Personalizing Head Related Transfer Functions for Earables

Zhijian Yang
University of Illinois at Urbana Champaign

Romit Roy Choudhury
University of Illinois at Urbana Champaign

ABSTRACT

Head related transfer functions (HRTF) describe how sound signals bounce, scatter, and diffract when they arrive at the head, and travel towards the ear canals. HRTFs produce distinct sound patterns that ultimately help the brain infer the spatial properties of the sound, such as its direction of arrival, θ. If an earphone can learn the HRTF, it could apply the HRTF to any sound and make that sound appear directional to the user. For instance, a directional voice guide could help a tourist navigate a new city.

While past works have estimated human HRTFs, an important gap lies in personalization. Today's HRTFs are global templates that are used in all products; since human HRTFs are unique, a global HRTF only offers a coarse-grained experience. This paper shows that by moving a smartphone around the head, combined with mobile acoustic communications between the phone and the earbuds, it is possible to estimate a user's personal HRTF. Our personalization system, *UNIQ*, combines techniques from channel estimation, motion tracking, and signal processing, with a focus on modeling signal diffraction on the curvature of the face. The results are promising and could open new doors into the rapidly growing space of immersive AR/VR, earables, smart hearing aids, etc.

CCS CONCEPTS

• **Computer systems organization** → **Embedded and cyber-physical systems**; • **Human-centered computing** → **Ubiquitous and mobile computing**; **Interaction techniques**.

KEYWORDS

Head Related Transfer Function (HRTF), Spatial Audio, Virtual Acoustics, HRTF Personalization, Earables, AR, VR

ACM Reference Format:
Zhijian Yang and Romit Roy Choudhury. 2021. Personalizing Head Related Transfer Functions for Earables. In *ACM SIGCOMM 2021 Conference (SIGCOMM '21), August 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3452296.3472907

1 INTRODUCTION

Humans can inherently sense the direction θ from which a sound arrives at their ears. The human brain essentially analyzes the time difference of the sounds across the two ears and maps this difference Δt to θ. If the mapping was *one-to-one*, then the estimation of θ would be easy. Unfortunately, the mapping is *one-to-many*, meaning that for a given Δt, there are many possible θs. Figure 1(a) shows an example where all points on the (red) hyperbola produce identical Δt at the ears. How can humans still disambiguate the direction θ? The answer lies in what is classically known as the *head related transfer function* (HRTF), explained next.

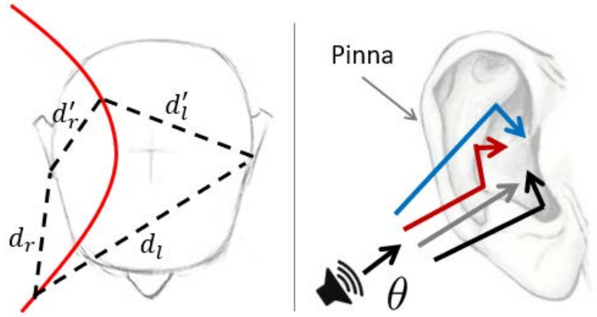


Figure 1: Humans identify sound direction through (a) time difference of arrival, and (b) pinna multipath.

Briefly, the sounds that actually enter the ear-canal is influenced by the shape of the human head and the pinna of the ear (shown in Figure 1(b)). The pinna produces micro-echoes to the arriving signal, while the 3D curvature of the head bends (or diffracts) the signals [16, 22, 29]. The net result is that the eardrum receives a sophisticated signal pattern that helps the brain disambiguate θ. In summary, one can view the head (including the pinna) as a *filter* that alters the signal depending on its angle of arrival θ. In frequency domain, this filter is called *head related transfer function* (HRTF).

Knowing HRTF for each θ opens new possibilities in spatial acoustics. An earphone could take any normal sound $s(t)$, apply the (left and right) HRTFs for a desired θ, and play the two sounds in the corresponding earbuds [25, 59]. The brain would perceive this sound as directional, as if it is arriving from an angle θ with respect to the head. Applications could be many, ranging from immersive AR/VR, to gaming, to assisted technology for blind individuals [52].

For instance, (1) users may no longer need to look at maps to navigate from point A to point B; a voice could say "follow me" in the ears, and walking towards the perceived direction of the voice could bring the user to her destination. Blind people may particularly benefit from such a capability. (2) A virtual-reality meeting could be held through immersive acoustic experiences. Members could pick their seats in a virtual meeting room and each member could hear the others from the direction of their relative configuration. (3) Gaming and other 3D applications would naturally benefit. Each musical instrument in an AR/VR orchestra could be fixed to a specific location around the head. Even if the

SIGCOMM '21, August 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8383-7/21/08…$15.00
https://doi.org/10.1145/3452296.3472907

head rotates, motion sensors in the earphones can sense the rotation and apply the HRTF for the updated θ. Thus, the piano and the violin can remain fixed in their absolute directions, offering an immersive user experience.

HRTF-guided spatial sounds are already available in products today [3, 6–8], however, important challenges remain open. One key challenge is in HRTF personalization [29, 58]. Today's products use a global HRTF template, i.e., the HRTF is carefully measured for one (or few people) in the lab and this "average" template is then incorporated across all products. Unsurprisingly, the spatial acoustic experience is known to be sub-optimal [27] and varies widely across individuals [5, 10]. The natural question is: *why not estimate personalized HRTFs for each user?*

To answer this, let us briefly understand today's method of estimating HRTF [22, 55]. A user, Bob, is brought to an acoustic echo-free chamber, seated at a special immovable chair, and fitted with a normal earphone. A high quality speaker then plays carefully designed sounds (e.g., a frequency sweep) from all possible angles θ and distances r around Bob's head. The ground truth for θ and r are accurately measured from ceiling cameras installed in the chamber. Finally, the recordings from the left and right ears are converted to the HRTFs for the corresponding $\langle \theta, r \rangle$ tuple. Estimating personalized HRTF at home would entail hundreds of *accurate* $\langle \theta, r \rangle$ measurements, while maintaining the exact head position. This is impractical even for the technology savvy individual.

This paper aims to estimate a user's personal HRTF at home by leveraging smartphones, arm gestures, and acoustic signal processing. The high level idea of our system, *UNIQ*, is simple. We ask a user to sit on a chair, wear her earphones, and then move her smartphone in front of her face (as much as their arms would allow). The smartphone plays pre-designed sounds that the earphones record; the smartphone also logs its own IMU measurements during the arm-motion. *UNIQ*'s algorithmic goal is to accept these 3 inputs — the earphone recordings, the IMU recordings, and the played sounds — and output the user's personal HRTF, $H_{(\theta,r)}$.

In estimating the personal HRTF, we face 2 key challenges: (1) The phone's location needs to be tracked with high accuracy as the phone is moving around the head. The IMU is inadequate for such fine-grained tracking, hence the acoustic communication between the smartphone and the earphone needs to aid the tracking algorithm. Unfortunately, since the acoustic signal propagation between the phone and earphone undergoes head-related diffraction and pinna-multipath, standard geometric models do not apply. This leads to a joint optimization problem, i.e., to solve for the phone's location, HRTF needs to be solved, and the vice versa.

(2) The above module solves the *near-field* HRTF. [1] However, the near-field HRTF is not ideal when the emulated sound source needs to be far away. Briefly, far-field sounds are almost parallel rays when they arrive at the two ears, which is not the case for the near-field. Since the HRTF varies as a function of the signal's incoming directions, the difference between near and far-field matters. Thus, the second challenge is to "synthesize" or "extrapolate" the far field HRTF based on the sequence of measurements from the near field.

UNIQ addresses these two main challenges by first modeling the 3D head-geometry using 3 parameters, applying diffraction on the parametric model, and deriving the expected signal equations at the ear. This expectation can now be compared against the acoustic measurements from the phone, along with the IMU readings that (partly) track the phone's motion. Together, *UNIQ* formulates a minimization problem, extracting the head parameters and the phone locations that best fit the model. With some additional refinements (such as discrete-to-continuous interpolation [40]), the near-field HRTF is ready. *UNIQ* then selects suitable components from the near-field HRTF to synthesize a physics-based model of far-field signals. This model is fine-tuned with the estimated head parameters to ultimately yield the far-field HRTF.

Finally, *UNIQ* shows an application of the far-field HRTF in estimating the angle of arrival (AoA) of ambient signals. This means when Alice is wearing her earphones, and someone calls her name, the earphones estimate the direction from which the voice signal arrived. Classical beamforming/AoA algorithms do not apply directly since the earphone microphones are now subject to diffraction and pinna multipath. *UNIQ* develops an HRTF-aware AoA estimation technique to enable these application-specific capabilities.

We implement *UNIQ* on off-the-shelf earphones and smartphones, and evaluate with 5 volunteers. Our success metric is two-fold: (1) We compare *UNIQ*'s personalized HRTF with the upper bound, which is the ground-truth HRTF accurately measured for each volunteer in our lab. (2) We also compare against the global or general HRTF available online; this is the lower bound for personalization.

Results show that our personalized HRTF is, on average, $1.75X$ more similar to the ground-truth HRTF than the global HRTF. The personalization extends improvements to all users, and is robust to various kinds of sounds such as music and speech. In the AoA application, we observe more than 20° average improvement when using the personalized HRTF over the global one. We believe our current method is a step forward in this long-standing problem of HRTF personalization [27, 58], made possible by the fusion of motion sensing and acoustics. Refinements are still possible as we describe in Section 7, however, in the context of this paper, the main contributions may be summarized as follows:

1. *To the best of our knowledge, this is among the earliest attempts to bring (motion + acoustic) sensor fusion to HRTF personalization.* We map the personalization problem to one in multi-modal localization and synthesis, and show that IoT-style architectures can usher new approaches.
2. *We model signal diffraction on the human head, solve for head parameters, and utilize it as a critical component in estimating the personal HRTF.* We develop a functional prototype that is convenient, practical, and relevant to emerging ideas in immersive AR/VR applications.

The rest of this paper will expand on each of these contributions, starting from groundwork and measurement, followed by system design, and evaluation.

[1] Normally, when the sound source is less than $1m$ from the head, it is considered to be in the "near-field". [4]

2 GROUNDWORK ON HRTF

This section sheds light on the 2 fundamental constructs of HRTFs: (1) the acoustic channel introduced by a user's pinna, and (2) diffraction caused by curvature of faces/heads. This should also help characterize the gap between the global and personal HRTF.

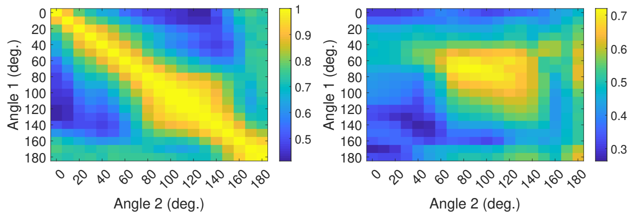
■ **Does the pinna's effect vary with angle of arrival, θ?** Recall that when a sound signal impinges on the pinna, it bounces and scatters in complex ways, reaching the ear-drum at staggered time instants. To test if this effect is sensitive to the angle of arrival θ, we ask a user, Alice, to wear an in-ear microphone on her left ear. We play short chirps from a speaker on the left side of Alice, so that the head's effects do not interfere with the microphone recording (we intend to only measure the impact of the pinna). The speaker is moved in a semi-circle, starting from the front of the nose ($\theta = 0°$) and ending at the back of the head ($\theta = 180°$), with measurements every $10°$. With 18 audio measurements, denoted $A(\theta)$, we now compute the cross-correlation c between $A(\theta_i)$ and $A(\theta_j)$, $i, j = \{1, 2, \ldots, 18\}$ as

$$c = max(f(\tau)) = max(\sum_{t=-\infty}^{\infty} A(\theta_i)(t) \cdot A(\theta_j)(t+\tau))$$

where τ is the relative delay between 2 audio signals.

Figure 2(a) shows the results. Evidently, the correlation matrix is strongly diagonal, implying that the pinna's impulse response is quite sensitive to θ, with almost a 1:1 mapping. This is consistent across our 5 volunteers, suggesting that the pinna indeed plays an important role in the human's ability to perceive directional sounds (at a resolution of $\approx 20°$).

■ **Does the pinna's effect vary across users?** The natural next question is whether the pinna's response varies across users for the same θ. For this, we cross-correlate the audio measurements from 2 users, $A_{Alice}(\theta_i)$ and $A_{Bob}(\theta_i)$, $\forall i$. Figure 2(b) shows the results. Clearly, Alice and Bob's pinnas do not match well, for example, Alice's recording (angle 1) at angle 80° corresponds well with Bob's recording (angle 2) at angle 140°. This means, when global HRTFs are used in ear-devices, the resolution for directional sounds can be no higher than $\approx 60°$, suggesting that the gap between global and personal is not negligible. Thus global HRTF obviously degrades user experience.



.

Figure 2: Pinna's effect: (a) Diagonal confusion matrix for the same user, across different angle of arrival, θ. (b) For different people, their pinna's transfer functions are markedly different.

■ **Do signals diffract on a person's face/head? Is diffraction distinct across users?**

Diffraction is the phenomenon where waves bend around the corners of an obstacle or through an aperture into the region of geometrical shadow of the obstacle/aperture [38]. From the physics of wave propagation (see detailed explanations in [9]), diffraction depends on the relative wavelength of the signal compared to the size of the object [9], as shown in figure 3. With larger wavelength, sound waves exhibit far more diffraction than, say, light or RF signals.

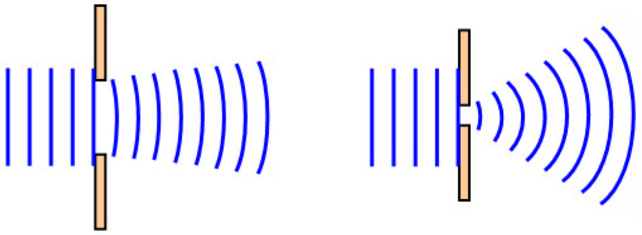


Figure 3: Diffraction illustration: a wave will propagate into the region of geometric shadow. The larger the wavelength compared to the aperture, greater is the diffraction [2].

Figure 4 illustrates an experiment to characterize diffraction on the human face, particularly due to the curvature of the cheek. We ask Alice to wear a *reference* microphone on her right ear; a second *(test)* microphone is pasted at 6 different locations on the left part of her face (starting with the tip of the nose and ending at the ear). An electronic speaker (shown on the user's right) plays a chirp and we calculate the chirp's time difference of arrival (TDoA), Δt, between the 2 microphones[2]. Multiplying speed of sound v with Δt, we get the difference in physical distance that the signal has traveled from the speaker to the 2 microphones: $\Delta d = v \cdot \Delta t$.

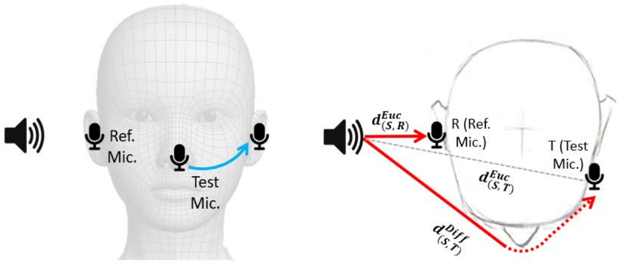


Figure 4: Experiment to test for signal diffraction on the curvature of the human head.

In parallel, using precise measurements from a camera, and a soft-tape that can bend along the curvature of the face, we obtain the following distances: the Euclidian distance from the speaker S to the reference microphone R, $d^{Euc}_{(S,R)}$, the Euclidian distance to the test microphone T, $d^{Euc}_{(S,T)}$, and the distance along the *diffracted-path* to the test microphone T, $d^{diff}_{(S,T)}$. The test for diffraction is now easy: Does Δd derived from audio recordings better match with the Euclidian path difference Δd^{Ecu} or the diffracted path Δd^{Diff}, where

$$\Delta d^{Ecu} = d^{Euc}_{(S,T)} - d^{Euc}_{(S,R)}$$

$$\Delta d^{Diff} = d^{Euc}_{(S,T)} - d^{Diff}_{(S,R)}$$

[2] This is possible because the 2 microphones are synchronized with a wire.

Figure 5 plots the results of matching. Evidently, Δd matches strongly with the diffracted path, especially as the test microphone moves further away from the reference. The results of this experiment were again consistent across multiple users, offering strong evidence that (1) audible sounds do not penetrate through the human head, and (2) modeling diffraction is critical for signal processing on human bodies.

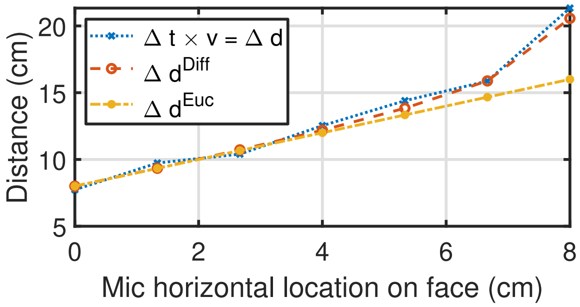


Figure 5: Acoustic and physical measurements are consistent; shows evidence that signals diffract along the curvature of the head.

Building on these basics, we turn to estimating personal HRTF and applying it to AoA estimation and beamforming.

3 SYSTEM SKETCH

This section outlines the key ideas, starting with near-field HRTF, then expanding to the far-field, and finally discussing an application of the estimated HRTF.

3.1 Near field HRTF

Consider a user moving her phone (in a circular trajectory) around her head, and her in-ear earphones recording the sounds transmitted by the phone. If we can accurately track the phone's location, then near field HRTF can be directly estimated. This is because the acoustic channel can be estimated from each location of the phone, and since location tells us the angle θ, the channel from each angle is now known. The per-angle acoustic channel is exactly the near-field HRTF. **Thus, to estimate the near field HRTF, the main challenge is in estimating the phone's location.**

While IMU sensors on the phone can help with localization, it is far too noisy for the accuracy levels needed with HRTFs. The main reason is well known, i.e., location estimation with IMUs requires a double integration on the accelerometer data, which causes the noise to grow multiplicatively. In light of this, *UNIQ* operates in the polar coordinates $< r, \theta >$. The intuition is to fuse the IMU's gyroscope data with the acoustic channel information – the gyroscope helps with inferring the angular component θ and the acoustic signal delays (between the earphones) help with estimating the distance r. Even though each is erroneous in its own way, we hope joint optimization will achieve accuracy and robustness.

While inferring geometric distances r from multiple microphones should be feasible, it poses unique problems in our case with earphones. Since the head and pinna filter the acoustic signal arriving at the ears, conventional techniques from array signal processing are no longer accurate. In other words, we need to model the head's *diffraction* effect on sound waves to make the recorded acoustic information usable. Additionally, we also need to cope with head parameters, which is obviously different across people, and will affect diffraction. In sum, we are faced with the problem of jointly estimating the phone location and diffraction-related head parameters, using a fusion of both IMU and acoustic information. This motivates our first module in Figure 6: **"Diffraction-aware Sensor Fusion"**.

This module gives us the near-field HRTF, but only at discrete angles around the head. To generalize to continuous angles, we input the discrete estimates into the **"Near Field HRTF Interpolation"** module. The interpolated output allows *UNIQ* to synthesize binaural sounds for any location *near* the user[3].

3.2 Far field HRTF

Now consider what happens when an earphone user wants to simulate sounds from the far field (e.g., a user listening to a piano in a virtual concert – the sound should appear to come from the far-away stage). Say this far field location is at an angle θ from the head. Even though we know $HRTF(\theta)$ from our near-field estimation, using this $HRTF(\theta)$ for far-field is non-ideal. This is because sound signals arriving at the ears from a nearby location at angle θ would be different from a far-away location at angle θ. As illustrated in Figure 7, far-field produces parallel rays while near-field produces non-parallel rays, causing different multipath, arrival times, and diffraction profiles at the 2 ears.

In view of this, *UNIQ* needs to model how parallel rays from angle θ would scatter/diffract on the head and arrive at the ears. Since the near-field HRTF has already modeled head and pinna multipath, we combine information from multiple $HRTF(\theta_j)$ to synthesize the far-field HRTF. We fine-tune this far-field HRTF by adjusting the delays and amplitude differences based on the head parameters learnt from the sensor fusion module. These operations make up the **"Near-Far Conversion"** module, which outputs the far-field HRTF. Combining near and far-field HRTFs, we can now create binaural sounds from any location around the user.

Finally, we develop a **"Binaural Angle of Arrival (AoA) Estimation"** module as an example application of far-field HRTF. We show how personalized HRTFs can estimate the direction of real ambient sounds with improved accuracy.

4 SYSTEM DESIGN

Figure 6 captures the system architecture. We begin this section with (1) Diffraction-Aware Sensor Fusion, which feeds into the (2) Near Field HRTF Interpolation module, as well as the (3) Near-Far Conversion module. The final output of *UNIQ* could then enable a number of applications; we discuss one example: "Binaural AoA Estimation".

4.1 Diffraction-Aware Sensor Fusion (DSF)

Once a user rotates the phone around her head, we have the IMU measurements and the microphone recordings. DSF's task is to

[3] Binaural sounds describe what a person would hear when a sound originates at some given location.

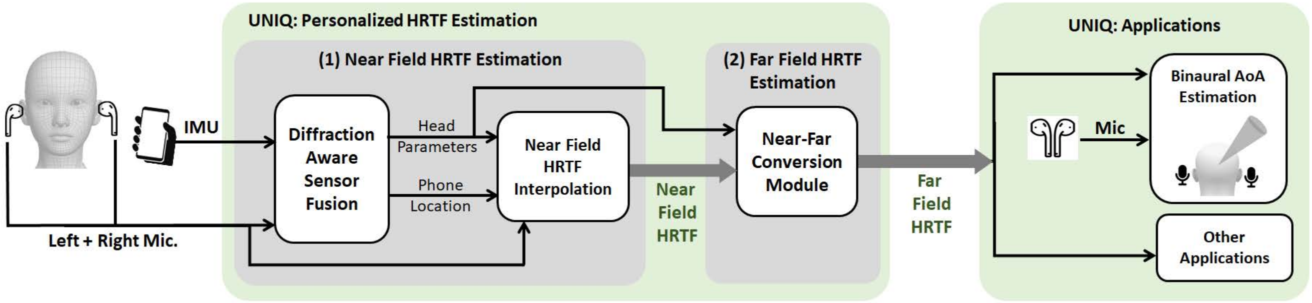

UNIQ: Personalized HRTF Estimation
(1) Near Field HRTF Estimation
(2) Far Field HRTF Estimation
UNIQ: Applications
IMU
Left + Right Mic.
Diffraction Aware Sensor Fusion
Head Parameters
Phone Location
Near Field HRTF Interpolation
Near Field HRTF
Near-Far Conversion Module
Far Field HRTF
Mic
Binaural AoA Estimation
Other Applications

Figure 6: System Architecture: *UNIQ* estimates both near and far-field HRTF taking inputs from the phone IMU and earphone microphone. The system pipeline is composed of 3 modules (diffraction-aware sensor fusion, near field HRTF interpolation, and near-far conversion) followed by an application that estimates binaural AoA from the personalized HRTF.

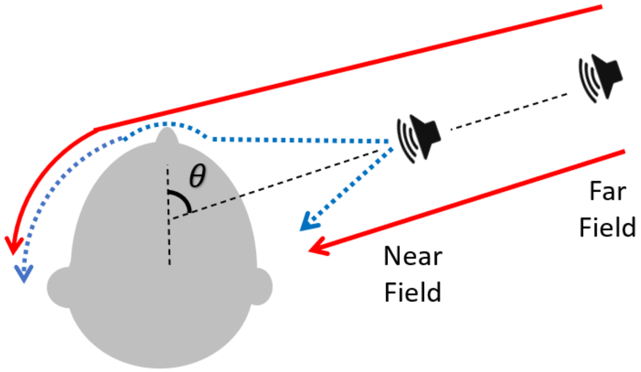


Figure 7: Illustration of near and far HRTF for angle θ.

accept these measurements as inputs and output both the head's geometric parameters and the phone's location. For this, let us model diffraction first.

Modeling Head Diffraction

Figure 8 shows a simplified version of signal diffraction on the head. To model this, we start by approximating the head shape as a conjunction of two half-ellipses, attached at the ear locations. This is necessary since the head is not symmetric between the front and back, hence spherical models have been avoided in literature [49]. The head shape can now be expressed through a 3-parameter set, $E = (a, b, c)$, where a, b, and c are the axis lengths of the two ellipses. Now, assuming the sound source is towards the right of the head, the signal would not penetrate through the head to arrive to the left ear, but would bend over/around the left cheek of the user (diffraction). With head parameters E known and for a given phone location P, we can estimate the time at which the diffracted signals would arrive at the two ears respectively.

Figure 9 shows the measured acoustic channel at the two ears for the above scenario (the channels are estimated by deconvolving the received signal with the known source signal). Clearly, the channel has multiple peaks (or taps) since the signal reflects on the face and these reflections also diffract. However, we are interested only in the first peaks at the two ears, since they are the ones that reliably capture the relationship between the phone and ear locations. This

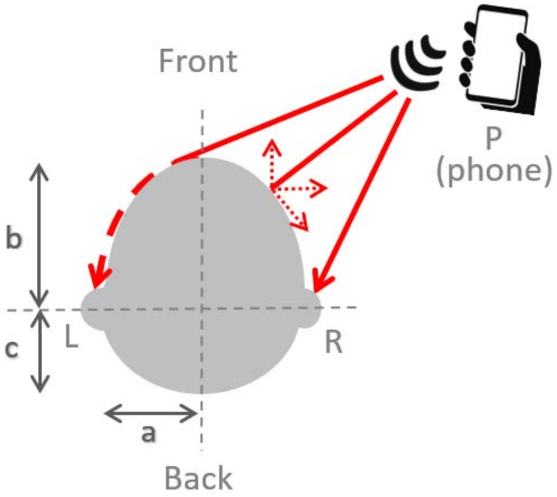


Figure 8: Sound waves arriving from phone at location P will diffract around the head before reaching the two ears.

is because the subsequent peaks in the channel are paths that arrive after reflecting on various points on the face, and while they may be useful to image the face, they are not necessary for our purposes of phone localization. Thus, *UNIQ* extracts the first peaks from the two channels and uses the relative delay Δt to connect the phone location and the head-shape in a common framework, as shown in equation 1:

$$\begin{aligned}\Delta t &= \text{relative delay for first peak in } h_L, h_R \\ &= f(\text{Diffraction}) \\ &= f(a, b, c, P)\end{aligned} \tag{1}$$

This serves as the basis for diffraction-aware sensor fusion.

Sensor Fusion Algorithm

Now, consider the IMU readings from the phone and the sound recordings from the in-ear microphone (the phone and the earphones are synchronized). *UNIQ* infers the phone's inertial rotation from the IMU's gyroscope, which translates to the phone's polar angle relative to the head. Of course, this still does not give the phone location (since the distance to the head is unknown).

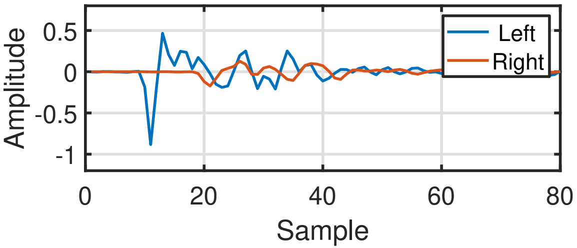


Figure 9: Channel impulse response: first tap corresponds to diffraction path

On the other hand, if the parameters $E = (a, b, c)$ are known, the relative delay from the acoustic channels can give phone location (with some ambiguity since 2 front/back locations can produce the same delay at the ears). Said differently, IMUs and acoustic channels do not individually solve the localization problem, but contribute adequate information to (over) determine the system of equations. This is exactly why sensor fusion helps – *UNIQ* jointly solves for head parameter and phone location through a fusion of IMU and acoustics.

The steps of the fusion algorithm can now be laid out:

1. As the smartphone rotates around the head, the IMU measurements are integrated to obtain the phone's orientation α. Since we ask users to face the phone's screen towards their eyes, α should be exactly equal to the polar angle θ (illustrated in Figure 10(a)). Over time, the phone orientation and the polar angle change, denoted as θ_i and α_i, $i = 1, 2, \ldots, N$.
2. Using the measured acoustic channels, and pretending we know the head parameters E, we can localize the phone and map it to the polar angle $\theta_i(E)$.
3. When the parameters E are correct, the α_i and θ_i should match $\forall i = 1, 2, \ldots, N$.
4. Due to noise in IMU and acoustics, we minimize the squared error $\|\alpha - \theta\|^2$ with decision variables as E:

$$E_{opt} = \underset{E}{argmin}\left(\sum_{i=1}^{N} \delta_i^2\right) = \underset{E}{argmin} \sum_{i=1}^{N} \Big(\alpha_i - \theta_i(E)\Big)^2 \qquad (2)$$

With larger N, i.e., more measurements from the user, the E_{opt} converges better.

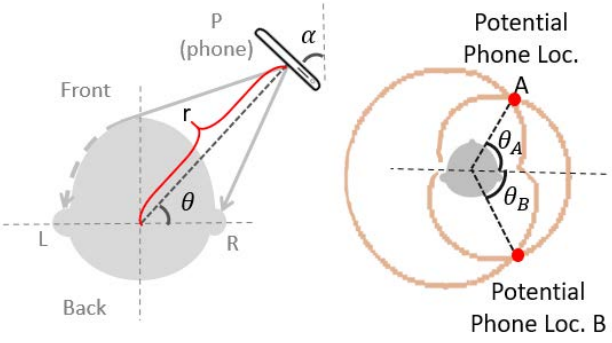


Figure 10: Near-field localization illustration. (a) Illustration of symbols. (b) Localizing phone using absolute diffraction path length from two ears.

Estimating Polar Angle $\theta_i(E)$ in Step 2 above:
Estimating α from IMU readings is a straightforward gyroscope integration. However, phone location and angle θ_i from the acoustic model is slightly more involved. Assume t_1 and t_2 are the diffraction path delays (first tap delays) for signals that arrive at the left and right ear, respectively. Now, assuming we already have the head parameters E, then we can draw 2 trajectories (as shown in Figure 10(b)). The first one is the trajectory of points from which the diffraction-based delay to the left ear is t_1. The second trajectory is the one from which the diffraction-based delay to the right ear is t_2. The phone's location must be at the intersection of these 2 trajectories. From the figure, we can observe that the two trajectories actually intersect at two points A and B, with polar angles $\theta_A(E)$, $\theta_B(E)$, and polar radius r_A, r_B. To disambiguate, we will pick the $\theta(E)$ that is closer to the IMU angle estimation α. By plugging $\theta(E)$ and α into the above Equation (2) and performing the optimization, *UNIQ* derives the optimal head parameter E_{opt}.

As a final step, we combine the IMU and acoustic localization results to obtain the estimated location of the phone as

$$P(\phi_i, r_i) = P\big((\theta_i(E_{opt}) + \alpha_i)/2, r_i\big) \qquad (3)$$

By indexing the measured HRTFs with the estimated phone locations, we complete the near-field HRTF estimation at discrete sample points. To obtain a continuous near-field HRTF, we employ interpolation.

4.2 Near field HRTF interpolation

It is difficult for a user to rotate the phone in continuous trajectories around their head. Thus, we allow users to position the phone at as many convenient locations as possible, and interpolate across other locations (shown in Figure 11). Interpolation is crucial because (1) downstream applications may intend to place sounds in any arbitrary location in the near-field; (2) as we will see soon, continuous near-field HRTF aids in synthesizing the far-field HRTF.

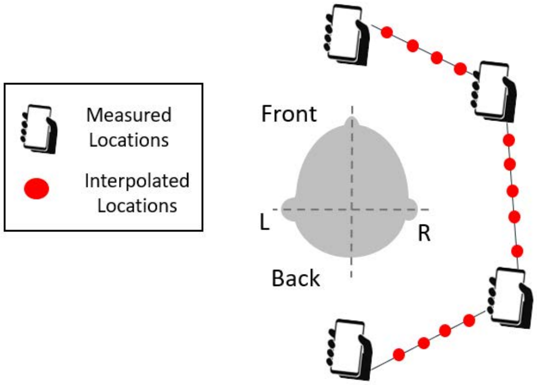


Figure 11: Near-field HRTF (linear) interpolation

The idea behind near-field HRTF interpolation is actually simple. If available measurements are from polar angles $\phi_1, \phi_2, \ldots\phi_N$ around the head, the interpolation module basically takes adjacent near-field HRTFs and linearly interpolates for all angles between ϕ_i and ϕ_{i+1}. Of course, the HRTFs from ϕ_i and ϕ_{i+1} need to be aligned carefully along their first taps before the interpolation; otherwise spurious echoes will get injected into the HRTF. To this end, we convert the HRTFs into the time domain impulse responses (i.e., HRIRs), align them, and interpolate. Finally, observe that for a given interpolated location L and HRTF H_L, we can partly assess the quality of interpolation (i.e., by modeling the diffraction from the known head parameters E and the location L). If the interpolated HRTF deviates from this model, we adjust the channel taps to match the expected time-difference and the amplitudes. These tuned channels for every angle $[0, 180]$ is converted back to the frequency domain, and declared as the final near-field HRTF.

By now, we have covered the system design for measuring the personalized near-field HRTF for a given user. Building on this, we will then show how we estimate the far-field HRTF from our near-field estimations.

4.3 Near-far conversion

Recall from Figure 7 that for a given angle θ, the near and far-field HRTFs are not the same. The goal of this module is to synthesize the far-field HRTF from near-field measurements. Our observation is that the far-field sound arrives to the ears as parallel rays (Figure 12), while the near field sound – behaving as a point source – emanates rays in all directions (Figure 13). This means the far field sound rays are actually contained in the near field measurements. The challenge lies in decomposing the near-field signals and extracting out the appropriate rays. An accurate solution to this problem is complex and computationally heavy because decomposing entails searching in a high dimensional space. We develop a heuristic based on first-order diffraction models and the physics of signal propagation. Our intuition is to understand directions from which far-field rays would arrive, and identify near-field locations that lie on those rays (see Figure 12). We elaborate with an example next.

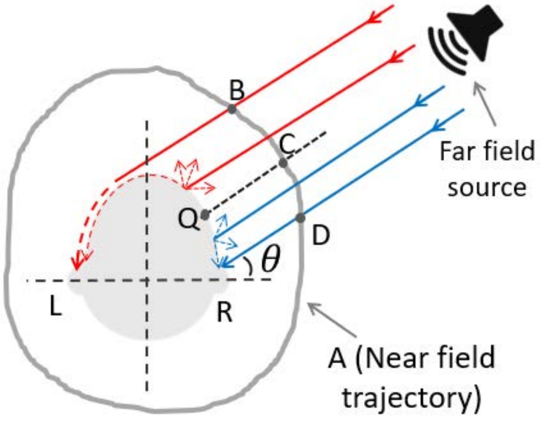


Figure 12: Near-far conversion: near-field HRTF on different part of trajectory A would contribute to far-field HRTF at different ears.

Figure 12 shows a roughly circular trajectory (A) on which we have estimated near-field HRTFs. Suppose we want to synthesize the far field HRTF arriving from angle θ as shown in the figure. The signal paths from the far-field, or rays, arrive in parallel, intersecting with the trajectory A at different locations (e.g., B, C, D) Now, let us define several "critical" rays: ray $B - L$ that arrives at the left ear, ray $D - R$ that arrives at the right ear, and ray $C - Q$ (also arriving from angle θ) is perpendicular to the tangent on the head at point Q. Our observation is that the physics of wave propagation dictates which rays will arrive at which ear. In other words, the incident signal will diffract along a direction that deviates least from its original direction. Hence, rays arriving on the left of Q (i.e., ones passing through the arc $[C, B]$) will diffract towards the left ear due to the curvature of the ellipse. Rays impinging the right of Q (i.e., passing through the arc $[C, D]$) will propagate towards the right ear. And signals on the outer side of B and D will not arrive at either ear.

Building on this intuition, observe that near-field HRTF measured from locations in arc $[C - B]$ can help synthesize the far-field HRTF at angle θ at the left ear. Similarly, near-field HRTF from arc $[C - D]$ would contribute to the far-field HRTF on the right ear. Thus, *UNIQ* approximates the far-field HRTF for the left ear as an average of near-field left-ear HRTFs from locations in $[C - B]$; for the right ear, average is from $[C - D]$. The method repeats for each value of $\theta \in [0, 180]$, meaning that B, C, and D would change accordingly.

Additional attempts on near-far conversion

While the above approach yields encouraging results, it is admittedly a heuristic. We have been exploring relatively deeper approaches, and while we have not succeeded yet, we discuss two of them here. We believe these are rich topics of future work.

Our approach is aimed at decomposing the components of near-field measurements – both diffraction and multipath from each arrival angle – and then aggregating a subset of these components to synthesize the far-field effect. Figure 13 aims to explain this systematically. When transmitting from the near-field, the sound source should be considered as a point source, emitting rays in different directions $\theta_1, \theta_2, ..., \theta_N$. Let us focus on a single point X_k on the near field trajectory. Our measured near-field HRTF for point X_k is essentially the sum of the effects from all the signal rays emanating from X_k, hence can be modeled as:

$$H_{near}(X_k) = \sum_{i=0}^{N} H(X_k, \theta_i) \tag{4}$$

Now if we want to synthesize far-field signals from direction θ_l, we need to select only the θ_l-bound rays from each of the points on the near field trajectory (as shown by the yellow arrows in Figure 13). We can write this synthesize process as:

$$H_{far}(\theta_l) = \sum_{i=0}^{M} H(X_i, \theta_l) \tag{5}$$

Evidently, if we can decouple the RHS of equation 4, and obtain $H(X_k, \theta_i)$ for any given k, i, then we can recombine and find the $H_{far}(\theta_l)$ in equation 5 (of course we still need to tune the delay of each ray based on geometry). Hence, the core research question pertains to correctly performing this decomposition.

■ **Attempt 1: speaker beamforming:** Modern smartphones have 2 speakers (one for the left channel and one for the right). If we can utilize these 2 speakers to create a time-varying beamforming pattern, this could help estimate $H(X_k, \theta_i)$. Specifically, denote the

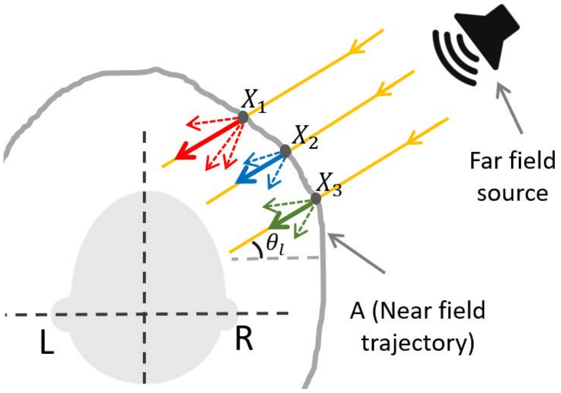


Figure 13: Near-far conversion attempts: if we can decouple near-field HRTF into rays, then far-field HRTF essentially needs to extract out one ray from each near-field location and recombine with appropriate weights.

beamforming pattern at one time instance as $w(\theta)$, which is a function of angle θ. Then we can rewrite Equation 4 as

$$H_{near}(X_k) = \sum_{i=0}^{N} w(\theta_i) \cdot H(X_k, \theta_i) \tag{6}$$

By creating time varying beamforming patterns $w_t(\theta)$ – by changing the relative phase and amplitude of the 2 speakers – we can generate multiple equations, one for each time instance. This could enable us to solve for $H(X_k, \theta_i)$. The difficulty, however, is that the 2 speakers are unable to create a spatially narrow beam pattern. This eventually leads to the system of equations being ill-ranked and causes large errors for the estimated $H(X_k, \theta_i)$.

■ **Attempt 2: blind decoupling:** The net effect of $H(X_k, \theta_i)$ on each signal ray has 2 components. First, the diffraction around the head creates a delay and attenuation. Second, the signal bounces from the pinna, creating an effect we call the pinna multipath. Hence, the net effect on each signal ray can be expressed as

$$H(X_k, \theta_i) = A_i \delta(\tau_i) * h_k \tag{7}$$

where δ is the Dirac delta function, τ_i is the ray's diffraction delay, A_i is signal attenuation, and h_k is the time domain pinna multipath channel ($*$ denotes convolution here). We plug Equation 7 to Equation 4, and we can have

$$H_{near}(X_k) = \sum_{i=0}^{N} A_i \delta(\tau_i) * h_k \tag{8}$$

Now, if we can estimate $\sum_{i=0}^{N} A_i \delta(\tau_i)$ and h_k separately, the decoupling can be solved. $\delta(\tau_i)$ can be estimated from diffraction geometry, but we do not know A_i and h_k. This becomes a blind decomposition problem. While sparsity opportunities could help solve this problem, we realize that our physics based signal model may be inadequate to capture the sophisticated real-world signal propagation patterns. We believe machine learning techniques are relevant here; we leave that to future work.

4.4 Interface to Applications

The near and far-field HRTFs estimated by *UNIQ* can now be exported to earphone applications as a lookup table. The table is indexed by θ, and for each θ_i, there are 4 vector entries:

$$\theta_i : \langle H_{near}^{left}, H_{near}^{right} \rangle, \langle H_{far}^{left}, H_{far}^{right} \rangle$$

Each HRTF is obviously a channel filter, so when an application intends to synthesize a binaural sound S from a desired location L, the application first determines if L is nearby or far-away, and the angle θ_i of the location L relative to the head. If L is far-away, then the application filters the sound as

$$Y_{left} = H_{far}^{left} S, \; Y_{right} = H_{far}^{right} S$$

The earphone now plays the two sounds, Y_{left} and Y_{right} on the left and right ears, respectively. The user perceives the sound to be coming from angle θ_i from a far-away location. We next present one potential application that can benefit from the estimated HRTFs.

4.5 Binaural Angle of Arrival (AoA)

Understanding the incoming direction of real ambient sounds (relative to the user's head) can enable smart earphones to fuel new applications. For instance, earphones could serve as hearing aids, and beamform in the direction of a desired speech signal; thus, Alice and Bob could listen to each other more clearly by wearing headphones in a noisy bar. In another example, earphones could analyze the AoAs of music echoes in a shopping mall and enable navigation by triangulating the music speakers. Now, to accurately estimate the AoAs of these ambient sounds, the earphones need to apply the HRTF (since conventional AoA techniques are not designed to cope with the HRTF distortions). This motivates HRTF-aware AoA estimation, with both unknown source signals (such as Alice and Bob's speech) and known signals (such as those from ambient acoustic speakers).

■ **Known source signals:** If the source signal is known, we first extract the acoustic channels from the left and right ears. To now estimate AoA, we look for the following 2 features from the channels: (1) the first tap relative delay between left and right channels, and (2) the shape of the time-domain channel. Observe that (1) is impacted by head diffraction and (2) is related to the pinna multipath, both embedding information about the signal's AoA. As mentioned in Section 2, both these features vary across humans. This is why the personalized HRTF is helpful here. We match these 2 features from our measured channel against our estimation $HRTF(\theta)$ — the θ that maximizes the match is our AoA estimate.

Mathematically, let t_0 be the relative first tap delay from our binaural recording, and $t(\theta)$ be the same relative delay but for the personal HRIR templates estimated for each θ. Also denote $c_L(\theta)$ and $c_R(\theta)$ as the correlation values for left/right channels with (left/right) HRIR templates for all θ. We define a target matching function T that contains both relative delay and channel correlation information:

$$T(\theta) = \lambda |t_0 - t(\theta)| + [1 - c_L(\theta)] + [1 - c_R(\theta)] \tag{9}$$

After training for the appropriate λ, we find the actual AoA by minimizing the target function.

■ **Unknown source signal:** For unknown source signals, we can no longer extract the 2 acoustic channels for left and right ears, making it difficult to find the relative first tap delay, or left/right channel shape.
However, we still have the opportunity to infer the first tap delay from the *relative* channels between the left and right ear-recordings – this can help estimate the AoA.
Of course, this is not straightforward since signals arriving at both ears contain a lot of pinna multipath, and thus have poor autocorrelation. This will cause multiple peaks in the relative channel, as shown in Figure 14. Let us assume each peak has a relative delay Δt_i. Based on our diffraction model, each relative delay Δt_i can further translate into 2 AoAs: $AoA_{i,1}$ and $AoA_{i,2}$ (one for front and one for back). Now our task is to find the true AoA from all the potential AoAs.

The key idea for disambiguating is to still utilize the time domain shape of the channel. Since we cannot extract the left or right channel, our key intuition is to compare the shape of the "relative" channel. Suppose the left ear recording is L, and right ear R, in the frequency domain. Then the relative channel is $\frac{L}{R}$. We can also calculate the relative channel for all angle θ in the personal HRTF template $\frac{HRTF_L(\theta)}{HRTF_R(\theta)}$. Ideally, for the correct θ, these 2 relative channels should match:

$$\frac{L}{R} = \frac{HRTF_L(\theta)}{HRTF_R(\theta)} \tag{10}$$

Since division are sensitive to errors when the denominator is small, we change Equation (10) into a multiplication form:

$$L \times HRTF_R(\theta) = R \times HRTF_L(\theta) \tag{11}$$

By plugging all the potential AoA_i's (inferred from relative channel peaks) into the above equation, and finding the one that gives the closest LHS and RHS, we identify the true AoA.

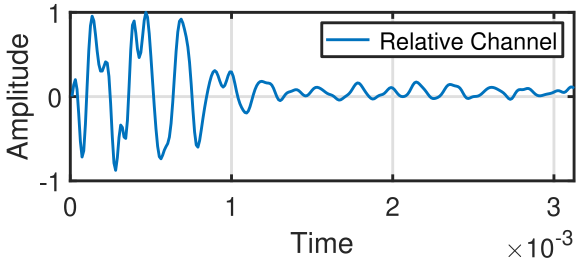


Figure 14: Relative channel between left and right ear: there are multiple channel taps due to poor signal autocorrelation.

By now, we have covered the key system design ideas. We will then show some system details.

4.6 Engineering and System Details

■ **System frequency response compensation:** Before performing HRTF measurements, the first step is to compensate for the frequency response of the speaker and microphone pair. This is important because any channel we estimated would intrinsically embed this frequency response inside. We estimate frequency response of the speaker and microphone pair by placing microphone co-located with speaker and play a flat amplitude chirp signal.

■ **Tackling room reflections:** The traditional approaches to HRTF measurement are conducted in echo-free acoustic chambers. Home users obviously do not have access to such "anechoic" chambers. However, we can eliminate room-level echoes as a pre-processing step in *UNIQ*. The idea is simple: when users rotate the phone around their heads, head diffraction and pinna multipath should arrive earlier than room reflections. We eliminate the latter channel taps to exclude room reflections.

■ **Automatically correcting user gestures:**
A user may not be able to rotate the phone around the head in the very first attempts; practical problems can occur such as the arms lowering, the phone spinning, etc. This can affect measurement and downstream accuracy. *UNIQ* identifies such cases by detecting that the estimated phone distance to head center r_i in Equation (3) is too small, or the overall error $\sum_{i=1}^{N} \delta_i^2$ in equation (2) is too large. This triggers a message to the user to redo the measurement exercise. With this, we are ready to move to system evaluation.

5 EVALUATION

Figure 15 shows our system setup. *UNIQ* is implemented on a Xiaomi [11] smartphone and a Sound Professionals earphone (model: SP-TFB-2) [1], which supports in-ear microphones. In-ear microphones are becoming popular and can improve the HRTF quality since the sounds will be recorded closer to the ear-drum. Since our phone does not have a front-facing speaker, we connect the audio output to a small external speaker. User wears the earphone and rotates the smartphone (with pasted speaker) around her head.

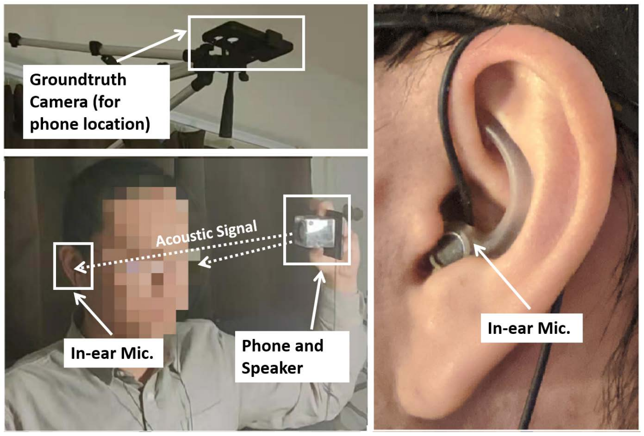


Figure 15: System prototype. Left: experimental setup. Right: zoom in to in-ear microphone

During the measurement process, we collect $100Hz$ IMU data from the phone, and $96kHz$ sound recording from the in-ear microphone. The speaker, microphone, and IMU are all synchronized. The data processing pipeline runs on MATLAB. The ground-truth data for smartphone (and head) locations are obtained from an overhead camera installed on top of the user's head.

Figure 16 shows the frequency response of our speaker microphone pair. The response curve is unstable below $50Hz$ and stabilizes reasonably over $[100Hz, 10kHz]$. This shows that our hardware is not anything special; in fact, expensive phones and headphones may exhibit better frequency response curves. Finally, given that human ears are insensitive to sound below $100Hz$ [13], any standard hardware platform should be adequate to run the *UNIQ* system.

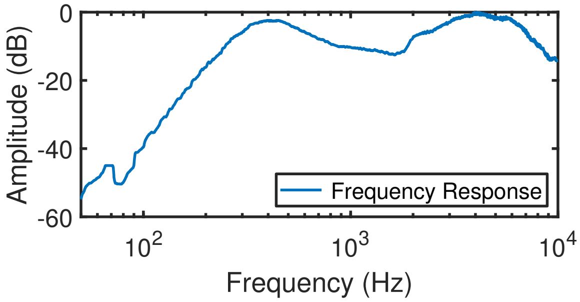


Figure 16: Frequency response of our speaker-microphone pair. Most hardware platforms exhibit such response curves, if not better [31].

5.1 Results

Phone Localization Accuracy

Figure 17 plots the phone localization angular error in near-field. The X-axis in Figure 17(a) plots the ground-truth polar angle of the phone as viewed from the overhead camera. The Y-axis plots *UNIQ*'s estimate of the polar angle as the user rotates her hand. Perfect accuracy would mean that the plotted points would like on the $X = Y$ diagonal line. Evidently, *UNIQ*'s localization is consistently quite accurate. Figure 17(b) plots the CDF; the median error is 4.8 degrees. The error is mostly due to the difficulty of ensuring the phone's center is perfectly facing the user's own head. Imperfection of the acoustic diffraction model also partly contributes to the errors, but less significantly. Only in rare cases, the phone's localization error climbs to 15 degrees because the volunteer's movement has deviated too much from the instructions. This adds to the downstream errors, however, wee include these cases since they are a part of real-world operating conditions.

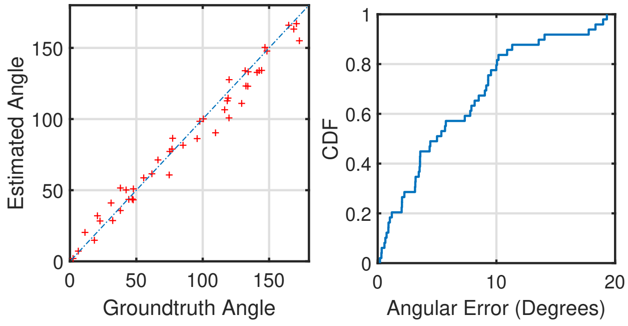


Figure 17: Phone's angular error for hand-rotation: (a) comparison with ground truth, (b) error CDF.

Personalized HRTF Estimates

The HRTF is a vector that completely embeds the head/pinna's acoustic impulse response. An objective way to evaluate HRTF estimate is to cross-correlate personalized HRTF vector with ground truth. This will reveal how closely *UNIQ* matches the truth. Further, plotting correlation between ground truth and global HRTF will also reveal the improvement of personal over global HRTF.

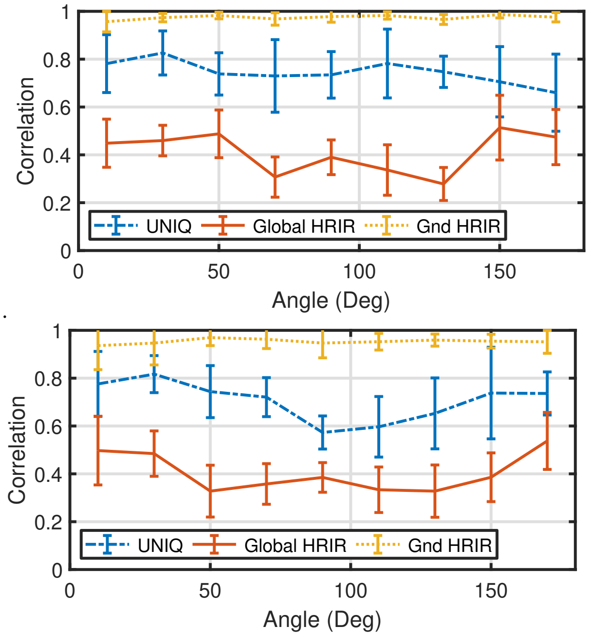


Figure 18: Cross-correlation between ground-truth versus *UNIQ*, global, and another measurement of ground-truth HRIR, for (a) left ear, (b) right ear.

Figure 18 shows the cross-correlations between estimated and general HRIR against the ground-truth HRIR (error bars represent standard division). We also show the cross-correlation between 2 separate measurements of ground-truth HRIR as a reference upper bound. Figure 18(a) plots for the left ear, and Figure 18(b) for the right; in both cases, the sound source was placed on the left of the head. Evidently, *UNIQ*'s estimated HRIR achieves an average correlation of 0.74 and 0.71 for the left and right ear, respectively. In contrast, the general HRIR can attain average correlation of 0.41 for both ears. This is a key result, illustrating that:

1. Global HRIRs significantly differs from personalized ones.
2. *UNIQ* considerably closes this gap (by a factor of $\sim 1.75X$)

For the right ear, our estimated HRTF exhibits higher accuracy when the angle is ≈0 or ≈180 degrees, but degrade around ≈90 degrees. This is because when the phone is at 90 degrees, the right ear microphone is exactly at the opposite side of the speaker, significantly suppressing the SNR of the received signal, resulting in lower accuracy. Higher quality earphones would certainly benefit in these cases.

Variation across Different Volunteers

Figure 19 shows the mean correlation for 5 volunteers (who wore the earphones and performed the smartphone rotation in front of their head). The two graphs – (a) and (b) – are again for the left and right ear, respectively. The personalization gain is consistent across all. Of course, *UNIQ* estimates the HRTF slightly less accurately for volunteers 4 and 5 compared to volunteers 1, 2, and 3. This is because when holding the phone, volunteers 4 and 5 moved the phone a bit too close to the back of their heads, due to their arm movement constraints, (even after automatic correction procedure of *UNIQ*), leading to sub-optimal estimates in the diffraction model, and downstream far-field estimations.

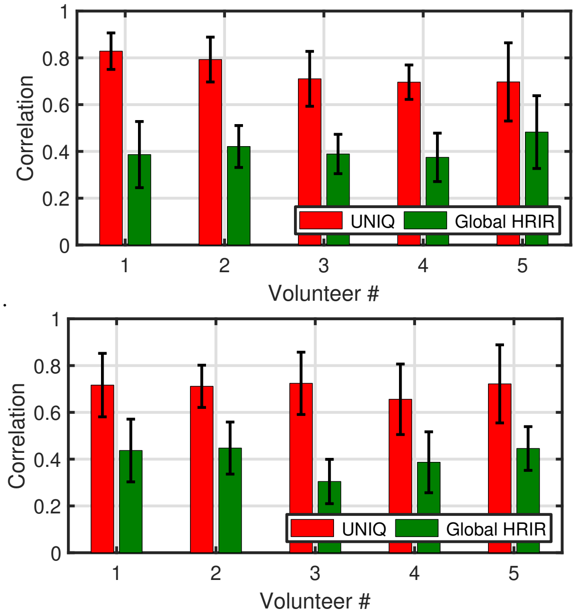


Figure 19: Average cross-correlation between estimated / global HRIR and the groundtruth across different volunteers for (a) left ear, (b) right ear

While the above results are statistical, Figure 20 zooms into few raw HRTFs in the time domain (called *head related impulse response*, HRIR). Specifically, the figure shows the (a) best case, (b) average case, and the (c) worst case estimation of *UNIQ*'s HRIR in comparison to the general HRIR. Evidently, across all 3 cases, our estimated HRIRs always decode the channel taps at correct locations; the general HRIR makes frequent mistakes. This is primarily due to *UNIQ*'s ability to capture per-user head and pinna multipath, which are obviously different from one human to another.

Application of HRTF to AoA

A more accurate HRTF implies that ambient sounds can now be better analyzed spatially, such as a hearing aid identifying the direction of an incoming sound. We put this to test by comparing the AoA error when applying the personalized HRTF from *UNIQ*, versus the global HRTF. We begin by playing a known source signal (from different locations in the far field) and estimating AoA.

Figure 21 plots the CDF of angular AoA error. With *UNIQ*'s personalized HRTF producing a median error of 7.8°, compared to global HRTF's median error of 45.3°. More importantly, the maximum error of personalized HRTF is 60° while the maximum for global HRTF is > 150°. This is because a global HRTF suffers considerably from "front-back ambiguity", i.e., it does not reliably differentiate between sounds arriving symmetric front and back angles, such as 45° north-east and 45° south-east. In fact, in 29% of our experiments, using global HRTF caused a front-back confusion.

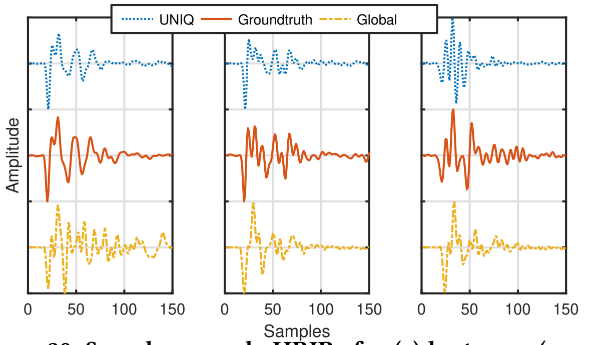


Figure 20: Sample example HRIRs for (a) best case, (corr = 0.96), (b) average case, (corr = 0.85), (c)worst case, (corr = 0.43). Global HRIRs almost always inferior.

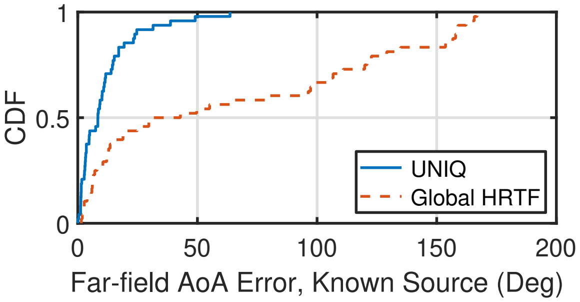


Figure 21: AoA estimation with personalized and global HRTF using a known source signal. Global HRTF performs poorly since measured signals deviate from HRTF estimate.

We repeat the above experiments with unknown source signals, such as when Alice calls Bob (and Bob is wearing a hearing aids or earphones). Alice's voice signal is unknown to Bob's device, however, the ear-devices can still decode Alice's direction better. We tested with a variety of "unknown" signal categories, such as white noise, music, and speech. Figure 22(a)-(c) shows the CDF of AoA error for each of these categories. The personalized HRTF offers consistent gains across all types of signals; the distribution has a somewhat heavy tail because, with unknown signals, the front-back ambiguity begins to affect *UNIQ* as well. The 80 percentile AoA error with personalized HRTF is within 20° for music and white noise. The improvement with speech is smaller because speech is dominated by lower frequencies, thus less sensitive to HRTF errors. Figure 22(d) zooms into the front-back cases, since these are crucial for real applications (we do not want Bob to hear a virtual voice that comes from a wrong direction). With *UNIQ*, the average front-back accuracy is 82.8% — white noise is highest at 87.2% and speech signals are lowest at 72.8%. This is because white noise spans over a large frequency range, offering more information about the acoustic channel; in contrast, speech signals are concentrated on base and harmonic frequencies, revealing less information about the channel. For global HRTF, the front-back accuracy is 59.8%.

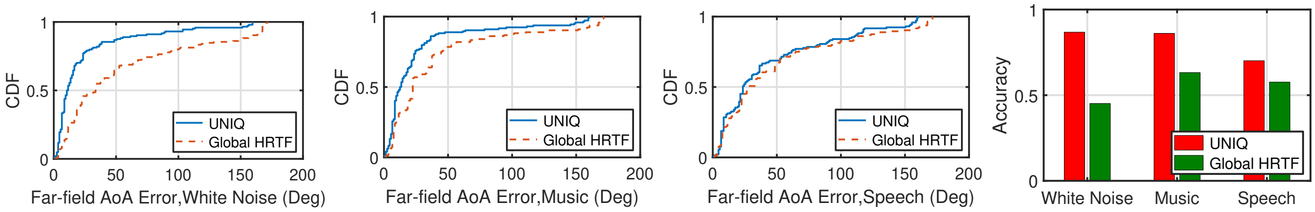


Figure 22: (a)-(c): AoA estimation error, (d): front-back identification accuracy for an unknown source signal.

6 RELATED WORK

■ **Smart earphones and earable computing:** With the development of mobile computing technologies [19], smart earphones are becoming popular these days. Past works have looked at enabling spatial audio[18, 21, 26, 51] and acoustic augmented reality [15, 54], step counting [44] and motion tracking [53], user authentication [20] and health monitoring [42] on sensor embedded smart earphones. This paper, however, adds another building block - personal HRTF to smart earphones. We believe this is a step towards even better functionality on future earables: e.g., more immersive spatial audio, and smarter beamforming, etc.

■ **HRTF personalization:** HRTF personalization has gained interest in recent years due to the development of VR and AR related technology. Traditionally people use large speaker arrays inside acoustic chambers to measure HRTF [17, 22], which is obviously not scalable. Some newer work [27, 29, 33, 36, 45, 58] tried to approach this problem without acoustic hardware from the pure signal processing perspective. They used acoustic simulation to generate the specific personalized HRTF for a given user from 3D scans of human head. These methods are reported to be slow and computationally heavy [28]. Moreover, obtaining an accurate 3D scan is also not easy [57]. Few attempts utilize mobile devices for HRTF measurement. [12] is the closest to our work. Their method, however, requires an external speaker placed on the table, and need to user to tie the smartphone onto the head, which is not portable. Moreover, their setup can be polluted by environmental multipath. Our approach, on the contrary, is novel, fast, and scalable. Users can get their personalized HRTF by simply rotating the phone around the head, in a couple of minutes.

■ **Acoustic/wireless sensing and sound source AoA estimation:** Acoustic/wireless sensing and sound source AoA estimation is a hot research topic in mobile, acoustic, and robotic community [23, 24, 34, 35, 37, 39, 41, 43, 47, 56]. Most past works require a mic array for sound AoA estimation [46, 48, 50]. [18, 30, 32] are the closest to our work, where the authors attempted to estimate sound AoA from artificially made robotic ears. Our problem is more challenging because past authors can design the robot head and ear entirely by themselves thus have full control and understanding of the accurate robot HRTF. In our case, we need to find the sound source AoA by extracting features from the not-so-accurate estimated HRTF, which brings about unique challenges.

7 LIMITATIONS AND OPEN PROBLEMS

■ **3D HRTF:** Our *UNIQ* prototype estimates the 2D HRTF for users. This may be acceptable, given that human ears exhibit relatively lower resolution in distinguishing elevation angles. Hence 2D may suffice for many applications. However, if an application desires 3D HRTF, extending *UNIQ* is viable – the user would now need to move the phone on a sphere around the head, and the motion tracking equations need to be extended to 3D. If this increases the tracking error, perhaps the phone camera can be utilized, enabling a fusion between motion, acoustics, and computer vision. We leave this to future work.

■ **Integrating Room Multipath:** As discussed earlier, *UNIQ* removes environmental reverberations through a pre-processing step in the time domain; this helps minimize the effect of room multipath on the estimated HRTF. However, rendering realistic 3D audio, especially in an indoor environment, requires that the room reverberations be embedded into the HRTF. Said differently, a real immersive experience can only be achieved by filtering the earphone sound with both the room impulse response (RIR) and the HRTF. Estimating RIR at home is an interesting but separate research question, outside the scope of this paper.

■ **User Experience and Externalization:** An estimated HRTF is accurate when the user is unable to correctly identify whether the sound she hears came from her earphone or an ambient loudspeaker. When she mistakes an earphone-played sound to be coming from the ambience, then the ideal goal of "externalization" is achieved. Of course, testing for externalization requires high quality earphone hardware and RIR integration. Moreover, optimization methods may be needed through human feedback, since externalization is also a complex function of human perception [14]. This paper shows that our estimated HRTFs are mathematically close to true HRTFs, but more work is needed to attain externalization.

8 CONCLUSION

The gap between global and personalized HRTFs remains an open problem. This paper ushers ideas from motion tracking and sensor fusion to partly close this gap. We show that simple arm gestures from users can offer valuable motion information, that in turn helps in modeling the user's unique HRTF parameters. As a side effect, we find that earphones can better estimate the AoA of ambient sounds. The results are promising and could underpin a range of immersive applications that are gaining relevance for the post-COVID future.

ACKNOWLEDGMENTS

We sincerely thank our shepherd Prof. Fadel Adib, and anonymous reviewers for their insightful comments and suggestions. We are also grateful to NSF (award numbers: 1918531, 1910933, 1909568, and 1719337), NIH (award number: 1R34DA050262-01), Google, and Samsung for partially funding this research.

REFERENCES

[1] 2015. The Sound Professionals. Retrieved Jan 26, 2021 from https://www.soundprofessionals.com/cgi-bin/gold/item/SP-TFB-2

[2] 2015. Wave Interactions and Interference. Retrieved Jan 24, 2021 from https://www.ck12.org/section/wave-interactions-and-interference-%3a%3aof%3a%3a-waves-%3a%3aof%3a%3a-ck-12-physical-science-for-middle-school/

[3] 2017. Beyond Surround Sound: Audio Advances in VR. Retrieved Jan 24, 2021 from https://www.oculus.com/blog/beyond-surround-sound-audio-advances-in-vr/

[4] 2017. Near-field 3D Audio Explained. Retrieved Jun 11, 2021 from https://developer.oculus.com/blog/near-field-3d-audio-explained/

[5] 2018. Simulating Dynamic Soundscapes at Facebook Reality Labs. Retrieved Jan 26, 2021 from https://www.oculus.com/blog/simulating-dynamic-soundscapes-at-facebook-reality-labs/

[6] 2019. Audio in mixed reality. Retrieved Jan 24, 2021 from https://docs.microsoft.com/en-us/windows/mixed-reality/design/spatial-sound

[7] 2019. Mach1 will provide spatial audio for Bose's AR platform. Retrieved Jan 24, 2021 from https://venturebeat.com/2019/12/18/mach1-will-provide-spatial-audio-for-boses-ar-platform/

[8] 2020. Apple brings surround sound and Dolby Atmos to AirPods Pro. Retrieved Jan 24, 2021 from https://thenextweb.com/plugged/2020/06/22/apple-brings-surround-sound-and-dolby-atmos-to-airpods-pro/

[9] 2020. Diffraction. Retrieved Jan 24, 2021 from https://en.wikipedia.org/wiki/Diffraction

[10] 2020. Inside Facebook Reality Labs Research: The Future of Audio. Retrieved Jan 24, 2021 from https://about.fb.com/news/2020/09/facebook-reality-labs-research-future-of-audio/

[11] 2020. Xiaomi United States. Retrieved Jan 26, 2021 from https://www.mi.com/us/

[12] 2021. DIY HRTF measurement using an iPhone. Retrieved Jun 11, 2021 from https://www.earfish.eu/sites/default/files/2018-01/DIY_earfish_iPhone_0.pdf

[13] 2021. Equal-loudness contour. Retrieved Jan 24, 2021 from https://en.wikipedia.org/wiki/Equal-loudness_contour

[14] Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, and W Owen Brimijoin. 2021. A framework for designing head-related transfer function distance metrics that capture localization perception. *JASA Express Letters* 1, 4 (2021), 044401.

[15] Jeffrey R Blum, Mathieu Bouchard, and Jeremy R Cooperstock. 2011. What's around me? Spatialized audio augmented reality for blind users with a smartphone. In *International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services*. Springer, 49–62.

[16] C Phillip Brown and Richard O Duda. 1997. An efficient HRTF model for 3-D sound. In *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 4–pp.

[17] Thibaut Carpentier, Hélène Bahu, Markus Noisternig, and Olivier Warusfel. 2014. Measurement of a head-related transfer function database with high spatial resolution. In *7th Forum Acusticum (EAA)*.

[18] Jorge Dávila-Chacón, Jindong Liu, and Stefan Wermter. 2018. Enhanced robot speech recognition using biomimetic binaural sound source localization. *IEEE transactions on neural networks and learning systems* 30, 1 (2018), 138–150.

[19] Hossein Falaki, Ratul Mahajan, Srikanth Kandula, Dimitrios Lymberopoulos, Ramesh Govindan, and Deborah Estrin. 2010. Diversity in smartphone usage. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*. 179–194.

[20] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using Ear Canal Echo for Wearable Authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.

[21] William G Gardner. 2005. Spatial audio reproduction: Towards individualized binaural sound. In *Frontiers of Engineering:: Reports on Leading-Edge Engineering from the 2004 NAE Symposium on Frontiers of Engineering*, Vol. 34. 113.

[22] William G Gardner and Keith D Martin. 1995. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America* 97, 6 (1995), 3907–3908.

[23] Reza Ghaffarivardavagh, Sayed Saad Afzal, Osvy Rodriguez, and Fadel Adib. 2020. Ultra-wideband underwater backscatter via piezoelectric metamaterials. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*. 722–734.

[24] Yasaman Ghasempour, Chia-Yi Yeh, Rabi Shrestha, Yasith Amarasinghe, Daniel Mittleman, and Edward W Knightly. 2020. LeakyTrack: non-coherent single-antenna nodal and environmental mobility tracking with a leaky-wave antenna. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 56–68.

[25] Michael M Goodwin and Jean-Marc Jot. 2007. Binaural 3-D audio rendering based on spatial audio scene coding. In *Audio Engineering Society Convention 123*. Audio Engineering Society.

[26] Michael M Goodwin, Jean-Marc Jot, and Mark Dolson. 2013. Spatial audio analysis and synthesis for binaural reproduction and format conversion. US Patent 8,374,365.

[27] Corentin Guezenoc and Renaud Seguier. 2020. HRTF individualization: A survey. *arXiv preprint arXiv:2003.06183* (2020).

[28] Nail A Gumerov, Ramani Duraiswami, and Dmitry N Zotkin. 2007. 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*, Vol. 1. IEEE, I–165.

[29] Hongmei Hu, Lin Zhou, Hao Ma, and Zhenyang Wu. 2008. HRTF personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics* 69, 2 (2008), 163–172.

[30] Sungmok Hwang, Youngjin Park, and Younsik Park. 2007. Sound direction estimation using artificial ear. In *2007 International Conference on Control, Automation and Systems*. IEEE, 1906–1910.

[31] C Jackman, M Zampino, D Cadge, R Dravida, V Katiyar, and J Lewis. 2009. Estimating acoustic performance of a cell phone speaker using Abaqus. In *SIMULIA Customer Conference*. 14–21.

[32] Cheol-Taek Kim, Tae-Yong Choi, ByongSuk Choi, and Ju-Jang Lee. 2008. Robust estimation of sound direction for robot interface. In *2008 IEEE International Conference on Robotics and Automation*. IEEE, 3475–3480.

[33] Lin Li and Qinghua Huang. 2013. HRTF personalization modeling based on RBF neural network. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3707–3710.

[34] Zhihong Luo, Qiping Zhang, Yunfei Ma, Manish Singh, and Fadel Adib. 2019. 3D backscatter localization for fine-grained robotics. In *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)*. 765–782.

[35] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic Ranging via Deep Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[36] 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)*. IEEE, 53–61.

[37] Yan Michalevsky, Aaron Schulman, Gunaa Arumugam Veerapandian, Dan Boneh, and Gabi Nakibly. 2015. Powerspy: Location tracking using mobile device power analysis. In *24th {USENIX} Security Symposium ({USENIX} Security 15)*. 785–800.

[38] Philip M Morse and Pearl J Rubenstein. 1938. The diffraction of waves by ribbons and by slits. *Physical Review* 54, 11 (1938), 895.

[39] Rajalakshmi Nandakumar, Krishna Kant Chintalapudi, Venkat Padmanabhan, and Ramarathnam Venkatesan. 2013. Dhwani: secure peer-to-peer acoustic NFC. *ACM SIGCOMM Computer Communication Review* 43, 4 (2013), 63–74.

[40] Takanori Nishino, Sumie Mase, Shoji Kajita, Kazuya Takeda, and Fumitada Itakura. 1996. *Interpolating HRTF for auditory virtual reality*. Ph.D. Dissertation. Acoustical Society of America.

[41] Chunyi Peng, Guobin Shen, Yongguang Zhang, Yanlin Li, and Kun Tan. 2007. Beepbeep: a high accuracy acoustic ranging system using cots mobile devices. In *Proceedings of the 5th international conference on Embedded networked sensor systems*. 1–14.

[42] Ming-Zher Poh, Kyunghee Kim, Andrew D Goessling, Nicholas C Swenson, and Rosalind W Picard. 2009. Heartphones: Sensor earphones and mobile application for non-obtrusive health monitoring. In *2009 International Symposium on Wearable Computers*. IEEE, 153–154.

[43] Swadhin Pradhan, Ghufran Baig, Wenguang Mao, Lili Qiu, Guohai Chen, and Bo Yang. 2018. Smartphone-based acoustic indoor space mapping. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–26.

[44] Jay Prakash, Zhijian Yang, Yu-Lin Wei, and Romit Roy Choudhury. 2019. STEAR: Robust Step Counting from Earables. In *Proceedings of the 1st International Workshop on Earable Computing*. 36–41.

[45] Niklas Röber, Sven Andres, and Maic Masuch. 2006. *HRTF simulations through acoustic raytracing*. Universitäts-und Landesbibliothek Sachsen-Anhalt.

[46] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice localization using nearby wall reflections. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[47] Tzu-Chun Tai, Kate Ching-Ju Lin, and Yu-Chee Tseng. 2019. Toward reliable localization by unequal AoA tracking. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 444–456.

[48] Jelmer Tiete, Federico Domínguez, Bruno Da Silva, Laurent Segers, Kris Steenhaut, and Abdellah Touhafi. 2014. SoundCompass: a distributed MEMS microphone array-based sensor for sound source localization. *Sensors* 14, 2 (2014), 1918–1949.

[49] Edgar A Torres-Gallegos, Felipe Orduna-Bustamante, and Fernando Arámbula-Cosío. 2015. Personalization of head-related transfer functions (hrtf) based on automatic photo-anthropometry and inference from a database. *Applied Acoustics* 97 (2015), 84–95.

[50] J-M Valin, François Michaud, Jean Rouat, and Dominic Létourneau. 2003. Robust sound source localization using a microphone array on a mobile robot. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 2. IEEE, 1228–1233.

[51] Lars Falck Villemoes and Dirk Jeroen Breebaart. 2012. Method and apparatus for generating a binaural audio signal. US Patent 8,265,284.

[52] Jeff Wilson, Bruce N Walker, Jeffrey Lindsay, Craig Cambias, and Frank Dellaert. 2007. Swan: System for wearable audio navigation. In *2007 11th IEEE international symposium on wearable computers*. IEEE, 91–98.

[53] Jens Windau and Laurent Itti. 2016. Walking compass with head-mounted IMU sensor. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5542–5547.

[54] Zhijian Yang, Yu-Lin Wei, Sheng Shen, and Romit Roy Choudhury. 2020. Ear-AR: indoor acoustic augmented reality on earphones. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[55] Guangzheng Yu, Ruixing Wu, Yu Liu, and Bosun Xie. 2018. Near-field head-related transfer-function measurement and database of human subjects. *The Journal of the Acoustical Society of America* 143, 3 (2018), EL194–EL198.

[56] Yanzi Zhu, Yibo Zhu, Ben Y Zhao, and Haitao Zheng. 2015. Reusing 60ghz radios for mobile radar imaging. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. 103–116.

[57] Harald Ziegelwanger, Wolfgang Kreuzer, and Piotr Majdak. 2016. A priori mesh grading for the numerical calculation of the head-related transfer functions. *Applied Acoustics* 114 (2016), 99–110.

[58] DYN Zotkin, Jane Hwang, R Duraiswaini, and Larry 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)*. Ieee, 157–160.

[59] Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis. 2004. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on multimedia* 6, 4 (2004), 553–564.