Head-related transfer function recommendation based on perceptual similarities and anthropometric features

Robert Pelzer, Manoj Dinakaran, Fabian Brinkmann, et al.

Citation: The Journal of the Acoustical Society of America 148, 3809 (2020); doi: 10.1121/10.0002884
View online: https://doi.org/10.1121/10.0002884
View Table of Contents: https://asa.scitation.org/toc/jas/148/6
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Sensitivity analysis of pinna morphology on head-related transfer functions simulated via a parametric pinna model
The Journal of the Acoustical Society of America 149, 2559 (2021); https://doi.org/10.1121/10.0004128

Perceptual implications of different Ambisonics-based methods for binaural reverberation
The Journal of the Acoustical Society of America 149, 895 (2021); https://doi.org/10.1121/10.0003437

A high-resolution head-related transfer function and three-dimensional ear model database
Proceedings of Meetings on Acoustics 29, 050002 (2016); https://doi.org/10.1121/2.0000467

Machine learning in acoustics: Theory and applications
The Journal of the Acoustical Society of America 146, 3590 (2019); https://doi.org/10.1121/1.5133944

Impact of face masks on voice radiation
The Journal of the Acoustical Society of America 148, 3663 (2020); https://doi.org/10.1121/10.0002853

Auditory perception stability evaluation comparing binaural and loudspeaker Ambisonic presentations of dynamic virtual concert auralizations
The Journal of the Acoustical Society of America 149, 246 (2021); https://doi.org/10.1121/10.0002942
Head-related transfer function recommendation based on perceptual similarities and anthropometric features

Robert Pelzer,1 Manoj Dinakaran,1,a) Fabian Brinkmann,1,b) Steffen Lepa,1 Peter Grosche,2 and Stefan Weinzierl1

1Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, D-10587, Germany
2Huawei Technologies, Munich Research Centre, Riesstrasse 25, D-80992 Munich, Germany

ABSTRACT:
Individualization of head-related transfer functions (HRTFs) can improve the quality of binaural applications with respect to the localization accuracy, coloration, and other aspects. Using anthropometric features (AFs) of the head, neck, and pinna for individualization is a promising approach to avoid elaborate acoustic measurements or numerical simulations. Previous studies on HRTF individualization analyzed the link between AFs and technical HRTF features. However, the perceptual relevance of specific errors might not always be clear. Hence, the effects of AFs on perceived perceptual qualities with respect to the overall difference, coloration, and localization error are directly explored. To this end, a listening test was conducted in which subjects rated differences between their own HRTF and a set of nonindividual HRTFs. Based on these data, a machine learning model was developed to predict the perceived differences using ratios of a subject’s individual AFs and those of presented nonindividual AFs. Results show that perceived differences can be predicted well and the HRTFs recommended by the models provide a clear improvement over generic or randomly selected HRTFs. In addition, the most relevant AFs for the prediction of each type of error were determined. The developed models are available under a free cultural license.

© 2020 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/). https://doi.org/10.1121/10.0002884
(Received 25 August 2020; revised 5 November 2020; accepted 18 November 2020; published online 18 December 2020)
[Editor: Ning Xiang]

I. INTRODUCTION

The influence of the human head, torso, and pinnae on acoustic signals arriving at the eardrum is described by head-related transfer functions (HRTFs; Møller, 1992). In binaural synthesis, individual HRTFs provide the information that the listener needs to perceive and localize sound accurately. Using nonindividual HRTFs causes an impaired perception of sound source localization, externalization, and timbre (Jenny and Reuter, 2020). The most precise approach to obtain individual HRTFs is by placing microphones in the ear canals and measuring the acoustic signals for a large set of source positions in an anechoic chamber (Brinkmann et al., 2019; Richter and Fels, 2019). Another approach is to run numerical simulations using individual high-resolution three-dimensional (3D) surface meshes (Dinakaran et al., 2018; Katz, 2001). However, both methods require elaborate equipment and specific skills to perform the measurements. Hence, it seems appealing to find other HRTF individualization approaches.

Previous work can be categorized into two different streams: approaches that predict the HRTF by exploiting correlations between anthropometric features (AFs) and HRTF idiosyncrasies on one hand and approaches where the listeners pick a HRTF from a pool-based perceptual similarity (Guezenoc and Séguiér, 2018).

A first approach for AF-based HRTF individualization was to align spectral HRTF features by frequency scaling of nonindividual HRTFs, which was shown to reduce localization errors (Bomhardt, 2017; Middlebrooks, 1999). However, this method has no effect on idiosyncratic features such as the detailed shapes of peaks and notches. Zotkin et al. (2002) used the nearest neighbor approach to show that localization performance and likeness increase when measuring AFs of a subject and selecting the best matching HRTF set from a database. Later, Zotkin et al. (2003) conducted HRTF individualization using a set of seven AFs and an additional low-frequency head-and-torso model. They concluded that low-frequency localization cues provided by the head-and-torso model are desirable for rendering with non-personalized HRTFs, whereas the individualization based on seven AFs did not always perform well. This list of seven AFs was further reduced to four AFs by Liu and Zhong (2016), based on correlations between pinna AFs and spectral distortion and a localization experiment. Xu et al. (2007) studied the influence of AFs using correlation and principal component analyses based on seven source directions and concluded that three factors containing head and shoulder measurements and two pinna angles explain most of the HRTF variance, and Bomhardt (2017) showed that localization with individualized HRTFs using the principal component analysis (PCA) approach is almost comparable.
to localization with individual HRTFs. Spagnol (2020b) found that the three AFs head width, head depth, and shoulder circumference are sufficient to find nonindividual HRTFs for which the predicted horizontal plane localization errors are smaller than the localization blur. Recent works indirectly exploited AFs using photogrammetry and deep learning based on images or videos of the head and pinnae either to predict the HRTF directly or to generate 3D head models from which the HRTF can be simulated (Kaneko et al., 2016; Mäkivirta et al., 2020; Miccini and Spagnol, 2020; Shahid et al., 2018). These approaches, however, require additional perceptual evaluations for benchmarking.

Fewer studies were concerned with direct perceptual individualization approaches. These approaches usually reduce the number of HRTFs that are presented to the subjects in order to speed up the perceptual selection procedure (Katz and Parseihian, 2012; Spagnol, 2020a; Xie et al., 2013). While the initial reduction can either be done based on HRTF similarity measures or perceptual ratings, the selection procedure might use a tournament mode to further speed up the selection (Iwaya, 2006; Voong, 2019). Katz and Parseihian (2012) showed that localization performance with HRTFs that were perceptually selected from a reduced pool of seven data sets is somewhere in the middle between that observed for individual HRTFs and the worst HRTF from the reduced pool.

While perceptual selection procedures require user feedback that might be prone to errors depending on the listening environment and ability of the user to distinguish nuanced differences between audio stimuli, the challenge of previously suggested AF-based individualization approaches lies in finding a link between the employed physical error measures and the listeners’ perception of the recommended nonindividual HRTFs. In the current study, we suggest an approach that combines these two streams by directly linking differences between AFs to perceptual differences between the corresponding HRTFs, thus, enabling a perception-based HRTF recommendation that does not require user feedback or physical error measures that need to be verified in additional listening experiments.

To this end, a listening test was conducted in which participants rated differences between individual and nonindividual HRTFs. This was done with acoustically measured and numerically simulated HRTFs for three selected perceptual qualities. In a second step, regularized random forest (RRF) machine learning models were developed to predict the perceptual differences from the underlying AFs and identify key AFs. The paper is organized as follows. Section II, outlines the methods, including the HRTF and AF database, the listening test design, and the statistical analysis. The results are detailed in Sec. III, which is followed by the discussion in Sec. IV.

II. METHOD

A. HRTF database

HRTFs were taken from the HUTUBS database, which is available online.1 The database contains acoustically measured and numerically simulated HRTFs on full spherical sampling grids of 96 subjects as well as acoustically measured headphone transfer functions (HP-TFs) of 2 headphone models and 25 AFs (cf. Table I and Fig. 1). Detailed information on the compilation and evaluation of the database is given in Brinkmann et al. (2019), and information on the AFs measurement is given in Dinakaran et al. (2016).

B. Perceptual testing

As the first step of our combined approach, we conducted a listening test in which the participants directly rated differences between their own HRTF (reference) and 15 different nonindividual HRTF sets for 2 source positions and 3 perceptual qualities, i.e., difference, coloration, and localization. The qualities were drawn from the Spatial Audio Quality Inventory (SAQI; Lindau et al., 2014) because of their relevance for various binaural applications whereby the localization was condensed from the three SAQI qualities related to horizontal, vertical direction, and distance, to limit the duration of the listening test. The 15 nonindividual sets were randomized in a way that all 90 selected sets of the HUTUBS database were tested after a set of 6 participants. As source positions, a frontal source with no elevation, a source with an elevation of 15°, and shifted to the left by 30° were selected. Separate ratings were acquired for measured and simulated HRTFs.

Different audio content was presented and deemed most suitable to elicit the selected SAQI qualities. For the rather holistic difference measure, a short anechoic female speech stimulus was used, considered as typical real-life content with a high temporal and spectral dynamic (8 s excerpt from a German poem). Broadband continuous pink noise was chosen to evaluate changes in the coloration, while a pulsed pink noise stimulus was chosen for the localization. The pulses were 0.5 s long with 0.3 s pauses and 0.02 s fade-in and fade-out (Majdak et al., 2010). All stimuli were played in a loop. The testing was done with a purely HRTF-based simulation to isolate the effect of the listener morphology and avoid interaction with the acoustic environment (room).

TABLE I. AFs from the HUTUBS database (Brinkmann et al., 2019).

| Identification | AFs | Identification | AFs |
|----------------|------|----------------|------|
| 1               | h    | 2              | d     |
| 2               | h    | 3              | d     |
| 3               | h    | 4              | d     |
| 4               | h    | 5              | d     |
| 5               | h    | 6              | d     |
| 6               | h    | 7              | d     |
| 7               | h    | 8              | d     |
| 8               | h    | 9              | d     |
| 9               | h    | 10             | d     |
| 10              | h    | 11             | d     |
| 11              | h    | 12             | d     |
| 12              | h    | 13             | d     |
| 13              | h    | 14             | d     |
| 14              | h    | 15             | d     |
| 15              | h    | 16             | d     |
| 16              | h    | 17             | d     |


1 Information on the compilation and evaluation of the database is given in Brinkmann et al. (2019).
The listening test was conducted using the MATLAB-based open-source listening test toolbox WhisPER (Ciba et al., 2014) to display the user interface, along with open sound control (OSC) messages to control the audio rendering and acquire the corresponding ratings. Dynamic auralization, accounting for the head orientation of the listener, was realized with a modified version of the SoundScape Renderer (SSR; Geier et al., 2014) to display the user interface, along with open sound control (OSC) messages to control the audio rendering and acquire listeners’ perceptual ratings in terms of (1) difference, (2) coloration, and (3) localization were predicted with RRF machine learning models (Breiman, 2001). The rationale of this was to get an estimate of each AF’s importance for the obtained perceptual differences and provide an indication of the performance of a possible AF-based HRTF recommendation system.

1. Data preparation

As predictor variables for machine learning, we calculated ratios between individual (i) and nonindividual (n) AFs, e.g., \( x_{1,i} / x_{1,n} \). To minimize predictor redundancy for the ear parameters \( d \) and \( \theta \), which are available for the left \( (l) \) and right \( (r) \) ears, AFs from the right side of the head were calculated as double ratios to only account for the asymmetry, e.g., \( d_{1,l,r} / d_{1,l,n} \). Because values for \( d_l \) and \( d_{10} \) did not substantially vary between the left and right ears, the corresponding asymmetry features were removed, resulting in 35 predictors for each model. Additionally, we added two dummy variables to each observation to flag whether a measured or simulated HRTF had been presented and to indicate the source position (frontal/lateral).

To acquire realistic learning performance benchmarks in spite of a limited sample size, we decided to employ a nested k-fold procedure for machine learning, following Vabalas et al. (2019). Thereto, we split our data stemming from 42 subjects into 7 equally sized folds. Because data were non-normal and folds were rather small, we applied a special stratification procedure to ensure that each fold would form a sufficiently comparable learning problem: We calculated the number of zero ratings on the target variable for each subject, then ordered subjects by this number, and partitioned resulting ordinal distribution into seven bins of size six. We then randomly drew one subject from each of the bins for fold one and did the same to construct folds 2–7.

C. Statistical analysis

As the second step of our combined approach, based on the acquired ground truth from the listening test, the

FIG. 1. (Color online) Definition of anthropometric measures listed in Table I and characteristic parts of the pinna. The crus helix slant height \( d_9 \) is marked by a dashed line to improve readability. The dot in (b) marks the center of the ear canal. The dots in (c) and (d) indicate the points that were used to estimate the pinna flare angle \( \theta_3 \) and the starting point for estimating the cavum concha depths \( d_6 \) and \( d_10 \). The stars in (d) show the points that were used to estimate the pinna height \( d_4 \) and pinna rotation angle \( \theta_1 \). The triangles and squares in (c) and (d) mark the end points for estimating the cavum concha depths.
As a result, we arrived at an optimally stratified sevenfold partition for each of the three target variables.

2. Nested k-fold machine learning

For feature selection and machine learning, the R package RRF (Deng and Runger, 2012), implementing guided regularized random regression forests was used in conjunction with the R package caret (Kuhn, 2008), which was employed for k-fold validation during model building. Nested k-fold learning was realized as illustrated in Fig. 2: Within seven different runs, we always used six of the folds—five folds for training, one fold for validation in a sixfold loop—for variable selection, grid-based hyperparameter tuning and model building. The seventh fold was completely held out from model training as a test set. The best combination of the hyper parameters mtry (number of variables randomly drawn in each split), regularization (size of the shrinking parameter), and ntree (number of trees in the forest) was brute-force searched within 96 combinations constructed from common values for regularization and ntree and 4 equally spaced values of mtry up to the maximum possible value given by the number of independent variables (mtry = \{18, 36, 54, 72\}, regularization = \{0.01, 0.05, 0.1, 0.5\}, ntree = \{50, 100, 250, 500, 750, 1000\}). After completing the first learning run, we gathered resulting prediction performance measures on the holdout data set. This comprised \( R^2 \) and normalized discounted cumulative gain (NDCG; see below), as well as the mean absolute error (MAE) between the actual and predicted user ratings. This procedure was repeated seven times with each fold becoming the holdout exactly once. In a last step, final benchmarks were then calculated as the average across all seven runs.

To estimate the performance for each of the three final models, we calculated the average explained variance \( R^2 \), as well as the average NDCG across subjects (Burges et al., 2005), resulting from model-predicted HRTF rankings for each participant, the average MAE, and finally the mean squared error (MSE)-based predictor importance, calculated from the out of bag (OOB) error (Breiman, 2001, Sec. 3.1) of the first nested run. Because our analytical interest was mainly focused on the performance on the upper ranks of results delivered by a hypothetical future HRTF recommender, we used the NDCG as the main decision criterion for choosing the final model variants and their optimal hyper-parameters,

\[
NDCG = \frac{1}{w} \sum_{p=1}^{N} \frac{2^{N-t(p)} - 1}{\log_2(1 + p)},
\]

where \( N \) is the number of rated HRTFs per subject, the predicted rank is \( 1 \leq p \leq N \), and \( 1 \leq t(p) \leq N \) is the true rank at position \( p \). The term \( N - t(p) \) denotes the relevance, which is largest for the best rated HRTF, and \( w \) is a weight to ensure that the NDCG becomes one when the predicted ranks equal the true ranks,

\[
w = \sum_{p=1}^{N} \frac{2^{N-t(p)} - 1}{\log_2(1 + p)},
\]

where \( t(p) \) are the true ranks sorted in ascending order. The measure can be interpreted like a ranking correlation coefficient that emphasizes the first ranks of a recommendation, which is why it is typically used in search engine competitions.

III. RESULTS

Forty-two subjects whose HRTFs are contained in the HUTUBS database (five female, median age 33 years old) participated in the listening test. Except for two, all subjects had listening test experience. The duration of the test, including instructions and training, was approximately 45 min.

A. Difference ratings

The ratings from the listening test are shown in Fig. 3 in terms of the median, interquartile range, and the 5 and 95 percentiles. For convenience, the raw ratings were scaled to the range between zero (no difference) and one (very large difference) by a division by three, whereby a zero rating denotes no perceivable differences between a specific individual and nonindividual HRTFs and a rating of one denotes

![Fig. 2](https://doi.org/10.1121/10.0002884)
very large differences. The results show that some of the randomly presented nonindividual HRTFs were perceptually identical to the individual HRTF with respect to the tested qualities, whereas others were perceived as being very different. This indicates that the HUTUBS database is sufficiently large to provide well-matching nonindividual HRTFs, in many cases, and sufficiently diverse to produce large differences between individual and nonindividual HRTFs, which is required for recommendation-based HRTF individualization.

Notably, median differences of the measured HRTFs are consistently larger than those of the simulated HRTFs, which might be related to additional sources of variance and errors, such as hair style, clothing, and positioning accuracy, during the measurements that are absent in the simulated HRTFs. In addition, perceived differences between source positions are smaller than differences between measured and simulated HRTFs, suggesting a larger consistency within measured and simulated data sets than between them. Last, perceived differences are generally larger for coloration than for the remaining two qualities. This might be an effect of the audio content for difference, where less critical speech material was used. Moreover, it suggests that correctly modeling the localization is less demanding than modeling the coloration because the noise content was used in both cases.

### B. Model evaluation

Obtained benchmarks for the machine learning model can be found in Table II. Perceived differences for localization and difference could be predicted quite well with an NDCG of approximately 75% and a MAE of 0.05 while the performance for the coloration model was considerably worse. Concerning the hyper-parameters, a forest of ntree = 500 trees turned out to be sufficient with varying variable selection (mtry) and regularization (reg) parameters for each model.

Whereas the NDCG and MAE indicate a good performance of the models, a comparison of predicted and rated differences is also of interest. To this end, we focused on the HRTFs with the best ratings per subject and compared them to the HRTFs that were recommended by the models. The results are shown in Fig. 4. The ratings of the best nonindividual HRTF across subjects and two source positions (dark box, solid outline) are close to the individual HRTF on average, and more than 75% of the ratings are below the overall median for all test conditions. Only in some cases, even the best HRTF was rated worse than the overall median, most likely because a well-matching nonindividual HRTF was not available in the randomly selected 15 data sets of the listening test.

The ratings of the recommended HRTFs based on the holdout data (dark box, dashed outlines, cf. Fig. 2) are comparable to the ratings of the best HRTF, which indicates the validity of the models. Only the dispersion of the data (IQR, 5–95 percentile range) tends to be slightly higher due to small recommendation errors.

The model predictions of the perceived differences (white box, solid outlines) are slightly larger than the actual ratings, i.e., the model seems to be a conservative estimator for the perceived differences between individual and nonindividual HRTFs.

Last, the predicted differences for the recommended HRTFs based on the entire HUTUBS database (white box, dashed outlines) were analyzed for all possible combinations of the 96 subjects. The median values decrease only slightly compared to the predictions based on the holdout data. However, the 95 percentile values that indicate the quality at the most unfavourable combinations of individual listeners and recommended HRTFs are significantly lower and consistently fall below the median in all cases. This shows that these unfavourable matches no longer occur when the entire database is available. Or, in other words, the HUTUBS database is large and diverse enough to ensure that in each case, an improvement is achieved by an appropriately selected HRTF.

### C. Predictor importance

For each of the three machine learning models, we calculated MSE-based predictor importance scores from OOB error of the first nested run, representing the increase in the
model fit directly attributable to each predictor (cf. Fig. 5). For easier readability, the predictor importances were normalized by scaling them to 100%. The number of predictors contained in the model is approximately equal for difference with 17 predictors and for localization with 20 predictors. In both cases, the number of predictors and their importance are approximately equally distributed across the three categories head and body features, ear features, and asymmetry features. In contrast, only seven predictors are included in the coloration model. In this case, no ear features predictor is included in the model.

Note that the MSE-based variable importance scores of dummy predictors (here, HRTF type, i.e., measured vs simulated) cannot be interpreted reasonably in tree-based models because they are based on permutations of empirical values. To this end, we calculated RRF models with different combinations of dummy predictors and used the NDCG differences between models to test the actual increase in the prediction accuracy due to this additional information. Compared to the baseline model without any dummy predictors, using information about the presented source (frontal, lateral) increased the NDCG by 8% on average, whereas using information about the HRTF type (measured, simulated) led to an average benefit of 16%. The best results were obtained by using both types of information in which case a NDCG increase of 23% was observed.

D. Primary research data

The ratings from the listening test and the statistical models are provided in the supplementary materials under a free cultural license, along with tables containing the AFs and their ratios. The models are provided as R statistics objects and accompanied by a script that shows their usage.

IV. DISCUSSION

The current study investigated the possibility of predicting perceived differences between individual and
nonindividual HRTFs based on AFs with the purpose of HRTF recommendation for individualization. To this end, perceptual differences were assessed in a listening test and predicted using RRF regression models.

A. Performance of the suggested approach

The model evaluation showed a high goodness of fit with NDCG and \( R^2 \) values between approximately 65% and 75% (cf. Table II). Accordingly, a good agreement between predicted and actual ratings from the listening test was observed (blue boxes in Fig. 4). The potential of the approach for HRTF individualization can be read from the model predictions based on the complete HUTUBS HRTF database, which were below the overall median ratings from the listening test for the tested qualities difference, coloration, and source position (white boxes with dashed outline in Fig. 4). Considering that the predictions tend to be conservative, i.e., overestimate the actually perceived differences, the true potential might even be slightly larger.

The comparison of our results to previous studies is not straightforward because, to the best of our knowledge, this is the first study that directly links differences in AFs to perceptual differences between pairs of HRTFs, thus, bypassing any technical HRTF measures. For this reason, a direct comparison of statistical parameters as \( R^2 \) would be inappropriate because perceptual ratings can be assumed to contain more noise than technical measures based on the HRTFs themselves. For example Jin et al. (2000) observed \( R^2 \) values between approximately 85% and 95% for regression models between anthropometrical data and HRTFs. In addition, we performed relative localization experiments that cannot be directly compared to absolute localization experiments as conducted, for example, by Katz and Parseihian (2012).

B. Key AFs

The identification of anthropometric key features from the predictor importance of our machine learning models provides valuable insights that could limit the amount of AFs that are required for future prediction models. The interpretation should be done with care because white middle-aged males are clearly overrepresented in the current sample, and a different sample might lead to, at least, slightly different results. Nevertheless, some of the trends of the predictor importances shown in Fig. 5 are still noteworthy. In general, features from all three categories (i) head and body, (ii) ear, and (iii) ear asymmetry are contained in the models. Although a clear tendency cannot be observed for the ear and asymmetry features, the head height and depth \( (d_{3,1}) \) and the pinna offset \( (d_{1,5}) \) are contained in all three models. The head width \( (x_1) \) and other features describing the shoulder or torso width are never used.

For the overall difference, ear features seem to be of particular importance with the cavum concha height \( (d_1) \) as the most prominent feature. This confirms the results of a study that made variations of several AFs using boundary-element method (BEM) meshes and identified cavum concha as the most prominent parameter for HRTF individualization (Ghorbal et al., 2017).

This is also true for the coloration with the difference that the asymmetry between the left and right ears seems to be more important with the cavum concha depth asymmetry \( (d_9) \) being the most important feature. This means that for perceptual HRTF similarity regarding coloration, differences between a listener’s individual and nonindividual HRTFs are best explained by the consistency between the left and right ear pinnae features.

At first glance, it seems surprising that the neck depth \( (x_8) \) is the most important feature for predicting the localization and the head width is not included despite its known effect on interaural time differences (ITDs) and horizontal plane localization (Algazi et al., 2001). However, the Pearson correlation showed that neck depth \( (x_8) \) strongly correlates with head width \( (r = 0.63) \) and correlates equally well or better than head width to other AFs influencing the ITD, such as head circumference \( (x_{16}, r = 0.45) \) and shoulder circumference \( (x_{17}, r = 0.70) \). Additionally, the neck depth \( (x_8) \) also shows correlations to pinna-related features that are weak but larger than correlations between the same features and the head width \( (x_1) \). It, therefore, seems likely that our machine learning approach identified neck width as a single key predictor for multiple AFs for the horizontal plane localization along with other features such as head height and depth \( (x_{2,3}) \) and pinnae offset \( (x_{4,5}) \). In addition, AFs related to the median plane localization, such as the cavum concha height \( (d_1) \) and pinna rotation angle \( (\theta_1) \), are also among the predictors.

C. Future work

For this exploratory study, we used AFs extracted from high quality 3D surface meshes, which could also be used to numerically simulate individual HRTFs (Katz, 2001). However, finding a best match within an existing database will always be the faster approach even if considering parallelized HRTF simulation in the cloud. Future work, for example, could investigate how well the suggested approach performs with fewer AFs or with features that can be obtained from images or videos instead of 3D surface scans (Mäkivirta et al., 2020; Torres-Gallegos et al., 2015).

The present work has demonstrated that a well-performing HRTF recommendation system based on perceptual error metrics and AFs using machine learning is possible. For more robust recommendations, future work should try to increase the size and the diversity of the database and include more than two source positions in the perceptual testing, e.g., positions that maximize the inter-individual variance should ideally be selected (Andreopoulou and Roginska, 2017). In addition, mechanisms should be implemented that continuously refine the system based on user feedback with, for example, absolute localization tests carried out in virtual reality environments to compare the performance of recommended and generic HRTF sets. In the long run, an open
application programming interface appears to be the best solution to continually gather ground truth in this way.

V. CONCLUSION

We presented a combined approach for HRTF recommendation that directly predicts the perceptually best matching HRTF based on the listeners’ AFs, whereas previous work either employed a manual perceptual selection procedure or predicted the HRTF using error measures whose perceptual relevance might not always be clear. On average, the predicted HRTFs performed almost as well as the perceptually best matching HRTF in terms of the three perceptual qualities difference, coloration, and localization, thus, showing its potential for HRTF individualization.

ACKNOWLEDGMENTS

We would like to thank the participants of the listening test for their time.

1See https://dx.doi.org/10.14279/depositonce-8487.
2See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0002884 for the ratings from the listening test and the statistical models, along with tables containing the AFs and their ratios.

AES Standards Committee (2015). AES69-2015: AES Standard for File Exchange—Spatial Acoustic Data File Format (Audio Engineering Society, Inc., New York, NY).

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). “The CIPIC HRTF database,” in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, pp. 99–102.

Andreopoulou, A., and Roginska, A. (2017). “Database matching of sparsely measured head-related transfer functions,” J. Audio Eng. Soc. 65(7/8), 552–561.

Bombhardt, R. (2017). “Anthropometric individualization of head-related transfer functions. Analysis and modeling,” Ph.D. thesis, RWTH Aachen University, Aachen, Germany.

Breiman, L. (2001). “Random forests,” Mach. Learn. 45(1), 5–32.

Brinkmann, F., Dinakaran, M., Pelzer, R., Grosse, P., Voss, D., and Weinzierl, S. (2019). “A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses,” J. Audio Eng. Soc. 67(9), 705–718.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). “Learning to rank using gradient descent,” in 22nd Int. Conf. on Machine Learning, Bonn, Germany, pp. 89–96.

Ciba, S., Brinkmann, F., and Lindau, A. (2014). “WhisPER. A MATLAB toolbox for performing quantitative and qualitative listening tests.”

Deng, H., and Runger, G. (2012). “Feature selection via regularized trees,” arXiv:1201.1587.

Dinakaran, M., Brinkmann, F., Harder, S., Pelzer, R., Grosse, P., Paulsen, R. R., and Weinzierl, S. (2018). “Perceptually motivated analysis of numerically simulated head-related transfer functions generated by various 3D surface scanning systems,” in 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada, pp. 551–555.

Dinakaran, M., Grosse, P., Brinkmann, F., and Weinzierl, S. (2016). “Extraction of anthropometric measures from 3D-meshes for the individualization of head-related transfer functions,” in 140th AES Convention, Paris, France, available at http://www.aes.org/e-lib/browse.cfm?elib=18277 (Last viewed 12/7/2020).

Geier, M., Ahrens, J., and Spors, S. (2008). “The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods,” in 124th AES Convention, Amsterdam, The Netherlands, available at http://www.aes.org/e-lib/browse.cfm?elib=14460 (Last viewed 12/7/2020).

Ghorbel, S., Auclair, T., Soladie, C., and Seguier, R. (2017). “Finn morphological parameters influencing HRTF sets,” in International Conference on Digital Audio Effects, Edinburgh, United Kingdom, available at https://hal.archives-ouvertes.fr/hal-01831314 (Last viewed 12/7/2020).

Guezenec, C., and Séguire, R. (2018). “HRTF individualization: A survey,” in 145th AES Convention, New York, NY, available at http://www.aes.org/e-lib/browse.cfm?elib=19855 (Last viewed 12/7/2020).

Iwaya, Y. (2006). “Individualization of head-related transfer functions with tournament-style listening test: Listening with other’s ears,” Acoust. Sci. Technol. 27(6), 340–343.

Jenny, C., and Reuter, C. (2020). “Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization,” JMIR Serious Games 8(3), e17576.

Jin, C., Leong, P., Leung, J., Corderoy, A., and Carlile, S. (2000). “Enabling individualized virtual auditory space using morphological measurements,” in IEEE Pacific-Rim Conference on Multimedia, pp. 235–238.

Kaneko, S., Suenaga, T., and Sekine, S. (2016). “DeepEarNet: Individualizing spatial audio with photography, ear shape modeling, and neural networks,” in AES Conf. on Audio for Virtual and Augmented Reality, Los Angeles, CA.

Katz, B. (2001). “Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation,” J. Acoust. Soc. Am. 110, 2440–2448.

Katz, B. F. G., and Parsehian, G. (2012). “Perceptually based head-related transfer function database optimization,” J. Acoust. Soc. Am. 131(2), EL99–EL105.

Kuhn, M. (2008). “Building predictive models in R using the caret package,” J. Stat. Software 28(5), 1–26.

Lindau, A., and Brinkmann, F. (2012). “Perceptual evaluation of headphone compensation in binaural recording based on non-individual recordings,” J. Audio Eng. Soc. 60(1), 54–62, see http://www.aes.org/e-lib/browse.cfm?elib=16166.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-I., Brinkmann, F., and Weinzierl, S. (2014). “A Spatial Audio Quality Inventory (SAQI),” Acta Acust. Acust. 100(5), 984–994.

Liu, X., and Zhong, X. (2016). “An improved anthropometry-based customization method of individual head-related transfer functions,” in 2016 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 336–339.

Majdak, P., Goupell, M. J., and LaBack, B. (2010). “3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training,” Atten. Percept. Psychophys. 72(2), 454–469.

Mäkivirta, A., Malinen, M., Johansson, J., Saari, V., Karjaleinen, A., and Vosough, P. (2020). “Accuracy of photogrammetric extraction of the head and torso shape for personal acoustic HRTF modeling,” in 148th AES Convention, Vienna, Austria.

Miccini, R., and Spagnol, S. (2020). “HRTF individualization using deep learning,” in IEEE Conference on Virtual Reality and 3D User Interfaces, Atlanta, GA, pp. 390–395.

Middlebrooks, J. C. (1999). “Individual differences in external-ear transfer functions reduced by scaling in frequency,” J. Acoust. Soc. Am. 106(3), 1480–1492.

Möller, H. (1992). “Fundamentals of binaural technology,” Appl. Acoust. 36(3), 171–218.

Richter, J.-G., and Fels, J. (2019). “On the influence of continuous subject rotation during high-resolution head-related transfer function measurements,” IEEE/ACM Trans. Audio Speech Lang. Process. 27(4), 730–741.

Shahid, F., Javeri, N., Jain, K., and Badhwar, S. (2018). “AI DevOps for large-scale HRTF prediction and evaluation: An end to end pipeline,” in AES Int. Conf. on Audio for Virtual and Augmented Reality (AVAR), Redmond, WA.

Spagnol, S. (2020a). “Auditory model based subsetting of head-related transfer function datasets,” in IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 391–395.

Spagnol, S. (2020b). “HRTF selection by anthropometric regression for improving horizontal localization accuracy,” IEEE Signal Process. Lett. 27, 595–594.

Thurlow, W. R., Mangels, J. W., and Ringe, P. S. (1967). “Head movements during sound localization,” J. Acoust. Soc. Am. 42(2), 489–493.

Torres-Galloges, E. A., Orduña-Bustamente, F., and Arambula-Casío, F. (2015). “Personalization of head-related transfer functions (HRTF) based
on automatic photo-anthropometry and inference from a database,” *Appl. Acoust.* 97, 84–95.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). “Machine learning algorithm validation with a limited sample size,” *PLoS One* 14(11), e0224365.

Voong, T. M. (2019). “Tournament formats as method for determining best-fitting HRTF profiles for individuals wearing bone conduction headphones,” in *23rd Int. Congress on Acoustics (ICA)*, Aachen, Germany, pp. 4841–4847.

Xie, B., Zhang, C., and Zhong, X. (2013). “A cluster and subjective selection-based HRTF customization scheme for improving binaural reproduction of 5.1 channel surround sound,” in *134th AES Convention*, Rome, Italy, paper 8879.

Xu, S., Li, Z. Z., Zeng, L., and Salvendy, G. (2007). “A study of morphological influence on head-related transfer functions,” in *IEEE Int. Conf. on Industrial Engineering and Engineering Management*, pp. 472–476.

Zotkin, D. N., Duraiswami, R., Davis, L. S., Mohan, A., and Raykar, V. (2002). “Virtual audio system customization using visual matching of ear parameters,” in *Object Recognition Supported by User Interaction for Service Robots*, Quebec City, Quebec, Canada, 2002, Vol. 3, pp. 1003–1006.

Zotkin, D. N., Hwang, J., Duraiswami, R., and Davis, L. S. (2003). “HRTF personalization using anthropometric measurements,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, pp. 157–160.