Evaluating HRTF Similarity through Subjective Assessments: Factors that can Affect Judgment

Areti Andreopoulou
Audio Acoustics Group, LIMSI - CNRS
andreopoulou@limsi.fr

Agnieszka Roginska
Music and Audio Research Lab, NYU
roginska@nyu.edu

ABSTRACT

This work investigates the associations between objectively measured distance metrics and subjective assessments of similarity in HRTF data. For this purpose two different means of matching users to HRTF sets were compared: a simple system computing correlations between personally collected HRTF data and a repository of 111 measured binaural datasets, and an HRTF user-preference study assessing the spatial quality of a subset of this data based on certain attributes. The purpose of this comparison is twofold: first, to investigate the presence of an association between HRTF distance and perceived spatial quality, and second, to identify factors that can affect subjective judgment. The results primarily highlighted the importance of binaural reproduction exposure and training for the appreciation and understanding of a virtual auditory scene. In addition, they offered a means of assessing the effectiveness of the utilized evaluation criteria as a function of user expertise.

1. INTRODUCTION

The accuracy of measured or modeled Head-Related Transfer Functions (HRTFs) can be evaluated either objectively based on a defined metric, or perceptually through a user study. While in the first case a well fitted dataset is the one that demonstrates the smallest possible variation from an originally measured set, in the latter it is the one that conveys an accurate and convincing spatial image to the users. Both alternatives have been extensively used in binaural audio research.

For methods evaluated objectively the discussion of similarity between two binaural filters becomes one of distance. Several different metrics have been suggested and the selection depends not only on the task, but mainly on the feature space. The most commonly used choices include the Euclidean or squared-Euclidean distance [1–3], the correlation distance [4–6], and the Mean Square Error (MSE) [7–9].

Unarguably, objective evaluation processes can be quick, as they mainly depend on the size of the data and the computational power of the analysis system. However, they rely on the assumption that there exist absolutely accurate HRTFs that can be used as a comparison to the rest of the data. They also reward perfect reconstruction, often assuming uniformity in the perceptual weights of spectral variation across frequency. Nevertheless, the brain has a certain degree of tolerance in HRTF variations, as studies have shown that the human auditory system has the ability to successfully adapt to altered spectral cues, given time [10]. Hence, perceptual criteria also need to be employed for a more conclusive evaluation process.

Subjective HRTF evaluation studies take the form of binaural localization, or user-preference tasks. In localization studies, users are requested to identify the apparent location of a virtual sound-source, presented through headphones, based on auditory information [4, 11–13]. In user-preference ones, participants, who may or may not be experts in binaural reproduction, are asked to subjectively evaluate the quality of different HRTF sets. The evaluation process can be based on a wide variety of criteria, ranging from spatial realism attributes, like externalization perception [14, 15], to spatial accuracy assessments, like the precision in the trajectory of a sound stimulus [16]. In addition, assessments may take the form of discrete or continuous scale responses.

Evidently, localization studies and user-selection procedures are complementary tasks evaluating HRTF spatial quality from different perspectives. When the end-goal is an accurate spatial reconstruction of an auditory scene, where it is essential that the location of the target sound source best matches apparent location of the reference one, subjective localization tests are necessary. For cases, however, when the goal is a convincing spatial impression of a virtual sound-scape, user selection studies may help reach the intended outcome faster.

This work attempts to approach the concept of HRTF similarity from a perceptual point of view, through a user-evaluation study. Its purpose is twofold: first to investigate the presence of an association between HRTF distance and perceived spatial quality, and second to identify factors that might affect or bias one's subjective judgment. Therefore, similarity between a HRTFs was quantified through two simple HRTF database matching implementations; one based on objectively computed correlation distances between datasets and another based on a user-preference elimination task. Both designs are described in the following sections, followed by a presentation and discussion of the study results.

2. HRTF DATABASE MATCHING

2.1 Post Processing

The designed algorithm operated on a repository of 111 HRTF datasets from the LISTEN [17], CIPIC [18], and FIU [19] databases. The following post-processing steps were applied on the data. Binaural filter pairs were normalized to eliminate the potential effects of amplitude on the task, shortened to 1.5 ms to include only the pinnae responses, and band-limited between 0.5 kHz and 16 kHz. The specific frequency range was selected because it was previously identified as the one containing the most predominant localization cues [6, 20, 21]. Each HRTF set was reduced to an optimal subset of binaural filters, which minimize distance between datasets belonging to the same group, while maximizing inter-group discrimination. This optimization, which results to at least 67% data reduction, was based on Linear Discriminant Analysis (LDA), and was presented and discussed in length in a previous publication [22].

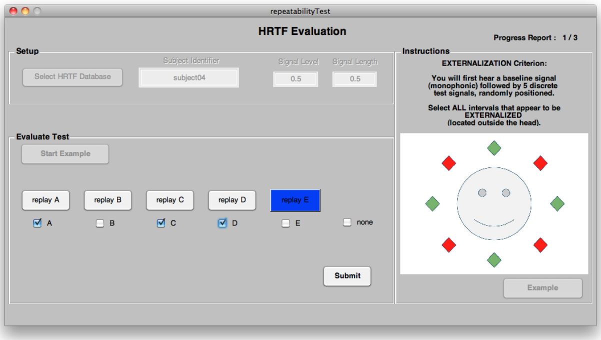
In brief, the LDA system was trained on the MARL database of repeated HRTF measurements [23], which consists of 40 datasets collected from four subjects, over the course of eight months. For the purposes of this analysis the data was divided into four labeled groups each containing HRTFs originating from the same subject, and was sent to a linear classifier. The classifier was trained based on a set of features (HRTF components), and their corresponding labels. Upon training, the algorithm returned a set of weights describing the extent to which each feature contributed to a successful classification. Data reduction was achieved by setting a perceptually evaluated threshold, and eliminating all components below it.

2.2 Databased Matching Implementation

The database matching algorithm was designed to compare sparse queries to an HRTF dictionary and return a ranked list of all available datasets, along with the corresponding percentage of similarity. The similarity estimation was based on aggregated correlation distances of the HRTFs' cepstrum. More specifically, a separate distance matrix was computed for each active location from the correlation distance between the decomposed DTFs. The overall similarity between datasets was calculated by averaging across the resulting matrices. Similar implementations for computing HRTF distance have been previously described in the literature [24].

2.3 Search Query

The personalized search queries for the matching algorithm were based on sparsely measured HRTF datasets. The recordings took place in the Spatial Audio Research Lab, a semi-anechoic space at NYU. Participants were sitting on an adjustable stool, and their alignment was monitored through a Polhemus Liberty electro-magnetic tracker. No support for their head, back and arms was provided. Five Genelec 8030a speakers were positioned in a spiral configuration at a distance of 1 m from the subjects' heads. The measurements were done with the blocked-meatus method, using custom-made miniature binaural microphones with Sennheiser KE - 4 capsules, in azimuth increments of 15°, at 5 elevations from $-30°$ to $30°$.



Progress Report : 1 / 3

Subject Identifier: subject04

Signal Level: 0.5

Signal Length: 0.5

☑ A □ B ☑ C ☑ D □ E □ none

EXTERNALIZATION Criterion:

You will first hear a baseline signal (monophonic) followed by 5 discrete test signals, randomly positioned.

Select ALL intervals that appear to be EXTERNALIZED (located outside the head).

Figure 1. Graphical Interface for collecting user responses in the HRTF preference task.

3. METHODS

3.1 Participant pool and Experiment Outline

Twenty people volunteered to take part in this study, all students of the NYU Music Technology graduate and undergraduate programs. Participants had reported having normal hearing. Volunteers were divided into two groups based on their level of expertise in binaural-audio reproduction. The first consisted of users who had some exposure to immersive audio concepts. Such experience ranged from a couple of relevant courses to several years of research in the field. The second consisted of people with no experience in the field, its concepts and terminology. The ratio of participants in each group was nine to eleven.

No training in binaural audio reproduction was offered to any of the users, except for the opportunity to familiarize themselves with the functionality of the interface. The reason behind this decision lies in the wide range of experience in the "informed" group. We acknowledge that participants whose familiarity with binaural audio reproduction was solely based on an academic course or the participation in a few binaural audio studies cannot really be considered a "experienced" users. Yet, their awareness can be closer to that of a trained subject. Hence, to fully explore the effect of binaural audio reproduction familiarity on user-preference decisions, no training was offered to participants in the "naive" group.

The duration of the study was approximately one hour and participants had the option of completing it during one, or two sessions. The first part consisted of a sparse HRTF measurement, and three personalized responses of the Sennheiser HD 650 open headphones, averaged to create a single binaural equalization pair. The second part included the HRTF preference/evaluation task.

3.2 HRTF Preference Task

3.2.1 Overview

The purpose of this task was not to evaluate the localization accuracy of different HRTF datasets, but rather to as-

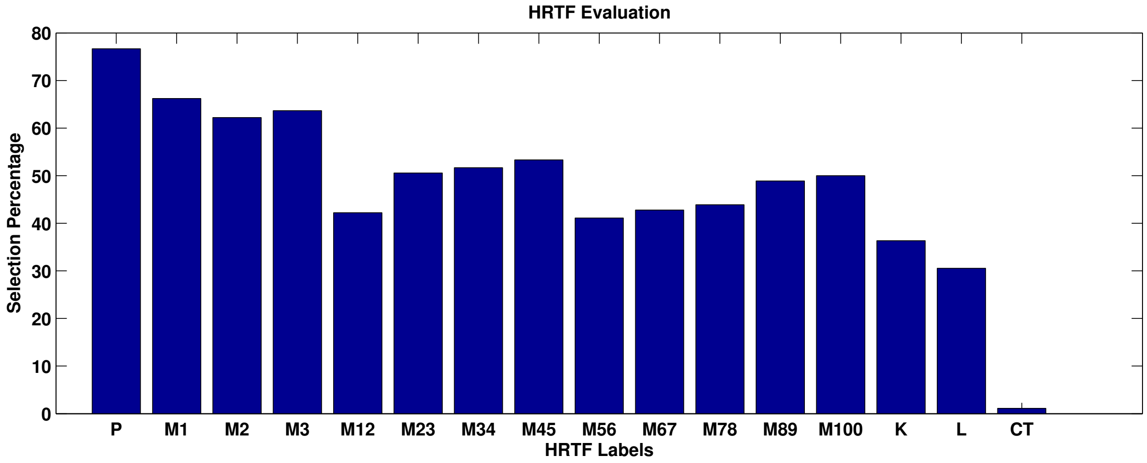


Figure 2. Aggregated user responses across all criteria and participants. P corresponds to the personally measured HRTF, Mi to the i^{th} HRTF in the returned ranked list, K to the KEMAR set, L to the least similar set, and CT to the catch trial.

sess their perceived spatial quality based on three criteria: externalization perception, front/back discrimination, and up/down discrimination. A collection of sixteen HRTF datasets was compiled for every participant, consisting of their personally measured dataset, the MIT - KEMAR set [25], a monophonic pseudo HRTF, used as a catch trial, and thirteen datasets selected across the ranked list of responses from the database matching implementation.

The following notation will be used across the rest of this paper to refer to the different HRTF classes used in the study. P will correspond to the personally measured HRTF, Mi to the i^{th} HRTF in the returned ranked list, K to the KEMAR set, L to the least similar set, and CT to the catch trial.

The CT was created from the first 128 samples of the 0° azimuth/elevation KEMAR binaural pair, with the filters cross-summed and repeated at various amplitude values. The stimuli were .5 sec pink noise bursts, presented to participants through the Sennheiser HD 650 open headphones. In order to minimize any bias in the responses potentially caused by ITD mismatches, all HRTFs were converted to minimum phase and the extracted ITD information were replaced by the individually measured ones. Headphone equalization was also applied to reduce the effect of the reproduction equipment on the evaluation procedure.

3.2.2 *Protocol*

The *HRTFPref* evaluation tool has been described extensively in several studies in the past [14, 22, 26]. In brief, the task consists of three stages, each having multiple trials. For every trial, participants are presented with a reference monophonic sound followed by a series of spatialized stimuli at various directions, and are instructed to select all HRTFs that meet the stage-specific criterion. Trials consist of a maximum of five intervals (HRTFs).

In order to eliminate variations in signal colorization, the reference sound is created by cross summing the left and right ear responses of the 0° azimuth & 0° elevation location of the current HRTF. HRTFs are presented multiple times in a given criterion, and only the ones selected more than 60% of the times advance to the next stage. Such a configuration results in an elimination task. The first stage of the study assesses the perceived spatial quality of a given HRTF based on externalization perception, the second on front/back discrimination, and the last on up/down discrimination. User responses were collected through a graphical interface designed in MATLAB 2010_b (Figure 1).

4. RESULTS

The first attempt to investigate the relationship between HRTF dissimilarity and perceived spatial quality is based on observations of the overall user-evaluations across the collection of HRTFs in the study. Figure 2 plots the aggregated user-preference across all criteria, and participants. On the plot HRTFs appear in a decreasing similarity order from left to right, with HRTFs closer to the personally measured set P (search query) appearing on the left on graph. The ranking of all datasets was controlled by the output of the designed HRTF database matching system.

The collected data indicates the presence of an association between HRTF rank and perceived spatial quality. As it can be seen, user responses follow a declining order between the top matches and the least similar HRTF classes, with the K and L sets receiving considerably lower scores than P and the top three matches M_1 - M_3. However, for HRTF classes between the two extremes (center of the graph) a lot more variation is observed, with HRTFs of lower ranks occasionally receiving better scores than higher ones. An example of such behavior is the increase in the scores between HRTF classes M_{78} and M_{89}.

Further observations arise when analyzing the user responses for each evaluation criterion separately. Figure 3 contains the aggregated user-preference responses per evaluation criterion, across all participants. By looking at the

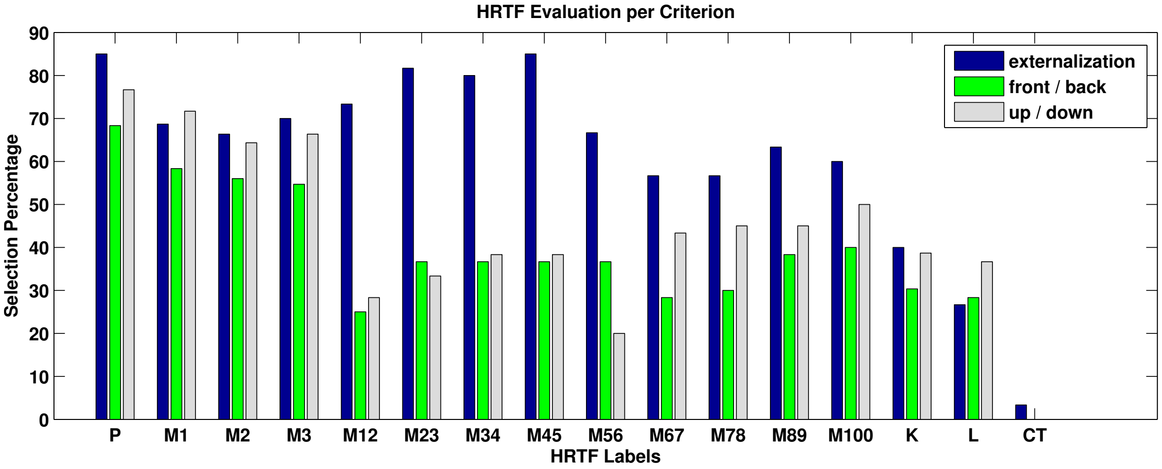


Figure 3. Aggregated user-preference responses per evaluation criterion, across all participants. P corresponds to the personally measured HRTF, Mi to the i^{th} HRTF in the returned ranked list, K to the KEMAR set, L to the least similar set, and CT to the catch trial.

figure, it appears that the externalization criterion, almost consistently, received the highest preference ratings. For some cases these ratings reached the same levels as the personally measured sets, or the top matches. This implies that participants of this study evaluated a wide variety of HRTFs as being equally convincing, in terms of externalization performance, to their measured sets. In addition, it is this criterion that seems to be driving the direct relationship between objectively measured HRTF distance and perceived spatial quality. As it can be seen on the graph, externalization evaluations demonstrate a stronger declining behavior between top matches and HRTFs further down in the ranked list.

On the contrary, the front / back and up / down discrimination evaluations seem to plateau at around 40% across all classes, except for the personally measured HRTFs and M_1 to M_3. This implies that spatially convincing movements of virtual sources in an up/down or front/back manner were consistently attributed to datasets very close to the measured HRTFs. This observation is in line with the binaural audio literature, demonstrating that, with a few exceptions, localization performance is optimal when users are listening through their own binaural filters.

In an attempt to interpret the cause of these observations user responses were divided in two groups according to the users' level of expertise: "experienced" and "naive". As discussed in 3.1, the experienced user group consisted of volunteers who had some exposure to immersive audio concepts, while the naive one of those with no experience in the field. As mentioned earlier no training was offered to the users, except for the opportunity to familiarize themselves with the experiment interface.

Figure 4 contains the aggregated user evaluations per criterion and familiarity group. The top graph holds the responses of the "informed", and the bottom of the "naive" user group. The most evident observation emerging from this data division, is the imbalance in the ratings between the two groups. It appears that experienced users consistently attributed higher ratings to every HRTF class across all criteria, fact which implies variations in the evaluation standards employed by each group. This imbalance is especially spotted in the front/back and up/down discrimination criteria. One possible explanation for that, could be the lack of visual cues, enhancing the presence of sound sources in the frontal hemisphere, Another factor could be the static character of this experiment, where subject head-movement did not affect the reproduced binaural scene, resulting in virtual sources moving along with one's head in every turn. Even though participants were encouraged to keep their eyes closed when listening to the stimuli, and to refrain from turning their heads, it is quite possible that these limitations made these two tasks more challenging to "naive" participants. For that user group this resulted in flat average ratings between 20% and 40% across all HRTF classes except for the personally measured sets.

On the contrary, the experienced participant group, exhibited more variation in the corresponding average selection rates, which appear to follow a declining trend as a function of distance from the measured set. In other words, HRTF classes with lower similarity ranks were evaluated positively less often. In general, for the data collected in this study, there appears to be some correlation between levels of expertise and perceived spatial quality. However, this observation was made on a very small participant pool and it is, therefore, subject to further investigation.

5. DISCUSSION

In binaural audio related research the two means of HRTF evaluation are localization and user preference tasks. The former is an objective method, where an effective HRTF set is the one that results to smaller or fewer localization errors, while the second is purely subjective and results to a set that satisfies the personal quality standards of a user. The need for so distinct methods of assessment arises from the realization that the level of accuracy needed in a virtual auditory space is task dependent. For example, in mission critical applications, where effortless and accurate

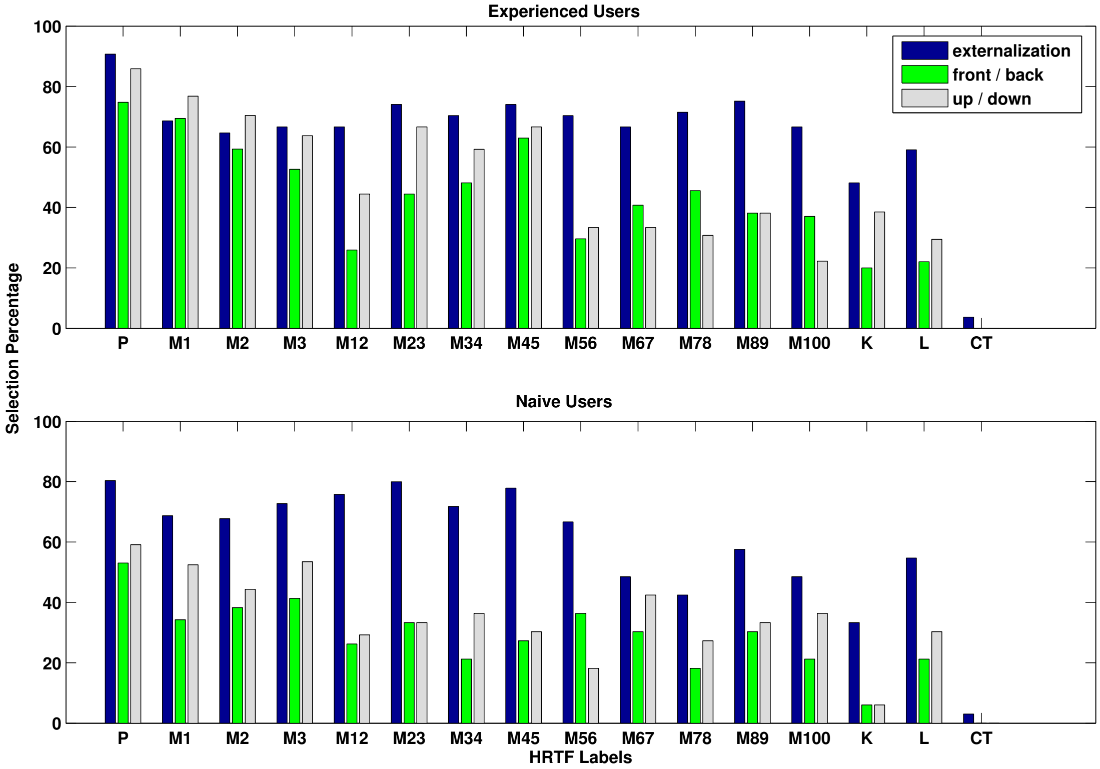


Figure 4. Aggregated user-preference responses per evaluation criterion and user familiarity. P corresponds to the personally measured HRTF, Mi to the i^{th} HRTF in the returned ranked list, K to the KEMAR set, L to the least similar set, and CT to the catch trial.

virtual reconstruction of one's auditory environment may prove vital, localization accuracy and adaptation time are the most meaningful means for HRTF evaluation. For applications in entertainment, however, an HRTF that meets the spatialization expectations of the user should be preferred for an optimal experience. Nonetheless, there hasn't been any formal proof that spatial accuracy can be an indication of enhanced perceived quality and vice versa, or a systematic approach to the appropriate criteria for subjective HRTF assessments. This paper investigated factors that may affect subjective judgment as a function of the utilized criteria and level of expertise.

The following main points arose from the analysis of the user responses. First, the externalization criterion does not provide sufficient information on the quality of binaural filters. Results indicated that especially "naive" participants tended to find the vast majority of HRTFs convincing with respect to this task, regardless of the level of decorrelation from their personally measured sets. Nevertheless, this was the only criterion in this study, whose levels appeared to have a direct relationship to HRTF dissimilarity measures. In other words, HRTFs more correlated to the personally measured sets received higher externalization ratings than the more dissimilar ones. This behavior was common across users regardless of their levels of expertise.

On the contrary, the up/down and front/back discrimination tasks offer a better understanding of the correlation between HRTF sets. As demonstrated earlier, HRTFs who have received a lower ranking by the database matching algorithm were also attributed lower scores in the preference task. However, this tendency seems to be stronger between "informed" users. Results depicted in Figure 4 showed that, unlike the experienced user group, the responses of the naive participants ranged from around 20% to 40% across all HRTF classes, except for the personally measured sets. This behavior suggests that people in this group were unable to perceive convincing front/back or up/down movement with any HRTF set but their own.

Such a finding highlights the importance of training and binaural audio reproduction exposure, when trying to understand the notion of moving sources, and, especially, when making general assessments about an HRTF's spatial quality. This observation is also supported by the difference in overall ratings across all HRTF classes between the two participant groups. Experienced user responses covered a wider range of ratings compared to the naive group ones, which, with the exception of the externalization criterion were compressed to a level around 30%.

Hence, spatial quality appreciation seems to be directly related to one's duration of exposure to binaural audio reproduction. This can be attributed to a number of factors: It is possible that the expectations of the "naive" users were less often fulfilled. Alternatively, users who had experience listening to, or working with binaural audio reproduction were accustomed to the sound-quality nuances and limitations, and their expectations were violated less often. It is also quite possible that this difference was a function of understanding rather than interpreting the concepts of the three criteria used for evaluation. Or, that the unappealing character of the pink-noise stimuli, even though common practice for binaural studies, was not conductive to an immersive experience for the "naive" participant group. This are all points that will be considered in future studies.

6. CONCLUSION AND FUTURE WORK

The results of this study highlighted the importance of binaural - audio nuances awareness, when assessing the spatial quality of presented media. By separating user responses according to their levels of expertise distinct ranking patterns arose for different HRTF classes, which imply that spatial quality appreciation may be directly related to binaural-audio reproduction exposure. Three criteria were evaluated in terms of their effectiveness in leading to the most appropriate HRTF dataset during a user-selection study. Externalization perception was found to be less effective in discriminating between data, but it was the only criterion whose ratings appeared to be related to objectively computed HRTF dissimilarity measures. The front/back and up/down discrimination tasks were found to be more effective in selecting spatially convincing HRTF datasets among "trained" but not "naive" users.

Future work includes the design of new evaluation studies, based on different criteria, and also the increase in the number of participants in the evaluation tasks. It is also of interest to further divide the group of experienced users into more refined subsets, and explore how different levels of expertise affect people's judgments.

7. ACKNOWLEDGEMENTS

This work was funded in part by the French project BiLi (Binaural Listening, www.bili-project.org, FUI - AAP14). Additional financial support was provided by New York University. The authors would like to thank the Music technology program at NYU for providing access to their equipment and facilities, as well as to all the volunteers who participated in out studies.

8. REFERENCES

[1] 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. IJCNN'05. Proceedings. 2005*, July, Ed., vol. 4. Montreal, Canada: IEEE, 2005, pp. 2041–2046.

[2] M. Queiroz, "Efficient Binaural Rendering of Moving Sound Sources Using HRTF Interpolation," *Journal of New Music Research*, pp. 37–41, 2011.

[3] 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.

[4] P. Bremen, M. M. van Wanrooij, and a. J. van Opstal, "Pinna cues determine orienting response modes to synchronous sounds in elevation." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, no. 1, pp. 194–204, Jan. 2010.

[5] F. Keyrouz, "Humanoid hearing: A novel three-dimensional approach," *Robotic and Sensors Environments (ROSE), 2011*, pp. 214–219, 2011.

[6] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *The Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 1583–1596, 2002.

[7] 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, 2008.

[8] W. Wahab Hugeng and D. Gunawan, "Improved Method for Individualization of Head-Related Transfer Functions on Horizontal Plane Using Reduced Number of Anthropometric Measurements," *Journal of Telecommunications*, vol. 2, no. 2, pp. 31–41, 2010.

[9] J. Leung and C. Carlile, "PCA Compression of HRTFs and Localization Performance," in *International Workshop on the Principles and Applications of Spatial Hearing*, Miyagi, Japan, 2009, pp. 5–7.

[10] 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, Sep. 1998.

[11] P. M. Hofman and A. J. Van Opstal, "Spectro-temporal factors in two-dimensional human sound localization." *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2634–2648, May 1998.

[12] M. Hofman and J. Van Opstal, "Binaural weighting of pinna cues in human sound localization." *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale*, vol. 148, no. 4, pp. 458–70, Feb. 2003.

[13] J. Jeppesen and H. Moeller, "Cues for Localization in the Horizontal Plane," in *118th Audio Engineering Society Convention*, Barcelona, Spain, 5 2005.

[14] A. Roginska, T. Santoro, and G. Wakefield, "Stimulus-dependent HRTF preference," in *129th Audio Engineering Society Convention*, San Francisco, CA, USA, 2010.

[15] B. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," in *Proceedings of the 2003 International Conference on Auditory Display*. Boston, MA, USA, 2003, pp. 1–4.

[16] 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.

[17] O. Warusfel, "http://recherche.ircam.fr/equipes/salles/listen/," 2003.

[18] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk Mountain House, New Paltz, NY, October 2001, pp. 99–102.

[19] N. Gupta, A. Barreto, M. Joshi, and J. Agudelo, "HRTF database at FIU DSP lab," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. Dallas, TX: IEEE, March 2010, pp. 169–172.

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

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

[22] A. Andreopoulou, A. Roginska, and J. P. Bello, "Reduced representations of hrtf datasets: A discriminant analysis approach," in *135th Audio Engineering Society Convention*, Oct 2013.

[23] A. Andreopoulou, A. Roginska, and H. Mohanraj, "A database of repeated head-related transfer function measurements," in *International Conference on Auditory Display (ICAD) 2013*, Lodz University of Technology, Poland, July 2013.

[24] B. Xie, C. Zhang, and X. Zhong, "A cluster and subjective selection-based hrtf customization scheme for improving binaural reproduction of 5.1 channel surround sound," in *134 Audio Engineering Society Convention*, May 2013.

[25] B. Gardner and K. D. Martin, "HRTF Measurements of a KEMAR," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[26] A. Roginska, G. Wakefield, and T. Santoro, "User Selected HRTFs: Reduced Complexity and Improved Perception," in *Undersea Human System Integration Symposium*, Providence, RI, 2010, pp. 1–14.