Hybrid multi-harmonic model for the prediction of interaural time differences in individual behind-the-ear hearing-aid-related transfer functions

Florian Pausch1,2,*, Shaima’a Doma1, and Janina Fels1

1 Institute for Hearing Technology and Acoustics, RWTH Aachen University, 52074 Aachen, Germany
2 Acoustics Research Institute, Austrian Academy of Sciences, 1040 Vienna, Austria

Received 5 May 2021, Accepted 2 June 2022

Abstract – Spatial sound perception in aided listeners partly relies on hearing-aid-related transfer functions (HARTFs), describing the directional acoustic paths between a sound source and the hearing-aid (HA) microphones. Compared to head-related transfer functions (HRTFs), the HARTFs of behind-the-ear HAs exhibit substantial differences in spectro-temporal characteristics and binaural cues such as interaural time differences (ITDs). Since assumptions on antipodal microphone placement on the equator of a three-concentric sphere are violated in such datasets, predicting the ITDs via Kuhn’s simple analytic harmonic model entails excessive errors. Although angular ear-canal offsets have been addressed in an extended Woodworth model, the prediction errors remain large if the frequency range does not comply with the model specifications. Tuned to the previously inaccurately modelled frequency range between 500 Hz and 1.5 kHz, we propose a hybrid multi-harmonic model to predict the ITDs in HRTFs and HARTFs for arbitrary directions in the horizontal plane with superior accuracy. The target model coefficients are derived from individual directional measurements of 30 adults, wearing two dual-microphone behind-the-ear HAs and two in-ear microphones. Model individualisation is facilitated by the availability of polynomial weights that are applied to subsets of individual anthropometric and HA features to estimate the target model coefficients. The model is published as part of the Auditory Modeling Toolbox (AMT, pausch2022) and supplemented with the individual features and directional datasets.

Keywords: ITD model, Hearing aids, Adult anthropometrics, Binaural technology, Virtual acoustics

1 Introduction

Implementation of conventional binaural synthesis requires knowledge of the sound field, generated by a free-field sound source, at the ear-canal entrance [1]. The spatial filtering of the sound field by the individual anatomy of a listener is described by HRTFs [2]. When wearing HAs, the sound field is captured by the microphones of an HA, with the corresponding spatial filters referred to as HARTFs or by similar terms [3–6]. It is well understood from systematic investigations that the HA-device style and the microphone positions have considerable effects on HARTFs and directional hearing [7–9]. In part, a reduced localisation performance in the horizontal plane when using behind-the-ear receiver-in-ear HAs can be attributed to modified binaural cues, and limited access to ITDs owing to the reduced bandwidth of the HA receivers [10].

Previous work demonstrated that HRTFs can be decomposed into a minimum-phase and an excess-phase component, the latter consisting of a linear-phase and an allpass term [11, 12]. Although some authors suggested the possibility to neglect the allpass term [13, 14], there is perceptual evidence that this simplification may be audible, while perceptually indistinguishable results can be achieved when replacing the allpass term by its interaural group-delay difference evaluated at 0 Hz [15]. The application of the simple harmonic analytic models by Woodworth [16] and Kuhn [17] to HRTF datasets allows to predict the ITDs in the horizontal plane, representing the linear-phase term when considered as mere time delays. In such models, the head is approximated by an ideal three-concentric sphere, whose radius can be estimated by a weighted linear combination of anthropometric features1 [18, 19], with feature-less ears arranged antipodally on the equator [16, 17]. Alternatively, the head geometry can be approximated by an ellipsoid [20, 21]. Numerical models [22–25] additionally

*Corresponding author: florian.pausch@akustik.rwth-aachen.de

1 In this article, the term features refers to both individual anthropometric features, and features related to the individual HA placement.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
allow to include refined pinna geometries, head asymmetries, and off-centre ear canals. The latter aspect has been partially addressed in an extended variant of Woodworth’s model, considering horizontal ear-canal angles larger than 90 deg [26], which also inspired a more recent geometric head model [27]. While the application of the two mentioned simple harmonic analytic models will lead to large errors, such refined models are likely better suited for estimating the ITDs in behind-the-ear HARTFs, since the HA microphone positions even further violate the assumptions of antipodally arranged reception points at the lateral head centres.

The development of ITD models typically relies on a reasonably large pool of acoustic measurement data and a collection of individual features [18, 21]. However, publicly available individual HARTF databases are rather sparse, cf. first to third columns of Table 1. Such databases provide a valuable basis for comparisons across HA devices, to highlight specific features and acoustic characteristics, and their variations between participants [28]. Further benefits extend to the development of acoustically transparent HAs [5] and individualised HA algorithms [29]. Individualisation of generic receiver directivities using individual anthropometrics will likely be more effective for HA users when based on ITD models derived from HARTF instead of HRTF datasets [30]. In binaural HAs, it is crucial to preserve the ITDs, which becomes a challenging task in presence of unwanted side effects of HA algorithms [31, 32]. In this context, the application of ITD models could serve as individual ITD reference data to continuously improve adaptive HA algorithms [33] and increase their robustness against physical imperfections, such as gain, position and phase errors in the HA microphone arrays [34], or detrimental acoustic conditions inflicted by noise sources or reverberation [35, 36].

Limiting factors of existing databases include low spatial resolution and selective angular coverage. This restricts their application to objective evaluations and perceptual experiments evaluating metrics for a discrete set of directions or a combination thereof with a static listener. Nonetheless, the development of direction-continuous ITD models requires spatially dense measurement data [24]. Similar demands apply to dynamic binaural real-time reproduction of virtual acoustic environments, where the listener’s unrestricted head movements are captured by a motion tracking system. Consequently, the auralisation system needs to continuously update the spatial transfer functions to maintain the position of a virtual sound source for a plausible listening experience [6, 37]. In such dynamic scenarios, the spatial resolution ideally lies below the minimum audible angle [38, 39]. If the binaural signals include simulated room acoustics a less dense grid of 2 deg × 2 deg is perceptually sufficient for auralising the direction-dependent reflection patterns at the listener position [40, 41].

In order to approach the target criteria of the aforementioned application areas, we specifically created a database of individual directional transfer functions. The transfer functions were measured at the front and rear microphones of two behind-the-ear HA prototypes, while attached to adult participants, and in-ear microphones at the blocked ear-canal entrances (IHTA-indHARTF [42]), cf. last column of Table 1. The measurements cover directions that are distributed on an equiangular 2.5 deg × 2.5 deg grid in azimuth and elevation, including elevation angles down to −70 deg. Based on the acoustic data obtained for equatorial directions, a hybrid multi-harmonic model (AMT: pausch2022 [43]) to predict the ITDs in the horizontal plane is proposed for each type of dataset [43], that is, for HRTFs, and front and rear HARTFs. The model is tuned to a frequency range between 500 Hz and 1.5 kHz, accounting for the range of highest sensitivity to ITDs between 700 Hz and 1.4 kHz [44], which is not covered with sufficient accuracy by Kuhn’s simple harmonic analytic model [17] and the extended Woodworth model [26]. To facilitate application to individualisation, we calculated optimal polynomial-regression weights to be applied on dataset-specific subsets of anthropometric and HA features for estimating the model coefficients. It is demonstrated to what extent the prediction performance of Kuhn’s analytic model is affected due to a violation of the specified microphone placement and the frequency range and how well the former can be intercepted by the extended Woodworth model. All three models are evaluated in terms of prediction errors, and estimation accuracy of maximum ITD values and arguments of the ITD maxima.

The remainder of this article is structured as follows: The first part ( Sect. 2) focuses on the creation of the acoustic database, including information on the study participants (Sect. 2.1), as well as the measurement apparatus (Sect. 2.2) and procedure (Sect. 2.3). After introducing the ITD-estimation method, the ITD characteristics of the measured acoustic datasets are analysed across participants, also addressing measurement reliability (Sect. 2.4). The procedure and repeatability of the manual feature-extraction approach are described and complemented by a correlation analysis (Sect. 2.5). Building upon the acoustic database and specific feature subsets, we present the rationale and design steps that led to the hybrid multi-harmonic model in the second part (Sect. 3). The results on model evaluation (Sect. 4) include assessment of structural model complexity (Sect. 4.1), comparative broadband and frequency-selective performance analyses (Sect. 4.2), and model-robustness analysis (Sect. 4.3). Finally, the main findings are discussed in consideration of perceptual implications, possible future evaluations, and limitations (Sect. 5), ending with conclusions (Sect. 6).

2 Database creation
2.1 Participants

We recruited N = 30 participants (8 female, 22 male) aged 28.9 ± 6.1 (μ ± σ, range: 22–43 yrs). Prior to the experiment, all participants were informed of their rights, the study purpose and the contents of the database.
Table 1. Comparison of previously measured individual HARTF databases with the presented database.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants and artificial heads</td>
<td>16 (6 f, 10 m), 3 artificial heads (HATS, KEMAR, 1 custom-made)</td>
<td>25 (11 f, 14 m), 1 artificial head (KEMAR)</td>
<td>41 (14 f, 27 m), 4 artificial heads (1 HATS measured twice, 3 custom-made)</td>
<td>30 (8 f, 22 m)</td>
</tr>
<tr>
<td>Age of participants (yrs)</td>
<td>20–36</td>
<td>21–60 (μ ± σ: 30.4 ± 7.3)</td>
<td>25–60 (μ ± σ: 41.6 ± 10.9)</td>
<td>22–43 (μ ± σ: 28.9 ± 6.1)</td>
</tr>
<tr>
<td>HA-device styles</td>
<td>3 ITE, 1 BTE</td>
<td>1 ITE</td>
<td>1 BTE</td>
<td>1 BTE</td>
</tr>
<tr>
<td>Number of microphones per HA device</td>
<td>ITE 1: 3; ITE 2: 2; ITE 3: 3; BTE: 3</td>
<td>Eardrum: 1; blocked ear-canal entrance: 1; insert headphones: 1</td>
<td>Blocked ear-canal entrance: 1</td>
<td>Blocked ear-canal entrance: 1</td>
</tr>
<tr>
<td>Placement and number of other microphones per ear</td>
<td>Eardrum: 1; blocked ear-canal entrance: 1; insert headphones: 1</td>
<td>Eardrum: 1; blocked ear-canal entrance: 1; insert headphones: 1</td>
<td>Blocked ear-canal entrance: 1</td>
<td>Blocked ear-canal entrance: 1</td>
</tr>
<tr>
<td>Sampling rate (kHz)</td>
<td>48</td>
<td>44.1</td>
<td>44.1</td>
<td>44.1</td>
</tr>
<tr>
<td>Filter length (samples)</td>
<td>256</td>
<td>356</td>
<td>472</td>
<td>256</td>
</tr>
<tr>
<td>Measurement directions</td>
<td>91 directions in total: 48 directions with 7.5 deg spatial resolution in the horizontal plane; 24 and 6 directions lying on horizontal rings at ±30 deg and 60 deg elevation with 30 deg and 60 deg azimuth resolution, respectively; 1 direction at the north pole; 8 additional directions in the median plane to obtain a vertical spatial resolution of 10 deg between ±30 deg; 2 additional sagittal plane directions each at azimuth 30 deg and 150 deg and ±15 deg elevation (4 in total).</td>
<td>87 directions in total: 48 directions with 7.5 deg spatial resolution in the horizontal plane; 24 and 6 directions lying on horizontal rings at ±30 deg and 60 deg elevation with 30 deg and 60 deg azimuth resolution, respectively; 1 direction at the north pole; 8 additional directions in the median plane to obtain a vertical spatial resolution of 15 deg above −30 deg elevation.</td>
<td>85 directions in total: 16 directions with 22.5 deg azimuth resolution in the horizontal plane; two horizontal rings at ±45 deg, each with 6 directions and 60 deg azimuth resolution; 1 direction at the north pole; 3 azimuth orientations of the participant at 0 deg, −15 deg, −30 deg, resulting in azimuth resolutions of 7.5 deg in the horizontal plane, and 30 deg in the two other horizontal rings.</td>
<td>9217 directions in total: Equiangular sampling with a spatial resolution of 2.5 deg × 2.5 deg in azimuth and elevation, covering all-round directions between elevation angles above and below −70 deg and 90 deg, respectively.</td>
</tr>
<tr>
<td>Measurement radius (m)</td>
<td>2.5–3</td>
<td>2.5–3</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Measurement environment</td>
<td>Anechoic chamber</td>
<td>Anechoic chamber</td>
<td>Acoustically treated room (250 ms mean reverberation time)</td>
<td>Hemi-anechoic chamber</td>
</tr>
<tr>
<td>Data formats</td>
<td>Custom MAT, SOFA [46]</td>
<td>Custom MAT, SOFA [46]</td>
<td>Custom MAT</td>
<td>OpenDAFF [47], SOFA [46]</td>
</tr>
<tr>
<td>License</td>
<td>CC BY-NC-SA 4.0</td>
<td>CC BY-SA 4.0</td>
<td>CC BY 4.0</td>
<td>CC BY-SA 4.0</td>
</tr>
</tbody>
</table>

Note: This table makes no claim to completeness. μ and σ represent mean and standard deviation, respectively. Participants are separated into female (f) and male (m). The listed HA-device styles include in-the-ear (ITE) and behind-the-ear (BTE) devices. Two commonly used artificial heads are referred to with the abbreviations HATS (TYPE 4128-C Head and Torso Simulator, Bruel & Kjær, Nærum, Denmark) and KEMAR (45BB Head & Torso, G.R.A.S. Sound & Vibration A/S, Holte, Denmark).
They subsequently provided consent to publish the collected data in a pseudonymous way. The personal data and experimental results were processed and archived in accordance to country-specific and European data-protection regulations [45].

2.2 Apparatus

All acoustic measurements were performed in a hemi-anechoic chamber with the dimensions 11 m × 5.97 m × 4.5 m (length × width × height) and a lower frequency limit of about 100 Hz [48].

The loudspeaker array consisted of 64 1-inch full-range loudspeakers (W1-2025SA, Tang Band, Taipei, Taiwan) which were built into identically designed closed 0.05 2 array segments, resulting in a frequency range of about 500 Hz to 18 kHz [49]. Covering elevation angles2 between 88.75 deg and −70 deg at a radius of 1.2 m, the loudspeakers were arranged in steps of 2.52 deg. A remote-controlled motor allowed to rotate the loudspeaker array around the participant for continuous HRTF measurements [50].

For the HARTF measurements, we used a pair of custom-made behind-the-ear receiver-in-ear HAs (GN ReSound, Ballerup, Denmark) without digitalsignalprocessor and with free access to the raw microphone signals [6]. Each device contained two omnidirectional micro-electro-mechanical-systems microphones (Knowles, Itasca, IL, USA), see Figure 1. The microphone signals were transmitted via slim cables with a diameter of 1.4 mm (Hi-Pro cable, Sonion, Roskilde, Denmark). In the current measurement, the receiver cables without an ear piece attached were used to position the devices behind the participant’s ear with negligible acoustic influence.

Conventional HRTFs were measured simultaneously for comparison purposes using in-ear microphones (KE3, Sennheiser, Wedemark, Germany), which were installed in power domes and placed at the ear-canal entrances. All cables were taped to the sides of the neck and to the participant’s clothing to ensure that both the domes and research HAs remain in place throughout the measurement, see Figure 4.

Based on the method of multiple exponential sweeps [51, 52], the measurement signal was generated between 20 Hz and 20 kHz at a sampling rate of 44.1 kHz with an individual sweep length of 215 samples using MATLAB™ (The MathWorks, Inc., Natick, MA, USA) and the ITA-Toolbox [53]. The measurement signal was transmitted via MADI protocol (RME HDSPe MADI, Audio AG, Haimhausen, Germany), DA-converted and amplified by two 32-channel power amplifiers with MADI interface (MA 32 D, KS Audio Research GmbH, Hettenleidelheim, Germany), and played back via the array loudspeakers. The captured signals were AD-converted (RME Octamic II, Audio AG, Haimhausen, Germany) and sent to the audio interface (RME Hammerfall DSP Multiface II, Audio AG, Haimhausen, Germany) via ADAT protocol for further processing in MATLAB™ [53].

Four infrared cameras (Flex 13, NaturalPoint Inc., Corvallis, Oregon, USA) were installed at the ceiling to capture the poses, i.e. the positions and orientations, of the rigid bodies attached to the participant and the loudspeaker array during the acoustic measurement. Synchronised recording and post-processing of the movement trajectories was facilitated using the software Motive (NaturalPoint Inc., Corvallis, Oregon, USA) and a dedicated MATLAB™ interface [53]. A 20-inch monitor (SyncMaster 205BW, Samsung, Seoul, South Korea) at 2.3 m distance in viewing direction provided real-time feedback on the participant’s pose. At the bottom of the displayed graphical user interface, see Figure 2, six boxes showed the errors by degrees of freedom relative to the target pose, with colour coding and numerical values in units of cm and deg for position and orientation, respectively. To visualise the current pose, a dynamic black cross-hair was displayed in the centre of the screen and contrasted with a static green cross-hair. Additional arrows indicated the suggested corrective
motions for alignment [49]. A tolerance range of ±1 cm and ±0.8 deg for position and orientation offset, respectively, defined the optimal pose range. When drifting beyond these thresholds, the smileys changed, while both frames and text turned yellow, and red for pose errors larger than 3.5 cm and 1.5 deg. In comparison, the feedback system used by Denk et al. [54] prescribed similar thresholds, with slightly stricter limits of ±0.6 deg for the yaw angle. More lenient thresholds were used in the orientation-only feedback system by Begault et al. [55], with lamps indicating the current orientation, discretised in steps of 1 deg.

Figure 3 shows the alignment errors averaged over time and participants, split into position and orientation components. While the black boxplots represent absolute median errors with respect to the target pose, the grey boxplots depict relative pose errors referenced to the individual mean pose. In addition to the intrinsic orientation-error components, the error with respect to the great-circle central angle shows the absolute angular error between ideal and actual loudspeaker directions given a misoriented participant. In the following, median errors with interquartile ranges (IQRs) are reported.

The relative errors are predominantly within the predefined tolerance ranges. This holds true for the relative orientation error, with a roll error of −0.1 deg (IQR: [−0.4, 0.3] deg), a pitch error of 0.1 deg (IQR: [−0.4, 0.6] deg), and a yaw error of 0.1 deg (IQR: [−0.3, 0.5] deg). The relative position errors amounted to offsets of 0 (IQR: [−0.3, 0.3] cm) in $x$-direction, 0 (IQR: [0, 0] cm) in $y$-direction, and 0 (IQR: [−0.5, 0.5] cm) in $z$-direction. Thereby, the feedback system effectively helped avoid drifting off the target pose over time, as typically observed in measurements without feedback [49], even if a fixed viewpoint is provided [56]. A larger absolute angular error of 1.5 deg (IQR: [0.9, 2.2] deg) is primarily caused by a lateral-position offset of 1.1 cm (IQR: [−0.6, 3.5]) cm in $x$-direction, and a front-back-position offset of 0.2 cm (IQR: [−0.7, 1.2] cm) in $z$-direction. Smaller absolute errors of approximately 0.3 deg within a range of ±0.5 deg are achievable when measuring seated participants with an additional headrest [54]. However, no interfering reflections from the thighs and a head rest are contained in the data using the presented measurement routine, and misalignment is compensated during post-processing, cf. Section 2.3.2.

2.3 Acoustic measurements

2.3.1 Procedure

Free-field reference measurements were performed in absence of the participants with the in-ear microphones without attached power domes positioned in the centre of the loudspeaker array. Similarly, the research HAs were oriented at a typical carrying angle and mounted in a way that the centre of the HA microphones and their connection axis corresponded with the participant’s target pose, i.e. the array centre with orientation in nominal viewing direction.

As part of the preparations of the individual receiver-directivity measurements, the interaural and longitudinal axes of the standing participants were adjusted using two cross-line lasers to coincide with the centre of the loudspeaker array, see Figure 4. To make the compensatory movements suggested by the feedback system more intuitive, the offset of the head-mounted rigid body to the centre of the interaural axis was corrected individually [49]. In a training session, the participants were familiarised with the graphical user interface and their task to realign with the presented target pose. Individual receiver-directivity measurements were subsequently conducted based on the method proposed by Richter and Fels [50], favouring minimal participant movement by rotating the loudspeaker array around the ideally static participant. Before the actual start of the measurement, the participants had 20 s to align themselves with the target pose. With a pre-defined angular speed of 1.3 deg/s after completed acceleration, a full rotation required 4.5 min, resulting in an azimuthal resolution of approximately 2.49 deg.
2.3.2 Post-processing

To eliminate unwanted reflections in the impulse responses after deconvolution [51, 52], an adaptive time window was applied on each of the raw binaural datasets per participant. After direction-dependent determination of the latest impulse-response start instances for all microphone signals, a right-sided Hann window (1 ms fade-out duration, starting 2 ms after the latest start instance) was used. 5.8 ms before the end of the fade-out, we applied a left-sided Hann window with 1-ms fade-in duration. We used the same time window on both the raw binaural impulse responses and the reference measurements. Subsequent cropping resulted in a final impulse-response length of 256 samples at a sampling rate of 44.1 kHz. The windowed spatial transfer functions were divided by the corresponding reference spectra, applying spectral regularisation [57] and band limitation between 100 Hz and 20 kHz. Due to the lack of energy below the lower cut-off frequency of the loudspeakers, the magnitude spectra were extrapolated towards 0 dB at 0 Hz for frequencies below 500 Hz for physical reasons [2]. Causality of the resulting impulse responses was restored by setting a pre-delay of 0.5 ms in all spatial transfer functions, while preserving binaural cues.

Further necessary post-processing steps [49] included a directional mapping of spatial transfer function via rotation matrices, accounting for time-dependent loudspeaker positions, and time-dependent participant poses relative to the nominal array centre and the loudspeakers. The second step in optimising the directional mapping involved an azimuthal rotation of the global coordinate system so that the zero-crossings of the individual ITDs, estimated as per the method described in Section 2.4 for all directions in the horizontal plane, were approximately zero at 0 deg and 180 deg azimuth [21]. This was necessary to compensate for a remaining azimuth offset of 0.4 ± 3.6 deg (μ ± σ), averaged across participants, between the coordinate system of the optical tracking system, and the coordinate-system reference set by the mechanical switch of the motor controlling the loudspeaker array. Finally, the estimated directions of incidence, distributed on a frequency-dependent non-uniform grid, were transformed to a frequency-independent equiangular grid with a resolution of 2.5° × 2.5° in azimuth and elevation angles, allowing easier data handling. For spatial interpolation, we estimated the complex spectra of the spatial transfer functions for directions of incidence within the convex hull of input points using a natural-neighbour interpolation of the real and imaginary parts separately [58].

Due to issues with local electrical crosstalk in the loudspeaker array, the measurement results at elevation angles of 75 deg and 10 deg had to be excluded prior to spatial interpolation during post-processing. Since it was not possible to entirely restore fine spatial details in the eliminated directions, the introduced errors were examined by comparing the spectra at \( \Theta = 45^\circ \) to their spatially interpolated counterparts, revealing a loss of predominantly high-frequency components. However, applying an adapted version of the inter-subject spectral difference [59] revealed a smaller error between the approximated and the original spectra, compared to the error between the original spectra and the spectra of neighbouring points at \( \Theta = 47.5^\circ \).

2.4 Interaural time differences

2.4.1 Estimation method

We define the ITD as the difference in times of arrival of a sound wave between the right and left ear, resulting in positive values for \( 0^\circ \leq \phi \leq 180^\circ \). The ITD characteristics of the HARTF datasets and their differences to HRTFs were analysed for directions in the horizontal plane, accounting for variations across participants. To avoid ITD misestimations due to noise, the post-processed measurements were band-limited using a low-pass filter with a cut-off frequency of 1.6 kHz, implemented as 10th order low-pass digital Butterworth filter. Direction-dependent ITDs were subsequently evaluated per participant and dataset in a frequency range between the lower loudspeaker cut-off frequency of 500 Hz, and an upper frequency limit of 1.5 kHz [15]. For reasons of increased robustness in the intended modelling-frequency range, we decided to evaluate the maximum of the interaural cross-correlation of energy envelopes (MaxIACC) instead of applying threshold-based methods, accepting slight sacrifices to perceptual relevance [60]. The resulting ITD-curve progressions allowed to derive trends in the overall shape, maximum ITD values, max{ITD}, and the arguments of the ITD maxima, arg max{ITD}, and were used for modelling purposes, see Section 3.

2.4.2 Measurement-based interaural time differences

In Figure 5, the mean measurement-based ITD estimations with standard deviations, averaged across participants, are shown per type of dataset for source directions in the horizontal plane.

A considerable variation in curve progressions and ITD maxima was observed across participants, on average amounting to 724.6 ± 24.6 μs (μ ± σ, range: 684.8–780.1 μs), 691.3 ± 19.8 μs (μ ± σ, range: 653.1–734.7 μs), and 674.1 ± 20.7 μs (μ ± σ, range: 630.4–721.1 μs) in HRTF, and front and rear HARTF datasets, respectively. The average maximum ITD values in HRTF datasets were found to lie between the ones reported for the CIPIC database (μ ± σ = 646 ± 33 μs) [19] and the KEMAR dataset (≈790 μs) [61]. The general decrease of ITD maxima in HARTF datasets can be explained by the resulting shorter path lengths due to the HA-microphone positions further off the head centre and increased horizontal HA-microphones angles \( \Theta_{HA} \) compared to the smaller horizontal ear angles \( \Theta_e \), see Figure 6 and Table 2.

3 Robust broadband ITD estimations were achieved except for ITD outliers in the HRTF dataset of participant 8 at azimuth angles between 42.5 deg and 50 deg. These ITD outliers were replaced by interpolated values using a piecewise-cubic Hermite-interpolating polynomial (function pchip) in MATLAB\textsuperscript{TM}. 
In HRTFs, the mean argument of the ITD maxima was estimated at 86.1 ± 2.6 deg (l ± r, range: 80–90 deg). For the front and rear HARTFs datasets, angularly shifted results of 81.3 ± 2.8 deg (l ± r, range: 77.5–87.5 deg) and 79.3 ± 2.3 deg (l ± r, range: 75–85 deg), respectively, can be reported. Reasons for deviations in terms of the arguments of the ITD maxima can be largely attributed to the deviating horizontal ear-canal and HA-microphones angle, HE and HA, respectively. Further analyses of ITD characteristics are provided in Section 4.2.

2.4.3 Reproducibility of measurement results

To examine the impact of slight variations in HA positioning on measured ITDs and overall reproducibility of the measurement results, one participant was measured acoustically four times (measurements I–IV), after repositioning the HA devices between runs. While the ear canals were blocked by the in-ear microphones used for the HRTF measurements in measurements I–III, measurement IV was conducted without in-ear microphones to investigate any potential effects of the closed ear canal on HARTFs, compared to an open one.

We generally obtained very similar ITD-curve progressions for directions in the horizontal plane. The mean deviations of 0.3 ± 6.5 μs (μ ± σ), 1.6 ± 7.7 μs (μ ± σ), and 1.2 ± 6.0 μs (μ ± σ) for HRTFs, front and rear HARTFs, respectively, are below the just-noticeable differences (JNDs) reported in Section 5.2.1, with a maximum error of ±31.7 μs. The different measurements further entailed consistent progressions in magnitude spectra with spectral differences also being for the most part below reported JND values [50]. Slight narrowband overshoots between 5 kHz and 7 kHz were only observed for HARTFs, and for frequencies below 1.5 kHz in all datasets. Most likely, the reason for the spectral similarity lies in the fact that the HA devices are placed behind the ears, with the HA microphones being shadowed by the pinna and therefore not being susceptible to wave interaction effects.

Measurement IV did not reveal any prominent deviations neither in the resulting ITDs nor in the magnitude spectra for frequencies above about 4 kHz [63], where monaural cues are to be expected, ruling out an effect of the ear-canal blockage.

2.5 Feature acquisition

2.5.1 Methods and definitions

For the extraction of individual features related to the body, head, pinna and HA placement, frontal and profile photos of each participant were taken. For this purpose, we mounted a ruler on a stand, next to which the participants stood. To minimise errors due to mismatched optical planes, the longitudinal axis of the participant was aligned with the ruler and the camera’s nominal line of sight using frontal and lateral crossline lasers. The centre of the HAMicrophone positions was additionally marked with white tape at the respective ear to be visible in the profile photos. The camera (X-Pro2 with XF35mmF2 R WR FUJINON, Fujifilm, Minato, Tokyo, Japan) with a resolution of 6000 pixel × 4000 pixel was placed at a distance of about...
3 m to capture the frontal view, left and right profile, and the seating position. We also captured detail views from each of the participant’s ears at about 50 cm distance, including a ruler for scaling. Inherent lens distortions and perspective distortions were minimised using image editing software (Capture One 21 Fuji fi lm v14, Phase One, Copenhagen, Denmark).

The large part of features was extracted manually [64] and includes available definitions [19] and new ones, specifying the individual placement of the HA devices, see Figure 6 and Table 2. Additional features were calculated using the geometrical formulas provided in Appendix A. The final set can be subdivided into head- and body-related features ($x_1$ to $x_{17}$), ear-related features ($d_1$ to $d_8$, $\Theta_1$, $\Theta_2$, and $\Theta_E$), and HA-related features ($d_9$ to $d_{14}$, $\Theta_3$, and $\Theta_{HA}$). Two different feature subsets were defined for the specific ITD-model implementation, see Section 3.4, to predict individual ITDs in HRTF and HARTF datasets.

**2.5.2 Correlation analysis**

Table 2 lists all measured and calculated features with means $\mu$, standard deviations $\sigma$ and percentage deviations $\chi$. The results for most features, averaged across ears, are in the range (±10%) of those reported by Algazi et al. [19], except for larger deviations in neck height $x_7$ (−47.6%), neck depth $x_8$ (18.3%), torso-top width $x_{10}$ (30.2%), seated height $x_{15}$ (34.2%), cymba-conchae height $d_2$ (13.6%, averaged across ears), fossa height $d_4$ (11.4%, averaged across ears), cavum-conchae Table 2. Statistical summary of the results on measured and calculated anthropometric and HA features.

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\chi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$ Head width</td>
<td>156.8</td>
<td>9.1</td>
<td>11.6</td>
</tr>
<tr>
<td>$x_2$ Head height</td>
<td>227.7</td>
<td>11.7</td>
<td>10.3</td>
</tr>
<tr>
<td>$x_3$ Head depth</td>
<td>199.8</td>
<td>10.4</td>
<td>10.4</td>
</tr>
<tr>
<td>$x_4$ Pinna offset down</td>
<td>15.3 (16.6)</td>
<td>6.5 (7)</td>
<td>84.6* (83.9*)</td>
</tr>
<tr>
<td>$x_5$ Pinna offset back</td>
<td>3.2 (0.5)</td>
<td>8.7 (7.1)</td>
<td>539.6* (2765)</td>
</tr>
<tr>
<td>$x_6$ Neck width</td>
<td>120.1</td>
<td>12.3</td>
<td>20.5</td>
</tr>
<tr>
<td>$x_7$ Neck height</td>
<td>42.4</td>
<td>13.8</td>
<td>65.1</td>
</tr>
<tr>
<td>$x_8$ Neck depth</td>
<td>128.7</td>
<td>20</td>
<td>31.1</td>
</tr>
<tr>
<td>$x_9$ Torso-top width</td>
<td>321.4</td>
<td>28.3</td>
<td>17.6</td>
</tr>
<tr>
<td>$x_{10}$ Torso-top height</td>
<td>192.2</td>
<td>25.4</td>
<td>26.4</td>
</tr>
<tr>
<td>$x_{11}$ Torso-top depth</td>
<td>260</td>
<td>35</td>
<td>26.9</td>
</tr>
<tr>
<td>$x_{12}$ Shoulder width</td>
<td>477.2</td>
<td>42.4</td>
<td>17.9</td>
</tr>
<tr>
<td>$x_{13}$ Head offset forward</td>
<td>−8.8</td>
<td>26.5</td>
<td>−600.6*</td>
</tr>
<tr>
<td>$x_{14}$ Height</td>
<td>1773.2</td>
<td>100</td>
<td>11.3</td>
</tr>
<tr>
<td>$x_{15}$ Seated height</td>
<td>1350.4</td>
<td>42.6</td>
<td>6.3</td>
</tr>
<tr>
<td>$x_{16}$ Head circumference</td>
<td>502.4</td>
<td>22.5</td>
<td>8</td>
</tr>
<tr>
<td>$x_{17}$ Shoulder circumference</td>
<td>1117.5</td>
<td>96.6</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Note: Distances and angles are provided in mm and deg, respectively. Mean and standard deviation are represented by $\mu$ and $\sigma$, respectively, and $\chi$ symbolises the percentage deviation 100 · $(2\sigma/\mu)$. Values in brackets for $x_4$ and $x_5$ represent the results for the right ear. Asterisks denote non-meaningful percentage deviations.
Figure 7. Visualisation of the upper correlation matrix between the features in Table 2, showing bivariate correlations that are significant at the 95% confidence level.

depth \( d_6 \) (91.5%, averaged across ears), and pinna-rotation angle \( \Theta_E \) (57.9%, averaged across ears), ignoring the non-meaningful deviations in pinna offset down \( x_1 \), pinna offset back \( x_5 \) and head offset forward \( x_{3,3} \). Possible reasons for the observed deviations include measurement uncertainties, and differences in sample-inherent characteristics such as age (CIPIC: 25.4 ± 8.6 yrs, range: 18–63 yrs), weight (CIPIC: 70.9 ± 14.8 kg, range: 39.5–117.9 kg; not available in the current database) and gender distribution (CIPIC: 16 f, 27 m). Exclusion of the aforementioned non-meaningful features (as well as HA-microphones offset up \( d_{1,3} \) and back \( d_{1,4} \)) resulted in average percentage deviations of 20.1 ± 15% \((\mu \pm \sigma)\) regarding head- and body-related features, and 36.9 ± 21.2% \((\mu \pm \sigma)\) and 34.9 ± 21.8% \((\mu \pm \sigma)\) regarding ear- and HA-related features of the left and right ear, respectively.

To identify interrelationships between features, averaged across ears, Figure 7 visualises the upper correlation matrix in hierarchical clustering order [65], only showing bivariate correlations that were significant at the 95% confidence level. No significant correlations were found for intertragal-incisure width \( d_2 \) and HA-microphones-to-scalp offset \( d_{1,0} \), whereas torso-top height \( x_{10} \) only resulted in an inconclusive significant correlation with \( x_{2,1} (\rho = .43) \), and was therefore removed. The most relevant of the top forty correlations are summarised below, largely in descending order of magnitudes in bivariate correlation coefficients \( \rho \).

Starting with ear-related and HA-related features, some of the strongest but also most obvious correlations were found between pinna offset back \( x_5 \) and horizontal ear-canal angle \( \Theta_E \) \((\rho \approx 1)\) as well as horizontal HA-microphones angle \( \Theta_{HA} \) \((\rho = .9)\), and between pinna-rotation angle \( \Theta_1 \) and frontal HA-microphones angle \( \Theta_3 \) \((\rho = .96)\). For obvious reasons, the horizontal HA-microphones angle \( \Theta_{HA} \) was strongly correlated with HA-microphones offset back \( d_{1,4} (\rho = .99) \), pinna offset back \( x_5 (\rho = .9) \), and horizontal ear-canal angle \( \Theta_E (\rho = .9) \). So were cymba-conchae height \( d_2 \) and pinna height \( d_6 (\rho = .77) \), as well as fossa height \( d_4 \) and pinna height \( d_6 (\rho = .73) \). Moderate correlations were present between cavum-conchae width \( d_5 \) and pinna width \( d_6 (\rho = .69) \). We additionally report strong negative correlations for HA-microphones offset up \( d_{1,1} \) with pinna-rotation angle \( \Theta_1 (\rho = -.75) \) and frontal HA-microphones angle \( \Theta_3 (\rho = -.7) \). No other meaningful negative correlations were present within the considered range of correlation magnitudes. Fossa height \( d_4 \) was naturally related to the HA behind-pinna depth \( d_{1,1} \), represented by a moderate correlation coefficient of \( \rho = .69 \).

The HA-microphones-to-ear-canal offset \( d_5 \) appeared to be strongly correlated with HA-microphones offset up \( d_{1,1} (\rho = .75) \), and moderately correlated with multiple ear dimensions, that is, pinna height \( d_6 (\rho = .61) \), cymba-conchae height \( d_2 (\rho = .61) \), pinna width \( d_6 (\rho = .6) \), and fossa height \( d_4 (\rho = .59) \). Cymba-conchae height \( d_2 \) was additionally moderately correlated with fossa height \( d_4 (\rho = .56) \) and pinna width \( d_6 (\rho = .51) \).

Regarding head- and body-related features, strong and plausible correlations were found between torso-top width \( x_9 \) and shoulder circumference \( x_{1,7} (\rho = .86) \), shoulder width \( x_{3,2} (\rho = .84) \), and head width \( x_1 (\rho = .77) \). Shoulder width \( x_{3,2} \) was strongly correlated with shoulder circumference \( x_{1,7} (\rho = .82) \) and moderately correlated with torso-top width \( x_{1,1} (\rho = .61) \) and neck width \( x_{10} (\rho = .57) \). Moderate correlations were also observed between head height \( x_2 \) and height \( x_{1,4} (\rho = .68) \), and head depth \( x_{3,1} (\rho = .67) \), the latter also showing moderate correlations with head circumference \( x_{1,6} (\rho = .59) \) and head width \( x_1 (\rho = .58) \).

2.5.3 Variability in feature extraction

To test the manual feature-acquisition approach for reproducibility, six participants extracted the feature subsets relevant for ITD modelling, cf. Section 3.4, from the photos of participant 18, one of whom performed two additional extractions. This approach allows to analyse inter- and intra-individual variations, which are summarised in Table 3. For head width \( x_3 \), head height \( x_1 \), pinna offset down \( x_5 \), and calculated HA-microphones offset up \( d_{1,3} \), the intra-individual extractions entailed smaller standard and percentage deviations than the inter-individual results. For the remaining features, both variations were in a comparable range, see Table 2. The effects of these variations on the ITDs predicted by Model 3 are analysed in Section 4.3.

3 Modelling interaural time differences

This section starts with a summary of the key aspects of the analytic ITD models by Kuhn [17], and Woodworth [16, 66] (in its extended version [26]). To address their shortcomings in view of the inherent ITD characteristics of HARTFs, see Section 2.4.2, and the mismatched target frequency range, a hybrid multi-harmonic ITD model is introduced, allowing to predict the ITDs in the horizontal
plane. In an intermediate step, the target coefficients of all three models are derived by fitting the ITD curves in the horizontal plane, the estimations of which are based on the corresponding measured directional datasets of the database presented in Section 2. Model individualisation is made possible by estimating these target model coefficients by application of weighted individual feature subsets. The method to obtain feasible weights, the specific model implementations using weighted individual feature subsets, and metrics and methods used for a comparative model-performance evaluation are introduced. All modelling steps and analyses were performed using R [67], RStudio [68] and MATLAB™ [53].

3.1 Modelling rationale

Kuhn’s analytic symmetric-sphere model [17], referred to as Model 1 with corresponding values of

\[ \text{ITD}_{\text{Model 1}} = \frac{3a_n}{c} \sin(\phi), \]

(1)

relies on the sphere-equivalent effective head radius \( a_n \) of the \( n \)th participant \((n = 1, \ldots, N)\) with horizontal ear-canal angles of \( \Theta_E = 90^\circ \). Model 1 is valid to predict the ITDs in the horizontal plane in HRTF data for low frequencies \( f \), i.e., \((ka_n)^2 \ll 1\), with \( k = 2\pi/c \) and \( c \) representing the wave number and the speed of sound, respectively. Given an average effective head radius of approximately \( \pi \approx 87.5 \text{ mm} \) [18, 69] and \( c = 343 \text{ m/s} \), the valid range of Model 1 corresponds to frequencies much smaller than about 624 Hz.

Also based on a symmetric sphere with horizontal ear-canal angles of \( \Theta_E = \pm 90^\circ \), the analytic model by Woodworth [16, 66], referred to as Model 2, predicts values of

\[ \text{ITD}_{\text{Model 2}} = \begin{cases} \frac{a_n}{c}(\phi + \sin(\phi)), & \text{for } 0 \leq \phi \leq \pi/2 \\ \frac{a_n}{c}(\pi - \phi + \sin(\phi)), & \text{for } \pi/2 < \phi \leq \pi. \end{cases} \]

Due to the assumed left–right symmetry, the first and second angular ranges also apply to the corresponding angular sectors in the right half space relative to the median plane, using negated azimuth angles \( \phi \), and a negative \( \pi \) in the term describing the second angular range. Model 2 notably underestimates the ITDs in the low frequency range but converges to better suitability than Model 1 for \( k\pi \gg 2.4 \), which corresponds to frequencies greater than 1.5 kHz, and allows optimal results for frequencies of 4 kHz and above [26].

Model 1 and Model 2 were applied to estimate the low- and high-frequency ITDs, respectively, in a generic HARTF dataset, measured from behind-the-ear HA’s [3]. However, the ITD curves, evaluated for the front and rear HARTF datasets, see Figure 5 in Section 2.4, differ in part considerably from the ITDs estimated for the HRTF datasets. Since especially the horizontal HA-microphones angle deviates from 90 deg in the former dataset type, the assumption of a front-back-symmetric head are violated, leading to ITD curves that are asymmetric at the frontal plane. We therefore implemented the extended Woodworth model, cf. equations (b1)–(b5) and Figure 4 in [26], referred to as Model 2+ with corresponding values ITD_{Model 2+}.

Plane-wave incidence was assumed since the measurement radius of the applied measurement system, see Section 2.2, satisfied assumptions about the distal region [70]. Model 2+ accounts for arbitrary horizontal microphone angles, \( 90^\circ \leq \Theta_E \leq 180^\circ \) and \( 90^\circ \leq \Theta_{HA} \leq 180^\circ \).1

Due to the lack of a model that combines the advantages of Model 1 and Model 2+, while also addressing the ITD-curve irregularities and additional deviations from a sine wave particularly in HARTF datasets, we propose a multi-harmonic model. Model 3 (AMT: pausch2022 [43]) predicts values of

\[ \text{ITD}_{\text{Model 3}} = \frac{a_n}{c}(\gamma_{1n}\sin(\phi) + \gamma_{2n}\sin(2\phi) + \gamma_{3n}\sin(3\phi)) \]

(3)

with the model coefficients \( \gamma_{an} \in \mathbb{R} \) \((q = 1, \ldots, 3; \text{dataset type})\) for the \( n \)th participant. The model is optimised for \( 0.8 \leq k\pi \leq 2.4 \), targeting the frequency range between 500 Hz and 1500 Hz, within which the ITDs are not predicted with sufficient accuracy by Model 1 and 2(+). All modelling stages described below considered the ITD results for \( L = 73 \) azimuth angles in the horizontal plane, covering \( 0^\circ \leq \phi \leq 180^\circ \) in steps of 2.5 deg.

3.2 Estimation of effective head radii

To obtain a common basis for model comparisons, we estimated the effective by-participant head radii \( a_n \) similarly as done in [18] but using Model 1 instead of Model 2, and a different ITD-estimation method, see Section 2.4. The corresponding measurement-based ITD data was fitted in R (package n1xb from the library n1mrt [71]). Model 1 was chosen for this estimation because it better matches the evaluated frequency range used for the ITD estimation and the measurement-based ITD-curve shapes in HRTF datasets. Due to the position of the HA devices and microphones, the resulting effective head radii in the HARTF datasets were estimated by adding the individual HA-microphones-to-scalp offsets \( d_{10} \), see Figure 6, corresponding with the geometric calculations of the horizontal HA-microphones angle \( \Theta_{HA} \), cf. equation (A3).

3.3 Estimation of model coefficients

Model 1 can be readily evaluated using these estimated effective head radii. Model 2+ [26] additionally required the calculation of across-ear averages of horizontal microphone angles, \( \Theta_E \) and \( \Theta_{HA} \), see Appendix A. To obtain the coefficients of Model 3, that is, \( \gamma_{1n} \) to \( \gamma_{3n} \), the measurement-based by-participant ITD curves were similarly fitted as done for the estimation of effective head radii per type of dataset, while retaining the previously estimated effective head radii. Therefore, Model 3 can be considered a hybrid model, since it is still based on a physically motivated effective head radius but takes advantage of an empirical curve-fitting approach.

In case no individual acoustic measurement data is available, a link to the individual features, see Table 2,

\[ ^1 \text{For angles } \Theta_E \text{ and } \Theta_{HA} \text{ smaller than } 90 \text{ deg, the ITD curves predicted by Model 2+ were mirrored along the frontal plane.} \]
Both error metrics were calculated for the L model implementation described below. Applying the same strategy to Model 3 failed due to its increased complexity and a curvilinear relationship between the features and the model coefficients \(a_n\) and \(\gamma_{1n}\) to \(\gamma_{3n}\). We instead used polynomial regression (see, e.g. [72]) to estimate the effective head radii \(a_n\) in Model 1 and Model 2, and the effective head radii \(a_n\) together with the \(q\)th by-participant coefficients in Model 3, as described in Appendix B. The performance of all models was evaluated using model coefficients that were estimated by means of polynomial-regression weights applied to a subset of features, as selected for the specific model implementation described below.

3.4 Specific model implementation

An empirical selection of \(M = 5\) features (HRTF dataset: \(\{x_1, x_2, x_3, x_4, x_5\}\); front and rear HARTF datasets: \(\{x_1, x_2, x_3, d_{13}, d_{14}\}\), averaged across ears, was chosen to support the prediction of maximum ITD values [18, 19] and the shift in the arguments of the ITD maxima [26]. Further corroboration of the feature selections roots in their largely loose interrelationships: In addition to the few moderate correlations for feature combinations within each subset, reported in Section 2.5.2, only weak \(|\rho| < .5\) or non-significant correlations \((p > .05)\) were found otherwise, except for an additional moderate correlation between \(x_2\) and \(d_{13}\) \((\rho = .55)\). To find a suitable polynomial complexity level and avoid over-fitting, the polynomial degree \(P\) was varied and consequently set in a way that the respective residual standard errors in the univariate and multivariate polynomial-regression models, cf. Appendix B, became minimal, suggesting \(P = 4\).

3.5 Model-performance evaluation

Individual model-prediction performance was quantified on the basis of the mean absolute error (MAE)

\[
\text{MAE} = \frac{1}{L} \sum_{l=1}^{L} |\text{ITD}_{\text{mod}}(\phi_l) - \text{ITD}_{\text{meas}}(\phi_l)|
\]

between model-based and measurement-based ITDs, \(\text{ITD}_{\text{mod}}\) and \(\text{ITD}_{\text{meas}}\), respectively, and the root-mean-square error (RMSE)

\[
\text{RMSE} = \sqrt{\frac{1}{L} \sum_{l=1}^{L} (\text{ITD}_{\text{mod}}(\phi_l) - \text{ITD}_{\text{meas}}(\phi_l))^2}.
\]

Both error metrics were calculated for the L evaluated azimuth angles, reporting means \(\mu\) and standard deviations \(\sigma\) of results averaged across participants.

To demonstrate the effect of the individual model coefficients on MAEs and RMSEs, the base variant of Model 3, consisting of \(a_n\) and \(\gamma_{1n}\) only, was gradually increased in complexity by sequentially adding the remaining coefficients \(\gamma_{2n}\) and \(\gamma_{3n}\), and compared to Model 1. In all complexity levels of Model 3, and for Model 1, the model coefficients were estimated using polynomial regression weights with a polynomial degree of \(P = 4\) that were applied on the respective feature subsets, see Section 3.4.

For a performance comparison between all models, that is, Model 1, Model 2, and the final Model 3, the correlation of measured and modelled ITD maxima was quantified by Pearson’s correlation coefficient \(\rho\), and the variance explained was evaluated by the adjusted coefficient of determination \(R^2\), describing the predictable proportion of variance by simple linear regression. To predict \(\Delta\max\{\text{ITD}\}\) and \(\Delta\arg \max\{\text{ITD}\}\), representing the errors between model-based predictions and the corresponding measurement-based estimations of \(\max\{\text{ITD}\}\) and \(\arg \max\{\text{ITD}\}\), respectively, six independent-sample one-way robust analysis-of-variance (rANOVA) models (package \texttt{tway} from the library \textnormal{WRS2} [73]) were formulated. We entered the factor Model at the levels Model 1, Model 2, and Model 3, and ran the analyses based on trimmed means (20% trimming level) with 10,000 bootstrap samples, separately for each dataset and error metric. For post-hoc testing, linear contrasts were conducted (package \texttt{lincon} [73]), reporting adjusted \(p\) values given multiple comparisons [74]. All statistical test results were interpreted at the 95% confidence level.

Direction-dependent deviations in model-based ITD predictions from measurement-based ITD estimations were analysed per dataset type by calculating the direction-dependent ITD differences, and subsequently averaging the results across participants.

The frequency-selective performance evaluation relied on evaluating the datasets of all participants for their ITDs in bands of \(\pm 100\) Hz around centre frequencies of 500 Hz, 1 kHz, 1.25 kHz, and 1.5 kHz. The results were compared to the corresponding model predictions, using the model coefficients as approximated by polynomial regression based on broadband measurement-based ITD estimations between 500 Hz and 1.5 kHz.

Since it is of particular interest how robustly Model 3 performs given variations in the model input data, that is, in the feature subsets introduced in Section 3.4, we evaluated its ITD-prediction performance using the inter- and intra-individual variations in feature subsets, summarised in Table 3.

4 Modelling results

4.1 Effect of model complexity

Figure 8 shows the effect of the individual model coefficients on the performance of Model 3 in terms of MAEs and RMSEs, in comparison to Model 1. In HRTF datasets, Model 3 delivered similar results as Model 1 when the
predictions were based on the first coefficient ($\gamma_{1n}$), and the first and second coefficients ($\gamma_{1n}$ and $\gamma_{2n}$), but on average performed substantially better, when increasing its complexity by including $\gamma_{3n}$, although exhibiting larger error variability. In front and rear HARTF datasets, adding $\gamma_{2n}$ already allowed a refined ITD adaptation in Model 3, which manifested itself in lower prediction errors compared to Model 1. The largest performance improvement in HARTF data-sets was observed after adding the second harmonic, that is, $\gamma_{2n}$ – a trend that continued after adding $\gamma_{3n}$, but was no longer as pronounced. Entering higher harmonics would likely only lead to marginal improvements. This said, the specific implementation of Model 3 represents a good compromise between complexity and achievable prediction performance.

### 4.2 Inter-model comparison

Figure 9 compares the measurement-based ITD curves with the ones predicted by the three ITD models. Direction-dependent deviations between the model-based ITD predictions and the corresponding measurement-based ITD estimations are represented by dashed-dotted (Model 1 and Model 3) and dashed (Model 2+) lines. The last three rows show the correlation between modelled and measurement-based ITD maxima, the differences in ITD maxima, $D^{\text{max}}_{\text{ITD}}$, and the differences in the arguments of the ITD maxima, $D^{\text{arg max}}_{\phi_{\text{ITD}}}$. Corresponding quantitative results are presented in Table 4. All reported correlations and predictors in linear regression models were significant ($p < .001$).

The results, comparing frequency-selective measurement-based and broadband model-based ITD estimations and predictions, respectively, across datasets and models, are shown for all participants in Figure 10. Due to the influence of noise in the narrowband evaluations, application of the chosen ITD estimation method resulted in discontinuous ITD curves between 25 and 60 deg azimuth (HRTF), 17.5 and 45 deg (front HRTF), and 15 and 45 deg (rear HRTF), mainly for the frequency bands with centre frequencies above 500 Hz, and only for some participants. For this reason, the ITD maxima and the arguments of the ITD maxima were evaluated in an angular range of $90 \pm 27.5$ deg when calculating the means and standard deviations, represented by error bars.

Noteworthy characteristics of the individual models are discussed based on various factors below.
Figure 9. (a) Measurement-based (light grey) and (b), (c) model-based broadband ITD estimations, colour-coded in grey, black and blue, respectively, evaluated for directions in the horizontal plane for HRTF and front (F) and rear (R) HARTF datasets. Deviations from measurement-based ITDs are shown as dashed-dotted (Models 1 and 3) or dotted lines (Model 2+), with standard deviations as shaded areas. (d) Scatter plots comparing measurement-based and model-based ITD maxima, fitted by linear regression lines. Box plots show medians and IQRs of differences in (e) ITD maxima and (f) arguments of the ITD maxima, with whiskers covering 1.5 times the IQR, and outliers being displayed as crosses. Horizontal black lines indicate non-significant (n.s.) mean differences at the 95% confidence level.
Table 4. Broadband ITD-model prediction errors, quantified by MAEs and RMSEs, averaged across participants. Bivariate correlation coefficients ρ quantify the correlation strength between measurement-based and model-based ITD maxima, further described by the corresponding linear-regression equations, \( \kappa \cdot \text{ITD}_{\text{meas}} + \lambda \), and adjusted coefficients of determination \( R^2 \). The last two columns present model-prediction errors regarding ITD maxima and the differences in the arguments of the ITD maxima.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>MAE ( \mu \pm \sigma ) (( \mu \pm \sigma ))</th>
<th>RMSE ( \mu \pm \sigma ) (( \mu \pm \sigma ))</th>
<th>max{ITD} ( \rho )</th>
<th>( \kappa )</th>
<th>( \lambda ) (( \mu \pm \sigma ))</th>
<th>( R^2 )</th>
<th>( \Delta \text{max}{ITD} \mu \pm \sigma )</th>
<th>( \Delta \text{arg max}{ITD} \varphi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRTF</td>
<td>1</td>
<td>38 ± 5.7</td>
<td>43 ± 6.5</td>
<td>.93</td>
<td>.72</td>
<td>161 ± .85</td>
<td>.85</td>
<td>42.7 ± 10</td>
<td>3.9 ± 2.6</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>131.3 ± 10</td>
<td>113.4 ± 67.4</td>
<td>.87</td>
<td>.67</td>
<td>90.2 ± .75</td>
<td>.75</td>
<td>146 ± 12.3</td>
<td>2.6 ± 5.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>18.1 ± 9.2</td>
<td>22.6 ± 9.5</td>
<td>.92</td>
<td>.92</td>
<td>72.2 ± .84</td>
<td>.84</td>
<td>13.7 ± 9.7</td>
<td>2.4 ± 2.5</td>
</tr>
<tr>
<td>HARTF (F)</td>
<td>1</td>
<td>65.1 ± 11.8</td>
<td>85.1 ± 14.7</td>
<td>.64</td>
<td>.83</td>
<td>152 ± .38</td>
<td>.38</td>
<td>34.6 ± 20.2</td>
<td>8.8 ± 2.8</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>81.6 ± 12.9</td>
<td>102.7 ± 11.7</td>
<td>.7</td>
<td>1.3</td>
<td>293.4 ± .48</td>
<td>.48</td>
<td>−108.9 ± 25.8</td>
<td>−6 ± 6.3</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>18.2 ± 5.7</td>
<td>22.6 ± 7</td>
<td>.85</td>
<td>.8</td>
<td>139.4 ± .7</td>
<td>.7</td>
<td>−2.4 ± 10.6</td>
<td>−1.9 ± 3.2</td>
</tr>
<tr>
<td>HARTF (R)</td>
<td>1</td>
<td>76 ± 12.7</td>
<td>100 ± 15.3</td>
<td>.63</td>
<td>.79</td>
<td>195.7 ± .38</td>
<td>.38</td>
<td>51.8 ± 20.5</td>
<td>10.7 ± 2.3</td>
</tr>
<tr>
<td></td>
<td>2+</td>
<td>72 ± 10</td>
<td>93.8 ± 10.2</td>
<td>.73</td>
<td>1.25</td>
<td>258.8 ± .51</td>
<td>.51</td>
<td>−91.7 ± 25</td>
<td>−4.1 ± 5.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>18.1 ± 5.9</td>
<td>22.6 ± 7.2</td>
<td>.87</td>
<td>.82</td>
<td>115.1 ± .75</td>
<td>.75</td>
<td>−6.9 ± 10.3</td>
<td>−2.4 ± 2.4</td>
</tr>
</tbody>
</table>

Note: \( \mu \) and \( \sigma \) represent mean and standard deviation, respectively.

4.2.1 Estimation of effective head radii

Fitting Model 1 to the measurement-based ITD curves and estimating the effective head radii by polynomial regression resulted in average values of 78.0 ± 2.2 mm (\( \mu \pm \sigma \), range: 74.5–93.6 mm) and 83.0 ± 2.9 mm (\( \mu \pm \sigma \), range: 78.6–91.6 mm) for the HRTF and HARTF datasets, respectively. The effective head-radius estimations for the HRTF datasets were slightly lower than the average value of 87 mm (range: 79–95 mm) reported by Algazi et al. [18], with higher average RMSEs, see Table 4, compared to their 32 \( \mu \)s (range: 22–47 \( \mu \)s).

4.2.2 Prediction of interaural-time-difference maxima

Application of the estimated effective head radii to Model 1 resulted in underestimated measurement-based ITD maxima in HRTF datasets, see Figures 9b, 9d and 9e, which tended to be more pronounced for higher ITD maxima. This was confirmed by a strong and significant correlation between \( \text{max\{ITD\}} \)-values, as well as the corresponding linear-regression result with a large adjusted coefficient of determination, see left panel of Figure 9d, and the linear-regression equation in Table 4. Reasons for the reduced \( \text{max\{ITD\}} \)-values can be traced back to the shape of the measurement-based ITD curves in HRTF data, which started to deviate from the sinusoidal form for the evaluated sub-bands with centre frequencies above 500 Hz, see left panel of Figure 10a. The additional development of dips between 0–50 deg and 120–180 deg likely further contributed to lowered broadband maximum ITD-model magnitudes. These phenomena also tempered with the predictions of ITD maxima in the lowest frequency band of HRTFs, which actually approaches its target frequency range. Interestingly, the radius correction in HARTF datasets resulted in regression lines above but, in case of front HARTFs, closer to the ideal diagonal, albeit with generally lower correlation coefficients and adjusted \( R^2 \)-squared values, see centre and right panel of Figure 9d. Deviations in \( \text{max\{ITD\}} \)-values between front and rear HARTF datasets further reflected the effect of a slight horizontal rotation of the connection line running through the HA microphones, cf. Figure 6, indicating different effective head radii which were, however, estimated by an average correction offset \( d_{10} \) to the HA-microphones centre.

In accordance with previous results [16, 21, 26], Model 2+ drastically underestimated the ITD maxima in all datasets using the estimated effective head radii, see Figures 9b, 9d and 9e. This was expected since the model is only intended to be used for \( \kappa d \gg 2.4 \). Approaching its valid frequency range, the estimation error gradually decreased towards higher frequency bands across datasets, cf. Figure 9b. The corrected effective head radii and the reduced magnitudes of ITD maxima in HARTF datasets additionally favoured smaller prediction errors, see centre and right panels of Figure 9d. However, the prediction consistency for HARTF datasets degraded, entailing steeper but less determined linear regression lines, and lower correlation factors, see Table 4.

Due to its increased complexity and flexibility, Model 3 was able to widely adapt to the ITD-curve-shape properties in all types of datasets. This allowed the dips in ITD curves, problematic for Model 1, to be considered, enabling generally more accurate ITD-curve fittings, which was reflected by \( \Delta \text{arg max}\{ITD\} \)-values closer to zero across datasets, see Figures 9d and 9e. Consistently high correlation factors and well-determined linear-regression lines, approaching the ideal diagonals, indicate superior performance of Model 3, see Figure 9d and Table 4. Although slightly underestimating the ITD maxima in the lowest frequency bands around 500 Hz, a very good prediction performance across the remaining evaluated frequency bands was observed, outperforming especially Model 2+, see Figure 10c.

Table 5 summarises the results from rANOVA and post-hoc tests, analysing trimmed-mean differences in \( \text{max\{ITD\}} \) and \( \text{arg max\{ITD\}} \), with indices of \( \Delta \) corresponding to the respective ITD models, see also Figures 9e and 9f. All rANOVAAs of \( \text{max\{ITD\}} \)-values turned out to be significant with large explanatory measures of effect
Post-hoc test results indicated significant trimmed-mean differences for all model comparisons, in all datasets. In HRTF datasets, the lowest mean model-prediction errors closest to zero were observed for Model 3, increasing significantly towards underestimation in Model 1, and particularly in Model 2+. In HARTF datasets, Model 3 again outperformed Model 1. While, compared to the prediction results in HRTF data, application of Model 1 resulted in higher median and mean errors in rear and front HARTF datasets, with a general tendency towards overestimation, Model 2+ showed reduced underestimation tendency for reasons discussed earlier. These results were supported in the same way by MAEs and RMSEs, see Table 4.

4.2.3 Prediction of the arguments of the interaural-time-difference maxima

Due to the approximately sinusoidal shape of the ITD curves in the HRTF datasets, all models performed comparably when predicting the arguments of the ITD maxima, yielding median and mean $\Delta \text{arg max} \{\text{ITD}\}$-values close to zero with rather small spreads but some outliers, see left panels of Figures 9b and 9f, and Table 4.

The mathematical form of Model 1 only allowed a unique argument of the ITD maximum at $\phi = 90^\circ$, resulting in an overestimation of arguments of measurement-based
ITD maxima in HARTF datasets, see centre and right panels of Figures 9b and 9f, and Table 4.

This limitation was specifically addressed by Model 2+, allowing a shift of the argument of the ITD maximum by taking account for a horizontal ear-canal and a HA-microphones angle, $\Theta_E$ and $\Theta_{HA}$, respectively. Although the predictions were shifted in the right direction, the measurement-based $\Delta \arg \max \{\text{ITD}\}$-values were underestimated by Model 2+, and the error exhibited a larger spread. Reasons for this can be sourced from the assumption of a spherical geometric model [26] for calculating $\Theta_E$ and $\Theta_{HA}$, see Figure 6 and equations (A1) and (A3). Calculations based on a refined ellipsoidal head [21] or extracting the angles based on photos from the top, or photogrammetric data [19, 75, 76] would result in smaller horizontal angles and likely lead to more accurate estimations in HARTF datasets using Model 2+. The observations for Model 1 and Model 2+ also apply to the results when comparing the model-based broadband ITD predictions to measurement-based narrowband ITD estimations, see Figure 10.

Model 3 eliminated the shortcomings of the two other models and resulted in $\Delta \arg \max \{\text{ITD}\}$-values close to zero across HARTF datasets, see Figures 9c and 9f and Table 4, when compared with the narrowband ITD estimations, see Figure 10c. Although the effective head radius was predefined, the measurement-based ITD curves could still be fitted precisely due to the remaining degrees of freedom in Model 3.

The already hinted performance similarity across models in HRTF datasets regarding $\Delta \arg \max \{\text{ITD}\}$ was corroborated by a non-significant rANOVA, see Table 5. Both rANOVAs of the results for HARTF datasets were significant with large $\xi$. All post-hoc test results were significant, except between Model 2+ and 3 in rear HARTF datasets, indicating a certain degree of prediction accuracy in Model 2+ when estimating the arguments of the ITD maxima, which, however, was accompanied by a large prediction variance. Based on the error distributions, cf. Figures 9f, and the results from the post-hoc tests, the best overall performance can still be attributed to Model 3, while observing largely significant over- and underestimation in Model 1 and Model 2+, respectively, which was also supported by the patterns in MAEs and RMSEs, see Table 4.

### 4.2.4 Direction-dependent prediction of interaural time differences

Figures 9b and 9c show the direction-dependent ITD-prediction errors by each model. All results provided in this section represent means and standard deviations, $\mu \pm \sigma$, averaged across participants.

Model 1 allowed to achieve fair ITD predictions in HRTFs datasets, exhibiting a nearly symmetrical error pattern at the frontal plane across the evaluated directions in the horizontal plane, see Figure 9b. On average, the ITD predictions in the azimuth ranges within 0–90 deg and 90–180 deg reached maximum deviations of 36.9 ± 24.5 $\mu$s (at 32.5 deg) and 70.1 ± 21.0 $\mu$s (at 142.5 deg), respectively, dropping to minimum prediction errors of −49.5 ± 17.6 $\mu$s (at 65 deg) and −44.3 ± 8.7 $\mu$s (at 105 deg). Due to the shift of $\arg \max \{\text{ITD}\}$ and the aforementioned curve irregularities in measurement-based ITD estimations in front and rear HARTF datasets, the model predictions within 0–90 deg changed to minimum and maximum errors of −45.6 ± 21.2 $\mu$s (at 40 deg) and 41.7 ± 23.8 $\mu$s (at 90 deg; rear HARTFs: −43.5 ± 22.6 $\mu$s at 35 deg, and 61.8 ± 25.5 $\mu$s at 90 deg), respectively, and resulted in prominent overestimations up to 170.4 ± 25.5 $\mu$s (at 125 deg; rear HARTFs: 197.5 ± 24.7 $\mu$s at 122.5 deg) within 90–180 deg.

Model 2+ consistently underestimated the measurement-based ITDs of HRTF datasets and also showed a rather symmetrical error pattern at the frontal plane, see Figure 9c. Deviation minima of −201.8 ± 25.9 $\mu$s and −198.8 ± 14.7 $\mu$s occurred at 62.5 deg and 112.5 deg, respectively, with a local maximum of −157.4 ± 16.8 $\mu$s at 90 deg. The ITD-deviation patterns shifted towards the same direction as did $\arg \max \{\text{ITD}\}$ in front HARTF datasets with slightly higher minima of −180.3 ± 16.7 $\mu$s at 42.5 deg (rear HARTFs: −172.5 ± 19.0 $\mu$s at 40 deg), and
Table 6. Effect of feature variations on MAEs and RMSEs in the ITD predictions of Model 3. The errors were calculated by evaluating the predictions of Model 3 using inter- and intra-individual variations when extracting the selected anthropometric and HA features from participant 18 (female) multiple times, see Table 3, in comparison to measurement-based ITD estimations. Deviations with respect to the mean MAEs and RMSEs, cf. Table 4, are indicated by $\Delta \mu$, with percentage deviations provided in brackets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Inter-individual extractions</th>
<th>Intra-individual extractions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu$</td>
<td>$\sigma$</td>
</tr>
<tr>
<td>HRTF</td>
<td>16.7/21.8</td>
<td>1.8/2.6</td>
</tr>
<tr>
<td>HARTF (front)</td>
<td>26.8/34.1</td>
<td>7.1/7.7</td>
</tr>
<tr>
<td>HARTF (rear)</td>
<td>22.2/27.7</td>
<td>6/7.7</td>
</tr>
</tbody>
</table>

Note: MAEs and RMSEs are provided in $\mu$s. $\mu$ and $\sigma$ represent mean and standard deviation, respectively.

$-138.1 \pm 30.8$ $\mu$s at 92.5 deg (rear HARTFs: $-116.7 \pm 31.2$ $\mu$s at 90 deg), and a local maximum of $-112.2 \pm 26.1$ $\mu$s at 75 deg (rear HARTFs: $-97.0 \pm 25.1$ $\mu$s at 75 deg). Noticeably lower model-prediction errors were present in both types of HARTF datasets for the evaluated directions above 120 deg.

Model 3 also entailed a roughly symmetrical error pattern at the frontal plane in HRTF datasets with the lowest fluctuation among all models, resulting in minimum prediction errors of $-39.7 \pm 18.1$ $\mu$s and $-38.0 \pm 13.6$ $\mu$s at 55 deg and 120 deg, respectively. A maximum model-prediction error of $13.9 \pm 9.8$ $\mu$s occurred at 85 deg. The prediction errors also remained consistently low in front HARTF datasets across all analysed directions with minimum deviations of $-31.8 \pm 16.8$ $\mu$s and $-20.1 \pm 14.9$ $\mu$s, appearing at 40 deg and 155 deg (rear HARTF datasets: $-25.1 \pm 19.6$ $\mu$s and $-19.3 \pm 13.6$ $\mu$s at 35 deg and 155 deg), respectively. Maximum errors of $17.8 \pm 28$ $\mu$s and $27.6 \pm 17.5$ $\mu$s at 17.5 deg and 120 deg (rear HARTF datasets: $12.9 \pm 26.7$ $\mu$s and $34.7 \pm 17.4$ $\mu$s at 12.5 deg and 117.5 deg, respectively) were observed.

4.3 Model robustness

Exhibiting comparable magnitudes, the lowest errors were observed when predicting the ITDs of HRTFs, regardless of using inter- or intra-individual variations in feature extractions. With regard to the results when applying inter-individual variations, the ITD errors increase for rare HARTF datasets, and most noticeable for front HARTF datasets. Similar tendencies can be observed given intra-individual feature variations, although the mean error increase is not as pronounced as in the former scenario.

5 Discussion

5.1 Inter-model comparison

In comparison to Model 1 and especially Model 2+, the results from the specific model evaluation collectively indicate that Model 3 is superior in terms of ITD-prediction accuracy, while addressing their shortcomings. This conclusion applies to direction-dependent ITD predictions, regardless of the type of directional dataset, to the estimation of maximum ITD-values and the arguments of the ITD maxima (except for ITDs in HRTFs), as well as to MAEs and RMSEs. Also when compared to the measurement-based ITD estimations in frequency sub-bands within the target frequency range, Model 3 delivered more consistent and precise prediction patterns. Although MAEs and RMSEs increase after considering inter- and intra-individual variations in feature extractions, Model 3 demonstrated robustness and still resulted in ITD predictions that...
were substantially below the best-case prediction errors of the other two models. A more pronounced error increase in predicted ITDs for front HARTFs indicates that particular care is advised when manually extracting the corresponding feature subset. Note that these ITD predictions, provided variations of the selected female feature subset, are likely biased due to the gender distribution among participants, and thus may have been pushed towards the upper range of the underlying error distribution.

5.2 Perceptual implications

5.2.1 Just-noticeable differences

For a better understanding of the perceptual implications, the observed ITD-prediction errors of the individual models need to be compared to the JNDs in ITDs known from the literature.

One of the earliest studies investigating the JNDs in listeners with normal hearing was conducted by Klumpp and Eady [77], reporting values as low as 9 μs. Similarly low values of 10 μs were observed when discriminating pulsed sinusoidal stimuli [78], and even lower ones of 6.9 μs in trained listeners (18.1 μs in untrained listeners) [79]. Aussal et al. [61] calculated a more relaxed threshold of 16 μs which, however, was not verified perceptually. Simon et al. [80] also reported higher perceptual JNDs of 33 μs and 68 μs at 30 deg azimuth and 90 deg azimuth, respectively, based on an alternative forced-choice left/right discrimination protocol. In perceptual experiments investigating the influence of simulated room reflections on JNDs using stimuli based on musical instruments [81], thresholds of 20 μs were determined, increasing by five to eight times higher values in reverberation.

Spencer et al. [82] measured notably higher average JNDs of 68 μs in low-frequency ITDs with high individual variability in participants with symmetric hearing loss, which were not significantly different from those with normal hearing in their study. Phase distortions and corresponding ITD errors are likely to be expected when such thresholds are measured via bilateral HAs that are not synchronised (for reviews, see [83, 84]). Poorest performance with JNDs of 150 μs and 100–350 μs has been observed in bilateral cochlear-implant users [85] and bimodally-fitted listeners [86], respectively.

The lowest reported JNDs generally impose only minimal tolerance of prediction errors in ITD models. Neither of the compared models in the present investigation would fulfil the highest perceptual standards in terms of MAEs and RMSEs when using the modelled ITDs of HRTFs for applications involving listeners with normal hearing. Using more relaxed JND values in the upper range of 16–68 μs, Model 1 and especially Model 3 would be suitable for modelling purposes. The upper JND limits are also largely complied with by Model 1, and entirely by Model 3, when accounting for direction-dependent ITD-prediction errors.

5.2.2 Outlook on perceptual model evaluation

In addition to the model assessment based on JNDs, future listening experiments are necessary to evaluate Model 3. Application of the required azimuthal rotation on the individual datasets during post-processing likely introduced an error in the resulting measurement-based ITD estimations, which consequently affected the modelling procedure and the model-based ITD predictions. It remains to be investigated whether this post-processing step is noticeable and has perceptual effects. In this context, it is of particular importance to systematically examine the influence of the observed ITD-prediction errors on spatial audio quality metrics, such as, for example, sound localisation in the horizontal plane. Variations of previously manipulated factors and their evaluation [9, 10] may be extended to include a weighting between the ITDs from HRTFs when utilising residual hearing, which are superimposed with the ITDs from behind-the-ear HARTFs [6]. To this end, it is worth investigating the influence of different types of ear pieces, which can be included as additional filters, or (binaural) HA algorithms. The results of such experiments involving young adults with normal hearing will likely allow to derive important baseline measures, and further corroborate the perceptual validity of the proposed ITD model.

5.3 Limitations

Head asymmetries are not addressed by any the compared models in their current mathematical forms, only allowing ITD predictions of zero at 0 deg azimuth. The attempt to introduce an identical phase-shift coefficient in all sine terms of Model 3 initially led to promising ITD predictions but resulted in robustness issues when varying the magnitudes in feature subsets. Applying the proposed ITD model using feature subsets measured from women may lead to increased prediction errors given the unequal gender distribution with a bias towards male participants in the presented database. However, we proposed an extensible model framework for the estimation of specific model coefficients, which can be adapted to new input data. For more general applicability, the model may subsequently be extended to include elevation angles [87, 88].

6 Conclusion

We presented a hybrid multi-harmonic model (AMT, pausch2022 [43]) for the prediction of ITDs in different...
types of spatial transfer functions, that is, HRTFs, and front and rear HARTFs, for arbitrary directions in the horizontal plane. The model coefficients were derived from the horizontal-plane datasets of the IHTA-indHARTF database [42], containing the individual receiver directivities from 30 adults, wearing two behind-the-ear research HAs and in-ear microphones at the blocked ear-canal entrance. Tailored to each type of spatial transfer function, optimal polynomial weights were calculated to facilitate the estimation of individualised model coefficients by scaling individual subsets of anthropometric and HA features. Objective evaluations against the simple and extended harmonic analytic models by Kuhn and Woodworth, respectively, revealed superior ITD-prediction performance of the proposed model when compared to the broadband, frequency-selective, and direction-dependent measurement-based ITD estimations in the evaluated frequency range, regardless of the considered type of spatial transfer function. It was demonstrated that the two analytic models cannot satisfactorily predict the (arguments of the) ITD maxima and the generally more complex ITD-curve shapes in HARTF datasets. In this context, the extended Woodworth model suffered from an oversimplified calculation of the horizontal HA-microphones angle based on a spherical head geometry, motivating the use of more accurate calculation or acquisition methods. In addition to the prediction accuracy of the proposed model, being within an acceptable range in view of perceptual requirements, robust ITD predictions are feasible even under varying input data, which further substantiates its applicability.

Acknowledgments

We thank all study participants, and Julian Burger for manually extracting the individual features from the photos. We also appreciate the constructive comments by Robert Baumgartner and the two anonymous reviewers. Special thanks go to Clara Hollomey and Piotr Majdak for the smooth model integration into the AMT.

Declaration of conflicting interests

We declare that there are no conflicts of interest. GN ReSound was not involved in any stage of this publication other than in providing the hardware.

Funding

The work of F.P. was supported by the European Union’s Seventh Framework Programme for research, technological development and demonstration (grant no. 607139, Improving Children’s Auditory Rehabilitation), and partially financed within the Horizon 2020 research and innovation funding programme (grant no. 101017743, SONICOM – Transforming auditory-based social interaction and communication in AR/VR). S.D. received funding from the German Research Foundation (grant no. 402811912, Individual Binaural Synthesis of Virtual Acoustic Scenes).

Data availability statement

The IHTA-indHARTF database is available for download via RWTH Publications at https://doi.org/10.18154/RWTH-2020-04267 [42]. Implementations of the presented models (pausch2022) and corresponding simulations (exp_pausch2022) are publicly available as part of the Auditory Modeling Toolbox (AMT, https://www.amtoolbox.org) [89] in the release of version 1.2.0 [43].

References


F. Pausch et al.: Acta Acustica 2022, 6, 34 19


Appendix A
Calculation of additional features

For simplification, the horizontal ear-canal angle $\Theta_E$ and the horizontal HA-microphones angle $\Theta_{HA}$, both with respect to the median plane, were calculated based on the assumption of a spherical head [26]. We estimated the effective by-participant head radii $a_n$ as described in Section 3.2. Projecting the pinna offset back $x_5$ on the sphere led to

$$\Theta_E = 90^\circ + \arctan \left( \frac{x_5}{\sqrt{a_n^2 - x_5^2}} \right).$$  \hspace{1cm} (A1)

The horizontal distance $\zeta$ between the HA-microphones centre and the ear-canal entrance was calculated by

$$\zeta = d_9 \cdot \sin(\Theta_E),$$  \hspace{1cm} (A2)

which was defined along the vertical plane running through the axis of the HA microphones, see Figure 6. We assumed the HA-microphones centre to be separated from the head centre by a distance $(a_n + d_{10})$. From the resulting triangle between the ear-canal entrance, the head centre and the centre of the HA microphones, the horizontal HA-microphones angle was determined by

$$\Theta_{HA} = \arccos \left( \frac{a_n^2 + (a_n + d_{10})^2 - \zeta^2}{2a_n \cdot (a_n + d_{10})} \right) + \Theta_E. \hspace{1cm} (A3)$$

using the law of cosines. The calculation of the HA-microphones offset up $d_{13}$ included the known pinna offset down $x_4$, yielding

$$d_{13} = d_9 \cdot \cos(\Theta_E) - x_4. \hspace{1cm} (A4)$$

Finally, the HA-microphones offset back $d_{14}$, describing the horizontal displacement relative to the head centre, was determined by

$$d_{14} = (a_n + d_{10}) \cdot \sin(\Theta_{HA} - 90^\circ). \hspace{1cm} (A5)$$

Appendix B
Multivariate polynomial regression

To estimate the effective head radius and the dataset-specific sets of model coefficients of Model 3, cf. equation (3), based on M features, a multivariate polynomial-regression model (see, e.g., [90]) with a polynomial degree of P was defined according to equation (B1)
\[
\begin{pmatrix}
a_1 & \gamma_{11} & \gamma_{21} & \gamma_{31} \\
a_2 & \gamma_{12} & \gamma_{22} & \gamma_{32} \\
\vdots & \vdots & \vdots & \vdots \\
a_N & \gamma_{1N} & \gamma_{2N} & \gamma_{3N}
\end{pmatrix}
= \begin{pmatrix}
1 & x_{11} & x_{12}^2 & \ldots & x_{11}^P & x_{21} & x_{22}^2 & \ldots & x_{21}^P & \ldots & x_{1N} & x_{1N}^2 & \ldots & x_{1N}^P \\
1 & x_{12} & x_{12}^2 & \ldots & x_{12}^P & x_{22} & x_{22}^2 & \ldots & x_{22}^P & \ldots & x_{1N} & x_{1N}^2 & \ldots & x_{1N}^P \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & x_{1N} & x_{1N}^2 & \ldots & x_{1N}^P & x_{2N} & x_{2N}^2 & \ldots & x_{2N}^P & \ldots & x_{1N} & x_{1N}^2 & \ldots & x_{1N}^P
\end{pmatrix}
\begin{pmatrix}
\beta_{01} & \beta_{02} & \beta_{03} & \beta_{04} \\
\beta_{111} & \beta_{112} & \beta_{113} & \beta_{114} \\
\beta_{121} & \beta_{122} & \beta_{123} & \beta_{124} \\
\vdots & \vdots & \vdots & \vdots \\
\beta_{1P1} & \beta_{1P2} & \beta_{1P3} & \beta_{1P4} \\
\beta_{211} & \beta_{212} & \beta_{213} & \beta_{214} \\
\beta_{221} & \beta_{222} & \beta_{223} & \beta_{224} \\
\vdots & \vdots & \vdots & \vdots \\
\beta_{2P1} & \beta_{2P2} & \beta_{2P3} & \beta_{2P4} \\
\vdots & \vdots & \vdots & \vdots \\
\beta_{MP1} & \beta_{MP2} & \beta_{MP3} & \beta_{MP4}
\end{pmatrix}
\begin{pmatrix}
\epsilon_{11} & \epsilon_{21} & \epsilon_{31} & \epsilon_{41} \\
\epsilon_{12} & \epsilon_{22} & \epsilon_{32} & \epsilon_{42} \\
\vdots & \vdots & \vdots & \vdots \\
\epsilon_{1N} & \epsilon_{2N} & \epsilon_{3N} & \epsilon_{4N}
\end{pmatrix}
\]

(B1)

See equation (B1) on top of this page

or, equivalently, as

\[
\Gamma = \mathbf{A} \mathbf{B} + \mathbf{E}. \tag{B2}
\]

The matrices \( \mathbf{A}, \mathbf{B} \) and \( \mathbf{E} \) contain the individual features, polynomial regression weights, and residual errors, respectively. Note that the effective head radii in both types of HARTF datasets differ from those in the HRTF dataset after accounting for the HA-microphones-to-scalp offset \( d_{10} \); cf. Section 3.2, as do the feature-subset selections, cf. Section 3.4. Minimising the sums of squared errors, \((\mathbf{\Gamma} - \mathbf{A} \mathbf{B})^\top (\mathbf{\Gamma} - \mathbf{A} \mathbf{B})\) regarding \( \mathbf{B} \), with \((\cdot)^\top\) representing the transpose, yielded \((MP + 1) \times 4\) optimal polynomial-regression weights [90]

\[
\mathbf{\hat{B}} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{\Gamma}, \tag{B3}
\]

which were estimated in R (packages \texttt{lm} and \texttt{polym} both from the library \texttt{stats}) for ordinary polynomials. Based on the resulting set of optimal weights, the ITD-model coefficients were approximated through

\[
\mathbf{\hat{\Gamma}} = \mathbf{A} \mathbf{\hat{B}}. \tag{B4}
\]

To obtain the model coefficients for one participant using the corresponding feature subset, equation (B4) reduces to

\[
\hat{y}_n = \mathbf{a}_n^\top \mathbf{\hat{B}}, \tag{B5}
\]

with \( \vec{\gamma}_n = (\vec{a}_n, \vec{\gamma}_1, \vec{\gamma}_2, \vec{\gamma}_3) \) and \( \mathbf{a} = (1, \mathbf{x}_{11}, \mathbf{x}_{11}^2, \ldots, \mathbf{x}_{11}^P, \mathbf{x}_{21}, \mathbf{x}_{21}^2, \ldots, \mathbf{x}_{21}^P, \ldots, \mathbf{x}_{1N}, \mathbf{x}_{1N}^2, \ldots, \mathbf{x}_{1N}^P)^\top \). For Model 1 and 2+, the multivariate regression model simplifies to a univariate one, with the estimated polynomial-regression weights contained in a vector \( \hat{\beta}_n = (\hat{\beta}_{01}, \hat{\beta}_{111}, \hat{\beta}_{112}, \ldots, \hat{\beta}_{1P1}, \hat{\beta}_{211}, \hat{\beta}_{212}, \ldots, \hat{\beta}_{2P1}, \ldots, \hat{\beta}_{MP1})^\top \), cf. first column of \( \mathbf{\hat{B}} \).

Cite this article as: Pausch F., Doma S., & Fels J. 2022. Hybrid multi-harmonic model for the prediction of interaural time differences in individual behind-the-ear hearing-aid-related transfer functions. Acta Acustica, 6, 34.