LOW-COST NUMERICAL APPROXIMATION OF HRTFS: A NON-LINEAR FREQUENCY SAMPLING APPROACH

Maurício do V. M. da Costa *

MTDML, Institute for Musicology and Music Pedagogy
University of Osnabrück
Osnabrück, Germany
madovalemade@uni-osnabrueck.de

Luiz W. P. Biscainho

DEL/Poli & PEE/COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
wagner@smt.ufrj.br

Michael Oehler

MTDML, Institute for Musicology and Music Pedagogy
University of Osnabrück
Osnabrück, Germany
michael.oehler@uni-osnabrueck.de

ABSTRACT

Head-related transfer functions (HRTFs) describe filters that model the scattering effect of the human body on sound waves. In their discrete-time form, they are used in acoustic simulations for virtual reality (VR) or augmented reality (AR), and since HRTFs are listener-specific, the use of individualized HRTFs allows achieving more realistic perceptual results. One way to produce individualized HRTFs is by estimating the sound field around the subjects' 3D representations (meshes) via numerical simulations, which compute discrete complex pressure values in the frequency domain in regular frequency steps. Despite the advances in the area, the computational resources required for this process are still considerably high and increase with frequency. The goal of this paper is to tackle the high computational cost associated with this task by sampling the frequency domain using hybrid linear-logarithmic frequency resolution. The results attained in simulations performed using 23 real 3D meshes suggest that the proposed strategy is able to reduce the computational cost while still providing remarkably low spectral distortion, even in simulations that require as little as 11.2% of the original total processing time.

1. INTRODUCTION

The main goal of 3D acoustic simulations is to represent acoustic sources and environments as naturally as possible to the listener. To this end, human morphology must be taken into account since it imposes specific spectral distortions on the incident sound, both in magnitude and phase, that can be interpreted by the brain to, for example, estimate the direction of arrival (DOA) and distance, or identify the timbre of sound sources [1, 2].

For instance, the sound produced by a source positioned to the left side of the listener hits the left ear first, and arrives delayed and muffled (i.e. with attenuated high frequencies) at the right ear due to the greater distance and the acoustic shading effect of the head, respectively. These effects originate two among the spectral cues that help us locate sound sources in space: the interaural time difference (ITD) and the interaural level difference (ILD), both frequency dependent. Note that when a given source is equidistant from both ears, e.g. exactly in front or behind the listener, ideally, the listener can only rely on (simultaneous) spectral distortions in level to estimate its position. It has been shown that small movements of the head have a great impact in helping estimate such spatial positions and avoid front/back confusion, which highlights the relevance of even subtle differences between the sounds captured by the listener's left and right ears [2].

Head-related transfer functions (HRTFs) describe the mentioned body's acoustic effect over sound being emitted from any direction in the 3D space until impinging the ear canal entrances or the eardrums. Their digital versions provide an interface between the virtual acoustic field and the listener, thus producing a pair of signals (one for each ear) that will be converted back to the analog domain and reproduced for the listener by means of some acoustic transducer, such as a pair of headphones [2].

Although general-purpose HRTFs have been widely used due to the difficulty of acquiring data from specific listeners in an everyday situation quickly, reliably, and comfortably, individualized HRTFs tend to produce more realistic acoustic simulations for virtual reality (VR) or augmented reality (AR) [2, 3, 4]. For this reason, the research community has been working for decades on ways to personalize HRTFs, having as goals the increase in the accuracy of the HRTFs produced and the ease of implementation of the corresponding techniques [1, 5, 2].

Among the various existing techniques to accomplish this task, one way to produce individualized HRTFs without the need for any audio equipment or acoustically treated environment is by numerically calculating the sound field around the subjects' 3D representation [2, 6, 7, 8]. In this approach, the discrete complex pressure is estimated in the frequency domain within regular frequency steps for a set of given positions in space. In terms of equipment, this strategy only requires a device to capture the user's geometry and a computer to process the data.

Some important advances in this area allow for faster calculations, such as the widely adopted boundary element method (BEM) accelerated with the fast-multipole method and coupled with the collocation method [2, 6, 7]. Following this line, a recent strategy proposed in [9] resorts to an automatic mesh-grading procedure as a way to reduce the meshes, optimizing them in terms of computational load. The mentioned approaches can be considered *numerical approximations*, since they reduce the need for computational resources at the cost of introducing some (manageable) spectral distortion.

Nevertheless, the computational cost required for simulating (or approximating) HRTFs is still considerably high: computing an HRTF for a single user might take several hours on a powerful contemporary computer. Hence, this procedure is not applicable in daily-life situations where only moderate computational resources are available for the task, e.g. when using smartphones, tablets, or

* This work was funded by Volkswagen Foundation (VolkswagenStiftung), Germany (grant no. 96 881).

personal computers. For this reason, reducing the cost of numerical simulations is still a goal to be pursued.

One aspect of numerical simulations that, to the best of the authors' knowledge, is not explored in the literature concerns the dependence of computing time on frequency: a considerable part of the processing is performed at the higher end of the frequency spectrum. As will be shown in Section 3, when simulating HRTFs in a frequency range up to $22.05\,\text{kHz}$ (i.e. at a sampling rate $r_\text{s} = 44.1\,\text{kHz}$), only roughly 10% of the total processing time is spent to simulate the frequency spectrum up to $10\,\text{kHz}$, on average. Although some details in the range above $10\,\text{kHz}$ can still contribute to localization accuracy, human hearing presents a *quasi*-logarithmic resolution in frequency, thus exhibiting poorer resolution at high-frequencies; this suggests it may not be reasonable to allocate 90% of the computational cost to the simulation of the upper-frequency range. For instance, perceptual audio coders have taken advantage of low human auditory resolution at high frequencies to drastically compress audio files while still providing transparent results, i.e. indistinguishable from the original uncompressed files [10].

In [11], the authors also investigate the use of a non-linear sampling approach, which follows the equivalent rectangular bandwidth (ERB) scale, for smoothing purposes. Nevertheless, that approach presents two main disadvantages compared to our proposal, namely: (i) the incapability of reliably simulating/estimating the phase, and (ii) the fact that it requires a higher computational cost than our approach, for a comparable spectral distortion, due to the unnecessarily high resolution used at low/mid frequencies.

In this context, the objective of this paper is to propose an alternative, non-linear approach to frequency sampling that could reduce the cost of numerical simulations while still preserving the relevant spectral details from a perceptual perspective, by performing a perceptually justifiable numerical approximation.

The rest of the text is organized as follows: in Section 2, a theoretical background of numerical simulation is disclosed, followed by a description of the proposed hybrid-resolution method; then, Section 3 presents the experiments conducted to investigate the trade-off between reducing the simulation time and increase the spectral distortion of HRTFs when using the proposed approach; at last, in Section 4, the conclusions of this manuscript are drawn along with a description of future work.

2. METHODOLOGY

2.1. Numerical Simulation of HRTFs

Computing individualized HRTFs via numerical simulation requires a 3D geometrical representation (a 'mesh') of the subject, which consists of a discrete and finite set of points in space forming triangular faces.[1] The simulation thus calculates the scattering effect of the incoming sound wave over the body geometry and the HRTFs are obtained as a set of complex pressure bins in frequency, computed for specific pairs source/receiver locations in the 3D space [2, 6, 7]. The sound sources are usually distributed on the surface of a sphere centered at the origin of the Cartesian space, within a grid that can, for instance, be uniform or distribute the locations according to the human perceptual spatial resolution [2]. The receiver positions are typically faces of the mesh that best represent the occluded ear canal entrances or some point inside the open ear canals, depending on what is to be modeled.

Considering that the mesh is positioned in such a way that the midpoint between its left and right ear canal entrances coincides with the origin of the 3D space and the head is in line with the x axis [2], i.e. the receivers are approximately on the y axis, the (left, right) HRTF pair can be described as

$$\begin{aligned} \text{H}_\text{L}[\mathbf{x}^*, f, s] &= \frac{p_\text{L}[\mathbf{x}^*, f, s]}{p_0[f]} \quad \text{and} \\ \text{H}_\text{R}[\mathbf{x}^*, f, s] &= \frac{p_\text{R}[\mathbf{x}^*, f, s]}{p_0[f]}, \end{aligned} \tag{1}$$

where p_L and p_R denote the sound pressure at the respective left and right receiver points, $\mathbf{x}^*$ denotes the sound-source location,[2] f denotes the frequency, and s indexes the s-th subject. Both p_L and p_R are normalized w.r.t. the reference sound pressure p_0, measured at the origin *in the absence of the head*.

All current methods for numerical calculation of HRTFs are based on solutions of the Helmholtz equation, which describes the propagation of sound waves in the free field around the object of interest [2]:

$$\nabla^2 p(\mathbf{x}) + \kappa^2 p(\mathbf{x}) = q(\mathbf{x}), \mathbf{x} \in \Omega_e, \tag{2}$$

where $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$ is the Laplacian operator in the Cartesian 3D space; $p(\mathbf{x})$ denotes the complex sound pressure at the location $\mathbf{x}$; $\kappa = 2\pi f/c$ denotes the wave number, which is calculated using the frequency f and speed of sound c; Ω_e denotes the exterior domain around the object described by the mesh; and $q(\mathbf{x})$ denotes the complex contribution of the sound source in the acoustic field around the object. The simulation of HRTFs is performed by solving this equation for frequencies in the audible spectrum with regular frequency steps. In this context, since the sampling procedure is performed in the frequency domain, it would be inaccurate to adopt the term *sampling period*. Thus, to avoid ambiguity in notation, we will denote the frequency step mentioned as F_s, and the sampling rate related to the sampling of a signal in the time domain as r_s.

There exist different methods to solve this problem, as mentioned, the most prominent being: the finite-element method (FEM), which solves the Helmholz equation considering the object or the spatial domain around it as a volume; the finite-difference time-domain method (FDTD), which follows a similar approach to the FEM, but in the time domain; and the boundary-element method (BEM), which uses a special set of test functions in the weak formulation of the Helmholtz equation, namely the Green's function, and offers the advantage of only considering the surface of the object. This solution allows for the use of speed-up strategies, such as the fast-multi-pole method (FMM), the collocation with constant elements, and the reciprocity approaches. In addition, the resulting linear system of equations can then be solved using an iterative equation solver [2]. A complete description of methods for numerical calculation is out of the scope of this work (for more information on this topic, see [2, 6, 7, 8, 12]). This paper will focus on working with the BEM, which is the fastest and most commonly used solver for this purpose.

[1] Meshes are not necessarily described by triangles, but, in the case of numerical simulations, since connecting three points is guaranteed to define a flat surface, this choice offers the computational advantage of having a unique normal vector.

[2] The source position can be described either directly in the Cartesian 3D space as a three-dimensional vector or as a distance in meters and a direction, e.g. using azimuth and elevation angles.

Since the simulation is performed for each frequency f independently, the solution for a given set of regularly spaced frequencies $f \in \mathcal{F}$, where

$$\mathcal{F} = \{f \mid 0 \leq f \leq r_s/2, f_{i+1} - f_i = F_s\}, \tag{3}$$

can be paralleled and then processed to produce HRTFs for the desired source positions $\mathbf{x}$. Such independence is useful not only for speeding up the process but also for allowing us to arbitrarily chose which frequencies to compute the pressure with. This will be exploited in our solution to mitigate the high computational cost of numerical simulations using the BEM by estimating fewer frequency bins at high frequencies and then interpolating the resulting spectra to restore the regular frequency grid required to properly describe HRTFs and HRIRs. The next section presents the non-linear sampling approach proposed.

2.2. Numerical Approximation of HRTFs: Hybrid Linear-Logarithmic Sampling

On top of all the techniques already available to speed up the simulation of HRTFs, we aim to further reduce its computational cost. To this end, we will exploit the following particular observation: the processing cost grows nearly exponentially with frequency. As a result, a high percentage of the computing time is concentrated on simulating very high frequencies, which contrasts with the decreasing resolution of human hearing with frequency [10]. This suggests that most of the computational burden required by the simulation methods could be reduced without noticeable effects.

In line with this rationale, we propose a sampling approach that yields a hybrid linear-logarithmic frequency resolution. The main idea is to divide the frequency spectrum into two bands around a given crossover frequency f_c, keeping for simulations the usual fixed frequency steps F_s in the lower band and adopting in the upper band a logarithmic frequency spacing, specified by a number B of bins/octave. More rigorously, the set of frequencies to be simulated can be defined as $\hat{\mathcal{F}} = \{\mathcal{F}_{\text{lin}}, \mathcal{F}_{\text{log}}\}$, where

$$\begin{aligned} \mathcal{F}_{\text{lin}} &= \{f \mid 0 \leq f \leq f_c, f_{i+1} - f_i = F_s\} \text{ and} \\ \mathcal{F}_{\text{log}} &= \{f \mid f = f_{\text{max}} 2^{-l/B}, 0 \leq l \leq B \log_2(f_{\text{max}}/f_c)\}. \end{aligned} \tag{4}$$

The maximum frequency f_{max} to be simulated is defined by $f_{\text{max}} = r_s/2$ and the quantity $\log_2(f_{\text{max}}/f_c)$ indicates the number of octaves N_{octs} of the superior spectrum. This ensures that f_{max} will be simulated and the remaining frequencies will be sampled decreasing exponentially from f_{max} to f_c. The proposed HRTF is then first simulated using this non-linear frequency sampling, producing

$$\begin{aligned} \hat{\text{H}}_{\text{L}}(\mathbf{x}^*, \hat{f}, s) &= \frac{p_{\text{L}}(\mathbf{x}^*, \hat{f}, s)}{p_0[\hat{f}]} \text{ and} \\ \hat{\text{H}}_{\text{R}}(\mathbf{x}^*, \hat{f}, s) &= \frac{p_{\text{R}}(\mathbf{x}^*, \hat{f}, s)}{p_0[\hat{f}]}, \end{aligned} \tag{5}$$

where $\hat{f} \in \hat{\mathcal{F}}$.

Naturally, adopting the logarithmic scale only for high frequencies avoids applying an unnecessarily high resolution for the lower part of the spectrum, which would increase the computational cost without even being beneficial in any case, since we assume the spectrum is already perfectly represented by a sufficiently small F_s. In general, due to the free choice of B, the lowest frequency sampled in $\mathcal{F}_{\text{log}}$ may not coincide with the highest frequency sampled in $\mathcal{F}_{\text{lin}}$, which is not a problem. In addition, as a guideline, we can use F_s as a lower bound (equivalent to a maximum resolution) for the descending sampling procedure in $\mathcal{F}_{\text{log}}$; as a consequence, linear sampling can end up being extended beyond f_c. In fact, the strategy proposed in this paper could have been implemented differently, e.g. with the user setting only F_s and B, and allowing for the transition to happen when the resolutions meet; nevertheless, we found it useful to allow the user to set f_c, guaranteeing that a desired frequency band is simulated in predefined fixed frequency steps, independently from the choice of B. This will be especially important for the phase estimation, as will be discussed later in this section.

An example of the proposed sampling approach can be seen in Figure 1, which illustrates an HRTF simulated with the original regular sampling ($F_s = 150$ Hz) compared with the proposed hybrid sampling ($f_c = 2.76$ kHz, $F_s = 150$ Hz, and $B = 9$ bins/octave). Since the distance between samples after f_c is progressively higher, at some point, the log resolution seems to be insufficient to properly describe the original curve, starting to produce aliasing. This can be clearly verified at very high frequencies (around 18 kHz) where peaks and valleys occur between the non-linearly spaced samples. The question is how much these spectral details matter *perceptually*. This will be explored in more detail later.

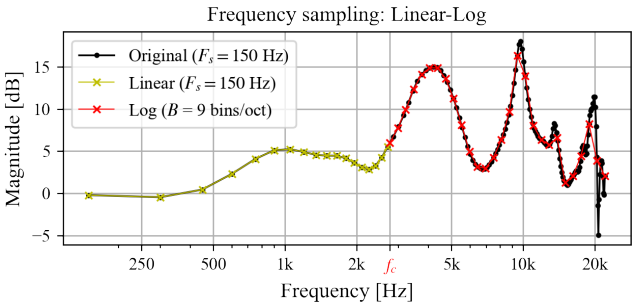

Figure 1: *Example of an HRTF simulated with the original linear sampling compared with the hybrid sampling proposed.*

Figure 2 illustrates an example of the computing time spent for each frequency bin in a real simulation using: the original regular frequency distribution ($F_s = 150$ Hz); and the proposed approach with $f_c = 5.51$ kHz, $F_s = 150$ Hz, and two different resolutions, for comparison ($B = 6$ and $B = 12$ bins/octave).

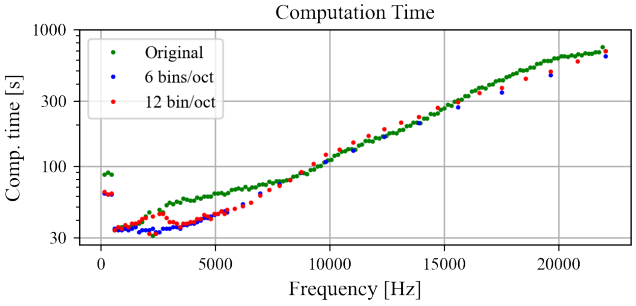

Figure 2: *Computing time varying with frequency. Original linear sampling (steps of* 150 *Hz), compared to the hybrid sampling approach (*6 *and* 12 *bins/oct).*

As can be seen in this example, which depicts a single simulation, there exists some fluctuation in processing time, even within the same frequency bins. This is expected, since computers perform several tasks in parallel, including calculating several different frequency bins sharing the same computational resources. Nevertheless, the same tendency of increasingly high computational cost with frequency is observed, regardless of the approach followed. However, since the proposed frequency distributions contain sparser bins at high frequencies, a reduction in the total computation time is expected.

To better compare the computing time for those simulations, it is worth plotting its accumulated value against frequency, as done in Figure 3, using the original linear sampling approach as a reference, with its total cost accounting for 100% of the time spent. The tendency of the accumulated computing time to grow exponentially (linearly in the logarithmic scale) towards high frequencies is evident; it also becomes clear that the resolution B used in the upper-frequency band changes the slope of the linear growth in the log scale. In this example, the HRTFs computed using the hybrid approach with $B = 6$ and $B = 12$ bins/octave cost, in total, around 12% and 20% of the linear approach, respectively.

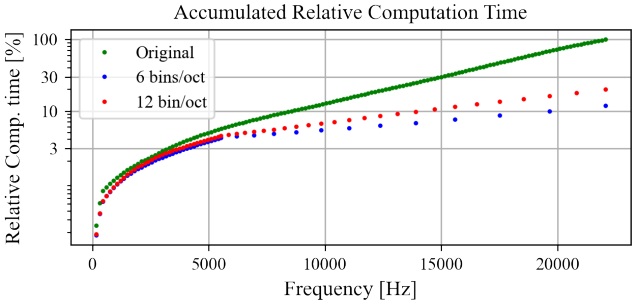

Figure 3: *Relative accumulated computing time varying with frequency. Original linear sampling ($F_s = 150$ Hz), compared to the hybrid sampling approach (6 and 12 bins/oct).*

It stands to logic that, on average, there should be an upper bound and a lower bound for the total computing time: it should never be higher than the reference, regardless of the choice of B and f_c, since F_s determines the maximum resolution possible, and F_s is also used in the reference; and it should also never be lower than the accumulated computation time at frequency f_c.

2.2.1. *Converting irregularly- to regularly-sampled HRTFs*

Certainly, some practical problems arise from using an arbitrary non-linear frequency resolution. It is required for an HRTF to have a complex set of samples in frequency evenly distributed throughout the whole frequency spectrum, allowing for the desired filtering procedure in which the HRTFs are intended to be used.

To this end, a simple linear interpolation procedure applied to the magnitude of the simulated spectrum $||\hat{\mathrm{H}}(\mathbf{x}, \hat{f}, s)||$ provides the regular frequency scale required, producing the magnitude spectrum $||\mathrm{H'}(\mathbf{x}, f, s)||$. Since the HRTFs are used in the digital domain, they will be henceforth denoted $\mathrm{H'}[\mathbf{x}, k, s]$, where $k \in \mathcal{K} \triangleq \{0, 1, 2, ..., K-1\}$ is the frequency index in the discrete frequency domain. Note that this interpolation is not able to restore or estimate parts of the spectra where information has been lost due to undersampling.[3] More sophisticated interpolation schemes might be explored in the future.

While the linear interpolation of magnitudes predicts reliable new samples, the periodicity of the phase makes estimating it between progressively more spaced samples a non-trivial problem. After trying different approaches, e.g. iteratively estimate the phase and correct its value using the unwrap procedure[4] on a frequency-bin basis, the best results were achieved by using the average group delay $d(\mathrm{H'})$ below f_c to linearly extrapolate the unwrapped phase. The average group delay is defined by

$$d(\mathrm{H'}) = \frac{1}{k_c} \sum_{0 \le k < k_c} \angle \mathrm{H'}[k+1] - \angle \mathrm{H'}[k], \tag{6}$$

where $k_c = \lfloor f_c / F_s \rfloor$ ($\lfloor . \rfloor$ denoting the floor function) is the index of the digital frequency related to f_c, and variables $\mathbf{x}$ and s have been omitted to simplify the notation.

Figure 4, illustrates the unwrapped phase of the same HRTFs shown in Figure 1. It can be observed that the higher the f_c, the lower the deviation in phase. However, this mismatch at high frequencies should have very little, if any, perceptual impact on the performance of the resulting HRTFs, especially since the phase at high frequencies is coherent with the average-group delay of the lower end of the spectrum.

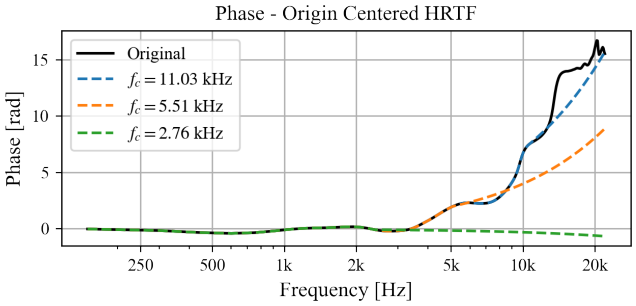

Figure 4: *Unwrapped phase of an HRTF: original regular sampling approach ($F_c = 150$ Hz); and the proposed hybrid approach ($F_c = 150$ Hz, $B = 9$ bins/octave, and f_c set to $11.03, 5.51, 2.76$ kHz).*

Since the pressure estimated at the ear canals is normalized by the pressure at the origin p_0, as mentioned in the previous section, the phase can assume negative and positive values. When it comes to generating HRIRs from the complex pressure simulated, a common procedure is to add a delay in such a way that the resulting filters become causal.

HRIRs of the same subject of the HRTF shown in Figure 1 can be seen in Figure 5, for different directions in the horizontal plane and different values of f_c; it was added a delay relative to the distance of 60 cm at speed-of-sound $c = 334$ m/s, i.e. approximately 1.8 ms. It can be observed that the impulse responses exhibit gradually higher energy concentration as f_c decreases, due to the increased frequency range whose phase gets linearized. As for the HRTF, an example of the spectra in the median plane is

[3] i.e., when the frequency spacing between the samples is below the Nyquist sampling frequency required for a lossless description of the sampled signal.

[4] Phase unwrapping algorithms aim to recover the true unwrapped phase signal by identifying and correcting phase discontinuities that occur in a phase signal wrapped between 0 and 2π.

shown in Figure 6 for the original simulation and an approximation using our proposal ($F_c = 150$ Hz, $B = 12$ bins/octave, and $f_c = 5.51$ kHz). Especially within the frequency range relevant for spatial cues (3-15 kHz), spectral detail is very well preserved.

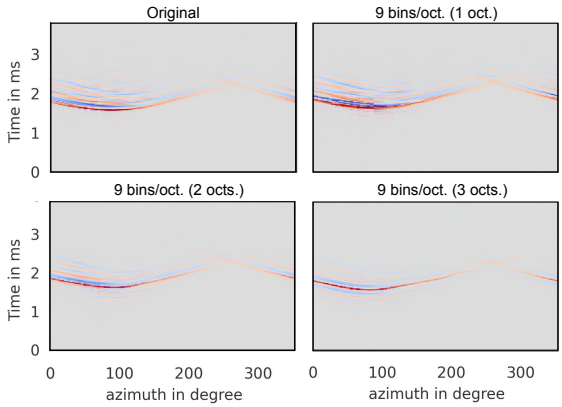

Figure 5: *Example of HRIRs of a subject (left ear) throughout the horizontal plane: original simulation ($F_c = 150$ Hz) and three examples of the proposed method ($F_c = 150$ Hz, $B = 9$ bins/octave, and $f_c = \{11.03, 5.51, 2.76\}$ kHz), respectively. Positive values are represented in blue, and negative values, in red.*

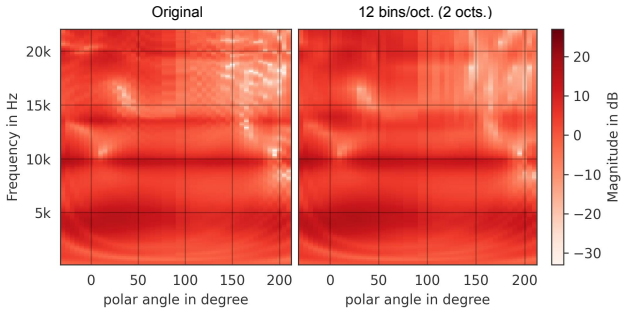

Figure 6: *HRTFs (left ear) throughout the median plane simulated with the original approach ($F_c = 150$ Hz) and with the proposed method ($F_c = 150$ Hz, $B = 12$ bins/octave, and $f_c = 5.51$ kHz).*

One thing to consider though is that, at high frequencies, wavelengths are comparable to the dimensions of the human ear, thus the incoming waves interact with the pinna in such a way as to produce steep peaks and notches that might be relevant to some degree (primarily up to 16 kHz [13, 2]) to produce spectral cues. For instance, the perception of elevation is highly dependent on those spectral cues, since the sound hits both ears virtually with no difference in time. Such characteristics of HRTFs help define important subject-dependent features that must be represented with some accuracy in the HRTFs to yield realistic results; a poor frequency resolution might underrepresent these spectral details.

Several experiments related to spectral smoothing have been conducted to determine to what degree HRTFs can be simplified before this can be noticed by listeners [14, 11, 15]. Thus, spectral smoothing might provide some indirect information with respect to the perceptual domain. For instance, using a fractional-octave smoothing window in the log spectrum can provide a realistic expectation of how much detail can be perceived in terms of spectral distortion, since the resolution of human hearing tends to follow an approximately logarithmic distribution. In Figure 7, one can see the same spectra illustrated in Figure 1, estimated for three different f_c, and smoothed using a third-octave rectangular window. As the resulting magnitudes are very similar, it is expected that the HRTFs sound very similar, if not indistinguishable.

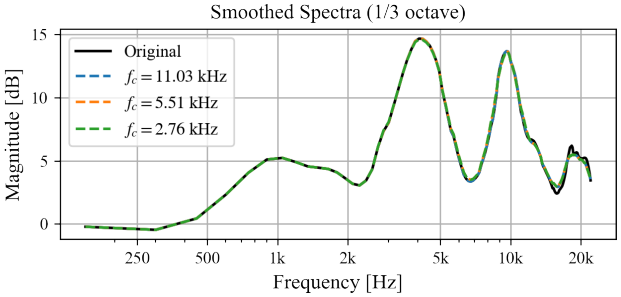

Figure 7: *Magnitude spectra of an HRTF smoothed using a third-octave filter: original regular sampling approach ($F_c = 150$ Hz); and the proposed hybrid approach ($F_c = 150$ Hz, $B = 9$ bins/octave, and $f_c = \{11.03, 5.51, 2.76\}$ kHz).*

Since spectral smoothing is a common post-processing procedure to remove measurement noise in HRTFs, this could also be included in the proposed approach, thus providing spectral curves that would smoothly follow the original simulated HRTFs. For example, a Hamming window with constant selectivity, or quality Q, could be used as a smoothing filter in the frequency domain, providing a smoothed curve more faithful to the original spectra the higher the resolution B compared to Q. In addition, the smoothing procedure would ensure smoother transitions between HRTFs of different directions, which is relevant whenever sound sources are not static.

In the case of the proposed approach, smoothing can be interpreted as a statistical estimation of the local energy in each frequency region, whose width varies geometrically in frequency. Such an estimate will be more precise the higher the resolution B.

3. EXPERIMENTS

The experiments performed in this paper consist of an acoustic analysis conducted on real 3D meshes acquired from 23 volunteers, over which different combinations of the parameters available for the simulations (the resolution parameters F_s and B, and the crossover point f_c) are tested. The spectral distortion and the total computing time are then presented for each case.

3.1. Spectral Difference Error and Computation Time Assessment

In order to compare the different versions of HRTFs, the Spectral Difference Error (SDE) is used. It can be computed for a specific frequency bin k by averaging the results over the D source positions $\mathbf{x}_d$ and the S subjects s as

$$\mathrm{SDE}[k] = \sqrt{\frac{1}{DS}\sum_{d=1}^{D}\sum_{s=1}^{S}\left(20\log_{10}\frac{||\mathrm{H'}[\mathbf{x}_d,k,s]||}{||\mathrm{H}[\mathbf{x}_d,k,s]||}\right)^2}, \quad (7)$$

where $||.||$ denotes the magnitude of the complex values. In turn, the overall SDE can be computed as

$$\text{SDE} = \sqrt{\frac{1}{DKS}\sum_{d=1}^{D}\sum_{k=0}^{K-1}\sum_{s=1}^{S}\left(20\log_{10}\frac{||\text{H'}[\mathbf{x}_d,k,s]||}{||\text{H}[\mathbf{x}_d,k,s]||}\right)^2}. \tag{8}$$

To produce the different simulations varying the available parameters, instead of simulating $\hat{\text{H}}$ for every single case, a preliminary study showed that estimating $\hat{\text{H}}$ from H via interpolations would be sufficiently accurate. Similarly, for the analysis of the computing time, a reference average time was calculated for each frequency bin using all available original simulations, with $F_\text{s} = 150\,\text{Hz}$. Then, estimating the cost in $\hat{f} \in \hat{\mathcal{F}}$ via interpolation over the average times in $f \in \mathcal{F}$ of the original simulations resulted in the percentage of the new average computing time w.r.t the reference, which allowed us to avoid possible fluctuations in computing time between different versions of the same HRTFs.

3.2. Participants

In total, 23 volunteers were recruited at the University of Osnabrück and compensated with 15 € after participation. They were, on average, *M* = 25.61 years old (*SD*= 4.77); 43.5% were female, and 56.5% were male; and all of them reported having normal hearing.

3.3. Ethical Approval

The experiments were all performed using 3D scans of voluntaries in accordance with the Declaration of Helsinki, with ethical approval obtained from Osnabrück University Ethics Committee (AZ.: 4/71043.5). In the process, the anonymity of participants and the confidentiality of their data were ensured. Participants were informed about the objectives and the procedure of the study as well as about their right to withdraw from the study at any time without adducing reasons or experiencing any negative consequences. All participants provided informed consent before participation in the study, which will also include future listening tests.

3.4. Acquisition and Preparation of the 3D Meshes

A 3D scan of the head and torso of each participant was obtained with the POP 3D scanner from Revopoint[5] with a resolution of 0.3 mm. The same procedure conducted in [16] was followed: The test subjects were asked to wear a nylon hair net to facilitate the scanning procedure since the hair is not modeled; then, point clouds were created and converted into 3D meshes; the scanning procedure lasted about 20 min and the results were verified visually to ensure the ears were captured without any artifacts and no major problems happened in the overall shape of the head and torso. Small artifacts caused by hair and clothes were neglected in this initial phase and dealt with later.

The meshes were then carefully treated for the numerical calculations in Blender.[6] First, artifacts and details related to hair and clothes in the meshes were smoothed out using editing tools in Blender. Such regions were transformed into flatter/smoother surfaces, contributing to saving computational resources by lowering the number of points in the mesh. The automatic mesh-grading procedure described in [9] was then used with the objective of optimizing the meshes in terms of computational burden. This procedure consists of assigning different resolutions to different regions of the mesh, based on both the degree of curvature and the distance from the ear canal. To determine the minimum and the maximum distances between points in the mesh, different values were tried, starting from 0.7 mm and 10 mm, respectively. The minimum distance was set to 1.25 mm, and the maximum to 18.75 mm, these being the largest distances that did not cause significant spectral distortion ($\text{SDE}[k] < 0.5\,\text{dB}$ below 16 kHz).

3.5. Simulation

After the pre-processing stage, numerical calculations were performed with the Mesh2HRTF [6, 7] library, an open-source code that simulates HRTFs and is integrated with Blender. The implementation of the proposed method was also entirely based on this library. The simulations were run for both ears using faces at the ear canal (which was occluded) as vibrating elements, with the frequency spectrum sampled every 150 Hz and within the range 0-22.05 kHz. The results were then sampled for 1550 spatial positions distributed on the surface of a sphere with a radius of 1.2 m centered on the participant's head.

In the experiments conducted, a relatively wide range of the adjustable variables was spanned. As before, the original simulations used the frequency resolution of $F_\text{s} = 150\,\text{Hz}$, which provides good quality sampling and can easily produce both 48 kHz and 44.1 kHz HRTF files. Simulations run to compare different values of F_s produced negligible differences (below 0.1 dB for the whole frequency spectrum) between the spectra. Although lowering F_s would make the proportional computation time saved dramatically higher (to our advantage), such a comparison would be unfair since a higher regular resolution is not necessary.

The values assigned to the parameters were then $F_\text{s} = 150\,\text{Hz}$, $B = \{1, 3, 6, 9, 12, 15, 18\}$ bins/octave; with f_c varying in octaves, related to $N_\text{octs} = \{1, 2, 3, 4, 5, 6\}$ octaves. As mentioned earlier, certain combinations of B and N_octs can cause the linear sampling to be extended beyond the crossover point to prevent the logarithmic resolution from exceeding the regular resolution defined by F_s. This will be clearly shown in the results reported in the next subsection.

3.6. Results

The results obtained with this variety of configurations are summarized in Table 1, in which the computation times are shown, and in Table 2, where the SDEs are presented.[7]

As can be observed in Table 1, the proposed hybrid-resolution approach is capable of saving large amounts of processing time. For instance, using $B = 6$ and only two octaves ($f_\text{c} = 5.51\,\text{kHz}$) logarithmically sampled, the total computational cost is expected to be, on average, only 12.9% of the cost of the original version.

Figure 8 illustrates the curves of the average accumulated computing times for the different versions of the proposed method in relation to the original one, thus showing lower proportional values the greater the savings in computing time, over frequency. B varies within the range $\{1 - 18\}$ for $N_\text{octs} = 2$ octaves down from the maximum frequency 22.05 kHz and $N_\text{octs} = 6$ octaves, which effectively sets the minimum f_c. As mentioned above, setting a

[5] Available from `http://www.revopoint3d.com` (last viewed: March 17, 2023).

[6] Available from `http://www.blender.org` (last viewed: March 17, 2023).

[7] Not all results are presented in the table, as to avoid repeating the relative total computation time of configurations with extended linear sampling.

Table 1: *Relative total computing time (%) for simulations using different hybrid lin-log resolutions (F_s = 150 Hz, B bins/octave) and varying the crossover point frequency f_c in octaves descending from the maximum frequency available (22.05 kHz). For a given choice of f_c, repeated values are omitted. Low distortion (SDE< 1.5 dB @0-15 kHz.) results are presented in boldface.*

	Hybrid resolution: $F_s = 150$ Hz, B [bins/oct.]						
N_{octs}	1 b/o	3 b/o	6 b/o	9 b/o	12 b/o	15 b/o	18 b/o
1 oct.	15.3%	17.2%	**20.3%**	**23.4%**	**26.6%**	**29.7%**	**33.0%**
2 octs.	6.9%	9.2%	**12.9%**	**16.7%**	**20.5%**	**24.3%**	**28.2%**
3 octs.	4.9%	7.4%	**11.5%**	**15.6%**	**19.8%**	**23.9%**	**28.0%**
4 octs.	4.2%	6.9%	**11.2%**	**15.5%**	**19.8%**	-	-
5 octs.	3.9%	6.7%	**11.2%**	-	-	-	-
6 octs.	3.8%	-	-	-	-	-	-

crossover point determines the linear behavior of the curves, whose inclination is controlled by B. When using minimum f_c (Figure 8 right), though, since the more linear parts of the curves start in different parts of the spectrum, they tend to share a common inclination, although a less linear behavior can be observed, especially at low frequencies.

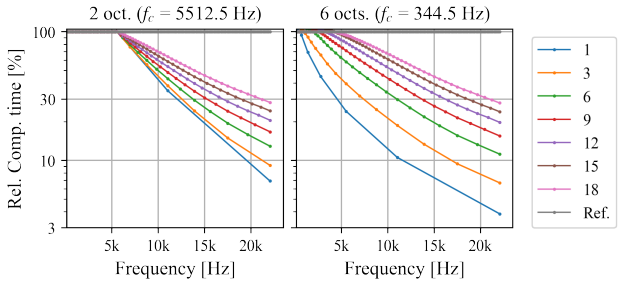

Figure 8: *Accumulated relative computing time for B within the range $1-18$ bins/octave, using $N_{octs} = 2$ octaves and $N_{octs} = 6$ octaves (minimum f_c) down from the maximum frequency 22.05 kHz.*

It is worth highlighting that, despite the great savings achieved with the proposed method, the growth rates of the computing time in frequency are still geometric (as illustrated in Figure 3), meaning that high frequencies still cost relatively much more than low frequencies. Since the average curve of the regular sampling approach tends to grow with a higher inclination, the relative saving rates are specific to the maximum frequency of the spectrum to be simulated. Naturally, the lower the maximum frequency, the closer the relative costs get, since the logarithmic resolution range becomes reduced and does not reach those regions of the spectrum where the frequency bins are more widely spaced. By way of comparison: if HRTFs were only simulated up to 16kHz, the full-range relative cost of 12.9% mentioned above would grow to $\approx 20\%$ when using $N_{\text{octs}} = 2$ and $B = 6$ (see Figure 8).

Table 2 shows, for each choice of B, the SDE for the minimum f_c possible, which provides the minimum computational burden, but also the maximum spectral distortion. For the previous example using $B = 6$, the maximum SDE calculated was 1.7 dB, spending 11.2% of the original computing time. Comparing the smoothed versions of the original and the proposed HRTFs using a third-octave smoothing window, the differences are even lower, as expected. For this case, the SDE_s (the 's' subscript denoting the smoothing procedure) was lowered to only 0.9 dB. For high values of B, such as 18 bins/octave, the SDE_s was as low as 0.3 dB, spending 28.0% of the original computing time.

Table 2: *Spectral difference error (SDE), in dB, for simulations using different hybrid lin-log resolutions (F_s = 150 Hz, B bins/octave). Values reported for the interpolated HRTFs (SDE) and their smoothed versions (SDE_s). The minimum f_c was used, as to provide the maximum SDE values.*

B [bins/octave]	1	3	6	9	12	15	18
SDE [dB]	4.2	2.5	1.7	1.4	1.2	1.0	0.9
SDE_s [dB]	3.5	1.6	0.9	0.6	0.5	0.4	0.3

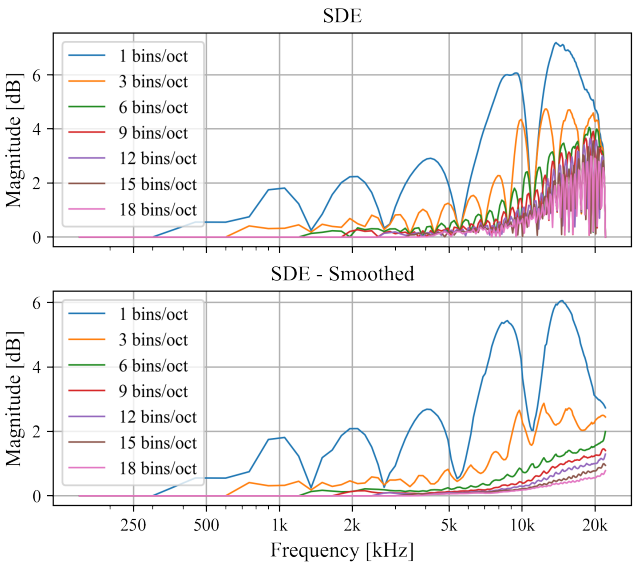

Figure 9: *SDE$[k]$ and $\text{SDE}_s[k]$ (spectra smoothed with a third-octave filter) for different resolutions.*

In order to analyze the spectral differences in more detail, SDE$[k]$ illustrates the distortion throughout the frequency spectrum, as can be seen in Figure 9 (top), where once again the minimum f_c is used to better illustrate the distortion occurred in the whole frequency spectrum. The distortion produced is primarily concentrated at high frequencies, as expected, because the sampling resolution is lowered with frequency. Even without considering the perceptual dimension, the distortion observed was considerably low for high values of B, peaking at around 3 dB at the very top of the frequency spectrum.

As can be expected, increasing the resolution B provides ripples with lower peaks in the SDE curves, and thus more transparent results. This can be seen in Figure 9 (bottom), where the SDE_s is shown for each configuration. Considering these curves, using $B > 6$ bins/octave creates SDEs below 1.5 dB, which we anticipate will provide perceptual transparency for all curves at least in the frequency range below 15 kHz. Some preliminary listening tests have shown that the just noticeable difference (JND) might lay between using $B = 3$ and $B = 6$ bins/octave. More rigorous listening tests using signal detection theory are planned to be carried out in the

future to assess the transparency of the HRTFs simulated with our method.

4. CONCLUSIONS

This paper presented a novel hybrid linear-logarithmic resolution approach for the numerical approximation of HRTFs. The proposed method requires remarkably less processing time than the regular sampling resolution usually employed. Experiments using 3D scans of 23 volunteers were conducted to compare computing time and spectral distortion for a wide range of resolutions. Preliminary experiments suggest that perceptual transparency may be achieved whilst saving around 89% of the processing time traditionally required. This solution is presented as a step towards everyday applications, lowering the computational power required for numerical simulations of HRTFs.

Since the original numerical simulations also present some spectral distortion in comparison to acoustic measurements, one concern to consider is the overall validity of the HRTFs obtained via the proposed numerical approximation, which further distorts the HRTFs to some degree. In future work, a subjective assessment will be conducted with the same volunteers from whom the 3D models were made to determine whether the subjects can notice a difference between simulations using their individualized HRTFs simulated with the traditional regular frequency resolution and using the proposed approach. This will involve different acoustic scenarios, sound sources with various characteristics, and dynamic changes in sound source position, thus covering many potential uses of the HRTFs. In addition, a more in-depth analysis will be carried out to determine the impact of the proposed procedure on localization accuracy and the spatial distribution of the spectral difference. Besides, it will be explored the use of upsampling procedures that could correct or restore spectral detail lost when simulating HRTFs using low spectral resolution.

5. ACKNOWLEDGMENTS

This work is funded by the Volkswagen Foundation (VolkswagenStiftung) Germany (grant no. 96 881).

6. REFERENCES

[1] Corentin Guezenoc and Renaud Seguier, "HRTF individualization: A survey," in *Proceedings of the 145th Audio Engineering Society International Convention*, New York, USA, October 2018.

[2] K. Pollack, W. Kreuzer, and P. Majdak, "Modern acquisition of personalised head-related transfer functions – an overview," in *Advances in Fundamental and Applied Research on Spatial Audio*, Brian Katz and Piotr Majdak, Eds. IntechOpen, London, United Kingdom, 1 edition, January 2022.

[3] Kazuhiro Iida, *Head-Related Transfer Function and Acoustic Virtual Reality*, Springer, 2019.

[4] Michele Geronazzo and Stefania Serafin, *Sonic Interactions in Virtual Environments*, Springer Nature Switzerland AG, Cham, Switzerland, 1 edition, 2023.

[5] S. Li and J. Peissig, "Measurement of head-related transfer functions: A review," *Applied Sciences*, vol. 10, no. 2, pp. 5014, July 2020.

[6] H. Ziegelwanger, W. Kreuzer, and P. Majdak, "Mesh2hrtf: Open-source software package for the numerical calculation of head-related transfer functions," in *Proceedings of the 22nd International Congress on Sound and Vibration*, Florence, Italy, July 2015.

[7] 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, July 2015.

[8] Brian F. G. 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. 2449, pp. 2440–2448, October 2001.

[9] T. Palm, S. Koch, F. Brinkmann, and M. Alexa, "Curvature-adaptive mesh grading for numerical approximation of head-related transfer functions," in *In Fortschritte der Akustik - DAGA 2021 : 47. Jahrestagung für Akustik*, Vienna, Austria, August 2021.

[10] Marina Bosi and Richard E. Goldberg, *Introduction to Digital Audio Coding and Standards*, vol. 721, Springer Science & Business Media, 2002.

[11] Laurence J. Hobden and Anthony I. Tew, "Investigating head-related transfer function smoothing using a sagittal-plane localization model," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, October 2015.

[12] Brian F. G. 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. 2449, pp. 2449–2455, October 2001.

[13] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, December 1974.

[14] Areti Andreopoulou and Brian F. G. Katz, "Comparing the effect of HRTF processing techniques on perceptual quality ratings," in *Proceedings of the 144th Convention of the Audio Engineering Society*, Milan, Italy, may 2018.

[15] Eugen Rasumowa, Matthias Blau, and Martin Hansen, "Smoothing individual head-related transfer functions in the frequency and spatial domains," *The Journal of the Acoustical Society of America*, vol. 135, no. 4, pp. 2012–2025, May 2014.

[16] M. Oehler, da Costa, M. V. M., M. Regener, and T. M. Voong, "Relevance of individual numerically simulated head-related transfer functions for different scenarios in virtual environments," in *Proceedings of the International Conference on Audio for Virtual and Augmented Reality 2022*, Redmond, USA, August 2022.