Evaluation of Individualized HRTFs in a 3D Shooter Game

Jonas Siim Andersen
Department of Architecture, Design & Media Technology
Aalborg University
Copenhagen, Denmark
jsan16@student.aau.dk

Riccardo Miccini
Department of Architecture, Design & Media Technology
Aalborg University
Copenhagen, Denmark
rmicci18@student.aau.dk

Stefania Serafin
Department of Architecture, Design & Media Technology
Aalborg University
Copenhagen, Denmark
sts@create.aau.dk

Simone Spagnol
Faculty of Industrial Design Engineering
Delft University of Technology
Delft, Netherlands
s.spagnol@tudelft.nl

***Abstract*—Previous research stresses the importance of Head-Related Transfer Function (HRTF) individualization approaches for accurately locating sound sources in virtual 3D spaces. However, in the realm of interactive experiences, methods for assessing whether individualized HRTFs bring a benefit to the player experience are rarely investigated. Methods to improve spatial audio rendering are needed now than ever since Virtual Reality (VR) is becoming a mainstream technology for interactive experiences. This paper proposes a method of using in-game metrics to test the hypothesis that individualized HRTFs improve the experience of both expert and novice players in a First-Person Shooter (FPS) game on a desktop environment. The FPS game provides players with a localization task across three different audio renderings using the same acoustic spaces: stereo panning (control condition), generic binaural rendering, and individualized binaural rendering. Collected metrics from the game include localization error, spatial quality attributes, and an extensive questionnaire. The individualized HRTFs for each participant were synthesized using a hybrid structural model. The model employs a deep learning architecture to synthesize a pinna-related response from a pinna image, and combines it with a measured generic head-and-torso response. The interaural time difference (ITD) is then adjusted to match that of an HRTF dataset subject minimizing a localization error metric. The results show that the 22 participants performed significantly better in the localization task with their individualized HRTF. Increased localization accuracy with respect to the generic HRTF was recorded both in azimuth and elevation perception, and especially in the case of expert game players.**

***Index Terms*—3D audio for gaming, HRTF individualization, First-Person Shooter**

I. Introduction

Gaming experiences are reaching high fidelity implementations that provide stunning audiovisual results. For the current generation of gaming technologies, there has been a focus on audio technologies, with Sony focusing specifically on 3D audio for their latest home entertainment system, the *Playstation 5*. To meet this end, they introduced the *Tempest 3D AudioTech*[1] engine, which is still being expanded upon. As part of future updates, engineers at Sony are investigating the possibility of bringing individualized HRTFs to the system. Furthermore, the popular game engine *Unreal Engine* (UE) is being updated to its fifth version, which will include improvements in the audio rendering such as ambisonic soundfields[2] to simulate a 3D soundscape. Therefore, consumer needs currently call for accurate yet inexpensive methods to implement high-fidelity 3D sound in games.

Katz et al. crafted a VR experience [1] whereby HRTFs selected by individual players were evaluated in an environment where the player had to localize virtual sound targets around them. Their research investigated whether the accuracy of the player is affected by the choice of HRTF as the game becomes more difficult to play. However, not much consideration was placed into scalable binaural systems that could be applied to games. Furthermore, HRTF individualization techniques are becoming increasingly mature. For example, newer and efficient models for retrieving personalized measurements using deep neural networks (DNNs) have been investigated [2]–[4] but are yet to be applied to gaming experiences. This paper proposes the application of an individualized HRTF model to a gaming experience in the form of a First-Person Shooter (FPS) video-game to investigate the hypothesis that individualized HRTFs can improve both user performance and the quality of their experience compared with generic HRTFs.

The paper is organized as follows: Section II explores relevant literature on binaural rendering for video games and on HRTF rendering and assessment. Section III introduces our HRTF individualization model, while sections IV and V

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 797850.

[1]https://blog.playstation.com/2020/09/01/devs-speak-how-ps5-consoles-ultra-high-speed-ssd-and-tempest-3d-audiotech-engine-will-enhance-the-future-of-gaming/

[2]https://www.digipen.edu/showcase/news/unreal-engine-5-audio-programmer-max-hayes-on-digipen-projects-that-prepped-him

showcase the design of the FPS game as well as the experiment to test the hypothesis that this paper sets forth. Finally the results are gathered, discussed and concluded in the last sections VI, VII, and VIII, respectively.

II. Related work

A. Understanding the importance of spatial audio in games

Spatial audio in games not only provides a more immersive experience for players, it also helps them navigating the in-game environments [5]. Within a 3D environment, a mix of visual and auditory cues proves to be most effective, with the latter giving the player a reference to the visual stimuli thereby confirming their observations.

Additionally, spatial audio in games can be employed in the sonification of in-game processes, whereby information stored within the game can be visually disconnected by instead presenting it through a series of explicit sounds that influence the player's choices. For instance, a game could inform players through sounds that they are low on health, so that when they hear something dangerous approaching them from the distance, they are prompted to play defensively [6].

When implementing interactive experiences such as video games, virtual reality, or augmented reality (AR), different models for spatial audio can be used, depending on the experience being conveyed as well as the available hardware specifications. The simplest form of spatial sound consists in attenuation-based curves which alter both the sound level and the left/right channel balance depending on the player's position relative to the audio source. The attenuation-based curve can be further enhanced by applying an HRTF, which enhances both horizontal and vertical auditory localization. Physics-based approximations of the virtual space within the virtual environment can also be realized with the support of techniques such as ray tracing or beam tracing [7].

In video games, sound localization cues have found widespread adoption. Competitive shooter games rely heavily on sound cues as a means for the player to locate enemies or objectives. Games such as *Counter-Strike: Global Offensive* and *Half Life: Alyx*, both developed by Valve, employ the Steam Audio API to both model sound propagation and perform binaural rendering. In 2017, as part of an update of Counter-Strike, a highly competitive shooter game, Valve introduced a generic HRTF that players can manually enable. Although initially met with skepticism, it has been eventually embraced by the players' community, thanks to its continuous improvement leading to a better experience for the users [6]. Conversely, Half Life: Alyx is a single-player experience played within a VR environment. In this case, the sound is instrumental for immersing the player into the in-game world, where head tracking technology allows for the interaction with agents such as non-playable characters outside of the player's field of vision.

The rather complicated process of measuring individual HRTFs prompted modern games with binaural audio implementation to exclusively rely on generic HRTF sets. Thus, the question remains on whether localization performances can be further improved by employing a customized HRTF based on the player's individual anthropometry.

B. Generic, individual, and individualized HRTF rendering

HRTFs derived from generic subjects such as dummy heads often result in localization errors and inaccurate spatial perception [8]. In fact, while generic HRTFs and other audio rendering techniques can approximate the interaural cues involved in horizontal localization, the monaural cues needed to discern vertical direction are highly dependent on the anthropometric characteristics of the individual ear [9].

In a work by Møller et al. [8], test participants were exposed to individual and generic HRTF-filtered stimuli. From their data, an increased number of error in the generic HRTF condition was recorded for both nearby and distant sound sources in the median plane, suggesting that an individualized binaural profile can improve sound localization performances. Similarly, Wenzel et al. [10] found an increase in front-back confusion and overall degraded elevation perception when using generic HRTFs, though they argue that their test participants maintained a solid grasp of directional information with generic HRTFs.

While individual HRTFs obtained through acoustic measurements provide the most accurate localization experience possible, they can prove quite impractical due to the need for dedicated facilities, tools, and the overall invasiveness of the procedure. Over the past decades, several HRTF individualization techniques have been devised in order to avoid the burden of conducting strenuous acoustical measurements with human subjects. These techniques consist in selecting, adapting, or synthesizing an HRTF set that best suits a given listener, on the basis of their anthropometry or perceptual feedback; an extensive review of HRTF individualization methods is provided by Guezenoc and Seguier [11].

C. Assessing spatial audio

While generic and individual HRTFs show interesting results in comparison, considerations should also be made into how the quality of the synthesized sounds is perceived when comparing different HRTF renderings. Nicol et al. [12] argue that the multitude of approaches to simulate real-life audio listening within virtual environments using spatial audio calls for a more in-depth look on how to assess the difference in the models, especially for binaural audio that has quite subjective results depending on the used HRTF set.

Oftentimes, HRTFs are solely evaluated on the basis of localization accuracy, thereby neglecting other properties such as timbre. The latter could be evaluated with the Basic Audio Quality (BAQ) model, where a degradation in the given signal can be measured with respect to a predefined reference. However, BAQ was developed to evaluate audio codecs rather than assessing HRTFs. Additionally, the traditional BAQ metrics do not account for errors or inaccuracies introduced by the binaural content, such as small errors in measuring individual HRTFs. There is no steadfast theoretical proof either, that an individual HRTF will provide the best possible audio experience for the user. What is required, instead, is a means

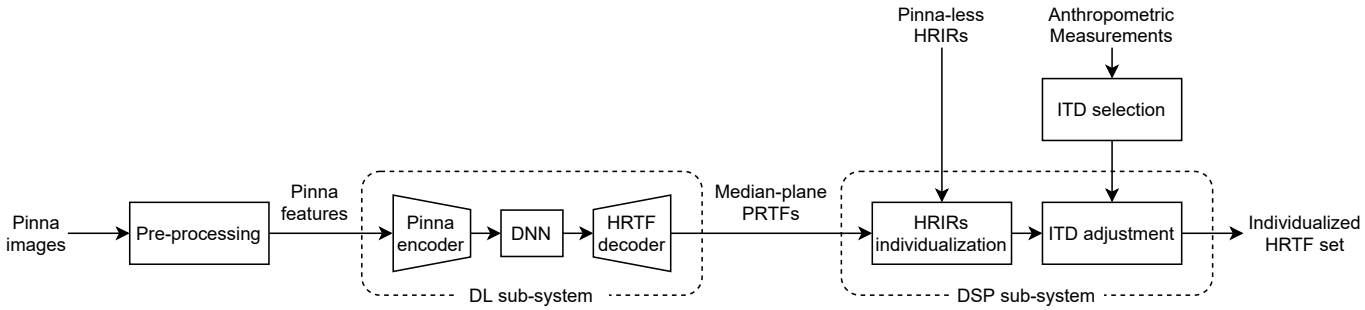

Fig. 1. HRTF individualization system. Figure reproduced from [13].

to evaluate binaural rendering systems that does not measure localization accuracy alone.

Nicol et al. propose a Quality of Experience (QoE) metric to measure the user's subjective stimuli and experiences within the system. In a gameplay scenario, the QoE could be measured by the actions committed by the player in different acoustic spaces. The user would for example expect a very reflective acoustic space, if they are making a lot of noise in an open hall. Other subjective data on the perceived quality of spatial sound reproduction may comprise attributes that can be ranked. One frequently used attribute is whether the sound reproduction is experienced naturally or artificially among listeners [14]. Further anchors that describe the attributes of spatial sound quality may include brightness, richness, externalization and preference [15].

III. HRTF INDIVIDUALIZATION

The technique employed to individualize the HRTFs evaluated herein is based on a recently proposed structural HRTF model [13]. The model combines deep learning (DL) and conventional DSP sub-systems, and features synthesized, selected, and measured components. Figure 1 provides an overview of these elements. The sub-system synthesizes pinna responses (PRTFs) from an image of the pinna or analogous 2D features, such as pinna edges. The DSP sub-system implements a structural model whereby an HRTF set comprising only of shoulder and head reflection effects is filtered using the aforementioned PRTF and processed to match the interaural time difference (ITD) of a subject from an HRTF database.

The system has been implemented in Python and MATLAB, and is capable of generating an individualized HRTF set in a format compatible with most modern binaural rendering engines[3]. Compared with existing structural modeling solutions, our model features a measured component and a PRTF synthesis technique based on DL. This latter component relies on the assumption that, using a 2-dimensional representation of the pinna, it is possible to derive features that correlate with the spectral characteristics of the PRTF. The solution has the advantage of being relatively easy to run while requiring a small amount of data from the user. The following subsections elaborate on the model.

[3]www.sofaconventions.org

A. PRTF synthesis

The DL sub-system comprises three building blocks: a *variational autoencoder* (VAE) used for deriving a compact representation, called z_{ear}, from 2D features such as pinna contours; a *conditional variational autoencoder* (CVAE) used for synthesizing a pinna response from its compact representation, called z_{hrtf}; a *deep neural network* (DNN) mapping the compressed representation z_{ear} into z_{hrtf}. These models are trained separately using their respective datasets and subsequently combined into a prediction script, capable of generating an individualized PRTF. The prediction script can be used independently from the training code base and only requires the pre-trained model weights to work. The following subsections cover the theoretical background as well as the training process for each of the models.

1) Pinna images autoencoder: This model derives a compressed encoding, known as *latent representation*, which can be later used as predictor for synthesizing HRTFs. We employed a variant of autoencoders called *variational autoencoders* (VAE). VAEs are probabilistic models mapping an input sample to a probability distribution and are trained to minimize a reconstruction metric as well as enforce an isotropic Gaussian distribution of the latent space. We trained the model on three distinct datasets, consisting of grayscale images, depth maps, and contours of the pinnae. The first two datasets were programmatically rendered from the 3D head meshes of 55 HUTUBS [16] subjects; the contours were extracted from the depth map dataset using a Canny edge detection algorithm. In order to augment the datasets, we introduced slight variations in the virtual camera angle and applied different types of noise.

2) Encoding of HRTFs: This second model is tasked with autoencoding HRTF magnitude responses with the purpose of reconstructing them from their compressed representation or generating new ones from arbitrary points in the latent space. Since HRTFs depend on both the anthropometry of the user and the elevation and azimuth angles under consideration, we use a variant of VAE called *conditional variational autoencoder* (CVAE), where the output data can be conditioned by a given spatial coordinate [17]. The training data consisted of pairs of HRTF logarithmic magnitude responses — derived from the impulse responses found in the SOFA files — and data labels. We trained the model on two variations of the HUTUBS

dataset: one containing HRTFs across the entire spatial grid, and one based on median-plane data only.

3) Prediction of encoded representations: The last model performs a simple supervised learning task, predicting the encoded representation of the HRTFs from the encoded representation of the 2D pinna features. This is believed to be possible because, in order to faithfully reconstruct its input, the pinna image VAE must encode information pertaining to the individual morphology of the pinnae within its latent dimensions, which can be used as predictors of the HRTF response. To train this network, we extracted latent vectors corresponding to each available HRTF and pinna image using the encoders of the previous two models, and fed them into the models along with the spatial coordinate of the target HRTF.

We tested several training strategies, such as different combinations of input features, spatial coordinates expressed with an interaural-polar system, and principal components of the encoder representation of the HRTF. The best-performing strategy was selected by evaluating the reconstruction performances, in terms of *spectral distortion*, on four unseen HUTUBS subject.

B. HRTF set generation

This part of the model consists in generating an entire HRTF set based on the previously synthesized PRTFs. Since the predictors used in the DL sub-system only relate to the pinnae, spectral features and binaural cues that are not caused by external factors have been identified and accounted for separately in the model.

To approximate the effect of head, torso, and shoulders, we use an artificial pinna-less subject from the VIKING dataset [18]. The subject consists of a KEMAR mannequin with its original pinnae removed and the slots filled with a silicone baffle. The resulting spectral features provide subtle localization cues at frequencies below 3 kHz and are therefore useful for localizing narrow-band sounds under that threshold [19]. The pinnae contribution is then applied to the pinna-less data in the time domain, using a minimum-phase IIR filter derived with the Yule-Walker method and matching the magnitude response of the PRTFs. The peaks and notches caused by the pinnae are known to exhibit little deviation across the horizontal direction [20], so only the median-plane PRTFs were used.

Finally, to improve localization across the horizontal direction, the ITD of the generated HRTFs is manipulated to match that of a best-fit HUTUBS subject using a selection algorithm [21]. In this case, a horizontal localization error metric is predicted on the basis of three anthropometric parameters corresponding to head width, head depth, and shoulder circumference. The metric is computed for all HUTUBS subjects, and the one minimizing the error is selected. We then extract the onset delays of the selected subject's HRIRs, interpolate them so as to match the spatial grid employed by the VIKING dataset, and apply them to the generated HRTFs.

For a more comprehensive explanation of the individualization method, as well as a discussion of its performances as derived objectively using spectral distortion metrics and a sagittal-plane localization model, please see [13].

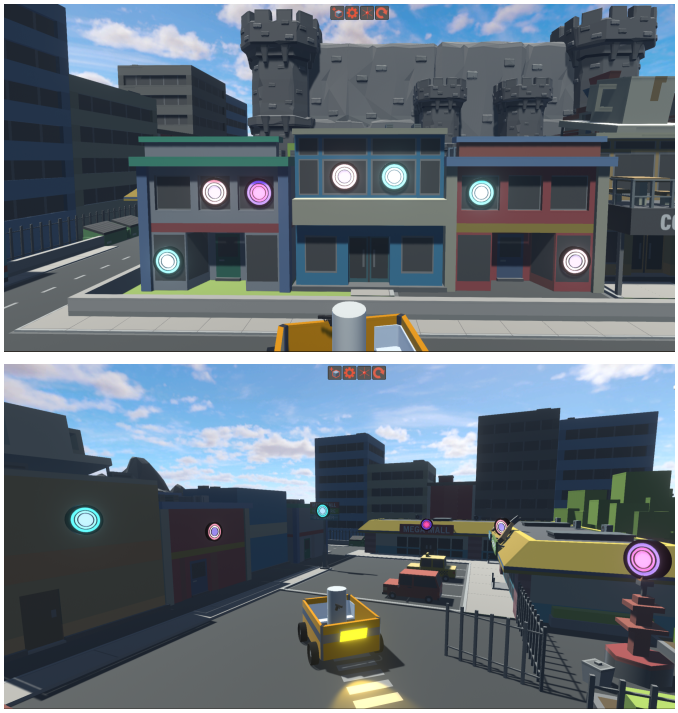

Fig. 2. The two clusters of targets in the City space.

IV. The 3D shooter game

Inspiration for the game was taken from the work previously mentioned in subsection II-A. The concept of a first-person shooter (FPS) game easily lends itself to assessing binaural localization due to the semantic nature of locating the enemy target and shooting it, which simplifies the localization task to a level of abstraction that everyone can understand. Furthermore, the first-person perspective is optimal for spatial audio, as the camera references the player's head movements. Therefore, a faithful acoustic environment is crucial for providing the player with an immersive gaming experience.

One of the main design criteria was ensuring a fun and simple experience. The player would navigate the environment from within a rail cart, following a predetermined path, where they can use the mouse or potentially a head-mounted display (HMD) to look around.

Three distinct acoustic spaces were designed. Within each of them, several clusters of shooting targets, anchored to different elements of the scenery, are displayed in sequence. One target is randomly selected from each cluster to play a looping Gaussian white noise burst with no reverberation applied. Both the acoustic spaces and clusters of shooting targets can be seen in Figures 2, 3, and 4.

Through auditory selection, the player is required to aim and shoot at the target they think is playing the sound, within a thirty seconds time limit. Visual feedback is provided depending on

Fig. 3. The two clusters of targets in the Castle space.

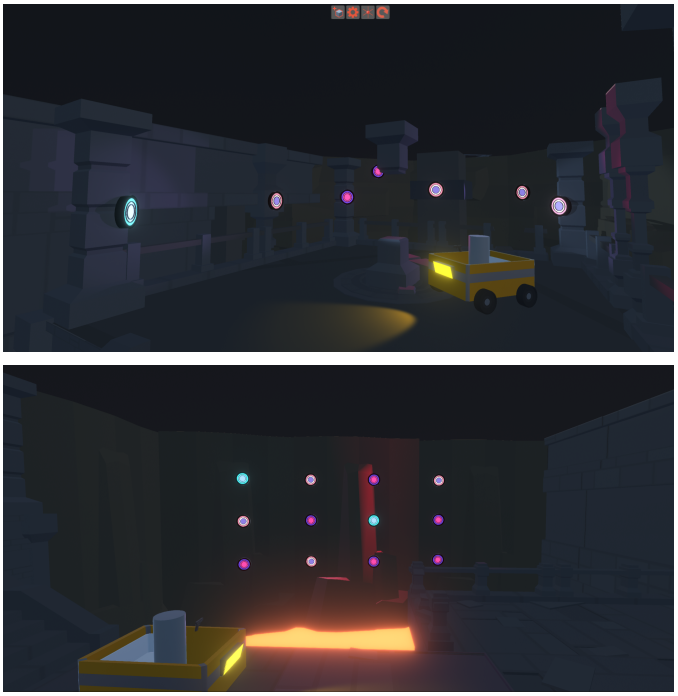

Fig. 4. The two clusters of targets in the Dungeon space.

whether they hit the correct target, a wrong target, or if time expired. Furthermore, a numerical score based on the number of correct targets hit is shown to the player, to entice motivation to complete the task successfully. In-game screenshots of the

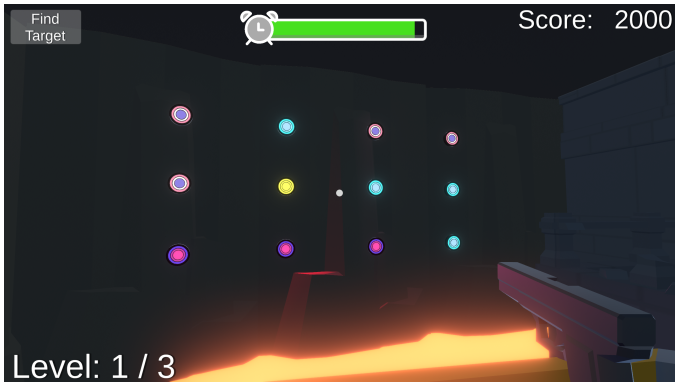

Fig. 5. In-game screenshot of the sound localization trial. Here the user has shot a target (marked by a yellow color).

Fig. 6. The result of the shooting trial is shown. Green color plus the UI marker means that the target was correctly chosen. Red color means that the target did not correspond to the spatialized sound.

shooting trial can be seen in Figures 5 and 6.

In addition, the acoustic spaces are populated with different environmental diegetic sounds [22] such as background city ambience, a bustling cafe, and a stationary car playing the radio at full volume. All of these sounds are rendered binaurally when the player moves across the environment in the cart. Depending on the current acoustic space, the amount of environmental reverberation is manipulated in order to match the visual scene. In the first acoustic space, consisting of a urban environment, an open reverberation profile is used — i.e. the sound is attenuated through air and bounces off cluttered surfaces. The second space, a castle, employs an open hall reverberation profile giving a higher amount of reflected sound. The third space, a dungeon, features a reverberation profile with substantial echo. In order to implement the changes in each acoustic space, a default parametric reverb filter provided by the Unity audio engine was applied. Parameters such as sound level, reflection delay, and room effect for both high and low frequencies were tuned to approximately simulate the acoustic spaces.

During the game design process, the Agile Software Development [23] strategy was employed to test various ideas and game mechanics. This was done by allowing five participants to test the game at different iterative steps focusing on different

core dynamics. At each step, a new sample of participants were recruited and their feedback collected through an open online questionnaire focused on extracting qualitative information. The feedback from the participants would then be used to inform the design for the next iterative step. The questionnaire was inspired by the conventions established by Fullerton on play-testing a game [24].

We conducted tests for two iterations of the game. In the first iterative test it was clear that the player lacked control and excitement. Moreover, the test participants expressed their concerns regarding the placements of the targets, which appeared too disconnected from the rest of the environment. Additionally, the players felt that the cart that moving too slowly, and expressed a desire to control the speed themselves.

For the second iterative test, the feedback was much more positive and suggested on adding more visual improvements, such as a score system and improving the visual feedback of the localization targets. Some participants expressed confusion over the color coding of the targets shot. Thus, a neutral yellow color was applied to the chosen target upon shooting, with the correct and wrong targets subsequently changing color to green and red, respectively. The player could also control the acceleration of the cart now, to play the game at a tempo to their liking. However, when approaching a target, the cart would slow to a default acceleration in order to make sure each player has the same experience during the target trial.

The game was developed using Unity 2019.4.1f1 bundled with the Steam Audio API which supports custom HRTF data in SOFA format. Additionally, royalty-free third-party elements were used to assist with building the environments, such as 3D assets and sound samples.

V. Experimental protocol

The experiment consists in completing multiple sound localization tasks construed as locating and hitting a sound-emitting target, distributed across the three acoustic spaces within the game described in the previous section. For each participant, three iterations — called stages hereafter — are performed, each featuring a different audio rendering condition.

The following subsections elaborate on the collection of participants as well as the procedure implemented within the game and followed throughout the experiment.

A. Participants

A total of 26 people signed up for the experiment, although 4 did not manage to complete it within the agreed time frame. Thus, the final sample size consisted of 22 participants, made up of $72.7\,\%$ males, $22.7\,\%$ females, and one participant who preferred not to specify. The age of the participants ranged from 22 to 30. None of them reported any hearing problem. Participants were also asked to rate how often they played video games. The distribution of participants not playing games often (casual players) and participants reporting to play games often (experienced players) turned out to be even, allowing a balanced comparison between the two clusters.

The participants were sampled using convenience sampling [25] whereby if they were interested in joining the experiment, they would receive a consent form as well as a guide on how to submit the anthropometric data needed to synthesize their individualized HRTF. Because of the COVID-19 pandemic and the related restrictions, participants were asked to run the experiment from home using their own devices.

Once the data for a given participant was gathered and their HRTF successfully synthesized and validated, they would be sent a personalized build of the game containing a generic HRTF (from the MIT KEMAR dataset [26]) and their individualized one. Since prior research [1] suggests that the order in which HRTFs are presented can affect localization tasks, it was also elected to divide the participants into two groups: the first group would experience the generic HRTF followed by the individualized one, whereas the second group would experience them in the opposite order.

B. Procedure

At the beginning of the experiment, the player is welcomed by a 2D scene featuring a song being played. The player is asked to wear their best available headphones, adjust the volume to a comfortable level, and confirm that the left and right channels of the headphones are placed correctly. Subsequently, the player is introduced to an initial tutorial stage, where a set of instructions is provided through a head-up display (HUD). These include the in-game controls for rotating the view, shooting, and changing the speed of the cart. The player is also explained the objective of the game, and is given the chance to shoot a sound-emitting dummy target. This stage uses a non-binaural profile (stereo panning).

Once the simple task of the tutorial stage is completed, the player is presented with the first stage of the game. The spatial audio model used here is a simple stereo panning profile and the task consists in shooting targets across the three spaces — City, Castle, and Dungeon. For each target hit by the player, the game stores three vectors, corresponding to the current position of the player, the position of the target hit by the player, and the position of the sound-emitting target. This data is used to calculate the vertical and horizontal error metrics presented in Section VI. The time elapsed from target presentation to shooting is also stored.

After the non-binaural stage, the audio engine is instructed to use the first of the two HRTF sets — generic or individualized, depending on the experimental group. The player is then spawned back at the beginning of the path and the second stage begins, with the same visual and environmental auditory layout as the first stage. Once this stage is completed too, the remaining HRTF set is loaded and the player enters the third and last stage. Performance data is collected exactly as in the first stage.

At the end of each stage, the player is presented with an in-game questionnaire asking them to rate the difficulty of localizing the targets, and to describe their audio experience using a number of attributes based on MUSHRA evaluation [27] as well as anchors inspired by previous research assessing

spatial audio quality [15], [28]. The attributes that players are required to rate on 7-point Likert scales are the following:

1) Instruction: Please rate the characteristics of the target (hissing) sounds you had to shoot throughout the level.
   - Inside my head / Outside my head
   - Difficult to localize / Easy to localize
2) Instruction: Please rate the characteristics of all the other sounds throughout the level.
   - Dark / Bright
   - Incoherent / Coherent
   - Synthetic / Natural
   - Low quality / High quality

After completing the last stage, the player is redirected to a Google Forms questionnaire inquiring about the overall experience. During this final questionnaire, background information such as age, experience with video games, and used headphone model were gathered, along with feedback on the overall experience and their self-reported favourite sound profile. To conclude the experiment, participants were asked to submit the JSON file that was generated by the application, containing all the performance data and the in-game questionnaires.

VI. Results

Once the data was gathered, it was possible to derive a measure of error for both azimuth and elevation localization. This was done by first calculating the orientation of the correct and hit targets relative to the player position by subtracting the player vector $\mathbf{v}_{\text{player}}$ from both the correct target vector $\mathbf{v}_{\text{true}}$ and the hit target vector $\mathbf{v}_{\text{hit}}$. We then proceeded to convert each vector $\mathbf{v} = \langle v_x, v_y, v_z \rangle$ into the azimuth and elevation angles θ and ϕ:

$$\begin{aligned} \theta &= \text{atan2}(v_z, v_x) \\ \phi &= \text{atan2}(\sqrt{{v_x}^2 + {v_z}^2}, v_y) \end{aligned} \tag{1}$$

Subsequently, the horizontal and vertical errors measures were computed as:

$$\begin{aligned} \theta_{\text{err}} &= |\theta_{\text{hit}} - \theta_{\text{true}}| \\ \phi_{\text{err}} &= |\phi_{\text{hit}} - \phi_{\text{true}}| \end{aligned} \tag{2}$$

The horizontal error was then wrapped in the range $[0°, 180°]$:

$$\theta_{\text{err}} = \begin{cases} 360° - \theta_{\text{err}}, & \text{if } \theta_{\text{err}} > 180° \\ \theta_{\text{err}}, & \text{otherwise} \end{cases} \tag{3}$$

Finally, the impact of front-back reversal was neutralized by applying this last transformation to the horizontal error:

$$\theta_{\text{err}} = \begin{cases} 180° - \theta_{\text{err}}, & \text{if } \theta_{\text{err}} > 90° \\ \theta_{\text{err}}, & \text{otherwise} \end{cases} \tag{4}$$

On top of these error metrics, we computed the time taken to hit each target, a well as the number of correct targets hit by each participant.

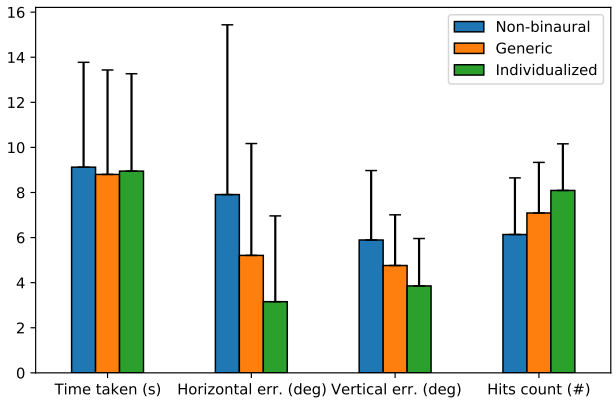

Fig. 7. Summary of all the localization metrics.

Figure 7 summarizes the results of each audio rendering method in terms of the aforementioned metrics. The most noticeable trend shown by the data is a progressive improvement in localization errors and hits count when going from non-binaural rendering, to generic HRTFs, and to individualized ones. The horizontal localization error is more than halved, going from 7.9° in the non-binaural case to 3.2° in the individualized HRTF case, while the vertical localization error gains 2° and the number of correct targets increases by 2 on average. Indeed, according to all the considered localization metrics, the non-binaural rendering offers the worst performances, the individualized HRTFs offer the best performances, and the generic HRTFs lie in the middle of the range. The average amount of time taken for each target stays approximately constant at around 9 seconds. Furthermore, along with an increase in performance, it is possible to notice a contraction in the standard deviation of the data, potentially indicating a consolidation of the performances.

TABLE I
P-values of paired sample t-tests for each metric and combination of rendering strategies

	Horizontal err.	Vertical err.	Hits count
Non-binaural, Generic	**0.015**	0.081	0.054
Generic, Individualized	**0.025**	0.091	**0.024**
Non-binaural, Individualized	**0.002**	**0.004**	**0.000**

In order to determine the statistical significance of the results described above, we performed a dependent t-test for paired samples, comparing the distribution of the data for each combination of rendering techniques. The p-values are shown in Table I. Numbers in bold highlight the cases where the null hypothesis, indicating that two related samples have identical average, can be confidently rejected. Predictably, the null hypothesis is consistently rejected whenever larger differences are present, such as in the case of the horizontal localization error. However, when considering the vertical localization, only the difference between non-binaural and individualized audio rendering appears to be statistically significant.

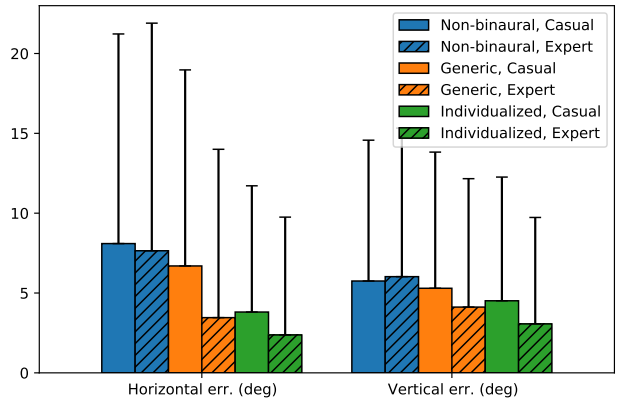

Fig. 8. Horizontal and vertical localization errors for expert (solid pattern) and casual (hatched pattern) players

We then considered the difference in performances between self-reported *expert* and *casual* video-games players amongst the participants, corresponding to groups of 10 and 12 subjects, respectively. On average, expert players performed better than their peers across almost every rendering strategy, reporting an horizontal localization error of $2.4°$, a vertical localization error of $3.1°$, and a hit rate of 72% for the individualized HRTF case on average. A comparison of the performances can be found in Figure 8.

Interestingly, expert players experience the greatest performance improvement in horizontal localization when going from non-binaural to generic HRTF rendering, while casual players appear to benefit more from individualized HRTFs. Conversely, when considering vertical localization error, a separate t-test between generic and individualized HRTFs infers a statistically significant difference in performances ($p = 0.036$) for expert players only.

Finally, we considered the results of the in-game questionnaire investigating the user's experience. These are summarized in Figure 9, where the mean rating for each question is shown, divided by rendering strategy, along with the standard deviation. Surprisingly, the non-binaural rendering strategy scored highest in terms of *externalization*, closely followed by the individualized HRTF. More predictable instead are the ratings for *ease of localization* which, similarly to the metrics presented earlier, progressively increase between non-binaural, generic, and individualized HRTF rendering.

With regards to the perceptual attributes of the auditory scene, the generic HRTF scored highest in *brightness*, *naturalness*, and *quality*, while the non-binaural rendering scored highest in *coherence*. The individualized HRTF appears to be notably worse in terms of *coherence* and *naturalness*. Nevertheless, except for *ease of localization*, all of the observed differences are quite small, amounting to half a point at best.

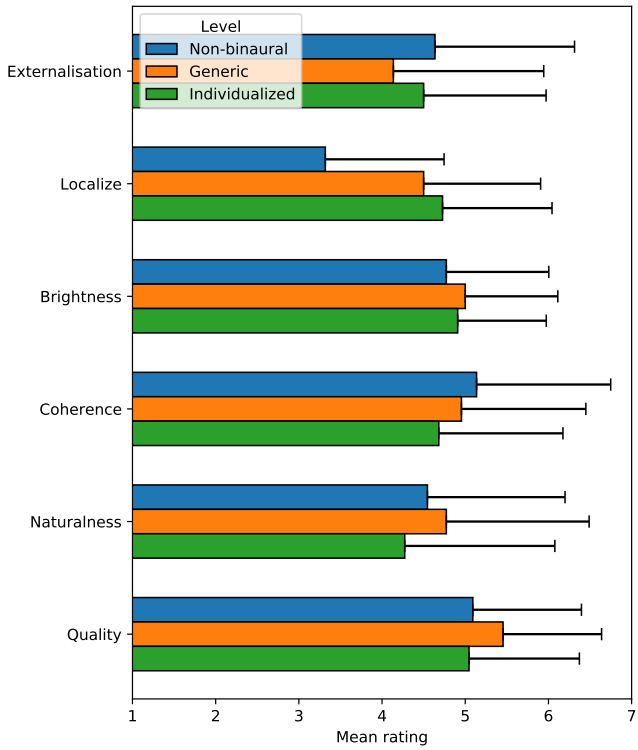

Fig. 9. Summary of responses of the in-game questionnaire.

VII. Discussion

A. Assessing the localization task

From the results gathered in the previous section, it appears obvious that the non-binaural instance had the worst results in terms of localization metrics. However, it would also be unfair to regard the results from the non-binaural instance to be of any significance in comparison, since it is the first instance that all the participants were introduced to. Accidents can easily occur as the participant is learning the mechanics of the game as well as the instructions of localization task. Instead, what can be derived from the non-binaural instance is the ease of understanding the dynamics of the game, which is crucial to the game design as well. In other terms, this preliminary instance proved effective in letting the players familiarize with the game.

From the analysis performed, we can see that generic and individualized HRTFs differ in horizontal localization performance as well as in the number of targets hit, with the individualized HRTF proving more accurate. This appears promising, although comparisons of individualized and generic HRTFs should be further tested within gaming experiences. On top of the vertical and horizontal error, additional localization metrics might be needed in order to highlight further significant differences.

It is important to bear in mind who would actually benefit from individualized HRTFs. Here we found that dividing test participants into casual and expert groups yields interesting

differences in results. When considering horizontal localization, our expert participants experience the highest improvement in performances when going from non-binaural rendering to a generic HRTF. This indicates that generic HRTFs could suffice the needs of expert players [29]. Conversely, casual players receive a similar performance gain when going from generic to individualized HRTF, suggesting that an accurate representation of individual azimuth cues could be paramount to ensure an immersive experience for casual users.

When considering vertical localization, the data portray a different picture. In this case, a similarly modest performance improvement is observed when going from non-binaural rendering to generic HRTFs, and from generic to individualized HRTFs, for both expert and casual players. However, this effect is only statistically significant for the expert group, suggesting that expert players may experience a more meaningful — though limited in magnitude — performance improvement when employing an individualized HRTF.

B. Assessing the experience

For the in-game questionnaire, it is hard to highlight any particular difference, except the non-binaural instance scoring particularly low in ease of localization. The non-binaural instance also ranks higher on the sound being externalized; however, this could be due to being the first instance that the user is subject to. If this was their first experience with a video-game offering 3D audio or if they had not paid attention to it before, presenting the non-binaural rendering as the first instance would put it at an advantage. Regarding ease of localization, we can see that the individualized HRTF scored highest, which supports the results from the performance metrics.

Concerning the other attributes, a number of test participants noted that some of the anchors seemed very vague to them, and that they were not used to the corresponding terminology. The MUSHRA anchors are also designed for more expert listeners, which could indicate that in this case they failed to properly make the users aware of what they were rating.

In the online questionnaire, 11 participants picked the generic HRTF as their preferred instance, with 9 picking their individualized HRTF. The remaining 2 picked the non-binaural instance. Some participants managed to describe how one type of HRTF was able to help them discern vertical angles. For instance, participant #7 reported: *"I picked Level 3 (i.e., generic HRTF) as I found it the easiest for localizing the target hissing sounds. But the soundscape also felt more empty (sic) in this level than in the other two"*. Instead, participant #9 stressed noticeable improvements with localizing vertical targets using the individualized HRTF: *"I think that was the closest where I was able to discern between sounds that are up or down. That was my biggest struggle"*. Importantly, the majority of the users found the game fun to play and easy to access.

C. Validity of the experiment

Although 22 users successfully completed the experiment, having the test done remotely could have introduced several issues. Most notably, each test participant used a different set of headphones, and therefore headphone compensation was not possible. In a worst case scenario, some of the headphones responses may not have accurately reflected the frequency content of the HRTFs [30].

Another issue with conducting the experiment remotely was that it was not possible to ensure the reliability of the game application on all platforms. In one case, the JSON database was not saved properly on an older version of OSX, causing one participant to be excluded from the experiment due to invalid data. Overall, the Windows builds worked more reliably due to it being the main platform the game was developed for. Nevertheless, in terms of game stability, users reported no crashing or any severe bug that hindered the experiment.

For the reasons above, it is believed that more accurate and reliable data can be acquired by performing the experiment physically at a lab using the same hardware and peripherals for all participants. This would also allow the experimenter to ensure that the collected anthropometric data is accurate.

VIII. Conclusions

In this paper, we investigated the effectiveness of individualized HRTFs generated using a recently introduced technique when used within an FPS game environment. As part of the experiment, a game was developed, where players were instructed to shoot a sound-emitting target. Throughout the game, multiple audio rendering schemes were applied, including a generic HRTF set and a non-binaural audio rendering method. Using the data collected throughout the experiment, we extracted localization error metrics based on horizontal and vertical errors, as well as quality-of-experience ratings for a number of relevant perceptual attributes.

Our results suggest that participants performed best with their individualized HRTF compared to a generic HRTF in terms of localization performance. When compared with generic HRTFs, the individualized ones resulted in increased localization accuracy on both the horizontal and vertical axes, with particular emphasis on the latter in the case of expert game players. Interestingly, we observed that the participants' opinion on their preferred HRTFs in terms of experience differs to the HRTF that prompted the best localization performance.

Nevertheless, further improvements can be made. Since some of the participants found the spatial attributes in the questionnaire vague, better communication regarding the spatial attributes should be considered, e.g. through a more thorough explanation, or by providing an example. This could include in-game sound references on what each attribute refers to or better descriptors.

Finally, it would be interesting to repeat the experiments with a more focused target of competitive players, where sound localization is integral to the gameplay. Such group may more aptly highlight the personal benefits of an individualized HRTF set in their gameplay experience.

REFERENCES

[1] D. Poirier-Quinot and B. F. Katz, "Impact of HRTF individualization on player performance in a VR shooter game II," in *AES International Conference on Audio for Virtual and Augmented Reality*, Aug. 2018.

[2] H. Hu, L. Zhou, H. Ma, and Z. Wu, "HRTF personalization based on artificial neural network in individual virtual auditory space," *Applied Acoustics*, vol. 69, pp. 163–172, Feb. 2008.

[3] G. W. Lee and H. Kim, "Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear," *Applied Sciences*, vol. 8, p. 2180, Nov. 2018.

[4] M. Zhang, X. Wu, and T. Qu, "Individual distance-dependent HRTFs modeling through a few anthropometric measurements," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2020)*, May 2020, pp. 401–405.

[5] M. Gröhn, T. Lokki, and T. Takala, "Comparison of auditory, visual, and audiovisual navigation in a 3D space," *ACM Transactions on Applied Perception*, vol. 2, pp. 564–570, Oct. 2005.

[6] J. Broderick, J. Duggan, and S. Redfern, "The importance of spatial audio in modern games and virtual environments," in *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 2018, pp. 1–9.

[7] M. Beig, B. Kapralos, K. Collins, and P. Mirza-Babaei, "An introduction to spatial sound rendering in virtual environments and games," *The Computer Games Journal*, vol. 8, Dec. 2019.

[8] M. F. Møller, Henrik Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: do we need individual recordings?" *Journal of the Audio Engineering Society*, vol. 44, no. 6, pp. 451–469, jun 1996.

[9] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Trans. Embedded Computing Systems*, vol. 9, no. 4, pp. 39:1–39:33, March 2010.

[10] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, pp. 111–123, Aug. 1993.

[11] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. 145th Conv. Audio Eng. Soc.*, New York, NY, USA, Oct. 2018.

[12] R. Nicol, L. Gros, C. Colomes, M. Noisternig, O. Warusfel, H. Bahu, B. Katz, and L. Simon, "A roadmap for assessing the quality of experience of 3D audio binaural rendering," *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014*, Apr. 2014.

[13] R. Miccini and S. Spagnol, "A hybrid approach to structural modeling of individualized HRTFs," in *Proc. 2021 IEEE Conf. Virtual Reality and 3D User Interfaces Work. (VRW 2021)*, Lisbon, Portugal, Mar. 2021.

[14] J. Berg and F. Rumsey, "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction," in *AES Convention 109*, Sep. 2000.

[15] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs : a case study of the SADIE II database," *Applied Sciences*, vol. 8, p. 2029, Oct. 2018.

[16] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A Cross-Evaluated Database of Measured and Simulated HRTFs Including 3D Head Meshes, Anthropometric Features, and Headphone Impulse Responses," *Journal of the Audio Engineering Society*, vol. 67, no. 9, pp. 705–718, Sep. 2019.

[17] R. Miccini and S. Spagnol, "HRTF Individualization using Deep Learning," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Mar. 2020, pp. 390–395.

[18] S. Spagnol, R. Miccini, and R. Unnthórsson, "The Viking HRTF dataset v2," Oct. 2020, DOI: 10.5281/zenodo.4160401.

[19] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *The Journal of the Acoustical Society of America*, vol. 109, no. 3, pp. 1110–1122, Feb. 2001.

[20] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *The Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3248–3259, Nov. 1996.

[21] S. Spagnol, "HRTF Selection by Anthropometric Regression for Improving Horizontal Localization Accuracy," *IEEE Signal Processing Letters*, vol. 27, pp. 590–594, 2020.

[22] A. Westerberg and H. Schoenau-Fog, "Categorizing video game audio: An exploration of auditory-psychological effects," in *Proceedings of the 19th International Academic Mindtrek Conference*, ser. AcademicMindTrek '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 47–54.

[23] A. Cockburn, *Agile Software Development: The Cooperative Game (2nd Edition) (Agile Software Development Series)*. Addison-Wesley Professional, 2006.

[24] T. Fullerton, *Game design workshop: a playcentric approach to creating innovative games*. CRC Press, 2019.

[25] T. Bjørner, *Qualitative Methods for Consumer Research: The Value of the Qualitative Approach in Theory and Practice*. Gyldendal Akademisk, 2016.

[26] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Computing, Tech. Rep., 1994.

[27] C. Mendonça and S. Delikaris-Manias, "Statistical Tests with MUSHRA data," in *AES Convention 144*, May 2018.

[28] S. Le Bagousse, M. Paquier, C. Colomes, and S. Moulin, "Sound quality evaluation based on attributes - application to binaural contents," in *AES Convention 131*, Oct. 2011.

[29] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang, "Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity," *Frontiers in Neuroscience*, vol. 12, p. 21, 2018.

[30] S. Spagnol, G. Wersényi, M. Bujacz, O. Balan, M. Herrera Martínez, A. Moldoveanu, and R. Unnthórsson, "Current use and future perspectives of spatial audio technologies in electronic travel aids," *Wireless Comm. Mob. Comput.*, vol. 2018, p. 17 pp., March 2018.