Full Papers

ON THE USE OF SUBJECTIVE HRTF EVALUATIONS FOR CREATING GLOBAL PERCEPTUAL SIMILARITY METRICS OF ASSESSORS AND ASSESSEES

Areti Andreopoulou

Audio Acoustics Group
LIMSI - CNRS
Orsay, France
andreopoulou@limsi.fr

Brian FG Katz

Audio Acoustics Group
LIMSI - CNRS
Orsay, France
brian.katz@limsi.fr

ABSTRACT

In the absence of a well suited measure for quantifying binaural data variations, this study presents the use of a global perceptual distance metric which can describe both HRTF as well as listener similarities. The metric is derived based on subjective evaluations of binaural renderings of a sound moving along predefined trajectories in the horizontal and median planes. Its characteristics and advantages in describing data distributions based on perceptually relevant attributes are discussed. In addition, the use of 24 HRTFs from two different databases of origin allows for an evaluation of the perceptual impact of some database-dependent characteristics on spatialization. The effectiveness of the experimental design as well as the correlation between the HRTF evaluations of the two plane trajectories are also discussed.

1. INTRODUCTION

Binaural audio is becoming an increasingly popular technology in the fields of auditory displays, entertainment, virtual reality applications etc. Nevertheless, it is often the case that binaural applications are not built around individual Head-Related Transfer Functions, which would allow for optimal user experience, but rather around non-individual data from pre-existing databases, a fact that can cause an unpredictable degree of spatial distortion for some users. The definition of an appropriate HRTF similarity metric that would explicitly describe that distortion has been a topic of interest in the field for many years.

One possibility for defining similarity between HRTF data is by quantifying variations in their spatial spectral and temporal cues. Several objective metrics have been used for this purpose, such as the average linear difference of Principal Component Analysis (PCA) weights of Directional Transfer Functions (DTF) [1, 2] or of logarithmic HRTF magnitudes [3], the correlation coefficients of logarithmic DTFs [4, 5], the Mean Square Error (MSE) of HRTF magnitudes [6], and the average Signal to Distortion Ratio (SDR) of logarithmic HRTF magnitudes [7]. The biggest limitation of these purely signal domain based approaches is that they can neither describe the perceptual implications of the quantified differences nor take into account brain plasticity, as studies have shown that the human auditory system is capable of successfully adapting to altered spectral cues, given time with passive adaptation [8] or more quickly with active training [9].

As an alternative, information about the quality of HRTFs can be extracted subjectively with the use of binaural localization or evaluation tasks. In the first case, participants are asked to identify the perceived location of virtual sound sources. Analysis of localization errors quantifies the degree of spatial distortion in the virtual space [8, 10]. In the second case, assessments of the perceived HRTF quality can be made through pair-wise comparison tasks [11], or via ratings using binary [12], fixed-point [13], or continuous scales [14]. Even though subjective evaluations cannot provide any information regarding the signal properties related to the assessments, they reflect the perceptual impact of a given HRTF on an individual's spatial experience. This information is more relevant for research investigating methods for optimizing user's binaural quality of experience.

In this context, this study presents a method for utilizing subjective evaluations as a means for creating global similarity metrics of HRTFs and listeners. Evaluations were obtained through a perceptual study of various binaural sound stimuli moving along trajectories on the horizontal and median planes.

2. EXPERIMENTAL PROCEDURE

2.1. Design

Fifteen people (five female) including one of the authors, selected from those contributing to the BiLi HRTF database [15], participated in this evaluation study. All were professionals or scholars in audio signal processing or audio engineering with extensive experience in binaural audio reproduction. The goal was the evaluation of the perceived quality of sound stimuli as a function of HRTF for known trajectories, specifically using an impulsive sound source moving along predefined trajectories in the horizontal and median planes.

Individual participant binaural renderings were created using 24 datasets from the publicly available LISTEN [16] and BiLi [15] HRTF databases. The corpus included the individual measured HRTFs of all participants (15 sets from BiLi, two from LISTEN) as well as a subset of seven LISTEN HRTFs, perceptually optimised as sufficient to cover a wide population range in a previous study [13]. In additional, two participants were represented by HRTFs present in both databases, resulting in a total of $15+7+2=24$ HRTFs. The presence of repeated HRTFs from the two databases allowed for an investigation into a) the consistency in participants' ratings for the two pairs of individual HRTFs and b) the possible role of database-dependant characteristics, discussed in [17], on the perceived quality of HRTF data.

This work was funded in part by the French FUI project BiLi ("Binaural Listening", www.bili-project.org, FUI-AAP14)

Both the LISTEN and BiLi databases contain anechoically measured blocked-meatus HRTFs comprising 187 and 1680 measured positions, respectively. LISTEN contains measurements at elevations from -45° to 90° at rough angular increments of 15°, while BiLi contains measurements on a Gaussian grid at elevations from -51° to 86° at rough angular increments of 6°. The obvious differences in the spatial resolution between the two databases rendered the creation of identical trajectory paths a very challenging task. In order to avoid potential artifacts introduced to the data through interpolation, the sound trajectories were created using exclusively measured locations. More specifically, the horizontal plane trajectories consisted of 12 angles from 0° to 330° in increments of 30°. For cases when no common coordinates existed the closest measured points were selected. The median plane trajectories created with the LISTEN HRTFs comprised 10 angles from -45° to 90° in increments of 15° while those created with BiLi comprised 11 angles of -45°, -27°, -15°, 0°, 15°, 27°, 45°, 62°, 74°, 86°, 106°, 118°, 135°, 153°, 165°, 180°, 195°, 207°, and 225°. Such differences of up to 4° were smaller that the reported localization blur at these elevations [18], and were therefore assumed to be perceptual irrelevant for the trajectory rendering.

All HRTFs were diffuse-field equalized, low-pass filtered at 20 kHz, and sample rate converted to 44.1 kHz. Potential DC offsets were removed and impulse responses were truncated using a 512-sample rectangular window in order to eliminate the presence of any room reflections. The window starting point was set to 20 samples ahead of the first detected onset as evaluated across the entire HRTF dataset. The onset was defined as the first sample greater than –10 dB relative to the peak value. In addition, the median global RMS value over all positions was selected as reference for level normalization across the all HRTF datasets.

As this study was primarily interested in the impact of spectral cues on personalized binaural listening, the Inter-aural Time Differences (ITDs) were kept consistent for each participant across the HRTF corpus. Individualized position dependent ITD values were estimated from every participant's measured HRIR using the centroid of the inter-aural cross-correlation method (CenIACCr) [19] and used as a reference for individualization. In contrast to previous studies where HRTFs were converted to minimum phase and the individually estimated ITDs were inserted as pure delays [13], this study maintains the full-phase component of the binaural filters. ITD adjustments were made by adapting the temporal alignment of the filter pairs for the remaining HRTF sets, such that the extracted ITD information coincided with the corresponding reference values.

2.2. Perceptual Evaluation

For the purposes of this study, a simple individualized auditory scene was created for each of the participants using the 24 HRTFs. The sound stimulus was 100 ms of Gaussian noise (50 Hz to 20 kHz) with 2 ms hamming ramps at onset and offset, chosen based on the results of [20]. Stimuli were presented sequentially along either of the two trajectories at the angular increments described above. 50 ms of silence was inserted between successive locations for both trajectories. The interface was designed in the MatLab programing environment in a self-manageable (unsupervised) manner. Detailed written instructions were provided explaining the goals of the test and offering information on how to run it and use its interface. Participants were instructed to complete the study in a listening room with an ambient noise level below 30 dBA using Sennheiser HD600 headphones and an RME Fireface 400 audio interface. Prior to the test, sound levels were calibrated to 80.5 dBA using a monophonic 1 kHz sine wave reference signal with the headphone placed on a baffled microphone.

The study consisted of two blocks, one for the horizontal plane trajectory and one for the median plane trajectory. The presentation order of the two trajectories blocks was fully randomized across participants. For each block, participants were presented with a single interface containing playback options for all 24 sound samples (one per HRTF) and were asked to rate the perceived spatial quality of each sample on a forced-choice 9-point rating scale with extreme limits defined as "poor" and "excellent". The quality assessment was relative to the written description of the two trajectory paths which was provided describing the current plane and the direction of sound movement along a fixed radius circle or arc. No further training was provided. Participants had full control over the experiment procedure; being allowed to play-back the sound samples in any order and listen to them repeatedly at will. The mean total test duration was 26 min (std 13 min).

One goal of this work is the creation of a perceptually determined space of HRTF data where observations regarding similarities between HRTF-related content can be defined based on subjective evaluations. The protocol was designed in such a way to allow additional investigations concerning spatial quality evaluations as a function of (a) test trajectory, revealing the potential impact of the selected positions under study through the use of separate ratings for horizontal and median plane trajectories, (b) HRTF database origin, through the use of two HRTF database sources, and (c) subjective repeatability, though the presence of 2 pairs of duplicate HRTFs, originating from the two different databases acquired roughly 10 years apart.

3. GENERAL OBSERVATIONS

3.1. Data normalization

Even though participants were encouraged to explore the full rating scale in the evaluation experiment, observation of the responses revealed that some refrained from utilizing the values at the two scale extremes. In order to compare responses across users, each individual participant block of ratings was normalized such that their local minima and maxima would extend to the global upper and lower limit values. This was achieved by subtracting from each user's response block the local minimum value and dividing by the local maximum, hence forcing all ratings to the range between 0 and 1. This normalization results in the conversion of the absolute quality ratings provided ("poor" to "excellent") to a relative rating scale, where the best and worst performing HRTFs for each participant are scaled to 0 and 1, respectively.

3.2. Use of rating scale

In contrast to previous studies employing binary [12], 3-point [13], or continuous rating scales [14], this work collected ratings on a fixed 9-point scale. The intention was to allow for higher resolution evaluations, while avoiding the potentially lengthy test-time for continuous rating scales due to participants focusing on very fine details. One of the primary interests was to investigate how participants utilized the scale. As seen in Figure 1, which presents the normalized rating of the 24 HRTFs according to horizontal and median plane trajectory ratings for 3 participants, there exists three categories: 1) those who utilized the full scale to rate the binaural samples for both trajectories (e.g. participant 5), 2) those who utilized somewhat less steps for ratings on one of the two planes (e.g. participant 6), and 3) those who used very few steps to rate the trajectories (e.g. participant 11 used 6 out of 9 steps for the horizontal plane and 4 out of 9 for the median plane).

A detailed overview of the scale steps used per trajectory can be found in Table 1. Even though interval scales do not dictate the use of

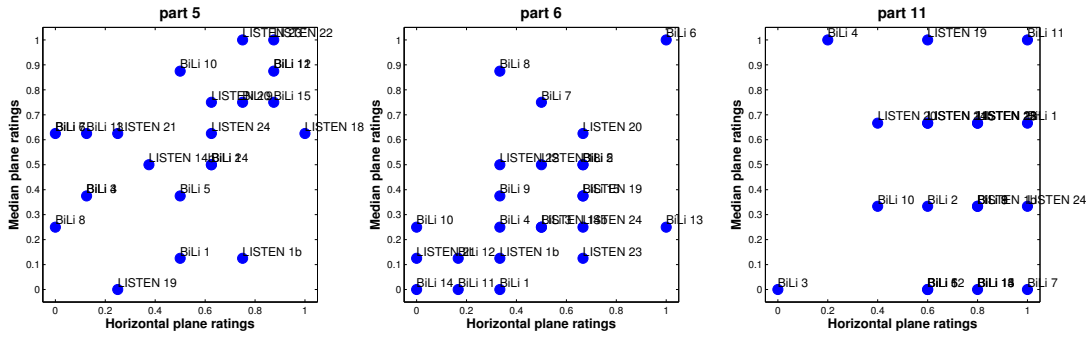

Figure 1: Distribution of HRTFs according to horizontal and median plane trajectory ratings for 3 participants.

Table 1: Distribution of the number of fixed-point steps employed by per trajectory across participants.

# of steps	Horizontal plane # of participants	Median plane # of participants
9	3	3
8	5	5
7	2	1
6	2	1
5	1	0
4	2	3
3	0	2

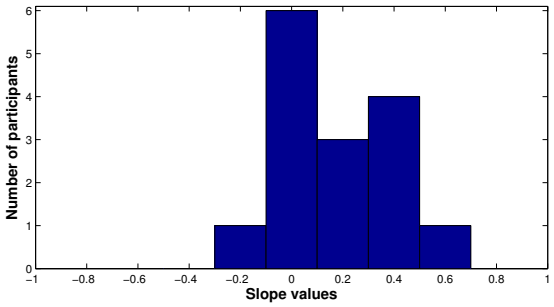

Figure 2: Distribution of linear fit slopes across the 15 participants.

the whole rating range, it is noted that up to 1/3rd of the participants utilized only 3 to 4 steps to rate either of the two trajectories. Past studies have highlighted that the use of too few scale steps reduces the discrimination between the evaluated data points [13]. An example of this can be seen at the normalized horizontal and median plane ratings of participant 11 in Figure 1. This issue can be addressed through alternate experiment designs which force participants to utilize the entire rating scale.

3.3. Horizontal and Median plane ratings association

While the spatial quality characteristics of the 24 HRTFs were evaluated separately for the horizontal and median plane trajectories, the presence, or absence, of similarity in the ratings is of significant interest. The expectation of a positive linear relationship between the two, where a given HRTF would receive very similar ratings in both planes, seems a rather intuitive hypothesis.

In order to test this hypothesis, a linear regression fit between the horizontal and median plane ratings was calculated for each participant using the least squares approach. The slope of the fitted line was used as as a metric of the relationship between the two ratings. A positive slope would imply that HRTFs were rated in an equivalent manner in both trajectories; a negative slope implies that high ratings for an HRTF in one trajectory would evoke poor rating in the other, while slope values in the region of 0 would indicate the absence of a clear relationship between ratings.

The distribution of the linear fit slopes across all 15 participants (see Figure 2) demonstrates that the hypothesis of a strong positive association between the responses for the two understudy trajectories cannot be fully justified. Only five participants exhibited a mild tendency for positively linking the two ratings ($0.3 < slope \leq 0.6$), but neither the slope values nor the number of participants were sufficient to confirm the hypothesis. The only exception is the single subject with a high slope value in the region of 0.6, who showed a tendency of rating HRTF data in a consistent manner along both trajectories. The majority of regression slopes were in the region of 0 ($|slope| \leq 0.3$), suggesting that the HRTF ratings for the two trajectories were not related in a intuitive manner. However, the presence of one participant presenting a modest tendency of inversely rating HRTFs between trajectories ($slope = -0.2$) was an interesting finding. In accordance with these observations, any subsequent analysis of participant responses will be conducted separately per trajectory.

4. CONSISTENCY IN HRTF RATINGS

4.1. Evaluation of individually measured HRTFs

Since participants were not informed of which trajectory sample(s) in the test stimuli corpus were created with their own HRTF(s), it is of interest to examine how they evaluated their individual data. To accomplish this, ratings were divided into three categories equating to "satisfactory" spatialization (top third of the rating scale), "mediocre" spatialization (middle third), and "unsatisfactory" spatialization (lower third). Figure 3 presents the rating results of the 24 HRTFs per participant, separately for the two trajectories, on the normalized scale from 0 to 1, color-coded according to the three result categories.

Upon inspection of the responses, the majority of the participants

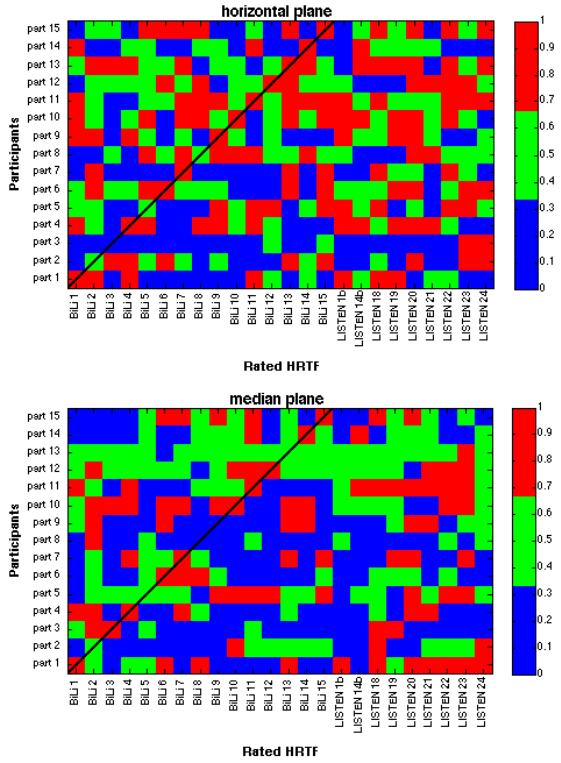

Figure 3: Normalized subjective quality ratings on a scale from 0 ("worst") to 1 ("best"). Participant IDs allow for identification of personal datasets, also indicated by the black unity line. BiLi$_1$ and LISTEN$_{1b}$ are associated to participant 1; BiLi$_{14}$ and LISTEN$_{14b}$ are associated to participant 14.

rated their personal HRTFs, as "satisfactory". Nevertheless, there exists cases for both trajectories where non-individual HRTF stimuli were rated higher than individual HRTF stimuli. More specifically, for the horizontal plane trajectory, 10 participants rated their own HRTF as "satisfactory", 4 as "mediocre", and 1 as "unsatisfactory", while for the median plane trajectory the corresponding numbers were 10, 3, and 2, respectively, with no correspondence between trajectories for subjects providing "poor" individual HRTF rating.

In addition, the presence of two HRTF sets in the stimuli corpus for participants 1 and 14 (BiLi$_1$ & LISTEN$_{1b}$ and BiLi$_{14}$ & LISTEN$_{14b}$) allows for an analysis of the impact of database-dependent characteristics on the evaluation of individual HRTFs. As seen in Figure 3, both participants (1 and 14) attributed "satisfactory" ratings for their HRTF pairs for both trajectories. No common preference was observed between the two database sources for the two subjects. This indicates that individual spatialization cues prevailed over potential database-dependent variations, rendering them somewhat perceptually irrelevant. The effect of such variations in the case of non-individual HRTFs as depicted in the ratings of the two HRTF pairs by the remaining participants is discussed bellow.

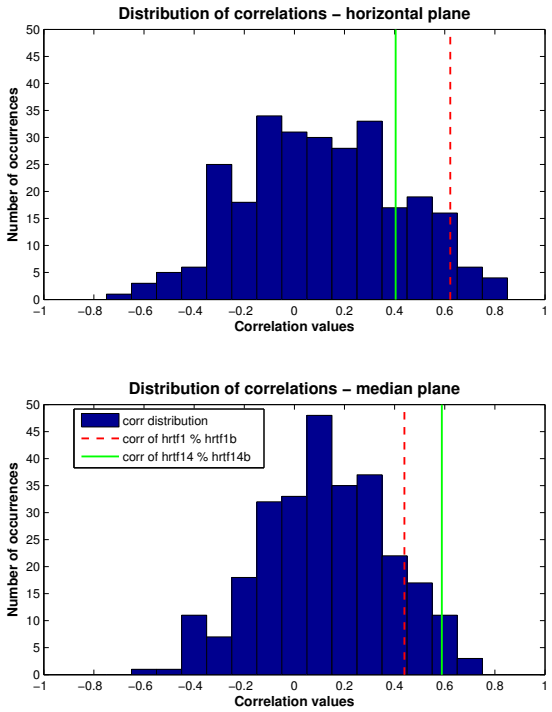

Figure 4: Distribution of correlation values of the participants' responses between all HRTF pairs per trajectory. The red dashed line indicates the correlation values between BiLi$_1$ and LISTEN$_{1b}$; the green line indicates those between BiLi$_{14}$ and LISTEN$_{14b}$.

4.2. Evaluation of non-individually measured HRTFs

The correlation of participant ratings was used as a metric for quantifying the possible impact of database of origin on the perceived spatial quality of non-individually measured HRTFs. More specifically, correlations of participant responses per trajectory were calculated between the two HRTF pairs. The resulting values were 0.62 (horizontal plane) and 0.43 (median plane) for BiLi$_1$ and LISTEN$_{1b}$, and 0.40 (horizontal plane) and 0.59 (median plane) for BiLi$_{14}$ and LISTEN$_{14b}$.

These results were compared to the resulting correlation values between the ratings of all possible HRTF pairs in the study. The distribution of results for both trajectories follows normal distributions (see Figure 4). Results for the horizontal plane have a mean of 0.1 (std 0.31); median plane results have a mean of 0.12 (std 0.26).

Results can be compared to the correlation values of the two repeated subject HRTF pairs. As can be observed, the correlation of rating results between BiLi$_1$ and LISTEN$_{1b}$ are 2 std from the mean in the horizontal plane and 1.65 std from the mean for the median plane. The corresponding values between BiLi$_{14}$ and LISTEN$_{14b}$ are 1.3 and 2.27 std, respectively. This finding implies once again that the common spatial auditory cues of these two HRTF pairs prevailed over any potential database-dependent characteristics in the data. However, it is apparent that such effects had a stronger impact in the case of non-individual as compared to individual HRTF data, where they were rendered perceptually irrelevant.

5. SUBJECTIVELY INDUCED SIMILARITY METRIC

In contrast to numerous studies which have used objective metrics for quantifying similarities in HRTF data [1, 3, 4, 5, 6], this work proposes a method for creating perceptually relevant spaces for HRTF and listener/participant distributions exclusively based on subjectively induced similarities. In such spaces, any data with associated evaluations will appear closer than the rest. For example, any HRTF datasets which have received similar ratings by the same participants will be closer in the space than those who have not.

5.1. Perceptually driven HRTF space

A spatial projection of the HRTF ratings was constructed by computing the trajectory dependent correlations of participant ratings between all HRTF pairs in the corpus as discussed in Section 4. The computed correlations were used to create a distance matrix for each trajectory representing the similarity between all HRTF pairs. In order to obtain a visual representation that would allow for a comparison with past research, Classical Multi-Dimensional Scaling (CMDS) was applied on the distance matrices to create a euclidean space. In this constructed space, HRTFs which were rated similarly by the same participants appear in close proximity. Past research using similar methods on HRTF spectral characteristics has demonstrated that such HRTF distributions exhibit strong database dependencies with HRTF sets originating from the same database tending to cluster together forming distinct groups, unless extensive post-precessing and standardization are performed [21]. However such database dependent variations do not appear to affect perceptual judgements in an equally drastic manner (see Section 4).

In order to evaluate how this proposed perceptually induced distance metric performs in the case of multi-database HRTF collections, the CMDS was performed on the correlation distance matrices and the rated HRTFs were projected on a 2-dimensional plane (stress = 0.1). As seen in Figure 5, the database of origin does not appear to play a role in the distribution of results for either trajectory. As such, this distribution preserves the cognitive impact of any possible database-dependent characteristics, rendering this representation more informative for various user evaluation tasks.

Upon inspection of Figure 5, there appears to be two extremes in the median plane distribution, BiLi$_2$ and BiLi$_{15}$. The interesting fact about these datasets is that their ratings were almost negatively correlated. The majority of participants who rated highly one of the two, rated very poorly the other. The particular relationship of these two HRTFs, which would otherwise have been hard to observe, is very obvious in this distribution by the position of the two HRTFs at opposite sides in the space.

Interestingly, the two pairs of HRTFs belonging to the same participants (BiLi$_1$ & LISTEN$_{1b}$ and BiLi$_{14}$ & LISTEN$_{14b}$), do not appear in immediate proximity to one another in the distribution. Since proximity in this space implies a global consensus in the ratings of two HRTFs, this finding implies that participants where not in complete agreement in their assessments of these datasets. Further analysis was deemed necessary in order to determine whether this observation has merit and is not just caused by a distortion of the space projection on a 2-dimensional plane.

Hierarchical clustering was applied to the two correlation distance matrices of HRTF ratings with the results depicted as dendrograms in Figure 6. The thresholds for cluster formations were set to 0.6 (horizontal plane) and to 0.57 (median plane), corresponding to the distance equivalents of the minimum correlation values of the two repeated HRTF pairs present in both databases as representing a reasonable limit for perceptual similarity (see Section 4.2). These threshold values resulted in the identification of 6 clusters and 2 singletons in the horizontal plane, and 8 clusters and 4 singletons in the median plane.

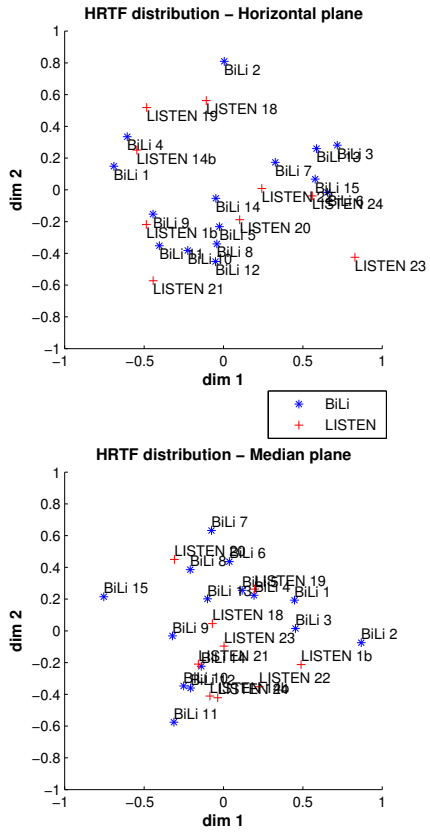

Figure 5: Distribution of HRTFs in a 2D perceptually induced space. Different markers and colors indicate database of origin.

The data clustering confirmed the observation that there was no consensus between the ratings of these HRTF pairs in space. This finding can have multiple interpretations: a) some participants did not identify an equivalence in the auditory cues of these HRTF pairs, b) database-dependent characteristics had a stronger impact on the evaluations for some of the participants, c) the observation is an artifact introduced in the data by the normalization procedure, or d) participant responses are inherently noisy to a large degree.

It can be noted that some studies have shown cases of participants in HRTF evaluation or selection tasks who exhibited preferences towards HRTFs from one database when data from multiple collections is combined [12]. It is therefore of interest to examine whether similar behaviors were present in this study. The difference is that the constructed HRTF space does not capture isolated preferences of participants, but presents general trends which could be reflected as single-database clusters of HRTF data.

For the horizontal plane HRTF distribution, If one excludes the singleton clusters, it can be observed that none of the remaining clusters contain HRTFs from only one of the two HRTF databases. This result indicates that for this part of the evaluation procedure there were no

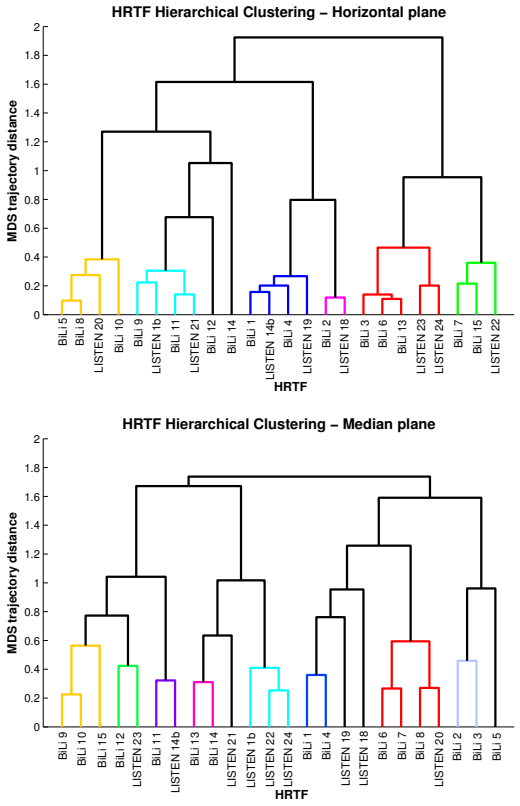

Figure 6: Dendrogram of HRTF data hierarchical clustering in the perceptually driven space based on horizontal (upper) and median (lower) plane trajectory rating similarities.

subgroups of participants who treated entries from one of the HRTF databases in a similar manner. Nevertheless, for the median plane HRTF distribution, there exist three 2-element clusters containing just HRTFs from the BiLi database. Increasing the threshold for cluster formation to a more tolerant value, for example to 1.0, resulting in the merging of small clusters into larger groupings resolves the issue of singleton clusters, but still results in the presence of a BiLi cluster comprising of HRTFs BiLi_2, BiLi_3, and BiLi_5. This observation suggests that potential database-dependent characteristics could have a stronger impact in the perceived quality of elevation changes (median plane trajectory evaluation) than for azimuthal changes (horizontal plane trajectory evaluations). Alternatively, this result could be an artifact of the limited size of the tested corpus and the inherent degree of variations in HRTFs.

5.2. Perceptually driven participant space

Similar to the generated HRTF spaces, perceptually driven participant spaces can also be created by computing correlations of HRTF rating distributions between all pairs of participants. In this space, similarities in subjective ratings over the entire data set can be employed to create a subjectively induced metric for similarity. In other words, participants who tended to rate the same HRTF corpus in a similar manner will be in proximity to each other. Participant clustering in this space could be an indication of spectral content similarity between participants' HRTFs,

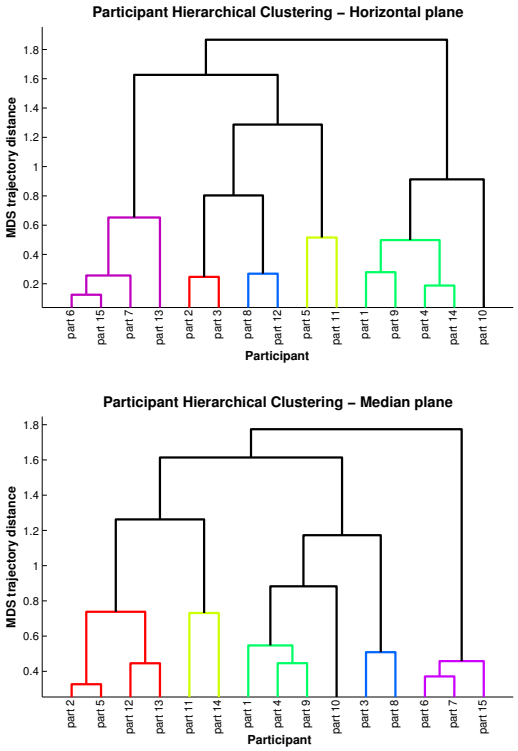

Figure 7: Dendrograms of participants' HRTF rating hierarchical clustering in the perceptually driven space based on horizontal (upper) and median (lower) plane trajectory rating similarities.

or of common standards for HRTF evaluation, implying that there might be groups of people sharing expectations for binaural audio renderings.

Figure 7 presents the hierarchical clustering results of the participants based on HRTF evaluations of the horizontal and median plane trajectory stimuli using the same criteria as previously defined. Both spaces produced 5 clusters and 1 singleton. A topic of interest in this space would be to investigate the consistency in participant clustering across the two trajectory planes, which would suggest that certain participant subgroups evaluated the HRTF corpus in a similar manner across trajectory. From the full list of participants, two clusters of three (participants 1, 4, & 9; and 6, 7, & 15) are consistently clustered together. It is also worth noting that participant 10 appears as a singleton in both planes. For the remaining participants, no pattern emerges between trajectories indicating that there exists similarities in the ratings of the HRTF corpus which vary as a function of trajectory. Comparison between participants' individual measured HRTF data could reveal possible common spectral characteristics in the same cluster, offering further insight into subject clustering results. Such analysis is the subject of ongoing work by the authors.

6. DISCUSSION

This paper presented a binaural evaluation study of sound stimuli created with 24 individualized HRTFs, including the participants' personally measured data. The task involved spatial quality assessments

of impulsive sound stimuli moving along predefined trajectories on the horizontal and median planes using a 9-point rating scale ranging from "poor" to "excellent".

Examination of responses revealed that some participants refrained from using the extreme scale values, rather concentrating their responses to the central region of the scale. In order to compare responses across participants, each individual block of ratings was normalized such that their local minima and maxima would extend to the global value limits of 0 and 1. The 9-point scale resolution was selected in an attempt to strike a balance between two previous studies, one using a 3-point [13] and one using a continuous scale [14] , with the aim of increasing the discrimination resolution while avoiding the excessive precision and associated experiment duration when using a continuous range. Upon observation of the collected ratings it was apparent that the provided resolution was not fully exploited by all participants. There existed cases in the evaluations of both trajectories where as little as $1/3$ of the interval range was utilized. The phenomenon was more frequent for median (33% of the participants) rather than for horizontal plane evaluations (13% of the participants). Even though the protocol did not dictate the use of all intervals for an evaluation, past studies have shown that the use of very few scale steps reduces the discrimination analysis ability between evaluated data points [13]. This issue could be resolved with alternative test protocol designs, and is the subject of future work.

According to the current protocol, participant's HRTF evaluation was based on personal internal references. These references can be assumed to be based on a combination of prior exposure to binaural listening and the textual description of the sound trajectories provided in the instructions, with the proportion of each contribution varying between participants.

Binaural evaluation studies typically differ from standard audio quality assessment mechanisms, such as the MUSHRA test, due to the inability to define global reference stimuli perceived identically across listeners, especially in the absence of individual HRTF stimuli. Such kind of designs have been used before in binaural testing but in different contexts, such as to examine HRTF rating repeatability [14]. Nevertheless, their use for quality assessments is not trivial due to the absence of reliable reference stimuli.

Previous quality evaluation studies have used global ratings, combined over different trajectories [13]. The current study investigated the rating patterns of participants for horizontal and median plane trajectories independently. Similarity between evaluations across the two tested trajectories for each participant were examined. The hypothesis that a correlation would exist between the two, such that a given HRTF would receive identical or very similar ratings for both trajectories, was not validated. In contrast, linear regression analysis of the responses revealed that the inter-trajectory evaluations were not linked in a intuitive manner for $2/3$ of the participants with some of them exhibiting negative correlations. This observation poses the question of the ability to select a single optimal non-individual HRTF for a given listener. With the discrepancies between ratings across trajectories, both azimuthal and elevational information are required in order to achieve complete assessments of HRTF data. How the different trajectory perceptions are weighted, if one would attempt to create a global quality assessment, in the auditory system remains to be determined. The influence of the test trajectory on the HRTF quality rating can no longer be ignored. In an application context, one could alternatively consider the need for task dependant trajectories, with the understanding that the selected optimal HRTF for one task may not be the same at that for an alternate task.

The inclusion in the corpus of HRTF sets of the same subjects originating from different databases allowed for a quantification of the perceptual impact of database-dependent characteristics in binaural data. In the case of individual HRTF ratings, owners had unknowingly attributed "satisfactory" scores to their own datasets (within the top 10% in the scale) across both trajectory planes, indicating that variations between these two databases can be considered perceptually irrelevant for individual HRTF measured data. In the case of non-individual measured data, the correlation between ratings was still apparent (greater than 1 std from the average of correlation distributions), but not equally strong. This fact implies that the common spatialization cue characteristics of the datasets prevailed over any potential database-depended characteristics in the collection for the majority of participants, but not in an equally manifesting manner as before. One possible explanation could be that database-dependent characteristics of an HRTFs can have a stronger impact in cases of non-individual HRTF assessments. Alternatively, it is possible that the HRTF corpus utilized in the study was too large for users to compare the characteristics of all the HRTFs to a high degree of detail. Perhaps a smaller data corpus, or an alternative test design of pairwise data comparisons, could have led to somewhat different evaluation results.

This work also discussed the construction and use of perceptually relevant spaces for data distributions based on subjectively induced similarities. In such spaces, data with similar evaluation patterns are in proximity to each other. One of the advantages of such spaces is that only data properties which affect assessment have an impact on the data distributions. For example, past research has demonstrated that HRTF spaces based on spectral content variations exhibit strong database dependencies, such that sets originating from the same database tend to cluster together [21]. However, such work did not consider the perceptual implications of these variations. On the contrary, the HRTF space utilized in this study highlighted the absence of database dependencies in the participant evaluations and distributed data only based on their evaluations. However, it should be noted that this observation has only been carried out for the two databases understudy, which exhibit rather small database-dependent variations [17].

The same principle of similarity was used to create a listener space where agreement in HRTF evaluations was the subjectively induced metric for participant similarity. Data clusters in this space indicate participants who evaluated the HRTF corpus in a similar manner. Consensus of the evaluation procedure could be regarded as consensus of binaural cue qualities, and hence a possible metric of individual HRTF similarity.

7. CONCLUSIONS

This work explored the use of a perceptual similarity metric for assessors and assessees. The discussion was based on the results of a binaural quality evaluation study assessing the quality of sound trajectories on the horizontal and median planes of 24 individualized HRTFs from two different databases. An overview of the responses as well as an evaluation of the metric qualities and the corresponding HRTF and participant spaces was provided.

Future work will involve exploration of alternate experiment designs that minimize the need for data normalization as well as methods for objectively approximating the perceptual similarity metric such that distributions could be created directly from HRTF data without the need for subjective evaluations. The stability of the experimental design also should be investigated through a repeatability study, assessing how much training is necessary for participants to provide reliable responses. Finally, it is of interest to further explore the potential perceptual impact of the database origin for a large set of databases.

8. REFERENCES

[1] F. Wightman and D. Kistler, "Multidimensional scaling analysis of head-related transfer functions," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Waisman Center, Wisconsin Univ., Madison, WI, October 1993, pp. 98–101. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=379987

[2] D. Schnstein and B. Katz, "HRTF selection for binaural synthesis from a database using morphological parameters." in *20th International Congress on Acoustics (ICA)*, Sydney, 23-27 August 2010, pp. 1–6. [Online]. Available: http://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ICA2010/papers/p266.pdf

[3] V. Lemaire, F. Clérot, S. Busson, R. Nicol, and V. Choqueuse, "Individualized HRTFs from Few Measurements: a Statistical Learning Approach," in *IEEE International Joint Conference on Neural Networks, 2005*, July, Ed., vol. 4. Montreal, Canada: IEEE, 2005, pp. 2041–2046. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1556214

[4] A. Andreopoulou, A. Roginska, and J. P. Bello, "Observing the Clustering Tendencies of Head-Related Transfer Function Databases," in *131st Audio Engineering Society Convention*, New York, NY, October 2011, pp. 1–10. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16065

[5] B. Xie and X. Zhong, "Similarity and cluster analysis on magnitudes of individual head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 3305–3305, 2012. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/131/4/10.1121/1.4708359

[6] T. Ajdler, L. Faller, C.and Sbaiz, and M. Vetterli, "Sound Field Analysis Along a Circle and its Applications to HRTF Interpolation," *Journal of the Audio Engineering Society*, vol. 56, no. 3, pp. 156–175, March 2008.

[7] X.-L. Zhong and B.-S. Xie, "Maximal azimuthal resolution needed in measurements of head-related transfer functions." *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2209–2220, Apr. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19354397

[8] P. M. Hofman, J. G. Van Riswick, and A. J. Van Opstal, "Relearning Sound Localization with New Ears," *Nature neuroscience*, vol. 1, no. 5, pp. 417–421, 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10196533

[9] G. Parseihian and B. Katz, "Rapid head-related transfer function adaptation using a virtual auditory environment," *The Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2948–2957, 2012.

[10] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization." *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–61, 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1564201

[11] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 340–343, 2006.

[12] A. Roginska, T. Santoro, and G. Wakefield, "Stimulus-dependent HRTF preference," in *129th Audio Engineering Society Convention*, San Francisco, CA, USA, November 2010, pp. 1–11. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15690

[13] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. EL99–EL105, 2012. [Online]. Available: http://link.aip.org/link/JASMAN/v131/i2/pEL99/s1\&Agg=doi

[14] D. Schönstein and B. Katz, "Variability in perceptual evaluation of HRTFs," *Journal of the Audio Engineering Society*, vol. 60, no. 22, pp. 783–793, 2012. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16552

[15] T. Carpentier, H. Bahu, M. Noisternig, and O. Warusfel, "Measurement of a head-related transfer function database with high spatial resolution," in *Forum Acousticum*. Krakow: European Acoustics Association, Sept. 2014, pp. 1–6. [Online]. Available: http://www.fa2014.agh.edu.pl/fa2014_cd/article/RS/R19_3.pdf

[16] O. Warusfel. (2003) Listen HRTF database. [Online]. Available: http://recherche.ircam.fr/equipes/salles/listen/

[17] A. Andreopoulou, D. Begault, and B. Katz, "Inter-laboratory round robin hrtf measurement comparison," *Selected Topics in Signal Processing, IEEE Journal of*, vol. PP, no. 99, pp. 1–12, February 2015.

[18] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners." *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–200, May 1990. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/2348023

[19] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3530–3540, 2014.

[20] M. J. M. Macé, F. Dramas, and C. Jouffrais, "Reaching to sound accuracy in the peri-personal space of blind and sighted humans," in *Computers Helping People with Special Needs*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7383, pp. 636–643. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-31534-3_93

[21] A. Andreopoulou and A. Roginska, "Towards the creation of a standardized HRTF repository," in *131st Audio Engineering Society Convention*, New York, NY, October 2011, pp. 1–6. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16096