Perceptually based head-related transfer function database optimization

Brian F.G. Katz, Gaetan Parseihian

▶ **To cite this version:**

Brian F.G. Katz, Gaetan Parseihian. Perceptually based head-related transfer function database optimization. Journal of the Acoustical Society of America, 2012, 131 (2), pp.EL99 - EL105. 10.1121/1.3672641 . hal-01780508

HAL Id: hal-01780508

https://hal.sorbonne-universite.fr/hal-01780508

Submitted on 4 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The Journal of the Acoustical Society of America

JANUARY 13 2012

Perceptually based head-related transfer function database optimization

Brian F. G. Katz; Gaëtan Parseihian

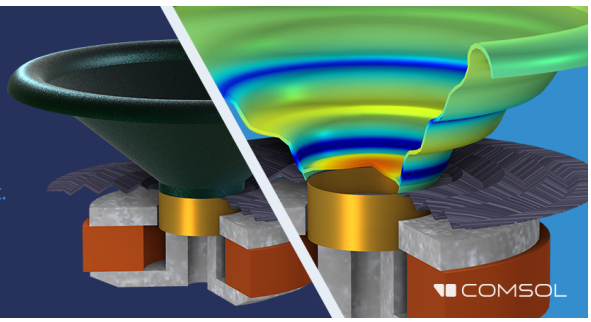
Check for updates

J. Acoust. Soc. Am. 131, EL99–EL105 (2012)
https://doi.org/10.1121/1.3672641

View Online
Export Citation
CrossMark

04 December 2023 21:26:17

ASA

Take the Lead in Acoustics
The ability to account for coupled physics phenomena lets you predict, optimize, and virtually test a design under real-world conditions – even before a first prototype is built.
» Learn more about COMSOL Multiphysics®


Perceptually based head-related transfer function database optimization

Brian F. G. Katz[a] **and Gaëtan Parseihian**
LIMSI-CNRS, BP 133, Universit Paris Sud, Orsay, France
Brian.Katz@limsi.fr, Gaetan.Parseihian@limsi.fr

Abstract: In the context of binaural audio rendering, choosing the best head-related transfer function (HRTF) for an individual from large databases poses several problems. This study proposes a method to reduce the size of a given HRTF database. Participants, 45 in total, were asked to rate the quality of binaural synthesis for 46 HRTFs. The lack of reciprocity in the ratings was noted. Results were used to create a perceptually optimized HRTF subset which satisfied all participants' judgments. The subset was validated using localization tests on a separate group of subjects with results showing reduced errors when subjects were given their best choice, rather than their worst choice HRTF.

© 2012 Acoustical Society of America
PACS numbers: 43.66.Pn, 43.66.Qp [JL]
Date Received: September 30, 2011 **Date Accepted:** December 1, 2011

1. Introduction

It is often the case in applications involving binaural synthesis that individual head-related transfer functions (HRTFs) are not available, and it is necessary to use non-individual HRTFs from existing databases. A number of publicly available HRTF databases exist (e.g., CIPIC,[1] LISTEN,[2] and Tohoku[3]), equating to hundreds of possible HRTFs for the potential user. The use of such databases for HRTF selection typically requires test subjects to evaluate massive numbers of HRTFs, a situation which is impractical. To address this issue, a method has been developed to reduce the database to a manageable number to allow for quick evaluations of just a few HRTFs by novice users. Previous studies have proposed methods for rapid HRTF selection using paired comparisons[4–6] but the process is still time consuming for large databases. In order to minimize subject selection time, large databases must be sorted or reduced. While signal comparisons of HRTF pairs can be performed, such methods have not been shown to always correspond to perceptual differences.[7] As such, this study presents a method for reducing a given database and for psychophysical selection of the best HRTF from the database. Separation of temporal and spectral cues in the HRTF was performed in order to focus attention on the spectral HRTF component, and not on inter-aural time differences (ITD).

2. HRTF database

The HRTFs for this study were taken from the publicly available LISTEN HRTF database.[2] This database contains anechoically measured blocked meatus HRTFs with 187 positions, elevations from $-45°$ to $+90°$ with positions at roughly $15°$ azimuthal spacings, and at a sample rate of 44 100 Hz. The "raw" HRTF measurement data (8192 length HRIR), not the diffuse field compensated, was used in order to minimize potential post-processing effects. As this study is primarily concerned with the spectral cues in the HRTF, and not inter-aural temporal differences, each HRTF was decomposed into its minimum-phase components, in order to represent the spectral cues, and excess phase components to represent the ITD, modeled as a pure delay.[8] Participants

[a]Author to whom correspondence should be addressed.

04 December 2023 21:26:17

in the study were chosen from those contributing to the HRTF database. Forty-six individuals who contributed to the database agreed to take part in this study (one individual later declined once the study commenced). During HRTF evaluation, the different minimum-phase HRTF components were combined with the individually calculated ITDs of each subject. Position dependent ITDs were estimated using the maximum of the inter-aural cross correlation (IACC).[8] This hybrid approach attempts, as a general approximation, to change the ears of each subject while keeping the individual's own head geometry constant.

3. Perceptual evaluation

A simple individualized virtual auditory scene was generated for each participant in the study for each of the 46 HRTFs. A repeated noise burst, 0.23 sec in duration and shaped with a Hann function window, was presented sequentially at fixed positions along two defined trajectories. The first trajectory was a circle in the horizontal plane with points at 30° spacings. The trajectory commenced directly left and followed two complete rotations around the subject. The second trajectory followed an arc in the median plane (azimuth = 0°) from elevation −45° in front to −45° at the rear with points at 15° spacings. The trajectory commenced in front, proceeded to the rear, and then returned along the same path to the front.

A grid of 46 selections was presented to each subject corresponding to each of the 46 HRTFs. A forced-choice three-point rating scale was used with the English labels: *bad*/*ok*/*excellent*. Subjects were presented with a textual description of the source trajectories and told which of the 46 HRTFs was their own HRTF. Subjects were allowed to listen to each sample repeatedly, and in any order, and were then asked to judge how well each rendering corresponded to the described trajectory. The duration of the test was approximately 35 min.

The judgment results for all subjects are presented in Fig. 1(a). The diagonal represents identity, and it is clear that the great majority of subjects rated their own individual HRTF as *excellent*. There are, however, three cases where this is not true.

The use of a perceptual test with subjects who contributed to the HRTF database offers the possibility of examining the symmetry of the selection process. In other words, one is able to inspect not only which HRTFs a subject is choosing, but also how that subject's HRTF has been rated by the respective owners of the rated HRTFs. Towards this purpose, the transpose of the selection results are shown, indicating when a subject was chosen as *excellent* by the corresponding HRTF's owner. There are 29 instances of *excellent-excellent* cross-referencing. In contrast, it is interesting to note that a total of 71 *excellent-bad* pairs are found (noted by the symbol ⊗ in the corresponding figure), indicating that HRTF preference is clearly not a symmetrical property.

4. Database reduction

This study proposes that perceptual performance fitness criteria be used for the database reduction method. For any generated subset, all participants should have at least one of their *excellent* rated HRTFs present. One approach for HRTF selection under consideration would be to choose the highest rated elements. Analysis of the *excellent* selections indicates that certain HRTFs were selected more often than others. The "top 20%" selected HRTFs comprise nine HRTFs: *AH*, *AR*, *AV, AZ*, *BE*, *BF*, *BL*, *BN*, *BQ*. Using this selection criteria, with a subset of these top nine HRTFs, 89% of the subjects would have at least one HRTF which they ranked as *excellent*.

However, in order to obtain 100% satisfaction for all participants (at least one *excellent*), the "top *X*%" subset selection method would require the "top 50%" of the original database, which is not a sufficient reduction in the database size. This subset generation method was therefore eliminated.

It can be postulated that if there are HRTFs in the subset which are very similar, then there exists a level of redundancy in the database. Therefore, it is of interest

04 December 2023 21:26:17

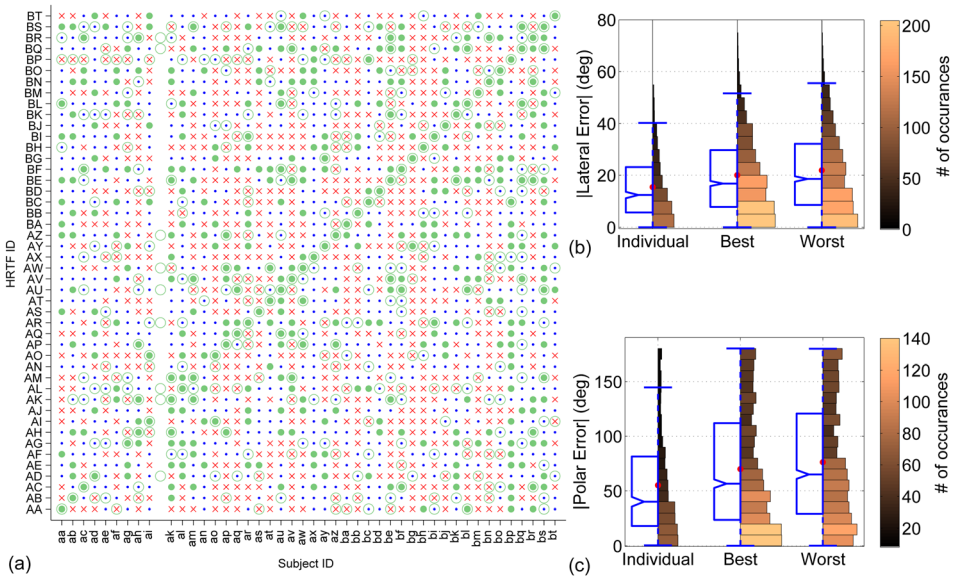


Fig. 1. (Color online) (a) Subjective HRTF rating test results: (×) bad; (·) ok; (•) excellent. (O) = trans-pose of excellent ($\bullet^T$). Lowercase identifiers represent subjects while uppercase represents the corresponding subject's HRTF. Note: subject *aj* did not participate but HRTF *AJ* was included in the test. Split boxplot-histogram of the magnitude of (b) lateral angular error and (c) polar angular error, by group. Boxplot and angular error scale at the left; histogram value legend color bar on the right. Mean values are represented by •.

to find the smallest subset possible that satisfies the maximum number of subjects, resulting in a perceptually pseudo-orthogonal HRTF database. To this end, an iterative method was employed to reduce the size of the subset, maintaining subject satisfaction as the fitness criteria. In contrast to the above mentioned method of subset selection by highest rating, this method aims to find the minimum number of HRTFs needed to satisfy the greatest number of listeners. The fitness criteria tolerance of 100% was maintained. The starting subset was based on the previously discussed "top *X*%" criteria, retaining the top 50% required for all subjects to be covered by at least one *excellent*. This removed HRTFs which were rarely well rated, and could be considered as test outliers or as "poor." From this subset, each HRTF was tentatively removed at random and the subset fitness score, the number of satisfied subjects, was calculated. The HRTF whose removal was associated with the highest score (meaning whose elimination had the least impact on the satisfaction level) was removed. As the perceptual database was based on a three-point rating, it often occurred that multiple HRTFs had the same maximum fitness score. In this event, the removed HRTF was selected at random from the potential candidates. This process was repeated until the fitness criteria tolerance, 100% satisfaction, became invalid and the resulting subset was registered.

No single solution exists *a priori* for this problem using this method. To obtain the smallest, or "optimal" solution, 50 000 complete trials of the algorithm were performed, each resulting in a subset. The mean subset size was 17.4, while the largest generated subset contained 22 elements, while one single result set was produced with the minimum size of only seven HRTFs: *AG*, *AJ*, *AR*, *AY, AZ*, *BN*, *BS*. This smallest subset was retained as the "optimized" subset.

It is noted that the three most often selected HRTFs (*BE*, *BF*, *BQ*) are not present in the final subset. Analysis of the gender distribution of the HRTF subset shows that 57% of those in the final subset were male, compared to a 65% male population in the database, indicating no specific advantage to using gender as a preselection mechanism.

04 December 2023 21:26:17

5. Database subset validation

The retained optimized database subset was evaluated using a new perceptual listening test. A total of 20 adult subjects (15 men, aged 20–60 years) served as paid volunteers, none had any known hearing deficit, and only few were familiar with virtual auditory displays (VAD). As described previously in Sec 3, HRTFs were decomposed into spectral components (representing spectral cues) and pure delay (ITD cues). As individual HRTFs were not available for ITD extraction, individual ITDs were synthesized for each participant. A PCA analysis of the estimated ITDs from the original database using the method of maximum inter-aural cross-correlation (Max-IACC) was performed and a linear regression model was generated for the principal competent (PC) weights using measured head morphological parameters. Head-circumference was selected as it provided the most stable regression and the least error with regards to measurement repetition variation. Applying the PC weights from the regression model to the principal components produces an estimated ITD based on the individual's head circumference.

5.1 HRTF subset selection

The new group of subjects performed an HRTF rating task, similar to that used in the previously described test, using only the optimized subset of seven HRTFs. In this instance, the two trajectories used in the above study were presented separately (horizontal then vertical) Rather than using the previously described three-point scale, each trajectory was judged using a continuous slider scale, from "bad" to "good" for each HRTF in the two planes. An overall judgment rating was taken as the sum of the two trajectory judgments. For each subject, their *Best* result, highest rank score, and *Worst*, lowest rank score, were tabulated.

The distribution of rankings (*best*/*worst*) for the seven HRTFs was HRTF: *AG*(15%/10%), *AJ*(5%/5%), *AR*(10%/5%), *AY*(5%/50%), *AZ*(40%/5%), *BN*(15%/5%), *BS*(10%/20%). While it is clear that *AZ* was selected significantly more often as *best*, and *AY* as *worst*, those same two HRTFs were still selected in the opposite category by some participants. These results clearly indicate that HRTF matching is individual, reinforcing the proposition that the use of a single generic HRTF to satisfy all users is not feasible.

In order to compare the quality of the HRTF subset for individual selection, participants were randomly divided into two groups with one group using their *best* ranked HRTF, and the other using their *worst*. HRTF performance was evaluated using a classical localization test and was compared with the performance of a control group consisting of four participants (one subject was common to the previous study) using their own individual HRTFs.

5.2 Localization task

Binaural sound sources were rendered using the LIMSI Spatialization Engine,[9] a real-time spatialization engine in Max/MSP based on full-phase HRIR convolution. Subjects, placed on a swivel stool in a quiet room, were equipped with a 6-DoF position/orientation tracked stereo headphone (model Sennheiser HD 570), a hand-located position-tracked ball, and a foot pedal. Subjects reported the perceived position of a static spatialized sound using a hand pointing technique validated by the foot pedal. This judgment elicitation technique has the ecological advantage of being egocentric and natural for the user and is known to be effective.[10] Each subject was instructed to orient themselves straight ahead and to keep their head fixed during the short sound stimuli presentation.

The stimulus, short to exclude head movement effects, consisted of a train of three 40 ms Gaussian noise bursts, 50 Hz–20 kHz with 2 ms hamming ramps at onset and offset and 30 ms of silence between each burst. This stimulus was chosen based on a preliminary study on the effect of repetition and duration for localization accuracy,

04 December 2023 21:26:17

showing improvement between 3×40 ms bursts vs 1×200 ms burst.[11] The overall level of the burst train was approximately 55 dBA measured at the ears. A total of 23 target positions (7 median plane, 6 horizontal plane), were randomly presented with five repetitions each. Subjects were naive with respect to the set of spatial positions. After presentation of the stimulus, subjects were instructed to point their hand (ball in hand) in the direction of the perceived sound source location and to validate the response with the foot pedal. The hand used to hold the ball could be changed at will during the experiment. The perceived registered position was calculated relative to the head position/orientation when the stimulus was initially played. No feedback was provided. Mean duration of this task was ten minutes.

5.3 Results

Localization error was evaluated using the interaural polar coordinate system,[12] transforming azimuth and elevation angles to lateral and polar angles. This roughly equates to a separation of temporal cues, related to ITD and represented by the lateral angle, from spectral cues, related to the HRTF and represented by the polar angle. As such, all front/back and up/down confusion errors are contained in the polar angle.

Localization errors in lateral and polar angles were analyzed with respect to the magnitude of the difference between the target angle and the perceived angle. Angular errors by group are shown in Fig. 1 using a combination *box-plot* or whisker-plot, presenting a statistical analysis of the data distribution, and *histogram,* allowing for an overview of the data distribution.

Inspection of the histograms highlights the error distributions. Lateral angle errors, shown in Fig. 1(b), followed a normal distribution with the mean values for *individual, best,* and *worst* HRTFs being 15.4°, 19.8°, and 21.8°, respectively. As the current study is concerned with the effect of the HRTF selection, not the ITD model, the resulting analysis shall focus on the polar angle responses.

Due to confusion errors, polar angle results do not follow a unimodal distribution, which implies that mean error values are not pertinent. Instead, a Kruskal-Wallis test was performed on the mean error over repetitions for each target position. Polar error distribution results were significantly better for the control group, where a comparison with the two groups using non-individual HRTFs showed a significant difference ($\chi^2(2,\ 21) = 22.68,\ p < 0.001$). Results of the two non-individual HRTF groups differed at a smaller significance level ($\chi^2(1,\ 22) = 5.99,\ p < 0.05$).

Previous studies have quantified localization errors by analyzing the proportion of front/back confusions.[13] A similar approach was employed here. The method used for classification is based on Ref. 13, where the error for each response was calculated relative to its actual position, and its position reflected across the plane of symmetry. If the error is reduced after reflection, it is considered a confusion error. This method can overestimate the number of confusions for positions near the symmetry plane. For this reason, a confusion detection exclusion zone around the symmetry plane is often used. A zone of 15° around each plane was defined, as per Ref. 14. Responses that improved through reflection across the coronal or frontal plane were classified as *front/back* confusions, the horizontal plane as *up/down,* and those that improved after both reflections were classified as *combined.* Responses that did not improve through reflections were termed *precision* errors.

A summary of the results of this analysis are shown in Table 1 for the different groups. The control group has a majority of errors considered *precision* (the goal), better than the other groups. The *best* HRTF group had better results, with less of each of the three types of confusion error, than the *worst* HRTF group.

A linear regression analysis was performed on the polar angle responses, correcting for confusion errors by applying the symmetry plane reflections. The mean and variance across subjects of the slope of the regression line and goodness-of-fit criteria r^2 for each group are shown in Table 1. While only small differences are observable, the slope of the *worst* group is farther from unity that the other two groups. Values for

04 December 2023 21:26:17

Table 1. Percentage of error according to type and mean linear regression analysis. Variances shown in parentheses.

	Precision	Front/back	Up/down	Combined	Regression slope	r^2
Individual	63 (14)	20 (6)	13 (7)	4 (3)	1.04 (0.021)	0.86 (0.003)
Best	46 (11)	32 (6)	15 (6)	6 (3)	0.93 (0.019)	0.83 (0.003)
Worst	38 (8)	35 (6)	19 (4)	8 (3)	0.91 (0.023)	0.83 (0.003)

r^2 are comparable for all groups. It is possible to conclude that the major effect between HRTF selection is the difference in the number of confusion errors.

6. Discussion

The purpose of this study was to present a method for the efficient selection of an HRTF from a large database. A database reduction method based on perceptual evaluation was established permitting the reduction of a public database from 46 to 7 perceptually different HRTFs while maintaining 100% satisfaction through a subjective listening test of 45 subjects. The lack of reciprocity in selection was noted and should be taken into account in future HRTF similarity comparisons. The subset and selection methods were validated by comparing localization test results between two groups of subjects with non-individual HRTFs; one with their highest subjectively ranked HRTF and one with their lowest. Results were referenced to tests performed on a control group using their own individual HRTF.

Errors between *worst* and *best* were notable through the presence of more *front/back* and *up/down* confusion errors for the *worst* HRTF group, which shows the efficiency of the database subset. Despite this difference, it is still apparent that users with individual HRTFs outperformed those with *best* selected HRTFs. In general, there was a large difference between subjects with individual HRTFs and subjects with non-individual HRTFs. Results of subjects using individual HRTFs are slightly poorer than those reported in previous studies with 12% *front/back* confusions[13] vs 20% in the current study, though the majority of subjects in the current study were novice users.

Previous studies using non-individualized HRTFs[15] have shown a 31% *front/back* confusion rate (29% of which are combined confusion errors) and 18% *up/down* confusion rate (55% of which are combined confusion errors). Separating the combined confusion errors, these results can be equated to an error type distribution of 22% *front/back* confusion and 8% of *up/down* confusion, comparable to the results of this study, using novice subjects, obtained by the *best* group: 32% *front/back* and 12% *up/down*. The exact protocol and method used to attribute error types differed from the method used in the current study making precise comparisons difficult.

Future studies are underway concerning localization improvement through repeated training sessions for subjects in the current study.

Acknowledgments

This work was supported in part by the French National Research Agency (ANR) through the TecSan program (Grant No. NAVIG ANR-08TECS-011) and the Midi-Pyrénées region through the APRRTT program.

References and links

[1]V. R. Algazi, R. O. Duda, D. P. Thompson, and C. Avendano, "The CIPIC HRTF database," in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY (October 21–24, 2001), pp. 99–102.

[2]IRCAM LISTEN HRTF database. http://recherche.ircam.fr/equipes/salles/listen/ (Last viewed December 17, 2011).

04 December 2023 21:26:17

[3]Tohoku HRTF database available at http://www.ais.riec.tohoku.ac.jp/lab/db-hrtf/ (Last viewed December 17, 2011).
[4]Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," Acoust. Sci. Tech. **27**(6), 340–343, (2006).
[5]H. Jo, W. Martens, and Y. Park, "Evaluating candidate sets of head-related transfer functions for control of virtual source elevation," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, Tokyo, Japan (October 8–10, 2010), 12 pp.
[6]J. C. Middlebrooks, E. A. Macpherson, and Z. A. Onsan, "Psychophysical customization of directional transfer functions for virtual sound localization," J. Acoust. Soc. Am. **108**(6), 3088–3091 (2000).
[7]F. Wightman and D. Kistler, "Multidimensional scaling analysis of head-related transfer functions," in *IEEE Digital Audio Workshop*, New Paltz, NY (October 17–20, 1993), pp. 98–101.
[8]P. Minnaar, J. Plogsties, S. Krarup Olesen, F. Christensen, and H. Mller, "The interaural time difference in binaural synthesis," in *Audio Engineering Society Convention*, Paris, France (February 19–22, 2000), Preprint 5133, 20 pp.
[9]LIMSI Spatialisation Engine, InterDeposit Digital No. IDDN.FR. 001.340014.000.S.P. 2010.000.31235.
[10]J.-M. Pernaux, "Spatialisation du son par les techniques binaurales: application aux services de telecommunications (Sound spatialization using binaural sound synthesis: Application to telecommuncation services)," Ph.D. dissertation, INGP Grenoble, 2003.
[11]F. Dramas, B. F. G. Katz, and C. Jouffrais, "Auditory-guided reaching movements in the peripersonal frontal space," Acoustics, **123**, 3723 (2008).
[12]J. C. Middlebrooks, "Narrow-band sound localization related to external ear acoustics," J. Acoust. Soc. Am. **92**(5), 2607–2624 (1992).
[13]F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. II: Psy-chophysical validation," J. Acoust. Soc. Am. **85**(2), 868–878 (1989).
[14]R. L. Martin, K. I. McAnally, and M. A. Senova. Free-field equivalent localization of virtual audio. J. Audio Eng. Soc. **49**(1/2), 14–22 (2001).
[15]E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using non-individualized headrelated transfer functions," J. Acoust. Soc. Am. **94**(1), 111–123 (1993).

04 December 2023 21:26:17