USAARL-JAOA-AI--2022-41

Behavioral Sound Localization for Video-Derived, Personalized Head-Related Transfer Functions (HRTFs)

Kirsti Connor, Shakti Davis, Hannah Wright, Gabriel Alberts,
Paul Calamia, Health Jones, & Christopher Smalt

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

Human Subject Use

In the conduct of research involving human subjects, the investigator(s) adhered to the policies regarding the protection of human subjects as prescribed by Department of Defense Instruction 3216.02 (Protection of Human Subjects and Adherence to Ethical Standards in DoD-Supported Research) dated 8 November 2011.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
08-09-2022	Author's Initial Manuscript	

4. TITLE AND SUBTITLE

Behavioral Sound Localization for Video-Derived, Personalized Head-Related Transfer Functions (HRTFs)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Connor, K.[1], Davis, S.[1], Wright, H.[1], Alberts, G.[1,2], Calamia, P.[1], Jones, H.[3], & Smalt, C.[1]

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Human Health & Performance Systems Group
MIT Lincoln Laboratory
Lexington, MA, USA

8. PERFORMING ORGANIZATION REPORT NUMBER

USAARL-JAOA-AI--2022-41

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Development Command
Military Operational Medicine Research Program
504 Scott Street
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

USAMRDC MOMRP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

[1]MIT Lincoln Laboratory; [2]Harvard Medical School; [3]U.S. Army Aeromedical Research Laboratory
Submitted to the Journal of the Audio Engineering Society.

14. ABSTRACT

Three-dimensional (3D) audio is commonplace in augmented or virtual reality environments, and is used to simulate the direction of arrival for sounds presented over headphones. The Head-Related Transfer Function (HRTF) enables 3D audio by characterizing how sound is received from a sound source in every direction. Optimal sound localization performance with HRTF-based 3D audio is thought to require personalized HRTFs because the localization cues are dependent on the size and shape of the head and pinnae. In this work, we propose a pipeline to generate HRTFs from high-resolution smartphone videos of the head and ears, and we validate the localization performance of the resulting HRTFs through human-subject behavioral experiments. To create a personalized 3D HRTF, a video was taken of a subject's head and torso, left ear, and right ear using a smartphone camera. Simulations of the acoustic pressure waves at the ear canals were performed in COMSOL using the boundary element method, taking approximately one hour per ear to complete. Preliminary virtual sound localization results were obtained on nine subjects under three listening conditions: open ear, a gold-standard generic HRTF, and the personalized HRTF derived from a smartphone video. Results show that our pipeline-derived HRTFs result in increased localization accuracy in comparison to the generic HRTF for some subjects.

15. SUBJECT TERMS

Personalized HRTFs, sound localization, spatial audio, 3D audio

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	12	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

14. Abstract (continued)
In general, subjects perform across a wide range of accuracy with the generic HRTF, whereas the personalized HRTF has significantly less variability across subjects. This trend is also seen with the percentage of very large errors (VLEs), i.e., errors above 45 degrees. Reduction in VLEs is critical in applications such as hearing protection devices, where such mistakes are currently common but can be particularly detrimental. These results indicate that further improvements to the pipeline have the potential to increase performance accuracy for all subjects.

Behavioral Sound Localization for Video-Derived, Personalized Head-Related Transfer Functions (HRTFs)

KIRSTI CONNOR[1], SHAKTI DAVIS[1], HANNAH WRIGHT[1], GABRIEL ALBERTS[1,2], PAUL CALAMIA[1], HEATH JONES[3], & CHRISTOPHER SMALT,[1]*

*Christopher.Smalt@LL.mit.edu

[1] *Human Health & Performance Systems Group, MIT Lincoln Laboratory, Lexington, MA, USA*
[2] *Speech and Hearing Bioscience and Technology Program, Harvard Medical School, Boston, MA, USA*
[3] *United States Army Aeromedical Research Laboratory (USAARL), Fort Rucker, AL, USA*

Three-dimensional (3D) audio is commonplace in augmented or virtual reality environments, and is used to simulate the direction of arrival for sounds presented over headphones. The Head-Related Transfer Function (HRTF) enables 3D audio by characterizing how sound is received from a sound source in every direction. Optimal sound localization performance with HRTF-based 3D audio is thought to require personalized HRTFs because the localization cues are dependent on the size and shape of the head and pinnae. In this work, we propose a pipeline to generate HRTFs from high-resolution smartphone videos of the head and ears, and we validate the localization performance of the resulting HRTFs through human-subject behavioral experiments. To create a personalized 3D HRTF, a video was taken of a subject's head and torso, left ear and right ear using a smartphone camera. Simulations of the acoustic pressure waves at the ear canals were performed in COMSOL using the boundary element method, taking approximately 1 hour per ear to complete. Preliminary virtual sound localization results were obtained on nine subjects under 3 listening conditions: open ear, a gold-standard generic HRTF, and the personalized HRTF derived from a smartphone video. Results show that our pipeline-derived HRTFs result in increased localization accuracy in comparison to the generic HRTF for some subjects. In general, subjects perform across a wide range of accuracy with the generic HRTF, whereas the personalized HRTF has significantly less variability across subjects. This trend is also seen with the percentage of very large errors (VLEs), i.e., errors above 45 degrees. Reduction in VLEs is critical in applications such as hearing-protection devices where such mistakes are currently common but can be particularly detrimental. These results indicate that further improvements to the pipeline have the potential to increase performance accuracy for all subjects.

0 Introduction

The human brain interprets the sounds in our everyday environment and gives us insight about the location in space where sounds have originated, as well as about the characteristics of the space we occupy [1, 2]. In situations where it is desirable to provide the same spatial information to the auditory brain in a virtual manner, the precise spectral and temporal characteristics of the sound reaching the ear must be simulated [3]. Receiving accurate virtual information about the sound in an environment can dramatically impact the efficacy of next-generation hearing aids [4], hearing protection [5], or the realism of experiences delivered in virtual or augmented reality (VR/AR), the applications for which are ever expanding [6, 7].

To simulate the direction of arrival for sounds presented over headphones (i.e., 3D audio) a head-related transfer function (HRTF) is necessary to recreate amplitude, phase and frequency of an acoustic wave received in each ear for a sound source emanating from any direction [3]. Optimal localization performance with HRTF-based 3D audio is thought to require personalized HRTFs because the localization cues have a complex dependence on the geometry of the head and pinnae [8]. HRTFs typically have been collected using a measurement process in an acoustically anechoic space, with in-ear microphones and loudspeakers arranged around the test subject [3, 9]. Even with rapid methods of data collection and processing for HRTFs covering a large number of directions, the measurement pro-

cess is not amenable to widespread use because of the size, complexity, and cost of the acquisition systems.

An alternative method for generating HRTFs, which has been widely studied, involves creating a 3D digital model of a subject's head (and possibly shoulders), using a numerical simulation technique to compute the acoustic field on the surface of the model, and extracting the simulated acoustic pressure from locations at or near the eardrums of the model from which HRTFs can be derived. Various techniques have been described for capturing the necessary geometry for the 3D model, including laser scanning [10, 11, 12], digital photography [13, 14, 8], and even magnetic resonance imaging [15]. The numerical simulations typically are done with the boundary element method (BEM) [11, 16, 10, 17] because the problem is well suited to a surface- rather than volume-based approach. However, versions of the Finite Element Method [18] and Adaptive Rectangular Decomposition [8] also have been considered.

An alternative approach to HRTF personalization involves the use of anthropometric measurements to guide the process of finding the best non-individualized HRTFs for a listener [19, 20]. This method ultimately may yield useful results, although it is dependent on a sufficiently large set of measured HRTFs which may make it impractical, and perceptual errors related to this approach are not well studied. In regard to behavioral performance in 3D audio tasks, measured personalized HRTFs have been studied by a number of groups [21, 9, 22], some of which show excellent localization performance under certain conditions. However, modeled HRTFs more often are validated either qualitatively or through numerical comparisons to measured HRTFs (using the same subject for both) rather than through behavioral tests [23, 16, 8]. A rare exception is described by [12], where localization performance was evaluated based on modeled HRTFs using both human subjects and a computational localization model. Their results from 3 subjects suggest that measured and modeled HRTFs provide similar localization ability as long as the latter are computed using 3D models with sufficient resolution.

In this paper, we propose a more accessible approach for generating personalized HRTFs that leverages high-resolution video of an individual and transforms the video into a digital 3D model to represent the individual's head, upper torso, and pinnae. We also validate localization performance of these customized HRTFs through human-subject behavioral experiments, which are sometimes done for acoustically-measured HRTFs [24] and rarely for model-based methods. If model-based HRTFs can be shown to be as effective as acoustically-measured HRTFs, their widespread use in applications such as hearing-protection devices will be possible due to the significant reduction in size, cost, and complexity of the equipment required to acquire them.

1 Methods

Our pipeline to produce and evaluate personalized HRTFs is shown in Fig. 1. At a high level, the process is three-fold: creating and preparing a mesh from a smartphone video, generating an HRTF from that mesh, and evaluating the HRTF with behavioral listening tests.

The first step is to capture short-duration, high-resolution videos of the subject's head, torso, and ears using a commercial smartphone. This video is then used to create a 3D triangle mesh of the head, upper torso, and pinnae using a commercially available mesh-generation software. The mesh is then manually scaled and oriented to appropriately represent the subject's head size in preparation for acoustic simulation. Next, the mesh is refined for simulation speed and accuracy using finite element software, which subsequently runs a BEM acoustic simulation on the mesh.

A personalized HRTF is then generated from the simulated complex pressure fields over a range of frequencies and azimuths and the results are stored in the open-source Spatially Oriented Format for Acoustics (SOFA) convention [25, 26] for further use.

Finally, the personalized HRTF is evaluated using behavioral listening tests on the subject. To evaluate the accuracy of the personalized spatial cues captured in the HRTF, the subject listens through headphones to a short sound at one of 24 azimuth angles and must indicate the direction of the sound using a pointer. These listening trials are performed for three testing conditions: the personalized HRTF, a generic HRTF, and an open ear case (no headphones).

1.1 Participant video

The first step in our 3D HRTF generation pipeline is to take high quality videos of a subject's head, torso, and ears. This can be accomplished with a smartphone that is capable of capturing 4K videos. In addition to a camera with sufficient video resolution, a simple uncluttered background and well-lit area are important for successful video collection. To control the lighting and background conditions, we set up a portable tent and distributed 10 sets of florescent lights around the subject.

Participants were provided a tight-fitting wig cap to wear over their hair to minimize loose hair around the ears and provide a smooth contour over the head. Stickers with coded targets were placed around each ear to facilitate alignment across multiple videos. To ensure proper scaling of the resulting mesh, we attached a calibration scale to the subject's forehead.

A total of five short videos are collected as the basis of the mesh generation: three videos are collected keeping most of the torso and head in the frame head, and two videos are collected with close-ups of the ears (one per ear). For the torso videos, the camera person walks slowly in a circle around the subject, holding the camera approximately 24 inches from the subject at three different heights: below the shoulders, even with the ears, and above the ears. Each height is important to capture the head, torso, and ears from different angles. Each of the torso videos is approximately 30 seconds in duration. The two ear videos are collected from a closer position, about 6 inches from the head,

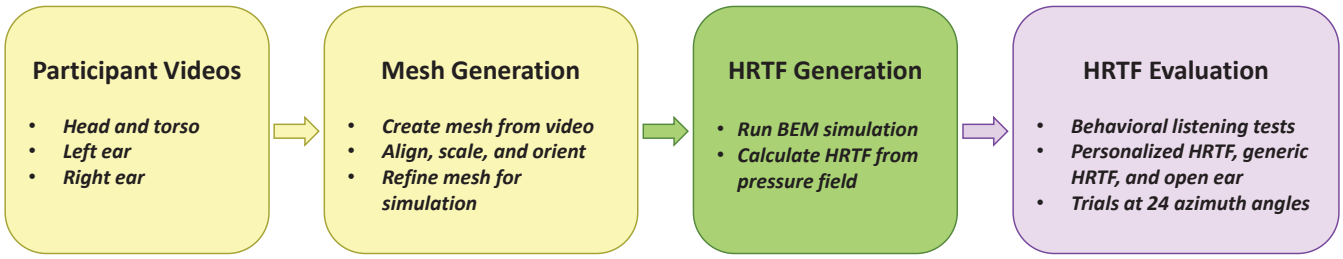

Fig. 1. Our pipeline for generating a smartphone-derived personalized HRTF. The mesh creation is performed in two main steps (yellow boxes): taking videos of the participant and generating a mesh from the videos. The HRTF is generated (green box) from an acoustic BEM simulation performed on the mesh. Finally, the HRTF is evaluated (purple box) via behavioral listening tests performed by the participant.

and the camera is slowly moved through a range of azimuth and elevation angles for a total duration of 45-60 seconds.

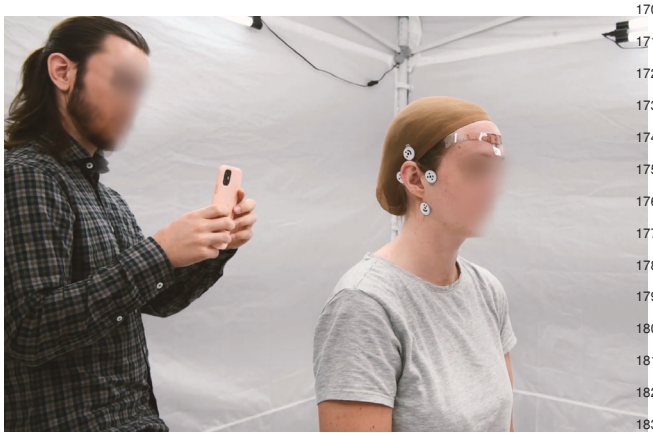

Fig. 2. Taking high-resolution videos of a subject using a smartphone.

1.2 Generation of a personalized mesh

1.2.1 Preliminary mesh

Once a suitable video of the subject is collected, we use commercial software to build a 3D representation of the person's head and shoulders. For the initial mesh generation, we use Agisoft Metashape© Professional. The software uses photogrammetry to triangulate points from multiple angles and build a dense cloud of the subject. We make use of the five separate videos in order to build a multi-resolution point cloud. Approximately 50 images are extracted from each torso video to form a relatively low density cloud of the head and shoulders. Additionally, approximately 150 images are extracted from each ear video and we generate much higher-density point clouds for each ear. Each point cloud is edited to remove extraneous points and the clouds are then merged into a single point cloud using the adhesive markers that were placed around the ears as alignment reference points. Following the merge, we build a preliminary medium-resolution mesh of the head, shoulders and pinnae consisting of approximately 500,000 triangular elements.

Merging the close-in ear models with the torso models provides a good balance between fidelity of the pinnae details, smooth head and torso, and manageable computational load. Using a low-density head/torso point cloud ensures that the overall geometry relating the two ears is correct (head size, symmetry, and placement of the ears) while the individual ear point clouds with higher density yield far more detail on the folds of the pinnae compared to the full-head video.

Next, we scale and orient the mesh. For this step, we take a single face-on image from the video that includes a calibration ruler on the forehead and project this color image onto the mesh to facilitate appropriately scaling the head. The mesh is then manually rotated and shifted to center the ear canals symmetrically along the y-axis with the head along the positive z-axis and the nose toward the positive x-axis.

The resulting scaled mesh (the "Preliminary Mesh" in Fig. 3) is then exported from Agisoft as a stereolithographic (STL) file.

1.2.2 Refined mesh

Following the generation of the preliminary mesh, we use COMSOL© to refine it into a well-defined, final mesh for the acoustic computations. COMSOL can create variable resolution (element density) across different regions of a mesh. It takes the preliminary mesh STL file as a geometry input and re-meshes the geometry with high quality elements, focusing on smaller, densely-sampled elements around the pinnae and larger, coarse elements over the head and torso. Fine-level details of the head and torso regions have little importance for the HRTF, and so using larger elements in these regions reduce computational time. Fine-level details around the pinnae are needed to capture geometric features that inform the HRTF.

The mesh sizing can be set to capture frequency-based details of the response by ensuring the maximum element size encompasses the wavelength of the incident frequency. Although the max element size can be adjusted for each frequency (thus producing a finer and finer mesh for each successive frequency up to 16 kilohertz), this adds significant computational time. For our purposes, we set the maximum element size such that the wavelength at 10 kHz is encompassed. This ensures that details below 10 kHz are captured well, which encompass the majority of sounds experienced by humans, while keeping computational time efficient. This max element size is used for all frequencies,

thus producing only one mesh. We set the minimum element size to 0.5 millimeters, which is 1/4 of the simulated microphone diameter region at each ear canal. This selection prevents unnecessarily small elements while still capturing enough detail over the pinnae geometry.

Since the number of elements in a mesh is proportional to the computational cost associated with running a BEM simulation, it is important to aim for a final mesh that is relatively small while preserving the geometrical details that inform the HRTF. Following the multi-resolution refinement, our final meshes average approximately 50k elements for the BEM simulations, which represents a 10-fold reduction in elements compared to the preliminary mesh.

Fig. 3 shows an example mesh before and after refinement. The images on the left are from the preliminary mesh (Agisoft-generated) and the images on the right are from the simulation-ready, refined mesh (COMSOL-edited). The preliminary mesh clearly has a significantly larger number of elements as compared to the refined mesh. The discrepancy in number of elements is largely due to the use of small elements throughout the entire preliminary mesh, whereas in the refined mesh, the head and torso have been represented by coarser elements while the pinnae elements remain at the finer scale. The close-up images of the ear have been color coded based on a calculated skewness metric of the elements, where green suggests low skewness (well conditioned for numerical analysis) and red indicates high skewness (significant potential for numerical instability). Regions with frequent yellow and red shaded elements indicate that the conditions needed for stable and accurate numerical computations may be violated. The close-up view of the left ear on the preliminary mesh shows that while the geometry of the ear is accurately represented, many of the elements are highly skewed and of low quality. Following the mesh refinement, however, the elements exhibit low skewness as indicated by the predominantly green color. This refined mesh has superior quality for stable, efficient, and accurate numerical computations.

The mesh refinement was performed for nine participants (17 total meshes) in the study. Key properties from the mesh generation are summarized in Table 1.2.2. Close-up images of the pinnae regions of the meshes for three subjects are shown in Fig. 4. These resulting pinnae show some surface imperfections such as extraneous bumps and small features behind the ears, but they generally exhibit smooth, well-defined folds for both ears. Minor residuals from the merging of the torso and ear dense clouds are apparent on some of the meshes, manifesting as raised or indented plateaus surrounding the ears. The overall uniqueness of each set of ears illustrates how different the geometry can be for each individual's ears and thus suggests the importance of personalized models for the HRTFs.

1.3 HRTF simulation

The refined mesh is then used to run an acoustic BEM simulation in COMSOL from which we calculate the personalized HRTF. In the mesh, we define two regions to represent the microphone receiver areas at the approximate entrance of the ear canal on either pinna. We use a 2-mm diameter patch for each region – a size shown to adequately capture key HRTF details such as nulls in the spectrum in a previous study [12]. We assign the material "Air" to the infinite void surrounding the mesh. The head/torso mesh is treated as a rigid boundary with no material properties, allowing the problem to be solved as a pure radiation problem. This makes the use of BEM particularly efficient.

As commonly done when simulating HRTFs, we use the reciprocity principle to greatly reduce computational time and cost in obtaining the full 3D pressure field. The reciprocity principle states that the source and receiver locations can be switched in an acoustic problem while still maintaining the same acoustic response [27, 28]. In this case, we place the source at the ear canals, while the "receiver" is evaluated along a circle (2D) or sphere (3D) at a specified radial distance from the center of the head. This approach allows sound characteristics associated from an arbitrary number of "receiver" positions to be extracted from a pair of simulations (one for each ear) at each frequency.

To run the simulation, we create a normal velocity boundary condition with a velocity of 1 meter per second on either the left or right ear canal (simulation needs to be done for both right and left ears). The solver sweeps over a range of frequencies from 100 hertz to 16 kHz at increments of 100 Hz. The simulation takes about 1 hour per ear to run. After the simulation is complete, for each frequency we extract the complex pressure at 1 degree increments on the azimuthal plane at a radius of 1 meter from the center of the head. In order to calculate the HRTFs, we first need to normalize the output pressure by the input pressure at the source (ear canal), thus removing the effect from the source on the output field. Next, the HRTF is calculated using the SOFA convention [25, 26] as a simple free field transfer function with a sampling rate of 48 kHz.

1.4 Behavioral listening tests

To evaluate the fidelity of spatial-audio cues incorporated in the video-derived personal HRTFs, we perform behavioral sound localization tests. In this task, volunteers listen to a broadband acoustic stimulus injected at a given azimuth angle. We use the audio recording of the cock slide of an AK-47 gun as a sound as the stimulus because of its military relevance as the sound of a threat signal and the short duration, broadband characteristic that allows for good localization under natural hearing conditions [29]. We apply the HRTF to the spectrum of the AK-47 audio

No. participants	9 (5 F / 4 M)
No. images per ear	143 ± 66
Meshes generated per participant	1.9
No. points in dense cloud	$1082k \pm 331k$
No. faces in final mesh	$52k \pm 12k$

Table 1. Summary of properties for video collection and mesh generation. The mean and standard deviation across subjects are reported, where applicable. The letter k indicates the number in thousands.

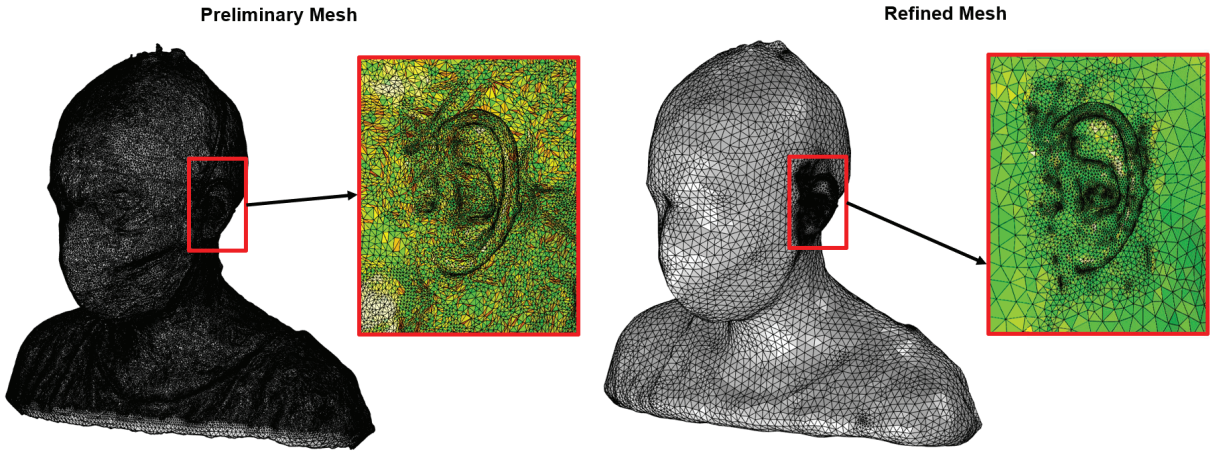

Fig. 3. Example meshes before (left) and after (right) refinement. The detailed insets of the ears show the assessed skewness (a mesh quality metric) for numerical calculations where green represents sufficient quality for stable calculations and yellow and red indicate conditions that may lead to instability.

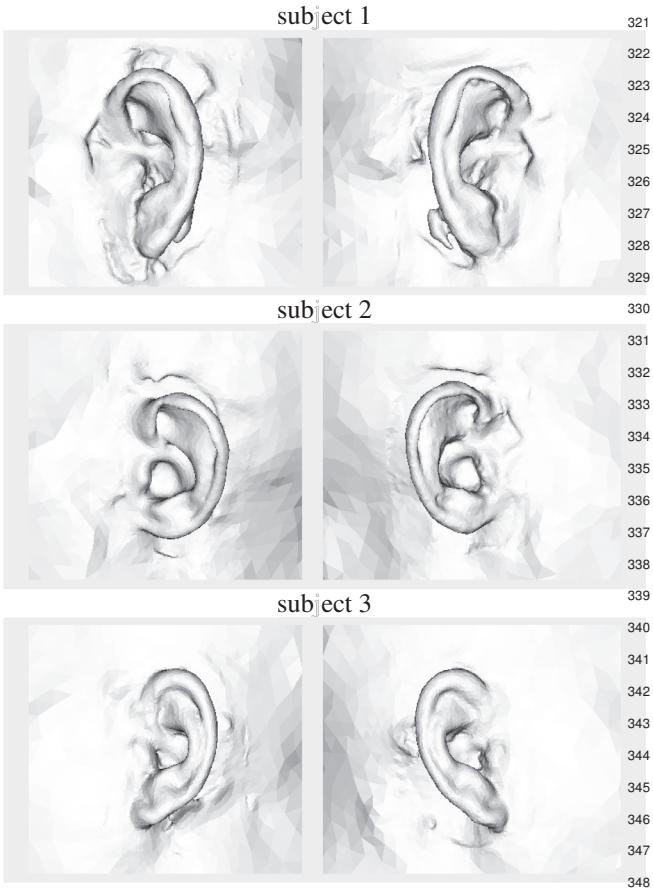

Fig. 4. Resulting ear meshes, after refinement, for three participants. Fine-detailed structure in the pinnae are preserved by this approach.

recording to impart spatialization cues. For this study, we limited the stimulus testing angles to the azimuth plane only at 15 degree increments, at 0 degrees elevation relative to the listener. The listener is asked to point a sensor (Polhemus FASTRAK®) in the direction from which they perceived the sound to originate (Fig. 5).

To benchmark the localization performance, we perform listening tests for each subject for three different spatialization conditions: open ear, generic HRTF (same for all subjects), and personalized HRTF. The open-ear condition serves as the control, allowing each listener to perform sound localization using their natural hearing. In the open ear condition, the stimulus is emitted from one speaker in a ring of 24 speakers surrounding the listener with an inter-speaker angle distance of 15 degrees. The exact location of each speaker is obscured by a black curtain. The ring of speakers is approximately 1 meter away from the subject's head (for more information, see our previous work on the experimental setup [29]). The two HRTF conditions are presented to the listener over headphones. For the generic HRTF condition, we chose an average of many HRTFs known to give good performance across a broad range of listeners, which was provided by the Air Force Research Laboratory (AFRL) [24] and is also included in the open-source OpenVale software package (https://gitlab.com/OpenVALE/).

To familiarize the subject with the process, a training block of 48 trials with the open ear condition is performed first. Then, for each condition (open ear, generic HRTF, personalized HRTF), the subject performs 5 blocks of 48 localization trials, for a grand total of 720 trials. Within a block, the azimuth angles are presented in random order. Between blocks, the order of the three conditions are randomized. Each behavioral listening session lasts approximately one hour with short breaks between blocks as needed. Localization performance is evaluated using two metrics: 1) the mean angle error (MAE) and 2) percent of trials with very large errors (VLEs). MAE is calculated as the mean absolute value of the true versus indicated azimuth in degrees (modulo 360 degrees) across all trials for a condition. A VLE is detected when the absolute value of the azimuth error is greater than or equal to 45 degrees. VLEs are reported as the percentage of trials that resulted in a VLE event for each condition.

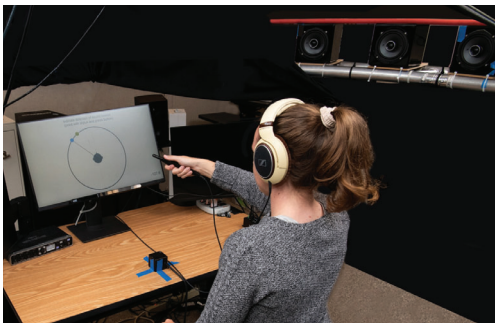

Fig. 5. A subject completes the behavioral listening test by listening to a spatialized stimulus and pointing a sensor at the perceived direction of arrival.

2 Results

2.1 HRTF characterization

Spatialization cues associated with an HRTF can be characterized by the interaural time difference (ITD) and interaural level difference (ILD). The ITDs are predominantly associated with the distance between the ears (interaural distance) and these time differences enable a listener to distinguish between laterally separated stimuli. The ILDs enable localization in elevation, notably allowing some ability to distinguish between front-back angles on an isometric cone where ITDs are equal. Fig. 6 plots (a) and (b) show the ITDs and ILDs from the video-derived personalized HRTFs for all subjects in the study, as a function of azimuthal angle. The ITD responses are very similar across all subjects, with only minor deviations at the extreme right and left angles. This is expected since the interaural distances are within a few millimeters of each other for all subjects. ILDs are shown within three different frequency bands and these spatial cues are seen to have more intersubject variability, particularly in the higher frequency bands. Fig. 6 plots (c) and (d) show the frequency-dependent magnitudes of the personalized HRTFs at zero-azimuth (straight forward), for the left and right ears of each subject in the study. Again, the frequency-dependent magnitudes show significant inter-subject variability including slight differences in the frequency and number of peaks and nulls. These individual differences in the ILD and spectra are thought to contain personalized spatialization cues that invoke the individual's innate cognitive abilities to distinguish where a sound is coming from.

An alternate method for visualizing the HRTF characteristics is to plot the magnitude as a function of azimuth angle and frequency. Examples of these 2D plots are shown in Fig. 7 for the generic HRTF (left) and one subject's personalized HRTF (right). While there are coarse similarities between the generic and personalized HRTFs, this view illustrates that there are distinct characteristics in a personalized HRTF, such as the dominant frequency bands at a given angle and subtle patterns of high-frequency peaks and nulls that are shaped by the details of the individual's pinnae. While asymmetry is not common, the 3D video method for personalized HRTFs also allows any asymmetries that may exist between the left and right pinnae to be translated into the HRTF model.

2.2 Behavioral listening tests

Behavioral data from three representative subjects is compiled in Fig. 8. Each plot shows the azimuth angle where the subject estimated the stimulus to originate as a function of the azimuth angle where the stimulus was designed to originate. For the open ear condition, this latter angle, labeled "Actual Azimuth" in the plots, refers to the angle of the one speaker in the 24-speaker ring that played the stimulus. For the two HRTF conditions, the HRTF at that angle was extracted and applied to the stimuli to impart distinct spatial cues. In all three conditions, this azimuth angle varied from 0 to 360 degrees in 15 degree increments. The estimated azimuth is reported to 0.1 degrees in accordance with the accuracy of the Polhemus tracker subjects used to point in the direction they perceived the stimulus to originate.

The dark gray line along y=x in each plot mark perfect responses where the estimated angle matches the designed stimulus angle of origin. The dashed gray lines mark responses where the subject committed front-back errors and mistakenly perceived the sound to originate in front of them when it actually was designed to originate behind them or vice versa. Each plot presents data from all 720 trails the subject completed, which are subdivided into the three independent conditions. Individual differences were noted with some subjects showing improved performance using the generic HRTF over the personalized HRTF, others showing moderate performance with both, and others who performed better using the personalized HRTF over the generic HRTF.

Results of the behavioral tests for a total of 17 sessions conducted on 9 distinct individuals are shown for the three conditions (open ear, generic HRTF, and personalized HRTF) in Fig. 9. The overall mean angle error (MAE) and standard error are summarized in the left panel, where each dot represents the MAE from one session and the size of the dot indicates the relative standard error within that session. As expected, MAEs are lowest when the sounds are played through the sound ring and received at an open ear. The mean azimuth errors in this control case are approximately 10° with very low standard error across subjects. This performance is consistent with open ear localization performance from other studies. When localizing a sound with the generic HRTF, the MAE increases to 18.2° and is highly dependent on the subject with a inter-subject standard deviation of 8.7°. When localizing a sound with the personalized HRTF, the resulting MAE = 18.5° is comparable to the generic HRTF condition, but performance is far more consistent across subjects with an inter-subject standard error of 5.4°. The reduced variability across subjects suggests the pipeline developed in this project (top right panel) produces a consistent quality of auditory cues for each individual. While the localization performance for the personalized HRTF based on the smart phone does not achieve performance of the open ear condition, Fig. 9b

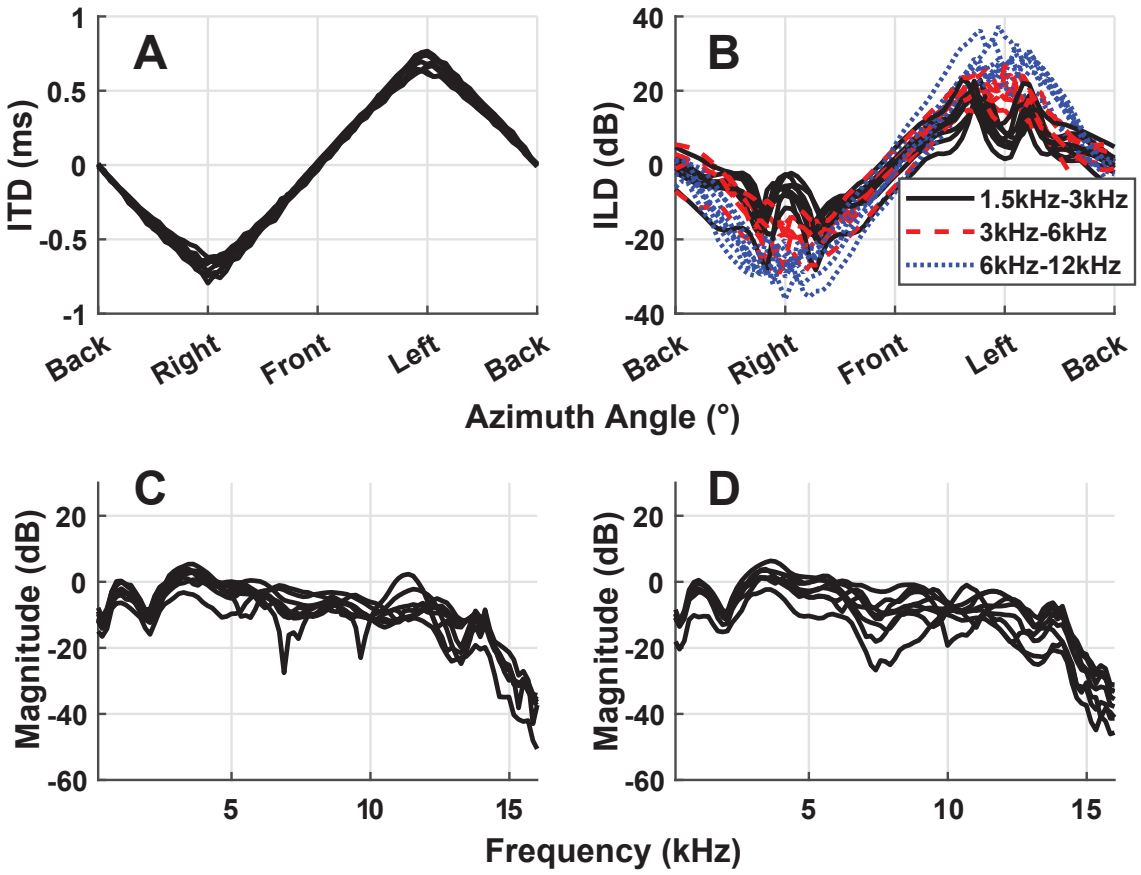

Fig. 6. Overlaid HRTF characteristics for the nine unique participants that make up the dataset. (a) ITDs, (b) ILDs in three audible frequency bands, and spectral magnitude plots for a zero-azimuth (forward) sound arriving at the (c) right and (d) left ear.

shows the that the MAE when VLEs are removed is comparable for all three conditions. The VLEs are typically representative of front-back reversal errors – a challenging error to overcome and common with most hearing protection devices. Fig. 9c shows the percent of trials resulting in a VLE for each subject and each condition. Approximately half of the subject sessions (9 out of 17) exhibited fewer VLEs using their personalized HRTF versus the generic HRTF. Even though some subjects do not experience improved VLEs with the personalized versus generic HRTF, it is encouraging that inter-subject variability is substantially reduced with the personalized HRTF, yielding consistent front-back errors across the full cohort. This suggests that any further improvements to the 3D modeling and HRTF generation pipeline has the potential to improve localization performance for all subjects.

2.3 Commercially-acquired personalized HRTF

To address the question of how our methodology compares against similar approaches in industry, we sent smartphone videos for three of our subjects to Genelec® Aural ID, a commercially-available service that advertises a similar methodology of using a smartphone video to derive a personalized HRTF [30]. We received Genelec® personalized HRTFs for three subjects in our study and repeated the behavioral localization tests to perform a direct comparison between the commercially produced personalized HRTF and our personalized HRTF. Localization performance for both personalized HRTFs was very comparable across the three subjects; one subject's result is shown in Fig. 10. The MAE across these trials was $18.9^\circ \pm 8.3$ for the commercial HRTFs and $18.3^\circ \pm 6.8$ for our HRTFs. Similarly, the percentage of trials resulting in VLEs was comparable with $8.8\% \pm 9.0$ for the commercial HRTFs and $8.3\% \pm 6.9$ for our HRTFs. For both metrics, we observed a slightly lower variability for our HRTFs than for the commercial HRTFs, while the mean performance for the two sources of personalized HRTFs was nearly identical.

3 Discussion

In this study, we generated video-derived, personalized HRTFs and evaluated localization accuracy through behavioral testing. Results from the study reveal several interesting findings. We now discuss these findings, address limitations in our current method, and discuss potential areas for future work.

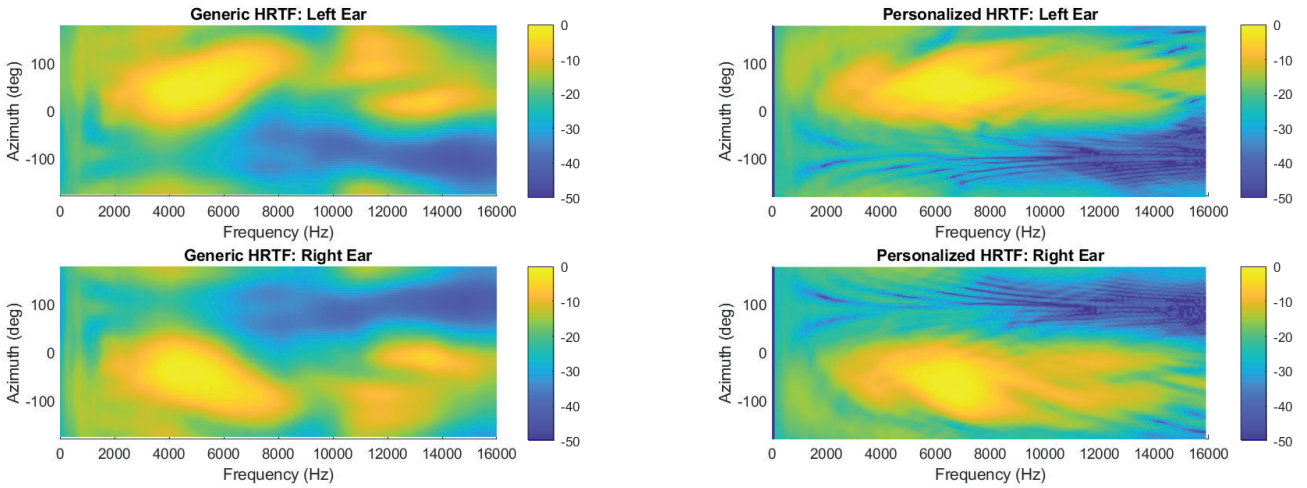

Fig. 7. Example magnitude plots for the generic HRTF (left) and one subject's personalized HRTF (right).

3.1 Findings from behavioral testing

Behavioral testing allows the full pipeline of Fig. 1 to be evaluated by human listeners at a system level. The localization performance results may be influenced by many factors including the video acquisition, mesh generation, simulation or the individual's cognitive ability to identify a sound source. We discuss possible dependencies on some of these factors here, beginning by considering qualitative inspections of the video-derived meshes.

Close-up images of the mesh pinnae for three of the nine subjects are shown in Fig. 4. Subjectively assessing the structure of the ears, the detailed folds of the pinnae appear to be smooth and well-defined; however, some structural defects are seen at the periphery such as bumpiness along the transition from the face to the ears, extraneous features behind the ears, or nonzero inset/offset of the ears into the head. These defects arise from factors such as low-light or shadowed regions behind the ears, sideburns or facial hair near the ears as captured in the video, and mis-alignment when combining the dense clouds from the ears with the head and torso. While noticeable to the eye, these mesh defects do not appear to hinder the performance associated with the HRTF. For example, subject 3 performs exceptionally well on the personalized HRTF with less than 1% VLEs, close to the average performance for open ear, as shown in Fig. 8(c), even though this subject has noticeable bumps in the mesh region behind their ears as shown in Fig. 4. Similarly, Subject 1 has a few extraneous features behind the ears as well as a small inset of the left ear relative to the head and some bumps in front of the left ear from facial hair and wearing a mask, yet their performance on the personalized HRTF is also better than average (2.5% VLEs compared to an average of 7% VLEs). Due to the small sample size here, we can only show anecdotal evidence of this hypothesis that the minor imperfections seen in the video-derived meshes are not limiting the performance. Further examination with many more subjects is needed to truly understand how the pinnae mesh quality impacts HRTF performance.

Next, we consider the influence of spectral HRTF characteristics on the behavioral tests. As illustrated by the ILD and spectral magnitude plots of Fig. 6, the personalized HRTFs show significant variability between subjects indicating that each individual is presented with unique spatialization cues for determining sound directionality. This diversity of HRTF spectral characteristics highlights the potential value of employing personalized HRTFs. Similarly, we observe high variability in the localization performance across subjects during behavioral testing. Fig. 8 shows the wide range of localization performance for three subjects assessing sounds produced with a generic HRTF where a common representation of the spatialization cues is presented to everyone. Two subjects perform relatively well with the generic HRTF ($\leq$ 5.0% VLEs), but Subject 1 has dramatically worse performance with nearly 20% VLEs compared to an average of 7% VLEs. Each subject also assesses localization using their personalized HRTF and we observe much lower variability under the personalized condition (all less than 5% VLEs) for these three subjects. This example demonstrates that while some people have the ability to achieve reasonable localization performance with a generic HRTF, other subjects are unable to adjust to the generic spatial cues, making it difficult to rely on a generic HRTF being sufficient for everyone. Thus, a reasonable strategy to assume when localization performance is important might be to generate personalized HRTFs from a smartphone video and allowing each subject to make their own choice between the personalized or generic HRTF.

Next, we delve further into the details of the violin plots of Fig. 9, which summarize the localization error results for all 17 testing sessions and reveal several key findings from our study. Eight of the nine subjects performed two listening sessions with two separate personalized HRTFs, for a total of 17 testing sessions. In Fig. 9(a), the MAE for the open ear condition shows that all subjects have very low MAE (average around 10°); this is the "ideal" case that all HRTFs strive to match. Neither the generic or

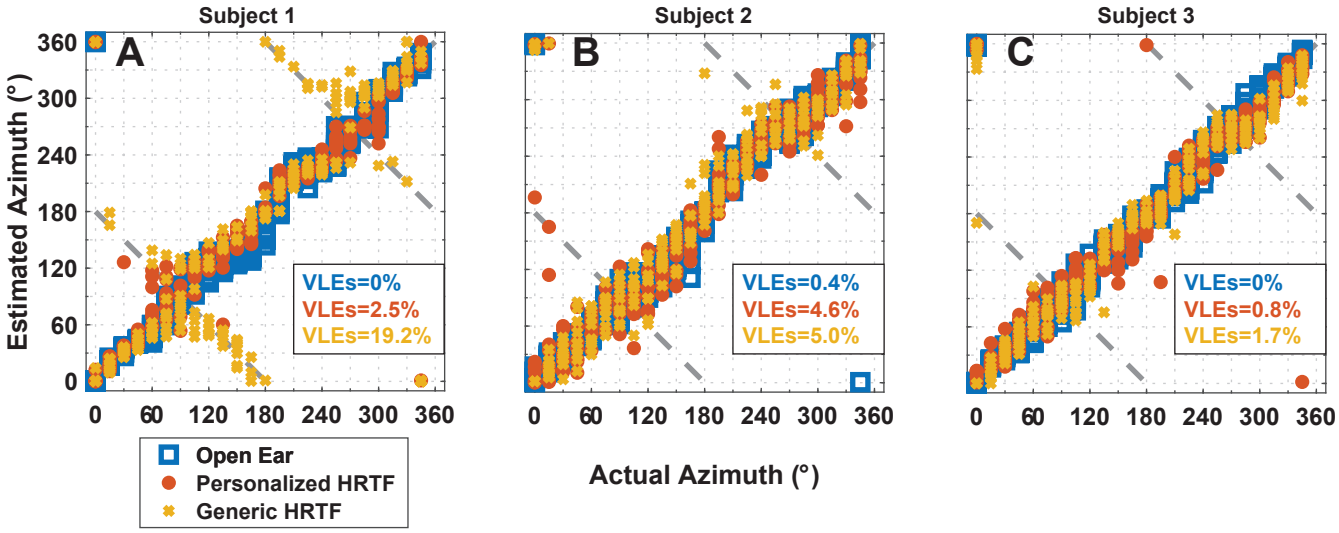

Fig. 8. Azimuth scatter plots for a) a subject who performs well with the personalized HRTF and poorly for the generic HRTF; b) two subjects that perform similarly with the generalized and personalized HRTFs.

personalized HRTFs approach the level of accuracy as the open ear case, with MAE averaging around 18° for both and much larger variability than the generic case. However, the variability (i.e., the vertical span of the shaded area) of the personalized HRTF is significantly tighter than for the generic HRTF. The high variability for the generic HRTF indicates that it is very unpredictable how an individual will perform with it, whereas the tighter variability for the personalized HRTFs suggests that it is more likely that a subject will perform closer to average with a personalized HRTF. This work was focused on generating a straightforward pipeline to produce personalized HRTFs. If this pipeline was improved by optimizing HRTF performance-related variables throughout the process, we would expect the process-related improvements to reduce the mean and, assuming that the low variability is preserved, such improvements could translate to better localization performance for all listeners.

In Fig. 9(b), we remove VLEs from the calculation of MAE for each testing session. The resulting violin plot shows a significantly lower average MAE for the generic and personalized HRTFs at around 12.5° for both. Also, variability for the generic and personalized HRTFs are now very similar and much tighter than in (a). This reveals something important: VLEs are the main drivers of the overall MAE for both HRTFs. This means that stimuli at large angle differences, such as front-back stimuli, are more difficult to distinguish with an HRTF than with open ear. This is seen also in Fig. 8; most angle errors occur along the dotted lines (front and back angles). Being able to determine whether a sound or threat is coming from in front of or behind you is critical in a scenario such as a military setting. Small angle errors might not be as problematic operationally since a person has other senses to fine tune directionality. Thus, reducing VLEs with HRTFs is critical to improve their utility. The result in Fig. 9(b) also corroborates the findings of the ITDs and ILDs. That is, the ITDs are very similar across subjects but the ILDs are variable, translating to the subjects performing similarly across azimuth but differently in elevation (front-back). This is exactly what we observe from the behavioral tests.

Fig. 9(c) explicitly shows the percentage of VLEs for each scenario. There are virtually no VLEs for the open ear tests. Again, the generic and personalized HRTFs have the same average VLEs (around 7%), but the personalized HRTFs show significantly reduced variability compared to the generic HRTF. For our subjects, the highest percent VLE in the generic case was almost 25%, but it was only half that at 12% for the personalized case. This shows that a subject using a personalized HRTF is much more likely to have lower VLEs, such as operationally critical front-back errors, than a generic HRTF. This highlights the value of these personalized HRTFs.

In order to see how our HRTFs compare to a those from a current commercial source, we sent the videos for three of our subjects to a company which produces video-generated personalized HRTFs. These three subjects performed behavioral testing sessions with both our pipeline-generated HRTF and the commercial HRTF. All three subjects performed very similarly with our pipeline-generated HRTF and the commercial HRTF. The results for one subject are shown in Fig. 10. In this case, the VLEs are equal for both HRTFs. This indicates that our pipeline-generated HRTFs are at least as accurate as current commercially-available HRTFs. If our pipeline was improved, our personalized HRTFs may produce HRTFs with higher accuracy over the current state-of-the-art.

3.2 Current limitations and future directions

Our current pipeline has certain limitations that could be improved upon to produce HRTFs with higher accuracy. We are not currently able to acquire anechoic, acoustically-measured HRTFs in our laboratory for each subject to compare with our video-derived HRTFs. Acoustically-

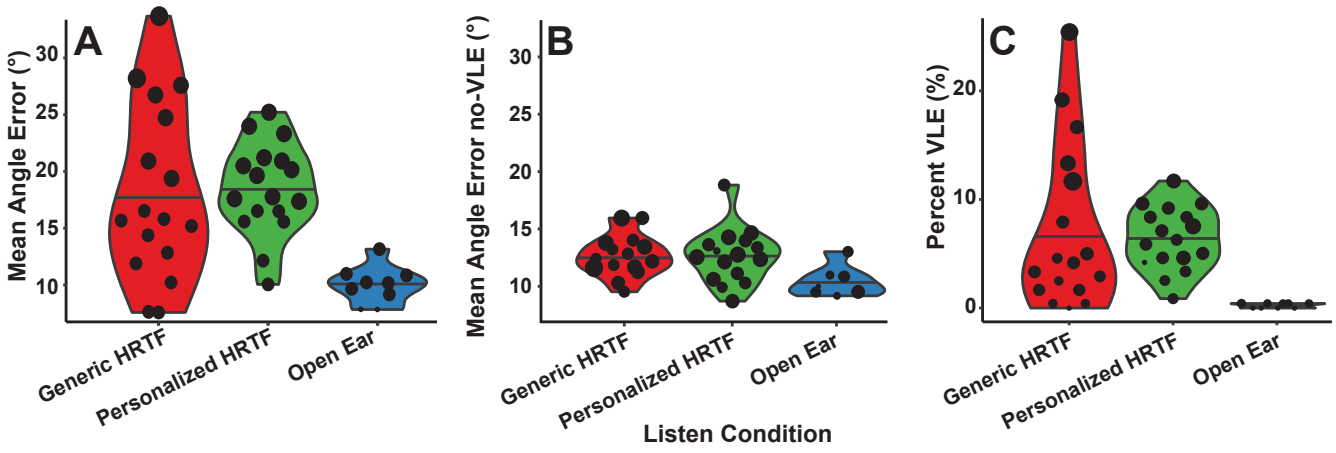

Fig. 9. Violin plots showing the distribution of errors across the behavioral tests for 17 testing sessions. There are 17 dots within each violin, representing the mean result for each testing session and the size of the dot representing the relative standard error in that session. The horizontal bar in each violin is the average of all of the dots in that violin. The width of the violin indicates the relative density of dots at each vertical location.

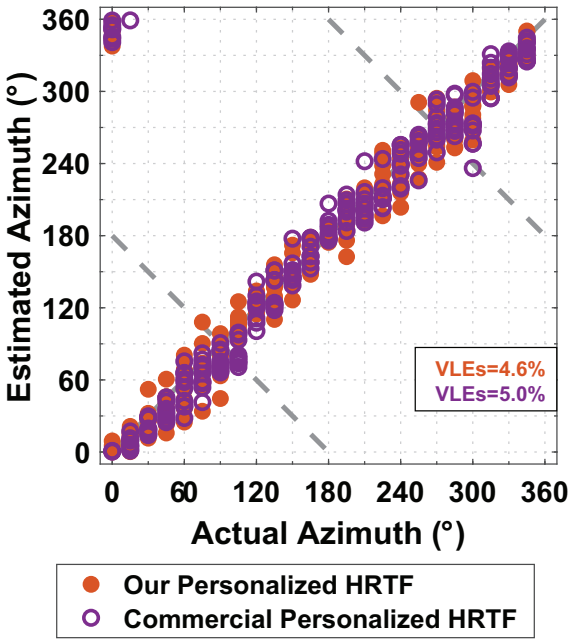

Fig. 10. One subject's performance comparison of two personalized HRTFs derived from the same set of videos. The two HRTFs are our pipeline-derived HRTF and a personalized HRTF acquired from a commercial source. The scatter plot of the behavioral localization results for the subject shows comparable localization performance for the personalized HRTFs. This result was also seen for two other subjects.

measured HRTFs are currently thought to produce the most accurate HRTF for a certain subject, so a comparison with acoustically-measured HRTFs could reveal more about the quality of our personalized HRTFs. Follow-on work at a facility with this capability is in progress, to compare our video setup with an acoustically-measured HRTF. We did not employ a headphone-correction transfer function of any kind, but chose headphones with a relatively flat frequency response out to 16 kHz.

Another current limitation of our pipeline is that we only have performed behavioral tests in the azimuth plane. Assessing our HRTFs by varying elevation would provide a more complete picture of the quality of our personalized HRTFs. In particular, observing their accuracy for large angle errors in directions other than front-back (such as up-down) would be important. A future direction for this work would be to perform 3D behavioral testing using OpenVALE software [24] in virtual reality.

One interesting feature we observed from our behavioral testing was high variability in performance between sessions for the same subject on the same generic HRTF. This indicates that there could be confounding variables in the test setting itself that limit the repeatably of how well a person can localize a sound. For example, how tired a person is, what time of day the test is taken, testing fatigue, or other environmental factors may produce quite different results between testing sessions for the same HRTF and person. This limitation actually inhibited our ability to compare potential correlations between pipeline variables and a person's performance on a personalized HRTF. We attempted to correlate, for example, the torso length in a mesh to a person's performance on its resulting HRTF. This is why eight subjects performed two testing sessions: for each subject, two personalized HRTFs were generated for meshes where torso length was the only difference between them. We were not able to find a clear correlation, and it could be because of the environmental factors affecting performance between testing sessions.

There are several factors in our pipeline which may significantly affect the accuracy of an HRTF. In the video taking process, these could include average and standard deviation of the image quality, length of videos, lighting and shadows, and sweep angle for the ears. In the mesh generation process, these could include number of tie points, dense cloud size for the torso and ears, length of torso present, final number of elements, minimum and maximum element size, mesh smoothness, pinna definition, and extra-

neous features. To improve the pipeline, a good idea would be to perform a follow-up study where these variables are explored in their relation to resulting HRTF behavioral performance. To mitigate the effect of between-session variability, subjects could test multiple conditions in a single test session. Identifying and improving the most important variables in our pipeline could significantly increase the resulting HRTFs' accuracy and thus utility.

4 Conclusion

To characterize 3D audio accurately for an individual, personalized HRTFs are believed to be necessary. Historically, cumbersome measurement-based methods were required to generate personalized HRTFs. In this paper, we present a prototype pipeline to generate personalized HRTFs efficiently from smartphone videos. We characterize their validity via behavioral sound localization tests, which has rarely been done for modeled HRTFs. Results from the behavioral tests show increased localization accuracy for our personalized HRTFs in comparison to a gold-standard generic HRTF. The personalized HRTFs show lower MAE variability, indicating that subjects reliably show good performance with a personalized HRTF, instead of the broad range of performance across subjects with the generic HRTF. Also, a significantly tighter variability of percent VLEs with the personalized HRTFs highlight their value, because reduction of VLEs is critical for HRTFs to be utilized operationally where mistakes like front-back errors can be particularly detrimental. Future work includes improving our pipeline via 3D behavioral tests, comparison to acoustically-measured HRTFs, and optimization of variables that may significantly affect HRTF performance.

5 Acknowledgments

The authors would like to thank Bob Dunn for his assistance with the acoustic presentation software, and Griffin Romigh for discussions on technical aspects of personalized HRTFs.

6 Disclaimer

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Department of the Army under Air Force Contract No. FA8702-15-D-0001.

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army/Navy/Air Force, Department of Defense, or U.S. Government position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army/Navy/Air Force endorsement or approval of the use of such commercial items.

7 REFERENCES

[1] D. R. Moore, "Anatomy and physiology of binaural hearing," *Audiology*, vol. 30, no. 3, pp. 125–134 (1991).

[2] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865 (2016).

[3] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321 (1995).

[4] P. Avan, F. Giraudet, and B. Büki, "Importance of binaural hearing," *Audiology and Neurotology*, vol. 20, no. Suppl. 1, pp. 3–6 (2015).

[5] P. Calamia, S. Davis, C. Smalt, and C. Weston, "A conformal, helmet-mounted microphone array for auditory situational awareness and hearing protection," presented at the *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 96–100 (2017).

[6] B. Xie, *Head-related transfer function and virtual auditory display* (J. Ross Publishing, 2013).

[7] M. Cohen and J. Villegas, "Applications of Audio Augmented Reality: Wearware, Everyware, Anyware, and Awareware," *Fundamentals of Wearable Computers and Augmented Reality,*, pp. 309–330 (2016).

[8] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha, "P-HRTF: Efficient personalized HRTF computation for high-fidelity spatial sound," presented at the *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 53–61 (2014).

[9] D. S. Brungart, G. Romigh, and B. D. Simpson, "Rapid collection of head related transfer functions and comparison to free-field listening," in *Principles and Applications of Spatial Hearing*, pp. 139–148 (World Scientific, 2011).

[10] N. A. Gumerov, A. E. O'Donovan, R. Duraiswami, and D. N. Zotkin, "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation," *The Journal of the Acoustical Society of America*, vol. 127, no. 1, pp. 370–386 (2010).

[11] B. F. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2440–2448 (2001).

[12] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 208–222 (2015), doi: 10.1121/1.4922518.

[13] M. Dellepiane, N. Pietroni, N. Tsingos, M. Asselot, and R. Scopigno, "Reconstructing head models from photographs for individualized 3D-audio processing," presented at the *Computer Graphics Forum*, vol. 27, pp. 1719–1727 (2008).

[14] S. Harder, R. R. Paulsen, M. Larsen, and S. Laugesen, "A three dimensional children head database for acoustical research and development," presented at the *Proceedings of Meetings on Acoustics ICA2013*, vol. 19, p. 050013 (2013).

[15] C. T. Jin, P. Guillon, N. Epain, R. Zolfaghari, A. Van Schaik, A. I. Tew, *et al.*, "Creating the Sydney York morphological and acoustic recordings of ears database," *IEEE Transactions on Multimedia*, vol. 16, no. 1, pp. 37–46 (2013).

[16] W. Kreuzer, P. Majdak, and Z. Chen, "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1280–1290 (2009).

[17] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2hrtf: Open-source software package for the numerical calculation of head-related transfer functions," presented at the *22nd International Congress on Sound and Vibration* (2015).

[18] T. Huttunen, E. T. Seppälä, O. Kirkeby, A. Kärkkäinen, and L. Kärkkäinen, "Simulation of the transfer function for a head-and-torso model over the entire audible frequency range," *Journal of Computational Acoustics*, vol. 15, no. 04, pp. 429–448 (2007).

[19] D. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," presented at the *2003 IEEE workshop on applications of signal processing to audio and acoustics (IEEE Cat. No. 03TH8684)*, pp. 157–160 (2003).

[20] H. Fayek, L. van der Maaten, G. Romigh, and R. Mehra, "On data-driven approaches to head-related-transfer function personalization," presented at the *Audio Engineering Society Convention 143* (2017).

[21] F. Wightman and D. Kistler, "Measurement and validation of human HRTFs for use in hearing research," *Acta acustica united with Acustica*, vol. 91, no. 3, pp. 429–439 (2005).

[22] G. D. Romigh and B. D. Simpson, "Do you hear where I hear?: Isolating the individualized sound localization cues," *Frontiers in Neuroscience*, vol. 8, p. 370 (2014).

[23] B. F. Katz, "Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements," *The Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2449–2455 (2001).

[24] G. D. Romigh, J. Ayers, J. Dube, and A. Horvath-Smith, "OpenVALE: An open-source virtual environment for auditory localization," presented at the *Proceedings of Meetings on Acoustics 173EAA*, vol. 30, p. 050015 (2017).

[25] J. Blauert, J. Braasch, J. Buchholz, H. S. Colburn, U. Jekosch, A. Kohlrausch, *et al.*, "Aural assessment by means of binaural algorithms- The AABBA project-," presented at the *Proceedings of the International Symposium on Auditory and Audiological Research*, vol. 2, pp. 113–124 (2009).

[26] "SOFA (Spatially Oriented Format for Acoustics)," (2022 May), URL `https://www.sofaconventions.org/`.

[27] A. D. Pierce, *Acoustics: an introduction to its physical principles and applications* (Springer, 2019).

[28] "Head and Torso HRTF Computation Application ID: 75011," URL `https://www.comsol.com/model/head-and-torso-hrtf-computation-75011`.

[29] C. J. Smalt, P. T. Calamia, A. P. Dumas, J. P. Perricone, T. Patel, J. Bobrow, *et al.*, "The Effect of Hearing-Protection Devices on Auditory Situational Awareness and Listening Effort," *Ear and Hearing*, vol. 41, no. 1, pp. 82–94 (2020).

[30] "Aural ID - Genelec," URL `https://www.genelec.com/aural-id`.

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command