OPEN ACCESS Freely available online

A. Andreopoulou and B. F. G. Katz, "Perceptual Impact on Localization Quality Evaluations of Common Pre-Processing for Non-Individual Head-Related Transfer Functions" *J. Audio Eng. Soc.*, vol. 70, no. 5, pp. 340–354, (2022 May).
DOI: https://doi.org/10.17743/jaes.2022.0008

Perceptual Impact on Localization Quality Evaluations of Common Pre-Processing for Non-Individual Head-Related Transfer Functions

ARETI ANDREOPOULOU,* *AES Member,* **AND BRIAN F. G. KATZ,** *AES Member*
(a.andreopoulou@music.uoa.gr) (brian.katz@sorbonne-universite.fr)

[1] *Laboratory of Music Acoustics and Technology (LabMAT), National and Kapodistrian University of Athens, Greece*
[2] *Sorbonne Université, CNRS, Institut Jean Le Rond d'Alembert, Lutheries - Acoustique - Musique, Paris, France*
(brian.katz@sorbonne-universite.fr)

This article investigates the impact of two commonly used Head-Related Transfer Function (HRTF) processing/modeling methods on the perceived spatial accuracy of binaural data by monitoring changes in user ratings of non-individualized HRTFs. The evaluated techniques are minimum-phase approximation and Infinite-Impulse Response (IIR) modeling. The study is based on the hypothesis that user-assessments should remain roughly unchanged, as long as the range of signal variations between processed and unprocessed (reference) HRTFs lies within ranges previously reported as perceptually insignificant. Objective assessments of the degree of spectral variations between reference and processed data, computed using the Spectral Distortion metric, showed no evident perceptually relevant variations in the minimum-phase data and spectral differences marginally exceeding the established thresholds for the IIR data, implying perceptual equivalence of spatial impression in the tested corpus. Nevertheless analysis of user responses in the perceptual study strongly indicated that variations introduced in the data by the tested methods of HRTF processing can lead to inversions in quality assessment, resulting in the perceptual rejection of HRTFs that were previously characterized in the ratings as the "most appropriate" or alternatively in the preference of datasets that were previously dismissed as "unfit." The effect appears more apparent for IIR processing and is equally evident across the evaluated horizontal and median planes.

0 INTRODUCTION

One of the biggest challenges in Virtual Auditory Display (VAD) applications is the definition of a reliable method for the identification of the most appropriate spatialization cues ensuring truly immersive and individualized auditory experiences for every user. The difficulty lies in the highly personalized nature, and not fully understood perceptual processing, of spatialization cues utilized by the human brain for the deduction of the surrounding environment properties. These cues, captured in what is termed the Head-Related Transfer Function (HRTF), can be acquired through acoustic measurements [1, 2] or analytical solutions based on an individual's anthropometry [3, 4]. The use of non-personalized HRTF data has been shown to lead to degraded spatial experiences [5–7].

Nevertheless, even for cases when personalized HRTFs are available, they are hardly ever used without any type of processing. Whether performed to ensure compatibility between datasets from different databases [8], reduce the complexity in high-resolution HRTF spectra to perceptually relevant information [9] and lower the high dimensionality of the data [10], or be used in VR applications [11] and/or over the Web [12], HRTF post-processing is a reality that alters some aspects of the original measured data. Most often, while such actions affect the frequency and phase spectra of the data in a destructive manner, they are designed to constrain such artifacts to be within levels that are considered "acceptable" based on past research.

Several decades ago, researchers demonstrated that the frequency resolution of the cochlea limits the spectral detail necessary for accurate sound localization and exter-

*To whom correspondence should be addressed: a.andreopoulou@music.uoa.gr

nalization perception, leading to the conclusion that much of the information in the HRTF data could, potentially, be smoothed without considerable degradation in spatial impression [13, 14]. Various approaches towards HRTF dimensionality reduction have been proposed since, most of which can be grouped under four categories: 1) minimum-phase approximations, 2) smoothing in constant absolute or relative bandwidths, 3) multivariate analysis, and 4) Infinite-Impulse Response (IIR) modeling. The current study examines methods 1 and 2 and method 4.

Phase is one of the very first HRTF components to be considered perceptually less significant under certain conditions. This is based on studies that have argued that minimum-phase plus delay models are adequate approximations of the HRTF phase information [15], so long as the low-frequency Interaural Time Difference (ITD) is appropriate [16]. Similarly spectral smoothing is another standard processing procedure applied on HRTFs for data simplification. From simple techniques, such as HRTF time-domain truncation (smoothing in constant absolute bandwidths) [17], to more perceptually informed methods of smoothing into constant relative bandwidths [9, 18, 19], the goal is to eliminate unnecessary spectral detail while maintaining all perceptually relevant information.

An alternative approach involves the decomposition and remodeling of HRTF data as a low-rank approximation. Using multivariate analysis, HRTFs can be reduced to their most significant/distinct characteristics [20, 10]. An example of such a method, frequently used and extensively evaluated, is Principal Component Analysis [21, 22], which can decompose data into a set of hierarchical basis functions and weights. The smaller the number of basis functions utilized for data reconstruction, the higher the degree of smoothing. A different, but equally effective, approach to HRTF simplification is that of modeling data by means of IIR filters. Low-order IIR spatialization filters have a computational advantage over conventional HRTFs because of their significantly smaller size. The amount of spectral detail maintained in the resulting data is related to the type and order of the modeling filters [23, 24].

Often the accuracy of such methods is evaluated objectively through comparisons between the original and smoothed data, using metrics such as the Mean Square Error, Signal to Distortion Ratio, and Spectral Distortion (SD) [21, 25, 22]. In such cases, the reconstruction of the smoothed data is considered acceptable when the calculated errors are within levels previously reported as perceptually or spatially irrelevant. For cases where the effectiveness of smoothing algorithms is evaluated subjectively, conclusions are often drawn based on audibility and localization accuracy.

It has been suggested that HRTF spectral characteristics can be significantly smoothed without perceptible changes in the location of the virtual sound sources [16] or audible data variations [24]. For example, [17] investigated the effect of binaural filter length on localization accuracy, using personally measured HRTFs. Results showed that HRTFs could be reduced up to 0.64 ms at a sampling rate of 50 kHz, without causing a statistically significant degradation in localization performance. Nevertheless, filters shorter than 1.28 ms were shown to lead to a significant increase in front/back confusions. [9] investigated the importance of high-frequency spectral detail in HRTF data based on audibility. Results indicated that above 5 kHz, ipsilateral data could be smoothed with a bandwidth up to 2 Equivalent Rectangular Bandwidths (ERBs) and contralateral data with a bandwidth up to 3.5 ERBs without audible artifacts. In a similar study, [26] proposed a smoothing methodology where HRTF magnitude responses were expressed in ERBs and smoothing was achieved by progressively discarding higher-frequency content. Smoothed data was evaluated using the sagittal plane localization performance model by [27]. The proposed smoothing thresholds led to an average polar error of $\approx 21.5^{\circ}$, which, based on past relevant research [16], was considered acceptable.

Nevertheless, while such modifications may not have an impact on spatialization accuracy, other properties of the resulting sound, such as spectral coloration, could be affected in a way that may distort the VAD in unpredictable yet perceptible ways [28]. In an attempt to understand perceived implications associated with HRTFs, [29] developed a set of eight perceptual attributes that can qualify the HRTF's contributions to a binaural audio experience from various perspectives.

The aim of the current study is to evaluate the impact of common processing and modeling methods on the perceived spatial quality of HRTF data. More specifically a subjective study was designed that collected repeated localization quality assessments of binaural stimuli moving along predefined trajectories.[1] The evaluated scenes were created using a set of non-individualized HRTFs, previously characterized as perceptually distinct, with or without any type of processing or modeling applied. Out of the diverse pool of processing techniques applied on HRTF data, two techniques were selected for evaluation in the present study because of their common use in binaural signal processing: IIR modeling and minimum-phase approximation with smoothing in constant absolute bandwidths.

The hypothesis of this study is that user assessments should remain roughly unchanged, as long as the range of signal variations between processed and unprocessed data lies within ranges previously reported as perceptually insignificant. If this hypothesis is validated, it could further support previous arguments that such techniques can be applied to binaural data without impacting the perceived spatial quality, in the condition where the resulting signal variations fall within the established ranges.

The presentation of the work is structured in the following manner. Sec. 1 details the HRTF datasets employed, normalization, and processing algorithms. Sec. 2 provides a Spectral Difference analysis of the impact of the tested processing methods as an objective baseline. Sec. 3 details the perceptual evaluation protocol and the rank evaluation metric. Sec. 4 examines the results regarding participant

[1] Preliminary elements of this study have been previously presented in the 144th Convention of the Audio Engineering Society [30].

stability, variability, and rating similarity between conditions. Secs. 5 and 6 offer a discussion and final conclusions of the study. Analysis of the test duration and rating scale usage for the full experiment are provided in APPENDIX A.

1 DATA PRE-PROCESSING

SEC. 1 details the criteria followed for the selection of the specific nonindividualized HRTF data collection, used in the designed perceptual study. It also discusses the methods selected for data standardization and processing.

1.1 HRTF Dataset

To evaluate the impact of pre-processing techniques on HRTF quality judgments, it is of specific interest to employ a set of HRTFs spanning a large range of quality while being of sufficient quality prior to processing to be considered undistorted. In order to provide a common test set to all participants while maintaining a reasonable-sized test sample irrespective of the number of participants, individual HRTFs of test participants were not considered.

The test set was assembled using a data corpus comprising HRTFs from the publicly available LISTEN [1] and BiLi [2] databases. The utilized data were manually selected based on the clustering results of a corpus of 24 HRTFs, projected on a perceptually relevant space exclusively created from subjective similarity quality judgements, such that 1) both databases were represented, 2) the assessed data was perceptually distinct (i.e., all data clusters were represented), and 3) no common HRTFs existed between the pre-test and main experiment, in order to avoid potential bias exposure from the pre-test stimuli [6, 31, 32]. This process resulted in a set of seven full-phase HRTFs (six from BiLi and one from LISTEN) for the pre-test and of 12 HRTFs (six BiLi and six LISTEN) for the main experiment.

1.2 Corpus Standardization

The measurement conditions of an HRTF dataset (measurement space, protocol, equipment used, etc.) have been shown to affect the collected data in the signal domain [8]. Beyond that, HRTF databases are captured in different sampling rates, spatial resolutions, and filter lengths, which is a fact that renders the creation of composite HRTF corpora a challenging task. Since the selected HRTF corpus for this study originated from two separate databases, certain data standardization steps were necessary to minimize the potential impact of the database of origin on the utilized data and ensure that user evaluations were not impacted by such variations.

More specifically HRTFs in the LISTEN database were full-phase, diffuse-field–equalized, 512-tap filters captured at a 44.1-kHz sampling rate, and HRTFs in the BiLi database were full-phase, free-field–equalized, 2,048-tap filters captured at a 96-kHz sampling rate. Therefore all selected datasets were diffuse-field equalized and low-pass filtered at 20 kHz, and the sample rates were converted to 44.1 kHz. Potential DC offsets were removed, and an equal-frequency resolution was maintained through the truncation of all HRTFs to 256 samples, using a dynamically aligned rectangular window. The window starting point was set to 20 samples before the first detected onset of each filter-pair. The onset via relative threshold detection was defined as the first sample greater than −10 dB relative to the global dataset peak value [33].

Comparisons between HRTF datasets from different databases of origin are often challenging because of differences in spatial resolution. While interpolation can be a possible solution, interpolated HRTFs are prone to inaccuracies, which could impact the results of evaluation studies, and should therefore be treated with caution. In order to avoid any such potential artifacts, assessed sound trajectories in this study were based exclusively on measured location data. For cases when no common coordinates existed, the closest measured points were selected.

The study focused on a subset of measured HRTF positions lying on the horizontal and median planes. Horizontal plane positions consisted of 12 angles from 0° to 330° in increments of 30°. Median plane positions of the LISTEN HRTFs comprised 19 angles from −45° to 225° in increments of 15°, and those of BiLi comprised the following 19 angles: −45°, −27°, −15°, 0°, 15°, 27°, 45°, 62°, 74°, 86°, 106°, 118°, 135°, 153°, 165°, 180°, 195°, 207°, and 225°. The resulting angular differences of no more than 4° are smaller than the reported localization blur at these elevations [34] and, therefore, were assumed to be perceptually irrelevant for the task at hand. Previous related studies, which have used the same median plane grids, further support this assumption because they have shown neither any evidence of database discriminability [32] nor any effect on participant performance [31]. Data pre-processing concluded with a level normalization across all HRTFs. The adjustment of each dataset was set according to the median global RMS value over all evaluated positions.

1.3 HRTF Processing

The effect of different types of non-personalized HRTF processing on human spatial perception was assessed through the comparison of three types of commonly used data representations: full-phase HRTFs, minimum-phase HRTF decompositions smoothed across constant absolute bandwidths, and IIR-modeled HRTFs. The minimum-phase method is of particular interest for its prevalence due to computational efficiency. Previous research [35] established the minimum-phase-plus-delay model as sufficient. However direct examination of these results shows the limitation of this approach outside the frontal region. This observation is further supported by [33] and [36], which have demonstrated the reduced spatial accuracy for peripheral positions via perceptual tests. The IIR method has also been widely used, predominantly in signal processing architectures with limited computing power (e.g., mobile devices) and in systems with a large number of sources, while offering smooth interpolation between measured points via alterations of the poles and zeros of the filter [37].

All HRTFs were processed according to the following descriptions, based on state-of-the-art practice:

- *Full Ph.*: Full-phase data represents the original reference corpus. The 256-tap filters had a frequency resolution of roughly 173 Hz.
- *Min Ph.*: Minimum-phase HRTFs, first oversampled by a factor of 64 and filtered using a band-pass filter with cut-off frequencies at 200 Hz and 20 kHz, were decomposed using the Hilbert-transform and truncated to 64-tap filters, thereby smoothing the spectral response, resulting in a rough spectral resolution of 698 Hz [38].
- *IIR*: IIR-based data were modeled using the Yule-Walker algorithm on minimum-phase approximations of the HRTF data. The frequency scale was modified prior to and subsequently after the Yule-Walker process using a Bark bilinear transformation (and its reverse) to accord the filter design with human hearing properties. HRTFs were modeled using six biquad (second order) filters, resulting in 36-tap approximations [39].

1.4 ITD Information

To limit the focus of this study on the impact of HRTF spectral smoothing on spatial perception, ITDs were kept consistent between datasets. Based on the results of [33] and [36], position-dependent ITD values were calculated using the centroid of the interaural cross-correlation method, averaged across datasets, and used as a reference for modification. For full-phase data, ITD adjustments were made by adapting the temporal alignment of the filter-pairs. For minimum-phase data and IIR reconstructions, ITDs were inserted by delaying the contralateral channel of each HRTF such that the resulting ITDs coincided with the corresponding reference values. As such, all HRTFs had the same ITDs for equivalent positions.

2 OBJECTIVE EVALUATION

Prior to perceptual testing, an objective signal domain assessment of the degree of spectral variations across the three evaluated data types (full-phase HRTFs, minimum-phase smoothed HRTFs, and IIR reconstructions) was carried out. The following spectral comparisons were limited to the ipsilateral content filters of the 12 horizontal plane–tested positions and both filters for the 19 median plane ones (Sec. 1.2). This choice was based on past research demonstrating an increased tolerance in variations of contralateral HRTF spectral detail for humans [9, 40].

To approximate human perception, HRTFs were smoothed in ERB bands. Spectral variations between pairs of the same HRTFs that have undergone different types of processing were measured using the SD metric (Eq. 1).

$$SD(i) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(20\,log_{10}\frac{|\,HRTF(k)\,|}{|\,HRTF'(k)\,|}\right)^2} \qquad (1)$$

where N denotes either the number of ERB bands—when SD is calculated per tested position—or the number of tested HRTF positions—when SD is calculated per ERB band. *HRTF(k)* and *HRTF'(k)* denote the baseline (full-phase) and a smoothed version of the same ipsilateral HRTF filter (either min-phase or IIR), respectively.

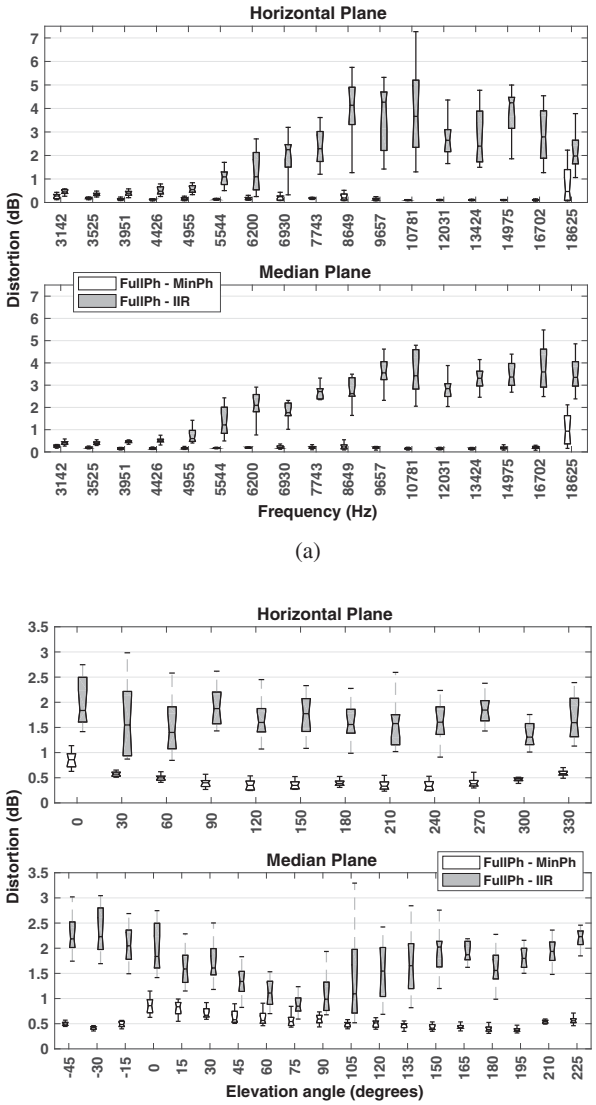

Fig. 1. Spectral Distortion (SD) distributions per Equal Rectangular Bandwidth (ERB) band averaged across all tested positions (a) and per tested position averaged across all frequencies (b) for the horizontal (upper) and median (lower) planes. Top and bottom box edges mark the 25th and 75th percentiles respectively, and whiskers extend to ±2.7 σ.

Fig. 1(a) shows the range of spectral distortion between the baseline (full-phase) and the minimum-phase and IIR data per ERB band. Variations were computed across tested positions leading to a single SD value per ERB band. Spectral variations of frequencies below 3 kHz were found to be very small in both planes (Table 1) and were, therefore, omitted from the figure to assist visibility of the results. As can be seen, the full range of SD values between full and minimum-phase data-pairs consistently lies below 1 dB, with the exception of the top ERB band. The increase in spectral variations in the very high–frequency content is attributed to variations introduced during data processing.

Table 1. Calculated Spectral Distortion (SD) per Equal Rectangular Bandwidth (ERB) band for frequencies <3 kHz.

	SD (dB)			
	Avrg	Std	Min	Max
Full Ph. / Min Ph. (Hor.)	0.49	0.08	0.11	1.30
Full Ph. / IIR (Hor.)	0.44	0.07	0.14	1.31
Full Ph. / Min Ph. (Med.)	0.58	0.07	0.19	1.38
Full Ph. / IIR (Med.)	0.57	0.07	0.21	1.36

On the contrary, the observed median SD values between full-phase and IIR data pairs are consistently higher than 1 dB for frequency bands above 6.5 kHz, reaching as high as 4.3 dB for the horizontal and 3.6 dB for the median plane data. Maximum variations were found to be as high as 7.3 dB at 10.7 kHz on the horizontal and 5.5 dB at 16.7 kHz on the median plane. These results indicate that the IIR reconstructions exhibit greater variation from the full-phase baseline, the potential impact of which needs to be subjectively evaluated.

It is also of interest to investigate whether any tested positions exhibit more variations than others, thereby biasing the perceptual evaluation task. Fig. 1(b) shows the range of SD values between the baseline and processed HRTFs per position. Variations were computed across ERB bands, leading to a single SD value per tested position. It can be noted that median variations for the minimum-phase data remain below 1 dB across all HRTFs with variations being higher for the frontal locations on or above the horizontal plane. The observed median distortion between the full-phase and IIR data is higher, reaching 1.9 dB ($SD_{max} =$ 3 dB) on the horizontal and 1.23 dB ($SD_{max} = 3.3$ dB) on the median plane. It should be noted that median SD values decrease as elevation angles increase, reaching a minimum for elevation angles 75° and 90° for filter-pair comparisons on the median plane.

According to [23], broad-band spectral distortions up to 1.9 dB on the frontal region did not lead to audible data variations. Hence the aforementioned spectral differences between the processed and reference (Full Ph.) data approximations across both planes are not expected to affect user ratings of minimum-phase data during the subsequent perceptual study but could have an affect on the evaluations of IIR-modeled data with variations exceeding the 1.9-dB threshold.

3 PERCEPTUAL EVALUATION

3.1 Experimental Procedure

A total of 23 individuals (eight female), all professionals and scholars in audio signal processing with extensive experience in psychoacoustic experiments, volunteered to participate in this study. All self-reported having normal hearing. The study was divided in two phases: the pre-test and main experiment. The purpose of the pre-test was twofold: it provided participants with the necessary familiarization to fully understand and perform the task in a timely manner while also assisting participant screening (in identifying people who could perform the required task in a consistent and reliable manner). Employing expert listeners to improve stability of the results with regard to the research question was preferable to lengthy training of all participants to the task (e.g., [41]). The main experiment assessed the perceived change in spatialization quality as a function of HRTF processing.

In both phases, the experimental procedure involved the assessment of binaural stimuli, moving along pre-defined trajectories on the horizontal and median planes, based on the conveyed accuracy of their movement. Horizontal plane trajectory movement started at 90°, directly on the left, and performed two counter-clockwise rotations around the listener's head at a constant distance. Median plane movement started at −45° elevation at the front and moved up, over, to the rear, and then back again to the front following the same arc over the listener's head twice. Participants were asked to rate the perceived spatial accuracy of each trajectory on a forced-choice nine-point rating scale ranging from "worst" to "best," based on written descriptions of the intended trajectory paths.

The pre-test and main experiment were divided in rounds consisting of two blocks, one for the horizontal plane trajectory and one for the median plane trajectory. For each block, participants were presented a single interface offering playback and rating options for all pre-rendered sound trajectories (one per HRTF processed dataset). Presentation order of blocks within each round and of HRTFs within each block were fully randomized. The study allowed participants to explore differences between sound trajectories at will, offering options for (re)playing the trajectory samples in any order, to facilitate comparisons. The only constraint concerned the use of the rating scale. The use of both scale extremes was required at least once per block, but the assignment of the same rating to multiple samples was allowed. This experimental procedure was selected because of its previously demonstrated ability in leading to stable and informative subjective HRTF assessments [31, 32].

Non-individualized binaural renderings were created using an impulsive sound stimulus and HRTFs from the data corpus discussed in Sec. 1.2. The stimulus was 100 ms of Gaussian noise (50 Hz to 20 kHz) with 2-ms Hann ramps at onset and offset, repeating sequentially for each position on the trajectories separated by 30 ms of silence.

The study was designed in MATLAB in such a way that the whole experiment could be completed by any participating individual without the need of a supervising moderator. A detailed description of the procedure and study goals and instructions on how to run the code and use the designed interface were provided. Participants were instructed to complete the study in a listening room with an ambient noise level below 30 dBA using Sennheiser HD 650 headphones and a Focusrite Scarlett audio interface, which were common to all participants. No headphone equalization was applied. Prior to the test, sound levels were calibrated to 80.3 dBA with the left headphone placed on a baffled microphone, using a monophonic 1-kHz sine-wave reference signal, ensuring the same presentation level for all test subjects and installations.

3.2 Pre-Test for Screening and Training: Repeatability

HRTF assessments using similar test designs to this one have been shown to be viable, albeit very demanding, due to the limited number of even expert participants able to produce reliable ratings across test repetitions ($\approx 50^{\circ}$) [31]. Hence user screening for the identification of individuals able to perform the assessment task in a repeatable manner is an essential step for such evaluation methods.

The pre-test comprised a variable number of rounds, each repeating the same experimental task: the evaluation of sound stimuli moving along horizontal and median plane trajectories on a nine-point scale. The system assessed the repeatability of each participant's ratings using the metric discussed in Sec. 3.4. When a participant produced repeatable ratings across both trajectories for three successive rounds, the pre-test phase terminated. Otherwise, repetitions continued until a maximum of seven rounds was reached. The pre-test was completed in a single sitting.

Based on the results of a previous study [31], test participants who are able follow this test-protocol in a consistent and reliable manner can do so because they can derive systematic evaluation strategies regardless of the number and type of non-individual HRTFs rated. Stated differently, people completing the pre-test successfully should be able to perform the main experiment test in a reliable and consistent manner. Hence only participants with repeatable responses across three successive evaluation rounds for both tested trajectories proceeded to the main experiment.

3.3 Main Experiment: Similarity

The main experiment comprised three rounds, each assessing HRTFs with a different type of processing (Sec. 1.3) on the conveyed spatial accuracy of stimuli moving on the same horizontal and median plane trajectories. Rounds appeared in a fully randomized order. The experiment task and control interface were identical to those of the pre-test. To minimize fatigue, test rounds were evaluated on separate days and compulsory breaks were inserted between the two test blocks of each round.

3.4 HRTF Rating Similarity Metric

For both the pre-test and main experiment, the evaluation of each participant's responses was based on the observed similarity between trajectory ratings of successive experiment rounds. Similarity of corresponding ratings between successive experiment-round pairs was calculated using Pearson's correlation coefficient r. When $r = 1$, the ratings of two successive experiment rounds for a given trajectory and participant are fully correlated, while when $r = 0$ and $r = -1$, ratings are uncorrelated and negatively correlated, respectively. Ratings of the perceived quality of the trajectory rendering relative to the given reference path provides an implicit measure of the global spatial quality assessment of localization for each HRTF [42, 43].

The sensitivity of the proposed metric to variations in the ratings, as a function of the magnitude of the change, and the number of ratings changed can be found in Table 2. These calculations are based on simulated data extensively discussed in [31]. As can be seen, this metric is relatively stable for small rating variations (up to ± 2 scale steps) but less tolerant to rating changes of larger magnitudes (≥ 3 scale steps), even when occurring rarely.

Two factors have been found to affect the results of the proposed metric: *Variability*[2] and *Stability*. Variability in HRTF evaluations is related to the range of r coefficients that result from the use of the similarity metric in pairs of HRTF ratings, and Stability is reflected in the values of those coefficients. Stability was calculated by computing the arithmetic mean across a series of r values of successive experiment rounds; Variability is calculated by computing the corresponding standard deviations. As shown in [31], Variability levels do not always imply equivalent Stability and vice versa. Hence both factors need to be taken into consideration in evaluating HRTF-rating repeatability [44].

Appropriate thresholds for the two factors were established based on a simulation study examining the sensitivity of the metric in rating variations of varying magnitude and quantity [31]. The selected values were $stb_{thld} = 0.75$ for Stability and $var_{thld} = 0.14$ for Variability, which allow for multiple rating variations of up to ± 2 scale steps and progressively fewer in the range ± 3 to ± 5 scale steps. A participant's HRTF ratings were considered repeatable if they were both stable ($stb_i \geq stb_{thld}$) and consistent ($var_i \leq var_{thld}$) in accordance with these thresholds. Permissible variations given these thresholds are indicated in Table 2. This metric was used in the analysis of ratings during the pre-test for the selection of participants for the main test and for analysis of the main test results.

4 RESULTS

4.1 Pre-Test: Similarity in Successive Ratings

As mentioned earlier, the pre-test involved the repeated assessment of the perceived spatial accuracy of sounds moving along predefined trajectories on the horizontal and median planes. The number of task repetitions (rounds), which involved the assessment of seven trajectories—and hence HRTFs—per plane, varied for each participating individual. The pre-test concluded either when a participant provided repeatable ratings according to the Stability ($stb_i \geq 0.75$) and Variability ($var_i \leq 0.14$) thresholds across three consecutive rounds for both planes or when a total of seven rounds were completed, whichever happened first.

Fig. 2 shows the repeatability of horizontal plane ratings for two pre-test participants on Stability over Variability. Each data point corresponds to three consecutive rounds. For example, Participant A completed seven experiment rounds without reaching repeatable performance. During rounds 3–5, the Variability of the ratings was within the acceptable range, but the corresponding Stability was under the defined threshold. Participant B, on the other hand,

[2] In past related publications, this factor was termed *Consistency* [31, 30]. Given that it reflects a metric that must be minimized, the term has been revised to *Variability*.

Table 2. Mean and standard deviation values of Pearson's r metric sensitivity to HRTF rating changes as a function of the size of change in the ratings (Δ_o) and number of ratings changing. Gray cells indicate variations within tolerance thresholds, which are discussed in Sec. 3.4.

# of ratings changing	Δ_o ±1	±2	±3	±4	±5	±6	±7	±8
1	0.99 (0.000)	0.97 (0.004)	0.94 (0.008)	0.88 (0.012)	0.81 (0.021)	0.72 (0.023)	0.62 (0.023)	0.50 (0.0)
2	0.99 (0.005)	0.94 (0.010)	0.87 (0.021)	0.76 (0.037)	0.61 (0.058)	0.42 (0.072)	0.20 (0.063)	-0.07 (0.0)
3	0.98 (0.003)	0.91 (0.015)	0.80 (0.035)	0.62 (0.063)	0.39 (0.098)	0.10 (0.111)	-0.24 (0.063)	-
4	0.97 (0.005)	0.88 (0.021)	0.72 (0.049)	0.47 (0.091)	0.16 (0.139)	-0.23 (0.142)	-0.63 (0.000)	-
5	0.97 (0.007)	0.85 (0.027)	0.63 (0.063)	0.31 (0.119)	-0.07 (0.169)	-0.53 (0.131)	-	-
6	0.96 (0.009)	0.81 (0.031)	0.53 (0.075)	0.13 (0.140)	-0.29 (0.177)	-0.80 (0.000)	-	-
7	0.95 (0.010)	0.77 (0.034)	0.41 (0.087)	-0.05 (0.148)	-0.49 (0.149)	-	-	-
8	0.94 (0.011)	0.73 (0.035)	0.28 (0.094)	-0.22 (0.126)	-0.67 (0.000)	-	-	-
9	0.93 (0.012)	0.68 (0.034)	0.13 (0.083)	-0.40 (0.000)	-	-	-	-

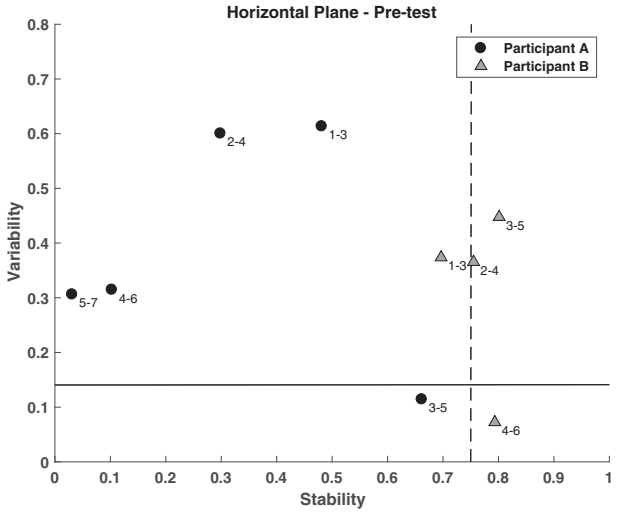

Fig. 2. Repeatability performance examples of two participants for the horizontal plane trajectory. Points represent Stability vs. Variability rates across triplets of successive rounds (rounds labeled for each point). Variability (solid line) and Stability (dashed line) thresholds are shown. Participant ratings are considered repeatable when results fall within the lower-right region delimited by the two threshold limits.

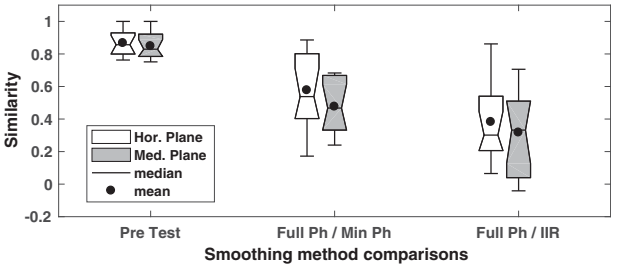

Fig. 3. Similarity distributions for the main test participants based on their ratings during the pre-test between the full-phase and minimum-phase data and between the full-phase and IIR data. Results are presented separately for the two tested trajectories.

reached repeatable ratings after six pre-test rounds. They started reaching *stable* ratings in round 4, but only achieved the selected *Variability* levels after six experiment rounds.

Out of the 23 recruited participants, 13 achieved repeatable ratings on both tested planes for the seven HRTFs across three consecutive rounds using the nine-point scale (57% success rate). This demonstrates the challenging nature of non-individualized HRTF rating procedures, even for experienced participants, and justifies the need for a pre-test in this experiment.

4.2 Main Experiment

The main experiment was divided into three rounds, each corresponding to a different type of HRTF processing (original full-phase, minimum-phase smoothed, and IIR reconstructions). Each round involved the assessment of binaural sounds moving along the horizontal and median plane trajectories discussed in Sec. 1.2, on a discrete nine-point scale. The same rating rules as for the pre-test applied to the main experiment. Each round comprised 12 HRTFs encoded with the same method.

4.2.1 Similarity of Quality Assessment Rankings

Similar to the pre-test, HRTF ratings per processing condition were analyzed using Pearson's r. Full-phase ratings were used as a common point for paired comparisons between processing methods, revealing how HRTF evaluations from each assessor changed as a function of processing relative to the original data. Repeatability performance of the 13 qualified participants in the last three rounds of the pre-test was also included in the analysis for reference. Resulting data distributions are shown in Fig. 3. The normality of the resulting distributions (pre-test, full-phase/minimum-phase, and full-phase/IIR, over both trajectory planes) was tested using a Lilliefors test. The test failed to reject the null hypothesis that the data was normally distributed ($\alpha = 0.05$) for all distributions.

The pre-test comparison of rating similarity r showed that all participants achieved very high scores across both planes. Median performances of 0.89 (minimum $r = 0.77$; maximum $r = 0.99$) for the horizontal and 0.85 (minimum $r = 0.75$; maximum $r = 0.99$) for the median plane strongly reconfirm the observation that all retained participants were highly capable of providing repeatable ratings across both trajectories.

By examining comparisons between the full-phase and minimum-phase data, the similarity distribution is much wider, ranging from $r = 0.17$ (almost uncorrelated data) to $r = 0.89$ (highly correlated data) for the horizontal plane trajectory and from $r = 0.23$ (almost uncorrelated data) to $r = 0.68$ for the median plane trajectory. The corresponding median correlations of 0.54 and 0.48 are much lower than those of the pre-test. According to Table 2, such low scores can be created by either multiple changes in ratings of $|\Delta_o| \leq 3$ scale points or at least one rating change of $|\Delta_o| \geq 7$ scale points. $|\Delta_o|$ variations $\geq$7 are notable because they imply a complete shift in the conveyed spatial quality of an HRTF processed in different ways. For example, it is possible for a full-phase HRTF to be rated as very good (score $\geq$7) with its minimum-phase smoothed equivalent as very poor (score $\leq$2) and vice versa.

Similar observations can be made for the similarity distributions between full-phase and IIR-modeled data. Median values are lower than those of other comparisons, with the horizontal plane trajectory being 0.30 (minimum $r = 0.07$, maximum $r = 0.86$), and the median plane being 0.32 (minimum $r = -0.04$, maximum $r = 0.71$). Additionally the presence of a number of uncorrelated data scores with values close to 0 in the median plane distribution implies that certain participants did not experience any correspondence of spatial quality between the full-phase and IIR-modeled data because such scores indicate multiple rating changes in the range of $|\Delta_o| \leq 4$ or at least one change as big as $|\Delta_o| \geq 8$ (Table 2).

It is possible to conclude that two distributions are significantly different with 95% confidence if their box-plot notches do not overlap [45]. As such, the following observations can be made regarding the distributions in Fig. 3:

- Differences in the rating similarity distributions between the horizontal and median plane ratings are not significant for any data group. This observation implies that the effect of HRTF processing on user spatialization does not seem to be more prominent for certain elevations.
- The similarity between the ratings of the full-phase HRTFs in the pre-test is significantly higher than those of the *full-phase vs. minimum-phase* and *full-phase vs. IIR* rating comparisons.
- Variations in rating similarity across trajectory planes are not significantly different between the processing methods.

4.2.2 Similarity Per Rating Scale Range

While monitoring participant rating similarity across the full nine-point scale offers an informative overview of the possible effects of HRTF processing on user assessments, it is of interest to focus on specific HRTF subgroups, such as those consisting of highly rated or rejected spatialization cues. Hence this section of the analysis focuses on the similarity between HRTF ratings within three spatial quality levels, *low*, *mid*, and *high*, thereby facilitating the separation of multiple small changes from occurrences of larger significant changes in ratings, which could both give rise to the same correlation value.

Table 3. Correspondence between the full-phase (reference) and minimum-phase/IIR spatial quality levels of the main test HRTFs computed in percentages for each trajectory plane.

		Horizontal Plane					
		Min-Ph.			IIR		
		Low	Mid	High	Low	Mid	High
Full-Ph.	Low	65.6	17.2	17.2	53.1	31.3	15.6
	Mid	46.5	32.6	20.9	34.9	32.6	32.6
	High	18.4	24.5	57.1	28.8	22.4	49.0
		Median Plane					
		Min-Ph.			IIR		
		Low	Mid	High	Low	Mid	High
Full-Ph.	Low	63.2	25.0	11.8	58.8	26.5	14.7
	Mid	41.7	33.3	25.0	39.6	29.2	31.2
	High	17.5	27.5	55.0	30.0	25.0	45.0

Using the full-phase HRTF ratings as reference, the percentage of HRTFs that were attributed the same or different spatial quality levels across all experiment rounds can be calculated. This analysis, summarized in Table 3, was performed as a function of trajectory planes, processing method pairs, and quality level. A noteworthy point arising concerns the relatively small percentage of HRTFs receiving the same spatial quality level assessment across the full-phase reference and either of the other two conditions (minimum-phase smoothed HRTFs and IIR reconstructions). The observed similarity levels lie between $\approx$29% and $\approx$66%, which practically means that, depending on the applied HRTF processing chain and trajectory plane under evaluation, participants changed their quality assessments of different HRTFs $\approx$34% to $\approx$71% of the time. The assessment change was quite prominent in the case of the mid-level spatial quality HRTFs, where the repeatability percentage was consistently $\leq$33.3% (chance level).

Of all possible assessment changes, the most notable are those concerning the two scale extremes, i.e., cases when the reference HRTF was evaluated as having "low" spatial quality and its processed versions as having "high" spatial quality or vice versa. For example $\approx$18% of the HRTFs evaluated as having "high" spatial quality in the full-phase condition were assessed as "low" in the minimum-phase condition for both trajectory planes. Similarly $\approx$29% (horizontal plane) and 30% (median plane) of the HRTFs evaluated as having "high" spatial quality in the full-phase condition were assessed as "low" in the IIR condition. The scores of this latter condition lie very close to chance (33.3%), indicating an increase in the complexity of the IIR-modeled HRTF evaluation task (see Table 3 for an overview of all possible assessment variations).

5 DISCUSSION

This work investigated the effect of different commonly employed HRTF processing algorithms on the perceptual evaluation of selected non-individualized HRTF data from the publicly available LISTEN [1] and BiLi [2] databases. Three versions of an HRTF subset database were compared: a) full-phase HRTFs, b) minimum-phase HRTF decompositions smoothed across 64 constant absolute bandwidths, and c) modeled HRTFs using six biquad IIR filters.

HRTF spectral smoothing and minimum-phase decomposition are the two most widely applied and thoroughly evaluated methods of HRTF processing. Several studies have proposed objective thresholds, "guaranteeing" the absence of audible artifacts and/or spatialization errors. For example, it has been suggested that minimum-phase plus pure-delay models are adequate approximations of the HRTF phase, not affecting the users' average localization accuracy. Nevertheless, even in early studies, there exists evidence of an increase in localization errors around the interaural axis [15] and of front/back confusions [16]. This increase in localization confusion within these spatial regions can also be reflected in the HRTF signal itself as greater data variability [33]. While the magnitudes of such errors/variations may not be large enough to trigger statistically significant results, they do suggest a shift in people's spatial impression, which must be further investigated.

In this study, spectral variations introduced by smoothing and HRTF modeling were quantified with the SD metric using the full-phase data as reference. For the minimum-phase data, smoothed in 64 constant absolute bandwidths, SD values were constantly $\leq$1 dB and hence were not expected to produce audible artifacts. For the IIR-modeled HRTFs, results of [23] have suggested that spectral variations up to 6.6 dB close to the interaural axis and up to 1.9 dB in the frontal area should not lead to audible variations. The median SD values of the IIR-modeled HRTFs in this study were consistently below these thresholds, indicating an overall successful data approximation.

Nevertheless, as shown in Fig. 1(b), the SD distributions of several tested positions on the horizontal and median planes extend beyond these thresholds, indicating that certain HRTF filters could potentially contain audible artifacts. Based on these results, it was hypothesized that user-assessments of processed HRTF data should remain roughly unchanged, since the range of variations between processed and unprocessed data lied within previously reported ranges. In order to investigate this hypothesis, a perceptual study was designed involving the repeated assessment of binaural stimuli moving along pre-defined trajectories on the horizontal and median planes using a forced-choice nine-point rating scale. The usefulness of user ratings as a mechanism for data assessment of non-individualized HRTF data has been established [42, 43], and the robustness of the specific protocol has been analyzed and discussed in [32].

The study was divided into a pre-test and main experiment. The pre-test provided a mechanism for participant task familiarization and reliability screening, increasing the robustness of the collected user responses through the use of expert listeners. Out of the 23 trained individuals who participated in the study, 57% (13 participants) achieved stable and consistent HRTF ratings according to the predefined repeatability thresholds across three task repetitions in the pre-test, thereby proceeding to the main test. The marginally higher percentage of repeatable participants in this study as compared with [31] and [32] can be attributed to their higher level of binaural audio expertise. Beyond participant selection, the results of the pre-test serve as a systematic documentation of the inherent difficulty and challenges in assessing the spatial quality of non-individual HRTF data. The uncertainty in the assessments of the participants, as reflected in their overall performance in the task, serves as strong evidence that steps for HRTF individualization are necessary to optimize user experience in VAD applications.

The effects of minimum-phase decomposition with spectral smoothing across constant absolute bandwidths and of IIR modeling on user HRTF quality assessments were studied using the ratings of the full-phase data as reference. Result analysis led to the following observations:

1. Distributions of Pearson's r coefficients between the full-phase/minimum-phase and the full-phase/IIR data were much wider with lower mean and median values than those of the pre-test condition. This finding strongly suggests that the applied processing algorithms impacted user assessment. Spectral variations on the minimum-phase data were consistently less than 1 dB across all tested positions and frequency bands. Therefore their impact on spatialization quality was expected to be minimal. Yet, based on the experiment results, it can be concluded that these spectral changes impacted the participants' overall spatial quality assessments in a way that could not have been predicted by past literature. On the contrary, the observed SD distributions between the full-phase and IIR-modeled data spanned beyond the range of imperceptibility as defined by [23], indicating that the spectral deviation of certain HRTF datasets could lead to audible artifacts, a fact which was also reflected in the experiment results.
2. Similarity rates of the full-phase/IIR comparison are lower that those of the full-phase/minimum-phase condition, implying that the tested IIR modeling affected HRTF spatial evaluation to a greater extent. This could be attributed to the increase in the SD values for frequency content above 8 kHz, which were observed to be as high as 7.3 dB, compared with the corresponding $SD \leq 1$ dB of the minimum-phase data. Such variations could have impacted the users' spatial assessment of the modeled HRTFs.
3. Differences in similarity levels between the horizontal and median plane ratings were not significant. This suggests that the influence of HRTF pro-

cessing on spatial quality evaluations is trajectory-independent.

While the evaluation of similarity between HRTF assessments across the full rating scale is appropriate for investigating the effect of post-processing modifications on the perceived spatial quality of HRTFs, it is also informative to focus on specific ranges in the rating scale. Admittedly most evaluation/selection protocols focus on the identification of the "best-rated" HRTFs, disregarding "mediocre" or "unfit" data [46, 47, 36, 11]. Additionally the use of a coarser rating scale has been shown to boost similarity in HRTF assessments [31]. Therefore the nine-point rating scale was down-sampled into three equal parts, reflecting *low*, *mid*, and *high* HRTF spatial quality categories, allowing examination of cross-category changes in ratings.

Using the full-phase ratings as reference, the percentage of HRTFs that were attributed the same or different spatial quality category was calculated. An outcome of this analysis concerned the difficulty of otherwise repeatable participants in providing stable ratings for mid-level HRTFs across all conditions and tested trajectory planes. As previously seen in [31], participant assessments were more stable when concerning "bad" (low spatial accuracy) or "good" (high spatial accuracy) HRTFs than "mediocre" (mid spatial quality) data. Stated differently, participants were more likely to change their assessments of mediocre HRTFs across repetitions. The percentage of "mediocre" HRTFs sustaining their level of spatial accuracy ranged between ≈29% (median plane full-phase/IIR) and ≈33% (median plane full-phase/minimum-phase).

This result is a strong indication that VAD applications, which mainly operate on non-individual HRTF data, can create confusing, if not contradicting, experiences to users. Within a collection of non-personalized data, the portion of mediocre HRTFs is not only indeterminable but also varies between individuals. A similar claim can be made for the use of such procedures on scientific studies as a means of binaural assessment, unless enough evidence is provided that participant responses are repeatable and reliable. Hence, both participant screening and repetitions of the assessment protocol are necessary to improve the validity of the collected data.

Additionally results revealed that signal modifications, even when lying below previously established thresholds of perceptual irrelevance, can shift one's assessment of spatial quality for a given HRTF. Up to ≈18% of the minimum-phase HRTFs and 30% of the IIR data evaluated as having "low" spatial quality received a "high" quality rating in the full-phase condition. Similarly up to ≈17% of the minimum-phase HRTFs and ≈16% of the IIR data evaluated as having "high" spatial quality received a "low" quality rating in the full-phase condition. Based on the aforementioned results, the hypothesis that user assessments should remain roughly unchanged, as long as the range of signal variations between processed and unprocessed data lies within ranges previously reported as perceptually insignificant, has to be rejected.

6 CONCLUSION

The dependence of a person's VAD experience on the appropriateness of the utilized binaural filters is a widely accepted fact. The use of personally measured, successfully individualized, or user-selected HRTFs has been shown to improve system fidelity and boost user experience. Nevertheless allowing people to bring and use their selected HRTFs on a given system poses certain technical complications. HRTFs are available in a variety of spatial resolutions, filter orders, and sampling rates, facts which introduce a need for post-processing and data homogenization. Yet, when users select an HRTF set with certain spatial and frequency resolutions and comprising certain signal information, how confident can one be that data post-processing will not affect their overall experience, preference, and HRTF evaluation?

This study discussed the impact of various commonly employed HRTF processing algorithms on the assessment of spatial rendering quality across a set of non-individualized binaural data. The work was based on the hypothesis that, when variations in signal content between unprocessed and processed data lie within previously established ranges of imperceptibility, user evaluations of processed HRTFs should remain roughly unchanged. This hypothesis was tested against a perceptual study, monitoring changes in repeated quality assessments of non-individualized binaural stimuli moving along pre-defined trajectories on the horizontal and median planes.

While the magnitude of the objective and subjective variations described in this paper are directly dependent on the specific implementations chosen to create minimum-phase and IIR versions of the tested data and cannot be generalized to all processes, the results highlight the effect of signal modifications on the perception of spatial quality across HRTFs of all quality levels. Despite the fact that objective evaluations suggested that the observed variations between the original and minimum-phase HRTFs lied within previously established ranges of acceptability, the experiment results demonstrated that they could completely shift one's assessment in a similar manner to that of IIR-modeled data, which carried spectral variations beyond the established thresholds of imperceptibility. More specifically it was demonstrated that the observed SD levels of both minimum-phase and IIR-modeled data could lead to the rejection of HRTFs previously rated as the "most appropriate" or alternatively to the selection of datasets previously dismissed as "unfit." The effect appeared more often for the tested IIR processing over minimum phase, but it was equally present for the two evaluated planes.

Such results further support the argument that HRTF processing and modeling algorithms should be applied with caution because they could have an impact on the perceived spatial quality of the data, which might not be predicted with current signal domain metrics. Different evaluation routines offering more elaborate assessments of the processed HRTF data need to be widely employed before the community develops a firm understanding of the potential impact of HRTF processing on human spatial perception.

One such approach would be the use of perceptual models that simulate localization performance [48, 49], though their use is not yet universally accepted. Additionally efficient participant screening and sufficient repetitions of experiment tasks are undoubtedly necessary in HRTF evaluation/selection tasks for the improvement of the validity of any collected data.

7 ACKNOWLEDGMENT

This work was funded in part by the French Fonds Unique Interministériel project BiLi ("Binaural Listening," www.bili-project.org, FUI-AAP14). Portions of this study were carried out at the Laboratoire d'informatiqu e pour la mécanique et les sciences de l'ingénieur (LIMSI) (UPR3251), Centre National de la Recherche Scientifique laboratory. The authors would like to thank the BiLi partners Arkamys and Orange Labs for their valuable assistance in HRTF post-processing. Portions of this work have been carried out in the context of the Sonicom project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement number 101017743.

8 REFERENCES

[1] O. Warusfel, "LISTEN HRTF Database," (2003), recherche.ircam.fr/equipes/salles/listen/ (accessed Apr. 20, 2022).

[2] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a Head-Related Transfer Function Database With High Spatial Resolution," in *Proceedings of the 7th Forum Acusticum 7th Forum Acusticum*, pp. 1–6 (Krakow, Poland) (2014 Sep.).

[3] B. F. G. Katz, "Boundary Element Method Calculation of Individual Head-Related Transfer Function. I. Rigid Model Calculation," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2440–2448 (2001 Oct.). https://doi.org/10.1121/1.1412440.

[4] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, "Approximating the Head-Related Transfer Function Using Simple Geometric Models of the Head and Torso," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2053–2064 (2002 Oct.). https://doi.org/10.1121/1.1508780.

[5] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123 (1993 Mar.). https://doi.org/10.1121/1.407089.

[6] B. F. G. Katz and G. Parseihian, "Perceptually Based Head-Related Transfer Function Database Optimization," *J. Acoust. Soc. Am.*, vol. 131, no. 2, pp. 99–105 (2012 Jan.). https://doi.org/10.1121/1.3672641.

[7] P. Stitt, L. Picinali, and B. F. G. Katz, "Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning," *Sci. Rep.*, vol. 9, paper 1063 (2019 Jan.). https://doi.org/10.1038/s41598-018-37873-0.

[8] A. Andreopoulou, D. R. Begault, and B. F. G. Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE J. Sel. Top. Sign. Process.*, vol. 9, no. 5, pp. 895–906 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2400417.

[9] B. Xie and T. Zhang, "The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency," *Acta Acust. united Acust.*, vol. 96, no. 2, pp. 328–339 (2010 Mar.). https://doi.org/10.3813/AAA.918282.

[10] A. Andreopoulou and A. Roginska, "Database Matching of Sparsely Measured Head-Related Transfer Functions," *J. Audio Eng. Soc.*, vol. 65, no. 7/8, pp. 552–561 (2017 Aug.). https://doi.org/10.17743/jaes.2017.0021.

[11] D. Poirier-Quinot and B. F. G. Katz, "On the Improvement of Accommodation to Non-Individual HRTFs via VR Active Learning and Inclusion of a 3D Room Response," *Acta Acust.*, vol. 5, paper 25 (2021 Jun.). https://doi.org/10.1051/aacus/2021019.

[12] A. Politis and D. Poirier-Quinot, "JSAmbisonics: A Web Audio Library for Interactive Spatial Sound Processing on the Web," in *Proceedings of the Interactive Audio Systems Symposium*, paper 16 (York, UK) (2016 Sep.).

[13] S. Carlile and D. Pralong, "The Location-Dependent Nature of Perceptually Salient Features of the Human Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 95, no. 6, pp. 3445–3459 (1994 Jun.). https://doi.org/10.1121/1.409965.

[14] A. Kulkarni and H. S. Colburn, "Role of Spectral Detail in Sound-Source Localization," *Nature*, vol. 396, pp. 747–749 (1998 Dec.). https://doi.org/10.1038/25526.

[15] D. J. Kistler and F. L. Wightman, "A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647 (1992 Mar.). https://doi.org/10.1121/1.402444.

[16] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, "Sensitivity of Human Subjects to Head-Related Transfer-Function Phase Spectra," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2821–2840 (1999 May). https://doi.org/10.1121/1.426898.

[17] M. A. Senova, K. I. McAnally, and R. L. Martin, "Localization of Virtual Sound as a Function of Head-Related Impulse Response Duration," *J. Audio Eng. Soc.*, vol. 50, no. 1/2, pp. 57–66 (2002 Feb.). www.aes.org/e-lib/browse.cfm?elib=11092.

[18] E. Rasumow, M. Blau, M. Hansen, et al., "Smoothing Individual Head-Related Transfer Functions in the Frequency and Spatial Domains," *J. Acoust. Soc. Am.*, vol. 135, no. 4, pp. 2012–2025 (2014 Apr.). https://doi.org/10.1121/1.4867372.

[19] J. G. Tylka, B. B. Boren, and E. Y. Choueiri, "A Generalized Method for Fractional-Octave Smoothing of Transfer Functions That Preserves Log-Frequency Symmetry," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 239–245 (2017 Mar.). https://doi.org/10.17743/jaes.2016.0053.

[20] M. Rothbucher H. Shen, and K. Diepold, "Dimensionality Reduction in HRTF by Using Multiway Ar-

ray Analysis," in H. Ritter, G. Sagerer R. Dillmann, and M. Buss (Eds.), *Human Centered Robot Systems*, Cognitive Systems Monographs, vol. 6, pp. 103–110 (Springer, Berlin, Germany, 2009). https://doi.org/10.1007/978-3-642-10403-9_11.

[21] H. Hugeng W. Wahab, and D. Gunawan, "Effective Preprocessing in Modeling Head-Related Impulse Responses Based on Principal Components Analysis," *Signal Process. Int. J.*, vol. 4, no. 4, pp. 201–212 (2010 Oct.).

[22] S. Takane, "Effect of Domain Selection for Compact Representation of Spatial Variation of Head-Related Transfer Function in All Directions Based on Spatial Principal Components Analysis," *Appl. Acoust.*, vol. 101, pp. 64–77 (2016 Jan.). https://doi.org/10.1016/j.apacoust.2015.07.018.

[23] A. Kulkarni and H. S. Colburn, "Infinite-Impulse-Response Models of the Head-Related Transfer Function," *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1714–1728 (2004 Apr.). https://doi.org/10.1121/1.1650332.

[24] G. Ramos and M. Cobos, "Parametric Head-Related Transfer Function Modeling and Interpolation for Cost-Efficient Binaural Sound Applications," *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 1735–1738 (2013 Sep.). https://doi.org/10.1121/1.4817881.

[25] Z. Liang, B. Xie, and X. Zhong, "Comparison of Principal Components Analysis on Linear and Logarithmic Magnitude of Head-Related Transfer Functions," in *Proceedings of the 2nd International Congress on Image and Signal Processing*, pp. 1–5 (Tianjin, China) (2009 Oct.). https://doi.org/10.1109/CISP.2009.5304273.

[26] L. J. Hobden and A. I. Tew, "Investigating Head-Related Transfer Function Smoothing Using a Sagittal-Plane Localization Model," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5 (New Paltz, NY) (2015 Oct.). https://doi.org/10.1109/WASPAA.2015.7336955.

[27] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications," in J. Blauert (Ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pp. 93–119 (Springer, Berlin, Germany, 2013). https://doi.org/10.1007/978-3-642-37762-4_4.

[28] H. G. Hassager, F. Gran, and T. Dau, "The Role of Spectral Detail in the Binaural Transfer Function on Perceived Externalization in a Reverberant Environment," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2992–3000 (2016 May). https://doi.org/10.1121/1.4950847.

[29] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual Attributes for the Comparison of Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632 (2016 Nov.). https://doi.org/10.1121/1.4966115.

[30] A. Andreopoulou and B. F. G. Katz, "Comparing the Effect of HRTF Processing Techniques on Perceptual Quality Ratings," presented at the *144th Convention of the Audio Engineering Society* (2018 May), paper 9920. http://www.aes.org/e-lib/browse.cfm?elib=19437.

[31] A. Andreopoulou and B. F. G. Katz, "Investigation on Subjective HRTF Rating Repeatability," presented at the *140th Convention of the Audio Engineering Society* (2016 May), paper 9597. www.aes.org/e-lib/browse.cfm?elib=18295.

[32] A. Andreopoulou and B. F. G. Katz, "Subjective HRTF Evaluations for Obtaining Global Similarity Metrics of Assessors and Assessees," *J. Multimodal User Interfaces*, vol. 10, pp. 259–271 (2016 Mar.). https://doi.org/10.1007/s12193-016-0214-y.

[33] B. F. G. Katz and M. Noisternig, "A Comparative Study of Interaural Time Delay Estimation Methods," *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. 3530–3540 (2014 Jun.). https://doi.org/10.1121/1.4875714.

[34] J. C. Makous and J. C. Middlebrooks, "Two-Dimensional Sound Localization by Human Listeners," *J. Acoust. Soc. Am.*, vol. 87, no. 5, pp. 2188–2200 (1990 May). https://doi.org/10.1121/1.399186.

[35] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller, "The Interaural Time Difference in Binaural Synthesis," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), paper 5133. www.aes.org/e-lib/browse.cfm?elib=9205.

[36] A. Andreopoulou and B. F. G. Katz, "Identification of Perceptually Relevant Methods of Inter-Aural Time Difference Estimation," *J. Acoust. Soc. Am.*, vol. 142, no. 2, pp. 588–598 (2017 Aug.). https://doi.org/10.1121/1.4996457.

[37] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," in *Proceedings of the International Computer Music Conference (ICMC)*, pp. 270–277 (Denton, TX) (2015 Sep.).

[38] M. Emerit, J. Faure, A. Guerin, et al., "Efficient Binaural Filtering in QMF Domain for BRIR," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), paper 7095. http://www.aes.org/e-lib/browse.cfm?elib=14080.

[39] F. Amadu and J. M. Raczinski, "An Efficient Implementation of 3-D Audio Engine for Mobile Devices," in *Proceedings of the AES 35th International Conference: Audio for Games* (2009 Feb.), paper 12. www.aes.org/e-lib/browse.cfm?elib=15166.

[40] K. Watanabe, R. Kodama, S. Sato, S. Takane, and K. Abe, "Influence of Flattening Contralateral Head-Related Transfer Functions Upon Sound Localization Performance," *Acoust. Sci. Technol.*, vol. 32, no. 3, pp. 121–124 (2011 May). https://doi.org/10.1250/ast.32.121.

[41] P. Majdak, M. J. Goupell, and B. Laback, "3-D Localization of Virtual Sound Sources: Effects of Visual Environment, Pointing Method, and Training," *Atten. Percept. Psychophys.*, vol. 72, no. 2, pp. 454–469 (2010 Feb.). https://doi.org/10.3758/APP.72.2.454.

[42] B. Katz and R. Nicol, "Binaural Spatial Reproduction," in N. Zacharov (Ed.), *Sensory Evaluation of Sound*, pp. 349–388 (CRC Press, Boca Raton, FL, 2019).

[43] F. Zagala, M. Noisternig, and B. F. G. Katz, "Comparison of Direct and Indirect Perceptual Head-Related Transfer Function Selection Methods," *J. Acoust.*

Soc. Am., vol. 147, no. 5, pp. 3376–3389 (2020 May). https://doi.org/10.1121/10.0001183.

[44] R. Nicol, L. Gros, C. Colomes, et al., "A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, pp. 100–106 (Berlin, Germany) (2014 Apr.). https://doi.org/10.14279/depositonce-17.

[45] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," *Am. Stat.*, vol. 32, no. 1, pp. 12–16 (1978 Feb.). https://doi.org/10.2307/2683468.

[46] Y. Iwaya, "Individualization of Head-Related Transfer Functions With Tournament-Style Listening Test: Listening With Other's Ears," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 340–343 (2006 Nov.). https://doi.org/10.1250/ast.27.340.

[47] A. Roginska, T. S. Santoro, and G. H. Wakefield, "Stimulus-Dependent HRTF Preference," presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), paper 8268. www.aes.org/e-lib/browse.cfm?elib=15690.

[48] P. L. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in J. Blauert (Ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pp. 33–56 (Springer, Berlin, Germany, 2013).

[49] R. Baumgartner, P. Majdak, and 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.). https://doi.org/10.1121/1.4887447.

A.1 COMPLEMENTARY DATA ANALYSIS

Beyond variations in the rating scores of HRTFs processed differently, there exist other information that can indicate how challenging each task was and highlight possible strategies that participants devised in order to complete the experiment. These are the use of the rating scale, that is, the number of different ratings out of the nine available that participants chose to use in order to rank HRTFs as part of the pre-test and main experiment, and the time they needed to complete each experiment round. The potential connection between this information and HRTF processing method is investigated below.

1 PRE-TEST: DURATION AND SCALE USAGE

The pre-test was completed in a single sitting with an average duration of 39.1 min (standard deviation: 17.4 min). During every round, the order of HRTFs and trajectories was fully randomized to avoid potential biases in the evaluation procedure. The normality of the data was tested using a Lilliefors test. The test failed to reject the null hypothesis that the data was normally distributed ($\alpha = 0.05$). Table 4 shows the average and standard deviation of the test duration across each of the seven rounds and two trajectory planes. As can be seen, the average duration of each round progressively decreases as the number of round-repetitions increases. This observation implies that, on average, participants spent more time during the first rounds familiarizing themselves with the task, procedure, and test-interface and, possibly, devising a strategy for HRTF rating. This appears to be true for both sound-trajectories under evaluation. This hypothesis is supported by a repeated measures two-way analysis of variance comparing the main effect of round (1–7) and plane (horizontal/median) and their interaction on the duration of each round. The effect of round was found to be significant [$\alpha = 0.05$; $F(6, 262) = 7.42$; $p < 0.001$], while the effect of trajectory plane was not [$F(1, 262) = 0.82$; $p = 0.38$]. The interaction of round and plane was not found to be significant [$F(6, 262) = 0.73$; $p = 0.63$]. A Bonferroni post hoc test showed that for the horizontal plane trajectory, the duration of round 1 was significantly different from those of rounds 3–7, while for the median plane the duration of round 1 was significantly different from that of rounds 6 and 7 ($\alpha < 0.05$).

Another factor that could potentially influence the repeatability of the users' assessments is that of the resolution of the utilized rating scale. The task comprised rating each binaurally rendered trajectory on a nine-point scale ranging from "worst" to "best." The experiment setup permitted participants to apply the same rating on multiple trajectories and use a subset of the available scale steps for their assessments, so long as they used the two scale extremes (1 and 9) at least once for each round. As a result, participants who chose to utilize a coarser scale (i.e., consistently using only three out of the possible nine rating values) to evaluate the sound trajectories would, in principle, have an advantage over other participants in terms of the rating repeatability. Therefore it is of some interest to examine the use of the rating scale as a function of pre-test round

Table 4. Mean and standard deviation of the duration (min) and number of scale steps used per pre-test round grouped according to trajectory plane and participant repeatability levels.

	Rounds						
	1st	2nd	3rd	4th	5th	6th	7th
Duration (Hor. - all parts.)	6.10 (8.36)	3.23 (1.65)	3.10 (1.93)	2.70 (1.39)	2.38 (1.43)	1.88 (1.00)	1.90 (1.21)
Duration (Med. - all parts.)	5.25 (2.90)	4.55 (3.15)	3.52 (1.37)	2.52 (1.06)	2.36 (1.32)	2.10 (1.15)	1.90 (0.96)
Scale Use (Hor. - Repet, parts.)	5.54 (1.27)	5.46 (1.39)	5.31 (1.25)	5.25 (1.58)	5.38 (1.51)	5.13 (1.46)	5.80 (0.84)
Scale Use (Med. - Repet, parts.)	5.62 (1.26)	5.23 (1.30)	5.46 (1.27)	5.50 (1.60)	5.25 (1.67)	5.38 (1.51)	5.60 (0.55)
Scale Use (Hor. - nonRepet, parts.)	6.00 (0.94)	5.40 (1.08)	5.20 (0.92)	5.70 (1.41)	5.30 (1.34)	5.00 (1.49)	5.50 (1.58)
Scale Use (Med. - nonRepet, parts.)	5.50 (1.43)	5.00 (1.56)	5.00 (1.41)	4.80 (1.69)	5.40 (1.84)	5.10 (1.66)	5.10 (1.85)

(1–7), plane (horizontal or median), and participant group (repeatable or non-repeatable).

As seen in Table 4, the average use of distinct rating-scale steps employed by "repeatable" and "non-repeatable" participant sub-groups did not change much as a function of experiment round and plane. The average use of rating scale across both tested planes for the repeatable participants ranged between 5.1 and 5.8 scale steps, while that for non-repeatable participants between 4.8 and 6.0. Such results show that the repeatability in the ratings of the qualified 13 participants was not related to the resolution of the utilized rating scale but rather reflected their ability to assess HRTFs in a consistent and reliable manner.

2 MAIN EXPERIMENT: DURATION AND SCALE USAGE

The analysis began by looking at the duration of each experiment round as a function of HRTF processing method and trajectory plane. A significant increase in duration could reflect an increase in the complexity of the task, which could then be traced back to the type of processing of the evaluated HRTFs. The normality of the data was tested using a Lilliefors test. The test failed to reject the null hypothesis that the data was normally distributed ($\alpha = 0.05$). Table 5 summarizes the duration spent in each round based on the type of HRTF processing. In general, durations were comparable, with the IIR rounds being consistently longer than the two others. A two-way repeated measures analysis of variance was performed to investigate the effects of trajectory plane and HRTF processing on the duration of each experiment round. Neither the effect of trajectory [$F(1, 78) = 0.54$; $p = 0.48$] nor of HRTF processing order [$F(2, 78) = 0.6$; $p = 0.56$] were found to be significant ($\alpha = 0.05$). None of the interactions between the independent variables were found to be statistically significant.

The use of the rating scale could also be an indication of the effect of HRTF processing on user evaluation. It is possible that, when given the choice, participants' use of the rating scale resolution would reflect the perceived spatial variation in the assessed corpus. A coarser rating scale could imply that a certain type of processing reduced spatial variations among otherwise distinct datasets. As shown in Table 5, the average amount of discrete scale steps used for the assessment of the 12 HRTFs in each experiment round did not vary as a function of trajectory or processing. On average, participants used approximately seven of the nine available scale steps to rate all trajectories, with the minimum being used for the minimum-phase data across both trajectories. As a result, it can be concluded that any perceived variations in spatial quality and/or accuracy is reflected on the user ratings of the evaluated data.

Table 5. Mean and standard deviation of the duration (min) and number of scale steps used in each experiment round grouped per HRTF encoding method and trajectory plane.

	HRTF processing		
	Full Ph.	Min. Ph.	IIR
Duration (Hor.)	8.40 (4.90)	7.52 (5.65)	9.51 (7.50)
Duration (Med.)	8.44 (4.23)	8.97 (5.93)	9.43 (7.45)
Scale Use (Hor.)	7.46 (1.76)	7.00 (1.15)	7.07 (1.32)
Scale Use (Med.)	6.77 (1.79)	6.69 (1.37)	7.38 (0.87)

THE AUTHORS

Areti Andreopoulou

Brian F. G. Katz

Areti Andreopoulou is an Assistant Professor in Music Technology at the Department of Music Studies and researcher in the Laboratory of Music Acoustics and Technology (LabMAT) at the National and Kapodistrian University of Athens, Greece. She has a Bachelor's degree in music studies from the National and Kapodistrian University of Athens (2005) and Master's (2008) and Ph.D. (2014) degrees in music technology from New York University. Her fields of interest include spatial audio, the design and evaluation of immersive environments, auditory displays, acoustics, and data sonification.

•

Brian F. G. Katz is a Centre National de la Recherche Scientifique (CNRS) Research Director at the ∂'Alembert Institute, Sorbonne Université, CNRS, and coordinator of the Sound & Space research theme. His fields of interest include spatial 3D audio rendering and perception and room acoustics. With a background in physics and philosophy, he obtained his Ph.D. in Acoustics from Penn State in 1998 and his habilitation à diriger des recherches (HDR) in Engineering Sciences from Université Pierre et Marie Curie (UPMC) in 2011. Before joining CNRS he worked for various acoustic consulting firms, including Artec Consultants Inc., ARUP & Partners, and Kahle Acoustics. He has also worked at LIMSI-CNRS and Institut de Recherche et Coordination Acoustique/Musique.