Llado Gonzalez, Pedro; Pollack, Katharina; Meyer-Kahlen, Nils

Toward a Standard Listener-Independent HRTF to Facilitate Long-Term Adaptation

Published in:
Journal of the Audio Engineering Society

DOI:
10.17743/jaes.2022.0134

Published: 02/04/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Llado Gonzalez, P., Pollack, K., & Meyer-Kahlen, N. (2024). Toward a Standard Listener-Independent HRTF to Facilitate Long-Term Adaptation. *Journal of the Audio Engineering Society*, *72*(4).
https://doi.org/10.17743/jaes.2022.0134

Toward a Standard Listener-Independent HRTF to Facilitate Long-Term Adaptation

PEDRO LLADÓ, KATHARINA POLLACK, AND NILS MEYER-KAHLEN

Abstract

Head-related transfer functions (HRTFs) are used in auditory applications for spatializing virtual sound sources. Listener-specific HRTFs, which aim at mimicking the filtering of the head, torso and pinnae of a specific listener, improve the perceived quality of virtual sound compared to using non-individualized HRTFs. However, using listener-specific HRTFs may not be accessible for everyone. Here, we propose as an alternative to take advantage of the adaptation abilities of human listeners to a new set of HRTFs. We claim that agreeing upon a single listener-independent set of HRTFs has beneficial effects for long-term adaptation compared to using several, potentially severely different HRTFs. Thus, the Non-individual Ear MOdel (NEMO) initiative is a first step towards a standardized listener-independent set of HRTFs to be used across applications as an alternative to individualization. A prototype, NEMObeta, is presented to explicitly encourage external feedback from the spatial audio community, and to agree on a complete list of requirements for the future HRTF selection.

1 INTRODUCTION

In applications such as virtual and augmented reality (VR/AR) and spatial music, virtual sound sources are rendered by convolving a source signal with a head-related transfer function (HRTF). Correct spatialization can be achieved by rendering the virtual sources with the user's own set of HRTFs; using another person's HRTF will inevitably lead to degradations [1]. Therefore, the obvious solution for providing most accurate spatialization is to always employ listener-specific HRTFs.

Listener-specific HRTFs can be obtained by performing acoustic measurements [2] or by creating a 3D-model of the listener's head and numerically computing their set of HRTFs [3, 4]. Measurements require an elaborate laboratory setup [2], which makes them unfeasible for most use cases. The most promising practical alternatives are to select the best fitting HRTF from a database by performing a short experiment [5] or to take pictures or videos of the user's head, which are then used to create a 3D model and to calculate their set of HRTFs [6]. All these application-dependent individualization strategies may affect the perceived quality or require some user-to-system adaptation [7] when moving from one application to the next.

Here, we wish to direct attention to an alternative solution: making use of the users' adaptation to a listener-independent set of HRTFs by standardizing one particular set. Numerous studies show that listeners adapt to a new set of HRTFs through continued exposure [8, 9, 10, 11, 12, 13, 14, 15, 16]. In the light of this research, it may be assumed that if a standardized set were to be used across applications, users may learn to use it as if it was their own, bypassing the problem of personalization. The goal of this initiative is not to prevent research or development of individualization techniques, but to improve the spatial perception of binaural rendering where no individualization is employed. Also, applications might exist, where an individualized HRTF is inevitable, e.g., optimized cross-talk cancellation.

Following this strategy, we introduce the Non-individual Ear MOdel (NEMO) initiative as a project towards a standardized listener-independent set of HRTFs. The path towards such standardization starts by choosing a specific set. We foresee that different opinions may prevail in the community regarding this choice. Thus, we provide a first list of requirements that lead to a tentative prototype. We have created a platform[1] to centralize the feedback and to provide easy access to the files and documentation. We wish to invite interested colleagues from research and industry to provide comments and perspectives on the requirements and the selection process.

2 BACKGROUND: ADAPTATION TO NEW HRTFs

Various studies have shown the adaptation abilities of human listeners to a new set of HRTFs over time, for example [8, 9, 10, 7] among others. In their review paper [12], Mendonça concludes that listeners improve their localization abilities with non-individual HRTFs over time, while, strikingly, local-

[1] nemohrtf.org

ization ability observed with their own HRTFs is not affected.

The main goal of the NEMO initative is to promote the use of a standardized listener-independent set of HRTFs to improve the perceived spatial representation after the adaptation process as much as possible. In [13] it was shown that post-adaptation performance can reach levels comparable to the performance observed when using one's own HRTFs. While in other studies, performance falls below this upper limit [9, 11], clear improvements were observed after adaptation [12] in any case. Especially promising results for long-term adaptation were found by Majdak et al. [13], where exponentially decaying trends in localization errors were reported throughout the training period, and by Stitt et al. [14], where it was found that the adaptation process continues as long as training is provided. Based on this evidence, we expect that the spatialization improves by consistent usage of a single set of HRTFs compared to being exposed to multiple sets, since the latter may hinder the learning process due to ambiguous information.

In principle, mere exposure to a new set of HRTFs is enough to induce adaptation [8, 9]. However, methods that involve training with feedback or active learning help to speed up the process [12, 15, 16]. Multiple studies have shown the beneficial effect of visual feedback [13, 10], but audio-only feedback seems to be effective [11] as well. Listening to spatial music would be considered exposure, whereas activities such as gaming in VR involve inherent feedback from visual cues or from motion-induced auditory cues, which would speed up the adaptation without the need for an explicit training phase. Further improvements may be obtained when active learning processes [12] are involved, which could naturally occur when players aim at improving their overall performance in a gaming application.

Besides the exposure or training regime and the listener-dependent ability to adapt to new cues, also the specific choice of a set of HRTFs has an impact on the adaptation process [9] and the post-adaptation performance, i.e., different sets have different *adaptability*, making the choice of a specific set an interesting matter of discussion. One could argue for the selection of a set that is very different to those of a large part of the population, as it has been shown that larger differences between a person's HRTFs and a new set favour adaptation [9]. Yet if the new set is sufficiently similar to the person's own HRTFs, it offers good pre-adaptation localization performance [9], so that adaptation is not crucial in many cases. Thus, an alternative approach could be to try to choose a set that offers good pre-adaptation performance for a large group of listeners. This idea seems to underlie the choice described in [2]. Such an approach is valid, but potentially biased by the selection of test subjects. A final way to reason for the selection of a particular set is the availability of spatial cues. Some HRTFs may provide ambiguous cues that hinder discrimination across elevation angles, while others may promote it by offering large differences between angles. We expect the availability of unambiguous cues to lead to good learning and post-adaptation performance.

3 REQUIREMENTS FOR A STANDARD DATASET

We have identified three areas of requirements that a standard dataset of HRTFs should fulfill: 1) usability, 2) flexibility and 3) adaptability.

3.1 Usability

To guarantee its usability, a standard dataset should be *open source*, to guarantee accessibility for all researchers, companies, developers and users. It should be provided in *several file formats*, such as the Spatially Oriented Format for Acoustics (SOFA) [17], .wav file or binary blob, for easy integration with existing applications. It should be offered in *several domains* such as in time, frequency and in spherical harmonics domains.Moreover, the data should be *well documented* and *up-to-date*. We hope to achieve this by maintaining the aforementioned website and repository.

3.2 Flexibility

We believe that NEMO should be *model-based*, in that the HRTFs are obtainable through numerical calculation, e.g., using [4, 18], or as analytical solution. This enables *extensibility* of the dataset, for example by larger sampling grids and positions in the near-field, or by simulating head-above-torso rotations. Being model-based, the set is easily *tuneable*, so that one can manipulate the model while maintaining the main characteristics, for example by removing the ear canal, or by adding a torso.

3.3 Adaptability

The dataset should provide *optimal post-adaptation performance*. One measure could be post-adaptation *localization* performance, i.e., the possibility of discriminating among directions, which we expect to be maximized by a set of HRTFs with the least ambiguous spatial cues. An example of a metric to assess this feature is describe in the next section.

[2] developer.oculus.com/blog/improve-spatial-audio-universal-hrtf-meta-quest

However, other performance metrics could be incorporated, for example related to *coloration or externalization*[19].

4 PROTOTYPE: NEMObeta

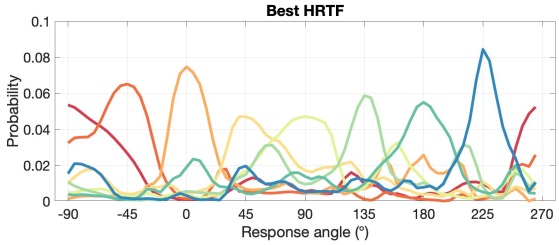
To launch the initiative, we introduce NEMObeta as a prototype. NEMObeta should be seen as an example of how the repository could look like when the actual set of HRTFs is selected. Also, it aims to motivate constructive criticism towards improving the requirements and the selection process.

Regarding usability, we investigated permissively licensed databases. As we believe that the best strategy to guarantee flexibility is to use model-based, numerically calculated HRTFs, we have further restricted our search to databases that include high quality 3D models, such as the databases HUTUBS [20], SONICOM [21], 3D3A [22], SADIE II [23], CHEDAR [24] and ITA HRTF [25]. The latter was ruled out for this first search, as it only contains models of the ears, and not of the entire head. CHEDAR is based on deformations of one set of ears, which may be an interesting way to obtain an HRTF with desirable features, but to date, there is no research showing which modifications are perceptually acceptable. All the remaining databases would be well suited, but for the search of NEMObeta, we decided to focus on the HUTUBS database, because it includes pre-computed model-based HRTFs, which we could use for the selection process described next.

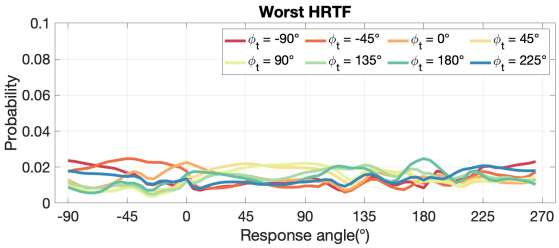
With regards to adaptability, we use median-plane localization as our exemplary attribute, to find the set with the least ambiguous cues. For this, an attempt was made to maximize model predictions of the possibilities for discriminating sound sources in elevation without training. We used the sagittal-plane localization model proposed in [26] to analyze the probabilities of discriminating across source positions. For each subject, i, and target angle in the median plane, $\phi_n \in [-90°, 270°)$, the probability mass vector, P_{i,ϕ_n}, was computed using HRTF$_i$ as both the template and the target, using the default parametrization. We define the *discriminability* D_i,

$$D_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\sigma^2\left(P_{i,\phi_n}\right)}, \qquad (1)$$

where N is the number of target angles in the median plane and σ^2 denotes the variance. The best set of HRTFs was selected by maximizing D_i. Figures 1a and 1b show the P_{i,ϕ_n} for the best (*pp55*) and the worst (*pp28*) sets at $N = 8$ different target angles, respectively.



(a) Best performing set of HRTFs, selected for NEMObeta



(b) Worst performing set of HRTFs, for comparison

Figure 1: Set of HRTFs selection. Probability mass vectors for the best (*pp55*) and the worst (*pp28*) performing sets of HRTFs in the HUTUBS database.

5 FUTURE WORK

This manuscript aims at highlighting the possible benefits of standardizing a set of HRTFs, which could potentially be used across platforms and applications. Now, we would like to receive external feedback to collect even more potential requirements, to satisfy as many researchers, companies, developers and users as possible. Also, the metrics to assess the adaptability, or how to approach future re-evaluations of the selected set are also important points of discussion among community members. We built a website to centralize all feedback and discussion.

After substantial feedback is received, we will formalize the collected requirements and will agree upon a particular dataset. As the last part of the selection process, a long-term study on the adaptation to a non-individual HRTF should be performed to asses the feasibility and the inter-subject differences at large scale. Then, we will work towards a standard with the goal of improving not only the final users' spatial representation, but also to ease the incorporation of spatial audio for all kinds of applications. Besides the benefits for all users of applications involving spatial audio, we hope that the initiative will open new research opportunities in studying adaptation to HRTFs.

References

[1] E. M. Wenzel, M. Arruda, D. J. Kistler, F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123 (1993), doi:doi.org/10.1121/1.407089.

[2] S. Li, J. Peissig, "Measurement of head-related transfer functions: A review," *Appl. Sci.*, vol. 10, no. 14 (2020), doi:doi.org/10.3390/app10145014.

[3] F. Di Giusto, S. van Ophem, W. Desmet, E. Deckers, "Analysis of laser scanning and photogrammetric scanning accuracy on the numerical determination of Head-Related Transfer Functions of a dummy head," *Acta Acustica*, vol. 7, p. 53 (2023), doi:doi.org/10.1051/aacus/2023049.

[4] F. Brinkmann, W. Kreuzer, J. Thomsen, S. Dombrovskis, K. Pollack, S. Weinzierl, P. Majdak, "Recent Advances in an Open Software for Numerical HRTF Calculation," *J. Audio Eng. Soc*, vol. 71, no. 7/8, pp. 502–514 (2023), doi:doi.org/10.17743/jaes.2022.0078.

[5] F. Zagala, M. Noisternig, B. F. Katz, "Comparison of direct and indirect perceptual head-related transfer function selection methods," *The Journal of the Acoustical Society of America*, vol. 147, no. 5, pp. 3376–3389 (2020), doi:doi.org/10.1121/10.0001183.

[6] K. Pollack, F. Brinkmann, P. Majdak, W. Kreuzer, "Von Fotos zu personalisierter räumlicher Audiowiedergabe," *Elektrotech. Inftech.*, vol. 138, no. 3, pp. 250–255 (2021), doi:doi.org/10.1007/s00502-021-00891-4.

[7] L. Picinali, B. F. Katz, "System-to-user and user-to-system adaptations in binaural audio," (2023).

[8] P. M. Hofman, J. G. Van Riswick, A. J. Van Opstal, "Relearning sound localization with new ears," *Nat. Neurosci.*, vol. 1, no. 5, pp. 417–421 (1998), doi:doi.org/10.1038/1633.

[9] M. M. Van Wanrooij, A. J. Van Opstal, "Relearning sound localization with a new ear," *J. Neurosci.*, vol. 25, no. 22, pp. 5413–5424 (2005), doi:doi.org/10.1523/jneurosci.0850-05.2005.

[10] S. Carlile, K. Balachandar, H. Kelly, "Accommodating to new ears: the effects of sensory and sensory-motor feedback," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2002–2011 (2014), doi:doi.org/10.1121/1.4868369.

[11] S. Carlile, T. Blackman, "Relearning auditory spectral cues for locations inside and outside the visual field," *JARO*, vol. 15, pp. 249–263 (2014), doi:doi.org/10.1007/s10162-013-0429-5.

[12] C. Mendonça, "A review on auditory space adaptations to altered head-related cues," *Front. Neurosci.*, vol. 8 (2014), doi:doi.org/10.3389/fnins.2014.00219.

[13] P. Majdak, T. Walder, B. Laback, "Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2148–2159 (2013), doi:doi.org/10.1121/1.4816543.

[14] P. Stitt, L. Picinali, B. F. Katz, "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning," *Sci. Rep.*, vol. 9, no. 1, p. 1063 (2019), doi:doi.org/10.1038/s41598-018-37873-0.

[15] C. Mendonça, G. Campos, P. Dias, J. Vieira, J. P. Ferreira, J. A. Santos, "On the improvement of localization accuracy with non-individualized HRTF-based sounds," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 821–830 (2012).

[16] G. Parseihian, B. F. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 2948–2957 (2012), doi:doi.org/10.1121/1.3687448.

[17] P. Majdak, F. Zotter, F. Brinkmann, J. De Muynke, M. Mihocic, M. Noisternig, "Spatially oriented format for acoustics 2.1: Introduction and recent advances," *J. Audio Eng. Soc.*, vol. 70, no. 7/8, pp. 565–584 (2022), doi:doi.org/10.17743/jaes.2022.0026.

[18] W. Kreuzer, K. Pollack, P. Majdak, F. Brinkmann, "Mesh2HRTF/NumCalc: An Open-Source Project to Calculate HRTFs and Wave Scattering in 3D," presented at the *Proceedings of the Euroregio BNAM Joint Acoustics Conference*, pp. 443–452 (2022).

[19] L. S. Simon, N. Zacharov, B. F. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632 (2016), doi:doi.org/10.1121/1.4966115.

[20] F. Brinkmann, M. Dinakaran, R. Pelzer, J. Wohlgemuth, F. Seipl, S. Weinzierl, "The HUTUBS HRTF database," (2019), doi:http://dx.doi.org/10.14279/depositonce-8487.

[21] I. Engel, R. Daugintis, T. Vicente, A. O. Hogg, J. Pauwels, A. J. Tournier, L. Picinali, "The SONICOM HRTF Dataset," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 241–253 (2023), doi: doi.org/10.17743/jaes.2022.0066.

[22] R. Sridhar, J. G. Tylka, E. Y. Choueiri, "A database of head-related transfer function and morphological measurements," presented at the *AES Convention 143* (2017 Oct.).

[23] C. Armstrong, L. Thresh, D. Murphy, G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, p. 2029 (2018 Oct.), doi:doi.org/10.3390/app8112029.

[24] S. Ghorbal, X. Bonjour, R. Séguier, "Computed hrirs and ears database for acoustic research," presented at the *AES Convention 148* (2020).

[25] R. Bomhardt, M. de la Fuente Klein, J. Fels, "A high-resolution head-related transfer function and three-dimensional ear model database," presented at the *Proceedings of Meetings on Acoustics*, vol. 29 (2016), doi:doi.org/10.1121/2.0000467.

[26] R. Baumgartner, P. Majdak, B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014 Aug.), doi:doi.org/10.1121/1.4887447.