde{H}) = \sqrt{\frac{1}{F}\sum_{f=1}^{F}\left(20\log_{10}\frac{\tilde{H}(k)}{\hat{H}(k}\right)^{2}} \tag{6}$$

We then report the median value across all azimuth and elevations in table 1. Our method is compared with several other methods as well. For the random method, we average the LSD when comparing the ground-truth HRTF with all other HRTFs in the RIEC



$\varphi = 0°$ $\varphi = 10°$ $\varphi = 20°$ $\varphi = 30°$

Figure 5: We plot the ground-truth HRTF and predicted HRTF for a given test subject for $\theta = 0°$ and 4 elevations. The HRTF that we create for the user closely matches the ground truth, even though the magnitude of some notches and peaks may not be exactly correct.

database. For a generic HRTF, we used the KEMAR HRTF [11] which contains measurements for a dummy head commonly used as a generic HRTF model. We also share the results reported in [15] and [50]. It's important to note that the method in [50] achieves it's reported results when the HRTF is measured at 1138 unique locations done actively by the user, and the method in [15] requires detailed anthropometric measurements of the ear, head, and torso.

5.3 Localization in a Virtual Auditory Display

To evaluate the effectiveness of our HRTFs in spatial audio applications, we created a virtual auditory display where sounds could be rendered spatially with dynamic head tracking. Our method aimed to replicate the experimental method described in [6]. A sound reproduction system was implemented in Unity and Steam Audio where a virtual sound was placed at an arbitrary location and played back to the listener through headphones. The listener could then move their head, with the sound location adjusting accordingly based on head tracking information sent to Unity via a UDP connection. Overall the system's latency from the head movement to the sound update was less than 30ms which is below the perceptual lag for binaural listening [36, 47]. Listeners were placed in a room with a grid of angular markings on 3 sides of them at 10° intervals for both azimuth and elevation. A white noise stimulus was played for a maximum of 5 seconds during which the listeners could make exploratory head movements within a maximum of 30° of the forward facing angle. They were then asked to indicate their perceived source location by pointing at the best grid location. Experiments were conducted for sources in the front hemisphere and back hemisphere separately with sources coming from random locations in $\theta \in [-70°, 70°]$ and $\varphi \in [-30°, 40°]$. A brief calibration period was used where the listener could see the ground truth location for the first 4 examples while making the exploratory head movements. Each subject then evaluated 20 random locations for each candidate HRTF.

Results are reported in Figure 6. For both total angular error and elevation error, listeners performed significantly better with our method ($p < 0.01$) compared to a generic HRTF. In addition, the localization error with our method was close to that of the anechoic ground-truth HRTF. We note that, although the mean azimuth error was better with our method and the ground-truth HRTF compared to a generic HRTF, it was not statistically significant ($p > 0.05$). We hypothesize that this is because ITDs are the primary method used by humans for azimuth inference, and both the ground-truth HRTF and our method contained generic ITDs with only personalized spectral features. Statistical significance was computed with an independent-samples t test between the two candidate distributions.

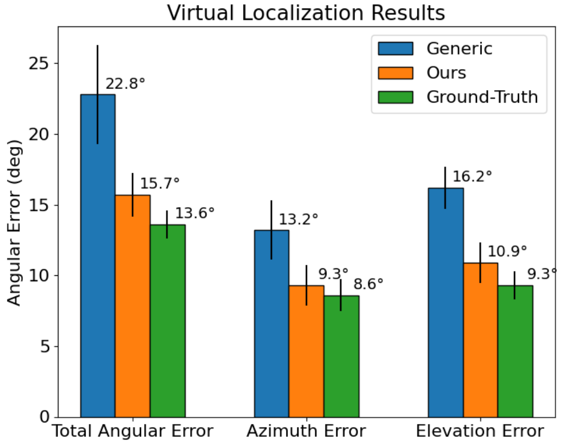


Figure 6: Localization results for the virtual auditory display experiment. Results are reported for 3 different experiments: a generic HRTF, the HRTF predicted using our method, and the ground-truth anechoic HRTF described in Section 4.2. For each experiment, we first show the total angle difference between the source and prediction. We then show the prediction error broken down by azimuth and elevation error. Results are averaged over all subjects and trials. Error bars shown are the first standard error of the mean.

Method	Front-back confusion rate
Generic	29.0% ± 5.4
Ours	**14.8% ± 4.6**
GT HRTF	9.6% ± 4.2

Table 2: Front-back confusion with rendered sounds. We report the percent of times the listeners made an error, along with the first standard deviation

5.4 Front Back Confusion

The last experiment conducted was a front-back confusion test using rendered sounds. A short white noise stimulus was rendered at a random location using a candidate HRTF and played back to the listener through headphones. The listener then had to predict whether the source was coming from the front or back hemisphere. The locations used were $\theta \in [-70°, 70°]$ in the front and back, and $\varphi \in [-30°, 40°]$. Like the previous experiment, the listener received the ground truth answer for the first 4 locations. However, unlike the previous experiment, the listener was not allowed to make exploratory head movements and had to predict front or back based on the rendering alone. Each subject then evaluated 30 random locations per HRTF before moving on to the next HRTF. The results are shown in table 2 which once again show a significant improvement ($p < 0.01$) when using our method compared to a generic HRTF.

6 LIMITATIONS AND CONCLUSION

Our method shows a strong ability to solve for a listener's HRTF using only binaural recordings of in-the-wild sounds and relative head tracking information. However, there are several limitations that need to be acknowledged.

First, our method was only demonstrated with a single stationary noise source. Such scenarios are limited in everyday settings, and solving for the HRTF with multiple sources or moving sources would present additional challenges. It would be necessary to localize moving sources and separate the contributions to the recording from multiple sources. Second, the user still has to actively localize the sources at the beginning of each recording, which presents an additional burden compared to a fully passive HRTF estimation method. This could be resolved by using a binaural localization method, and erroneous localizations could be compensated through outlier detection methods. Finally, the microphones in commercial earbuds are often not exactly at the ear canal entrance. The effect of the earbud on the HRTF would need to be taken into account through careful measurements of the earbud system. Despite these limitations, our method for HRTF estimation has immense potential as wireless earbuds proliferate among everyday users. We show strong performance on a variety of real-world user studies, and we hope that our method can be incorporated into commercial earbud systems in the near future.

REFERENCES

[1] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. 2001. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*. IEEE, 99–102.

[2] Apple. [n. d.]. *Listen with Personalized Spatial Audio for AirPods and Beats*. https://support.apple.com/en-us/HT213318 Accessed on: June 1, 2023.

[3] Apple. 2023. *AirPods (3rd generation)*. https://www.apple.com/airpods-3rd-generation/specs/ Accessed on: June 1, 2023.

[4] Apple. 2023. *CMHeadphoneMotionManager*. https://developer.apple.com/documentation/coremotion/cmheadphonemotionmanager Accessed on: June 1, 2023.

[5] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. arXiv:1907.05047 [cs.CV]

[6] zamir ben hur, david alon, philip w. robinson, and ravish mehra. 2020. localization of virtual sounds in dynamic listening using sparse hrtfs. *journal of the audio engineering society* (august 2020).

[7] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M. Seitz. 2022. ClearBuds. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. ACM. https://doi.org/10.1145/3498361.3538933

[8] Tzu-Yu Chen, Tzu-Hsuan Kuo, and Tai-Shih Chi. 2019. Autoencoding HRTFs for DNN based HRTF personalization using anthropometric features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 271–275.

[9] Chan Jun Chun, Jung Min Moon, Geon Woo Lee, Nam Kyun Kim, and Hong Kook Kim. 2017. Deep neural network based HRTF personalization using anthropometric measurements. In *Audio Engineering Society Convention 143*. Audio Engineering Society.

[10] Klaus Diepold, Marko Durkovic, and Florian Sagstetter. 2010. HRTF Measurements with Recorded Reference Signal. In *Audio Engineering Society Convention 129*. Audio Engineering Society.

[11] Bill Gardner, Keith Martin, et al. 1994. HRFT Measurements of a KEMAR Dummy-head Microphone. (1994).

[12] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

[13] Google. 2023. Mediapipe. https://github.com/google/mediapipe Accessed on: June 1, 2023.

[14] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. Attention Mesh: High-fidelity Face Mesh Prediction in Real-time. arXiv:2006.10962 [cs.CV]

[15] Hongmei Hu, Lin Zhou, Jie Zhang, Hao Ma, and Zhenyang Wu. 2006. Head related transfer function personalization based on multiple regression analysis. In *2006 International conference on computational intelligence and security*, Vol. 2. IEEE, 1829–1832.

[16] Tomi Huttunen, Eira T Seppälä, Ole Kirkeby, Asta Kärkkäinen, and Leo Kärkkäinen. 2007. Simulation of the transfer function for a head-and-torso model over the entire audible frequency range. *Journal of Computational Acoustics* 15, 04 (2007), 429–448.

[17] Tomi Huttunen, Antti Vanne, Stine Harder, Rasmus Reinhold Paulsen, Sam King, Lee Perry-Smith, and Leo Kärkkäinen. 2014. Rapid generation of personalized HRTFs. In *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society.

[18] Yuki Ito, Tomohiko Nakamura, Shoichi Koyama, and Hiroshi Saruwatari. 2022. Head-Related Transfer Function Interpolation from Spatially Sparse Measurements Using Autoencoder with Source Position Conditioning. arXiv:2207.10967 [cs.SD]

[19] Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. The cone of silence: Speech separation by localization. *Advances in Neural Information Processing Systems* 33 (2020), 20925–20938.

[20] Erno Langendijk and Adelbert Bronkhorst. 2000. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America* 107 (02 2000), 528–37. https://doi.org/10.1121/1.428321

[21] Song Li and Jürgen Peissig. 2017. Fast estimation of 2D individual HRTFs with arbitrary head movements. In *2017 22nd International Conference on Digital Signal Processing (DSP)*. 1–5. https://doi.org/10.1109/ICDSP.2017.8096086

[22] Logitech. 2023. *Personalized Spatial Audio with Head Tracking*. https://embody.co/pages/gaming-logitech Accessed on: June 1, 2023.

[23] Piotr Majdak, Yukio Iwaya, Thibaut Carpentier, Rozenn Nicol, Matthieu Parmentier, Agnieszka Roginska, Yôiti Suzuki, Kankji Watanabe, Hagen Wierstorf, Harald Ziegelwanger, et al. 2013. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society.

[24] James C Makous and John C Middlebrooks. 1990. Two-dimensional sound localization by human listeners. *The journal of the Acoustical Society of America* 87, 5 (1990), 2188–2200.

[25] A. Meshram, Ravish Mehra, and Dinesh Manocha. 2014. Efficient HRTF computation using adaptive rectangular decomposition. *Proceedings of the AES*

International Conference 2014 (01 2014).

[26] Alok Meshram, Ravish Mehra, Hongsheng Yang, Enrique Dunn, Jan-Michael Franm, and Dinesh Manocha. 2014. P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 53–61. https://doi.org/10.1109/ISMAR.2014.6948409

[27] John C Middlebrooks. 1999. Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1480–1492.

[28] A. Mohan, R. Duraiswami, D.N. Zotkin, D. DeMenthon, and L.S. Davis. 2003. Using computer vision to generate customized spatial audio. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, Vol. 3. III–57. https://doi.org/10.1109/ICME.2003.1221247

[29] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. 1995. Head-related transfer functions of human subjects. *Journal of the Audio Engineering Society* 43, 5 (1995), 300–321.

[30] Mike Peterson. 2021. Apple AirPods, Beats dominated audio wearable market in 2020. https://appleinsider.com/articles/21/03/30/apple-airpods-beats-dominated-audio-wearable-market-in-2020 Accessed on: June 1, 2023.

[31] Sound Professionals. 2022. SP-TFB-2 – Low noise in-ear Binaural microphones. https://soundprofessionals.com/product/SP-TFB-2/ Accessed on: June 1, 2023.

[32] Jonas Reijniers, Bart Partoens, and Herbert Peremans. 2017. DIY Measurement of your Personal Hrtf at Home: Low-Cost, Fast and Validated. *journal of the audio engineering society* (october 2017).

[33] Jonas Reijniers, Bart Partoens, Jan Steckel, and Herbert Peremans. 2020. HRTF Measurement by Means of Unsupervised Head Movements With Respect to a Single Fixed Speaker. *IEEE Access* 8 (2020), 92287–92300. https://doi.org/10.1109/ACCESS.2020.2994932

[34] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. 2022. Deep Impulse Responses: Estimating and Parameterizing Filters with Deep Networks. arXiv:2202.03416 [cs.SD]

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[36] Narayan Sankaran, James Hillis, Marina Zannoli, and Ravish Mehra. 2016. Perceptual thresholds of spatial audio update latency in virtual auditory and audiovisual environments. *The Journal of the Acoustical Society of America* 140, 4 (2016), 3008–3008.

[37] Ben Schoon. 2023. *Samsung Galaxy Buds 2 Pro can now record 360-degree binaural audio for videos from your phone.* https://9to5google.com/2023/01/12/samsung-buds-binaural-audio-recording Accessed on: June 1, 2023.

[38] Viktor Seib, Benjamin Lange, and Stefan Wirtz. 2020. Mixing Real and Synthetic Data to Enhance Neural Network Training – A Review of Current Approaches. arXiv:2007.08781 [cs.CV]

[39] Sony. 2023. *360 Reality Audio.* https://electronics.sony.com/360-reality-audio Accessed on: June 1, 2023.

[40] University of Southampton. [n. d.]. HRTF Mesaurement System. https://resource.isvr.soton.ac.uk/FDAG/VAP/html/facilities.html

[41] Rahulram Sridhar and Edgar Y Choueiri. 2017. A method for efficiently calculating head-related transfer functions directly from head scan point clouds. In *143rd Audio Engineering Society Convention 2017*.

[42] Rahulram Sridhar, Joseph G Tylka, and Edgar Choueiri. 2017. A database of head-related transfer functions and morphological measurements. In *Audio Engineering Society Convention 143*. Audio Engineering Society.

[43] Christian J. Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. 2021. Filtered Noise Shaping for Time Domain Room Impulse Response Estimation from Reverberant Speech. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 221–225. https://doi.org/10.1109/WASPAA52581.2021.9632680

[44] Kanji Watanabe, Yukio Iwaya, Yôiti Suzuki, Shouichi Takane, and Sojun Sato. 2014. Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoustical science and technology* 35, 3 (2014), 159–165.

[45] Elizabeth M Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. 1993. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123.

[46] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. WHAM!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160* (2019).

[47] Satoshi Yairi, Yukio Iwaya, and Yôiti Suzuki. 2008. Influence of large system latency of virtual auditory display on behavior of head movement in sound localization task. *Acta Acustica united with Acustica* 94, 6 (2008), 1016–1023.

[48] Kazuhiko Yamamoto and Takeo Igarashi. 2017. Fully perceptual-based 3D spatial sound individualization with an adaptive variational autoencoder. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–13.

[49] Zhijian Yang and Romit Roy Choudhury. 2021. Personalizing Head Related Transfer Functions for Earables. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (Virtual Event, USA) *(SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 137–150. https://doi.org/10.1145/3452296.3472907

[50] Navid H Zandi, Awny M El-Mohandes, and Rong Zheng. 2022. Individualizing Head-Related Transfer Functions for Binaural Acoustic Applications. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. IEEE, 105–117.

[51] Manlin Zhao, Zhichao Sheng, and Yong Fang. 2022. Magnitude modeling of personalized HRTF based on ear images and anthropometric measurements. *Applied Sciences* 12, 16 (2022), 8155.

[52] Bowen Zhi, Dmitry N. Zotkin, and Ramani Duraiswami. 2022. Towards Fast And Convenient End-To-End HRTF Personalization. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 441–445. https://doi.org/10.1109/ICASSP43922.2022.9746315

[53] Harald Ziegelwanger, Wolfgang Kreuzer, and Piotr Majdak. 2015. Mesh2hrtf: Open-source software package for the numerical calculation of head-related transfer functions. In *22nd International Congress on Sound and Vibration*.

[54] D.Y.N. Zotkin, J. Hwang, R. Duraiswaini, and L.S. Davis. 2003. HRTF personalization using anthropometric measurements. In *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*. 157–160. https://doi.org/10.1109/ASPAA.2003.1285855

[55] Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis. 2002. Customizable auditory displays. Georgia Institute of Technology.