HRTF Selection by Anthropometric Regression for Improving Horizontal Localization Accuracy

Simone Spagnol, *Member, IEEE*

Abstract—**This work focuses on objective Head-Related Transfer Function (HRTF) selection from anthropometric measurements for minimizing localization error in the frontal half of the horizontal plane. Localization predictions for every pair of 90 subjects in the HUTUBS database are first computed through an interaural time difference-based auditory model, and an error metric based on the predicted lateral error is derived. A multiple stepwise linear regression model for predicting error from inter-subject anthropometric differences is then built on a subset of subjects and evaluated on a complementary test set. Results show that by using just three anthropometric parameters of the head and torso (head width, head depth, and shoulder circumference) the model is able to identify non-individual HRTFs whose predicted horizontal localization error generally lies below the localization blur. When using a lower number of anthropometric parameters, this result is not guaranteed.**

Index Terms—**Anthropometry, auditory model, HRTF, ITD, sound localization.**

I. Introduction

When used for binaural rendering, head-related transfer functions (HRTFs) yield the most accurate localization results when they are individual [1]. Unfortunately, collecting accurate individual HRTFs requires considerable effort on both the user's and the experimenter's side. Although it is possible to numerically simulate individual HRTFs from a head mesh [2], [3], perceptual studies that validate numerically simulated against acoustically measured HRTFs are rare [4]. A further alternative consists in synthesizing HRTFs from a small number of anthropometric measurements of the listener's head, ears, and/or torso [5], [6]. However, this latter approach suffers from the lack of a full and thorough understanding of the mechanisms involved in spatial sound perception [7]. Providentially, the rising availability of public HRTF data makes it possible for a listener to evaluate several different non-individual HRTF sets out of hundreds of candidates and select the best fitting one [8].

In this context, objective metrics for fast selection or subsetting of HRTF sets are required in order to speed up the process. As opposed to subjective HRTF selection [9]–[11], the use of a small number of anthropometric parameters for estimating the relative fitness of a non-individual HRTF set compared to another is especially attractive because of the little effort it takes on the user's behalf. This approach was first pursued by Zotkin *et al.* [12] who proposed selecting the HRTF set that best matches an anthropometric data vector of the pinna. While this approach yielded a marginal improvement in elevation localization performances compared to generic non-individual HRTFs, it ignored horizontal localization, for which the Interaural Time Difference (ITD) of the spectral components below 1 kHz generally plays an important role [13].

Different analytical solutions for individual ITD estimation and adaptation, such as spherical and ellipsoidal head models, are available in previous literature (for a review see [14]). Despite their usability, these models suffer from evident discrepancies between modeled and human interaural differences [15]. Subjective ITD selection procedures [16] are available as well. However, no previous study to this author's knowledge explicitly focused on objective ITD-based HRTF selection from anthropometry. Accordingly, this work proposes a linear regression model for estimating localization error in the frontal horizontal plane as predicted from an auditory model from inter-subject anthropometric differences. The regression model is then used to select and evaluate non-individual HRTFs from a recently released HRTF database for a pool of database subjects.

II. Methods

A. The Data

This work uses the recently released[1] HUTUBS HRTF database [17]. The database includes both acoustically measured and numerically simulated HRTFs in SOFA format[2] as well as full anthropometric measurements for 90 different human subjects.[3]

This work considers the ground-truth, acoustically measured HRTF data. HRTFs were measured in an anechoic chamber with a sampling rate of 44.1 kHz and saved as 256-sample impulse responses. These are available with a resolution of 10° in elevation and a variable resolution in azimuth yielding an almost constant great circle distance between neighboring points. Since

Manuscript received January 29, 2020; revised March 19, 2020; accepted March 24, 2020. Date of publication March 30, 2020; date of current version April 30, 2020. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant 797850. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg.

The author is with the Aalborg University, 2450 Copenhagen, Denmark (e-mail: ssp@create.aau.dk).

Digital Object Identifier 10.1109/LSP.2020.2983633

[1] `http://dx.doi.org/10.14279/depositonce-8487`

[2] `http://www.sofaconventions.org`

[3] The database includes a total of 96 HRTF sets, two of which are artificial head measurements (IDs 1 and 96), one is a repeated measurement of a single subject (ID 88), and three do not have anthropometric data available (IDs 18, 79, and 92). These six sets were not considered in the following analysis.

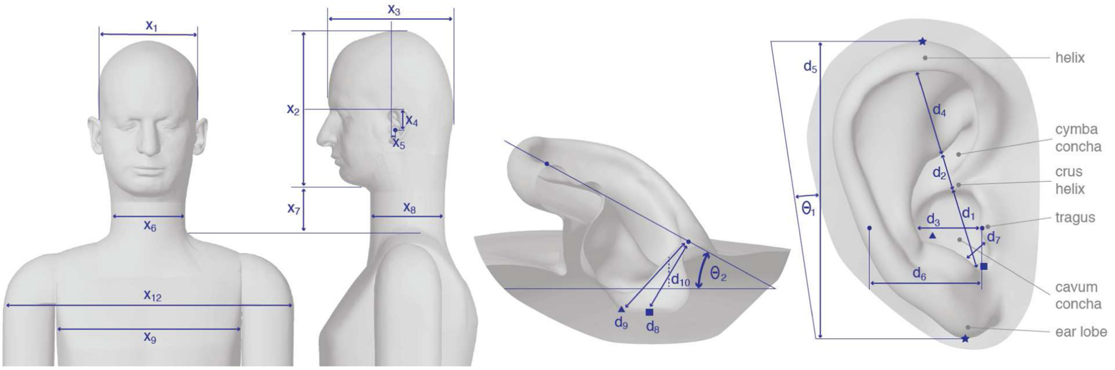

Fig. 1. Definition of anthropometric measures in the HUTUBS database. Not represented in the figure are the head circumference x_{16} and the shoulder circumference x_{17}. Figure reproduced from [17].

in this work we are interested in horizontal localization, we focus on the horizontal plane, where HRTFs are sampled in 10° azimuth steps.

Of all available anthropometric measurements, we consider those related to the head, torso, and pinnae, for a total of 36 measurements per subject (12 for head and torso, 12 for the right pinna, and 12 for the left pinna) which form the *anthropometric vector*. These are illustrated in Fig. 1. Although head width and, to a lesser extent, head depth are known to be the main predictors for interaural cues [18]–[20], the remaining measures were included in the feature selection stage for fine tuning our regression model.

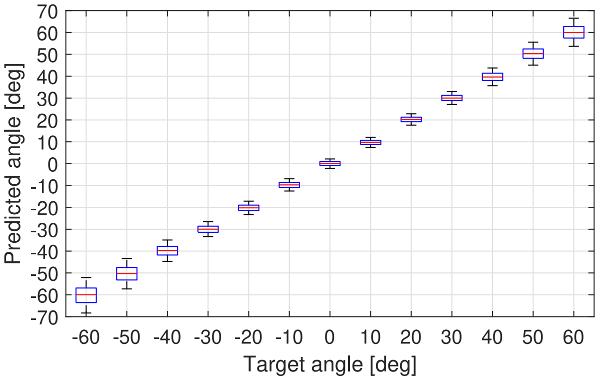

Fig. 2. Distribution of localization predictions by target angle over all virtual HUTUBS listeners and non-individual HRTFs. Central box mark: median, box edges: 25th/75th percentiles, whiskers: 5th/95th percentiles.

B. *Localization Predictions*

The auditory model used to predict localization with non-individual HRTFs is the binaural model by Wierstorf *et al.* [21] included in the Auditory Modeling Toolbox.[4] This model was originally conceived to enable localization predictions for a particular wave field synthesis setup; however as acknowledged by the authors, since the model provides predictions in agreement with real listening test data [21] it can be used to create localization maps for other setups.

The auditory model requires as input a lookup table describing the mapping of ITD cues from a reference HRTF set belonging to one specific virtual listener to azimuth angles. This is done by first computing through the binaural model by Dietz *et al.* [22] the ITD for each available azimuth angle in the frontal half of the horizontal plane and for each channel in a 4th-order all-pole gammatone filterbank (employing 12 bands between 200 and 1400 Hz with 1-ERB spacing and filter width), and then fitting twelve 12th order polynomials for the ITD-to-azimuth mapping. Accordingly, lookup tables for the 90 HUTUBS subjects were calculated for azimuths in the $[-90, 90]^\circ$ range (19 angles, right to left).

Then, the auditory model estimates the perceived direction for a given binaural stimulus. In this work, binaural stimuli were obtained by convolution of a 100 ms white noise signal with every available HRTF in the $[-60, 60]^\circ$ azimuth range (13 angles). The reason for considering such a restricted range is due to a flattening effect of the ITD for sources to the sides of the listener (which is responsible for the localization blur to be an order of magnitude worse than that for frontal sources [23]), which complicates the achievement of a proper fit of the ITDs and their corresponding azimuths [21].

For the prediction of the perceived direction of a binaural stimulus, the auditory model first calculates the ITD values over time as before. Then, for each of the twelve auditory channels, ITDs are transformed into azimuths by use of the lookup table. If the absolute ITD value in an auditory channel is larger than 1 ms, this channel is disregarded. Afterwards, the median azimuth value across auditory channels is taken as the predicted direction. If the angle in an auditory channel differs by more than 30° from the median, it is considered an outlier and skipped, and the median is re-calculated.

By using this auditory model we can estimate the perceived horizontal localization for each available pair of HRTF sets. In other words, we can simulate the horizontal localization performance of a virtual listener described by his/her individual HRTF set using another HRTF set. Accordingly, 90 (reference individual HRTF sets) ×89 (non-individual HRTF sets) ×13 (target azimuth angles) $=$ 104130 localization predictions were calculated. Fig. 2 shows the distribution of these predictions

[4] `http://amtoolbox.sourceforge.net`

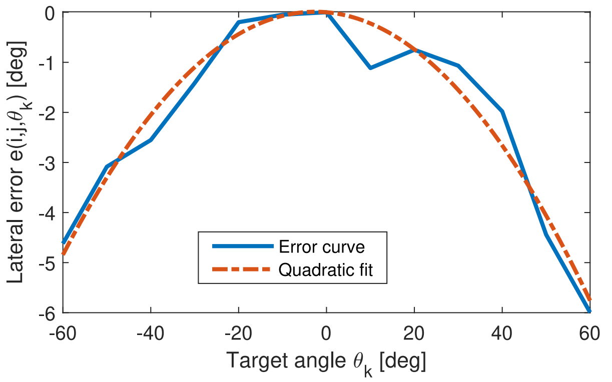

Fig. 3. Lateral error for a representative subject pair (virtual listener ID 52 and non-individual HRTF ID 7) and the relative best fitting quadratic function. This subject pair represents a predicted underestimation of lateral angle judgments, i.e., a negative error trend.

divided by target angle, where the increased localization variance for lateral sources can be appreciated.

C. Error Metric

In order to evaluate the perceptual fitness of one HRTF set to another, the following error metric is calculated. For each perceived direction prediction $p(i, j, \theta_k)$, where i is the reference subject, j is one of the other subjects ($j \neq i$), and θ_k is one target angle in the $[-60, 60]^\circ$ azimuth range, we first define the corresponding *lateral error*

$$e(i, j, \theta_k) = \text{sgn}(\theta_k) * [p(i, j, \theta_k) - \theta_k - p(i, j, 0)] \qquad (1)$$

where subtraction by the 0-degree prediction accounts for frontal asymmetry,[5] and the purpose of the sign function is to assign positive or negative error values to over- or underestimated lateral directions respectively, independently of the hemisphere.

As exemplified in Fig. 3, the general trend of e across target angles is that of a parabola. This is due to the increasing error for higher absolute values of the angle. Therefore, a good single-value indicator for the overall *error trend* across angles is the coefficient of the square term of the best fitting parabola, which controls the direction of concavity (i.e., whether target angles are underestimated – as in the above example – or overestimated) and the steepness (i.e., the increase rate of the error). We therefore fit $\mathbf{e}_{ij} = [e(i, j, \theta_1), \ldots, e(i, j, \theta_{13})]^\intercal$ across target angles $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{13}]^\intercal$, expressed in radians, with the best fitting quadratic function of the kind $\mathbf{e}_{ij} = a_{ij}\boldsymbol{\theta}^2 + b_{ij}\boldsymbol{\theta}$ in a least-squares sense. Here we force the y-intercept to zero because, by definition, $e(i, j, 0) = 0$ for every i and j. Formally, we solve the system of linear equations $\mathbf{V}\mathbf{c}_{ij} = \mathbf{e}_{ij}$ where $\mathbf{V}$ is the Vandermonde matrix of $\boldsymbol{\theta}$ with the constant column removed and $\mathbf{c}_{ij} = [a_{ij}, b_{ij}]^\intercal$, and save the matrix $\mathbf{A} = \{a_{ij}\}$ of error trend coefficients.

D. Regression Model

The 90 HUTUBS subjects are randomly split into a training set ($N = 70$ subjects), used for tuning the regression model, and a test set ($M = 20$ subjects), kept aside as independent pool of subjects for the evaluation stage. The assumption that every training and test set anthropometric feature comes from normal distributions with equal means is verified with a two-sample t-test. A leave-one-out cross-validation scheme is implemented on the training set: in turn, each one of the N subjects is set aside (subject k), a series of regression models is built on the remaining pool of $N - 1$ subjects, and then validated on the one left out.

TABLE I
SELECTED FEATURES AND REGRESSION COEFFICIENTS OF THE FINAL MULTIPLE LINEAR REGRESSION MODEL

Features	Intercept	Δx_1	Δx_3	Δx_{17}
Coefficient	0.378	-2.135	-0.625	-1.845

The standardized differences between the anthropometric vectors of each subject pair $\{i, j\}$ of the pool are used as input records for the regression models, thus catching their anthropometric similarity, while the error trend coefficient a_{ij} is used as target variable (output value). Similar records are computed for subject k as standardized[6] differences between his/her anthropometric vector and that of each subject l of the pool, using as corresponding target variable a_{kl}. In this manner, for each cross-validation fold the regression models are trained on $(N - 1) \times (N - 2) = 4692$ records, and validated on $N - 1 = 69$ records.

The input records are used to perform a series of multiple linear stepwise regressions, where the feature selection procedure aims at discarding features that are irrelevant predictors and may instead cause overfitting. The method consists in initializing a constant regression model and then iteratively including each of the non-selected features, evaluating its performance on the validation set using an appropriate metric, and selecting as predictor the feature resulting in the best performance, until a stopping condition is reached. In our case, the performance metric is the coefficient of determination (r^2) between predicted and target values for subject k, and the stopping condition occurs when no additional feature significantly improves r^2 by more than 0.01.

Each cross-validation fold comes with a number of (possibly different) selected features; these are generalized by preserving just those that appear in at least two thirds of the folds. The new subset of features is then used to train a multiple linear regression model on the whole training set, including all the N subjects for a total of $N \times (N - 1) = 4830$ records. Lastly, the performance of this final regression model is evaluated on the test set, using as records all possible comparisons of test versus training subjects ($M \times N = 1400$ records).

III. RESULTS AND DISCUSSION

The features selected through the cross-validation procedure are the differences in x_1 (head width), x_3 (head depth), and x_{17} (shoulder circumference). Table I reports the coefficients of the multiple linear regression model built on the training set,

[5] Alignment of the HRTF set to the perceived frontal direction can be done perceptually prior to non-individual HRTF fruition.

[6] The same scale and offset factors of the training data are used to standardize validation/test data.

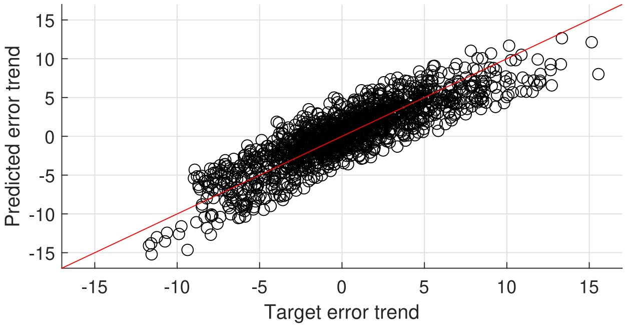

Fig. 4. Actual versus predicted error trend in the test set. The diagonal line represents perfect fit.

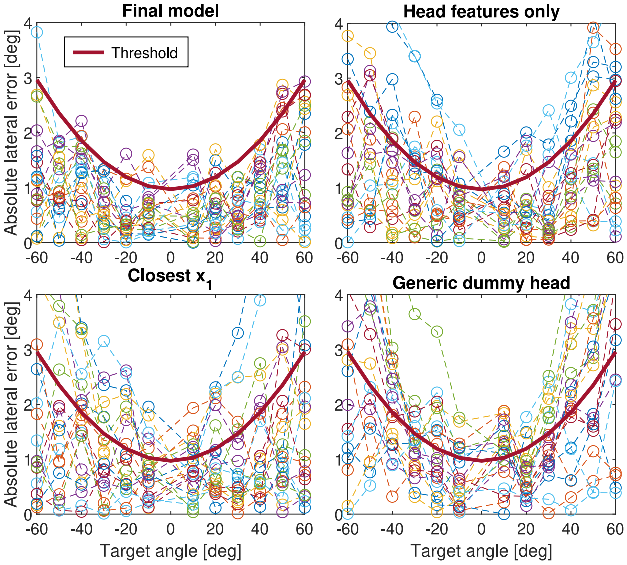

Fig. 5. Absolute lateral error for test set subjects with non-individual HRTFs: (1) best fitting HRTF according to the final regression model (top left), (2) best fitting HRTF according to an alternative regression model on the head dimensions only (top right), (3) HRTF selected by closest head width (bottom left), and (4) FABIAN dummy head HRTF (bottom right). Solid curves represent the approximate localization blur threshold.

that highlight (1) the prominent role of x_1, constantly selected as first predictor in all folds, and (2) the consistent direction of anthropometric differences. Intuitively, listening through the HRTFs of a subject with smaller/larger anatomical structures (positive/negative Δ's) results in under-/overestimation of lateral angles. The scatterplot in Fig. 4 shows the actual versus predicted error trend coefficients in the test set where the regression model scores $r^2 = 0.755$, a result that compared to the training set $r^2 = 0.768$ assesses its robustness and low variance.

As such, the proposed regression model can be used for both discarding predicted high-error non-individual HRTFs or selecting as best fitting HRTF the one minimizing the absolute value of the error trend coefficient. For the latter case, for every test set subject we calculate the best fitting training set HRTF according to the regression model and plot the absolute lateral error according to the auditory model as in Fig. 5 (top left panel). We can therefore verify whether the error lies below a psychoacoustic threshold representing the horizontal-plane localization blur. The threshold was derived from the results by Mills (Fig. 6 in [24]) by approximately averaging the thresholds for 500 Hz and 1 kHz and fitting them to a quadratic function. From Fig. 5, it can be seen that the predicted best fitting HRTF according to the regression model scores a low rate of lateral errors and that in most cases these do not exceed the threshold of more than 0.5 deg.

As expected (see Section II-A), head width and depth were selected as significant error predictors. By contrast, the inclusion of shoulder circumference as significant predictor could be seen as counterintuitive, and might be due to moderate correlations with head width ($r = 0.68$ in the HUTUBS database) and depth ($r = 0.38$). However, recent work [15] found that the inclusion of the torso in an idealized head model is responsible for the addition of small ripples to the ITD that increase its frequency-dependent structure and improve agreement with ground-truth acoustical data especially at middle lateral azimuths.

In order to check for the influence of the shoulder feature on the regression model used for HRTF selection, we build an alternative linear regression model on the two head features only and calculate the best fitting training set HRTF according to this alternative model for every test set subject as before. As additional control conditions, we also consider the HRTF of the training set subject with the closest head width to each test set subject as well as the FABIAN dummy head HRTF set (subject ID 1). Absolute lateral error predictions with the best fitting HRTF according to the regression model with head features only, the HRTF selected by closest head width, and the dummy head HRTF are reported in the three other panels of Fig. 5. These show how neither the alternative selection methods, nor the generic HRTF can guarantee the level of performance of the final regression model.

IV. Conclusions

The results of this work suggest that it is possible to perform objective ITD-based HRTF selection by using just a few anthropometric parameters of the head and torso. It has to be acknowledged that the assumptions underlying the ad-hoc metrics and auditory model, used for both training and testing, do not guarantee that listeners in the HUTUBS dataset would be perceptually satisfied with the selected HRTF. Future work will therefore focus on assessing localization accuracy with the selected HRTF sets through subjective perceptual tests.

This study focused on localization in the frontal part of the horizontal plane alone. On one hand, given the similarities in the ITD for frontal and rear sources [25], the results can be extended to localization in the rear semiplane as well. On the other hand, key perceptual attributes for HRTF comparison which mainly reside in spectral cues such as front/back position, elevation, and externalization [26], [27] were not considered here. While the proposed method cannot guarantee HRTF matching in these additional perceptual dimensions, it can still be used for selecting ITD and combine it with an alternatively chosen spectral part of the HRTF [28].

Acknowledgment

The author would like to thank Dr. Erica Tavazzi for insightful advices and support.

REFERENCES

[1] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?" *J. Audio Eng. Soc.*, vol. 44, no. 6, pp. 451–469, Jun. 1996.

[2] B. F. G. Katz, "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation," *J. Acoust. Soc. Amer.*, vol. 110, no. 5, pp. 2440–2448, Nov. 2001.

[3] H. Ziegelwanger, P. Majdak, and W. Kreuzer, "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization," *J. Acoust. Soc. Amer.*, vol. 138, no. 1, pp. 208–222, Jul. 2015.

[4] C. Guezenoc and R. Seguier, "HRTF individualization: A survey," in *Proc. 145th Conv. Audio Eng. Soc.*, New York, NY, USA, Oct. 2018, Art. no. 10129.

[5] P. Bilinski, J. Ahrens, M. R. P. Thomas, I. J. Tashev, and J. C. Platt, "HRTF magnitude synthesis via sparse representation of anthropometric features," in *Proc. 39th IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 4501–4505.

[6] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein, "A manifold learning approach for personalizing HRTFs from anthropometric features," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 559–570, Mar. 2016.

[7] S. Spagnol *et al.*, "Current use and future perspectives of spatial audio technologies in electronic travel aids," *Wireless Comm. Mob. Comput.*, vol. 2018, pp. 1–17, Mar. 2018.

[8] S. Spagnol, "Auditory model based subsetting of head-related transfer function datasets," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2020, pp. 1–5.

[9] B. U. Seeber and H. Fastl, "Subjective selection of non-individual head-related transfer functions," in *Proc. Int. Conf. Auditory Display (ICAD03)*, Jul. 2003, pp. 259–262.

[10] Y. Iwaya, "Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears," *Acoust. Sci. Tech.*, vol. 27, no. 6, pp. 340–343, 2006.

[11] B. F. G. Katz and G. Parseihian, "Perceptually based head-related transfer function database optimization," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. EL99–EL105, Feb. 2012.

[12] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 553–564, Aug. 2004.

[13] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1648–1661, Mar. 1992.

[14] R. Bomhardt, "Anthropometric individualization of head-related transfer functions: Analysis and modeling," Ph.D. dissertation, Inst. Technical Acoust., RWTH Aachen Univ., Aachen, Germany, Jul. 2017.

[15] T. Cai and B. Rakerd, "Computing interaural differences through finite element modeling of idealized human heads," *J. Acoust. Soc. Amer.*, vol. 138, no. 3, pp. 1549–1560, Sep. 2015.

[16] K. McMullen, A. Roginska, and G. H. Wakefield, "Subjective selection of head-related transfer functions (HRTFs) based on spectral coloration and interaural time differences (ITD) cues," in *Proc. 133rd Conv. Audio Eng. Soc.*, Oct. 2012, Art. no. 8770.

[17] F. Brinkmann, M. Dinakaran, R. Pelzer, P. Grosche, D. Voss, and S. Weinzierl, "A cross-evaluated database of measured and simulated HRTFs including 3D head meshes, anthropometric features, and headphone impulse responses," *J. Audio Eng. Soc.*, vol. 67, no. 9, pp. 705–718, Sep. 2019.

[18] G. F. Kuhn, "Model for the interaural time differences in the azimuthal plane," *J. Acoust. Soc. Amer.*, vol. 62, no. 1, pp. 157–167, Jul. 1977.

[19] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a spherical-head model from anthropometry," *J. Audio Eng. Soc.*, vol. 49, no. 6, pp. 472–479, Jun. 2001.

[20] A. Lindau, J. Estrella, and S. Weinzierl, "Individualization of dynamic binaural synthesis by real time manipulation of the ITD," in *Proc. 128th Conv. Audio Eng. Soc.*, May 2010, Art. no. 8088.

[21] H. Wierstorf, A. Raake, and S. Spors, "Binaural assessment of multichannel reproduction," in *The Technology of Binaural Listening*, ser. Modern Acoustics and Signal Processing, J. Blauert, Ed. Berlin, Germany: Springer, Jan. 2013, ch. 10, pp. 255–278.

[22] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, May 2011.

[23] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, 2nd ed. Cambridge, MA, USA: MIT Press, Oct. 1996.

[24] A. W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Amer.*, vol. 30, no. 4, pp. 237–246, Apr. 1958.

[25] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2841–2853, May 1999.

[26] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual attributes for the comparison of head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 140, no. 5, pp. 3623–3632, Nov. 2016.

[27] S. Spagnol, "On distance dependence of pinna spectral patterns in head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 137, no. 1, pp. EL58–EL64, Jan. 2015.

[28] M. Aussal, F. Alouges, and B. F. G. Katz, "HRTF interpolation and ITD personalization for binaural synthesis using spherical harmonics," in *Proc. 25th UK Conf. Audio Eng. Soc.*, Mar. 2012, pp. 04-1–04-10.