Audio Engineering Society

Convention Paper 10654

Presented at the 154th Convention
2023 May 13–15, Espoo, Helsinki, Finland

Examining the minimum detectable bi-lateral variation of generic Head-Related Transfer Functions

Shaimaa Doma[1] and Janina Fels[1]

[1]*Institute for Hearing Technology and Acoustics, RWTH Aachen University*

Correspondence should be addressed to Shaimaa Doma (sdo@akustik.rwth-aachen.de)

ABSTRACT

This work aims at deriving a minimum required resolution for optimization of head-related transfer functions (HRTFs). It builds on existing metrics, used to numerically evaluate HRTF differences, as well as on a model estimating just noticeable differences (JNDs) for uni-lateral variation of HRTFs. Integrating this model, as well as descriptors for both monaural and binaural cue differences, a three-alternative forced choice experiment is set up to investigate JNDs for bi-lateral variation of HRTF sets. Rather than introducing manual changes to the spectra, an exchange between magnitude spectra of generic HRTF sets is employed, while controlling for multiple conditions related to the descriptors. The probability of distinguishing between the stimulus pairs is linearly modeled using different subsets of numerical descriptors. A model integrating two monaural descriptors, 'issd' and 'mfcd', achieves the best performance, compared to the rest. It shows a tendency for slight improvement when combined with an estimate of the detectability of changes in interaural cross-correlation.

1 Introduction

As virtual audio applications increasingly offer individualized spatial audio solutions, the aim for personalized head-related transfer functions (HRTF) is becoming more widespread. Here, a variety of different HRTF acquisition methods, ranging from acoustic measurements to numerical simulations and less complex approximation methods, poses more or less feasible options, depending on the available facilities and hardware. As can be expected, these methods possess varying degrees of detail loss, manifesting as spatial and/or spectral cue distortion.

In optimizing these methods, the goal is to improve quality, i.e., to increase the similarity to a target HRTF, while minimizing measurement or computational effort. This poses the question at what point further attempts of optimization are no longer perceivable and, accordingly, no longer of benefit.

Numerous studies have attempted to use numerical descriptors of the error between a given and a target spectrum (e.g., [1, 2]), yet mostly without drawing a direct connection between perception and the used set of descriptors.

A multitude of models has been introduced, relating numerical differences to specific perceptual properties, such as localization (e.g., [3]) or spectral coloration [4]. These models are feasible for use with arbitrary HRTF spectra and are not restricted to describing the effect of manually introduced HRTF differences (so-called "degradation"). However, the latter restriction is often given for studies on just noticeable differences (JNDs),

which typically address the detectability of changes as described by some degradation parameter, e.g. a smoothing factor [5]. This dependency is limiting, as knowledge from such JND studies can hardly be applied to arbitrary HRTF differences - at least not in a direct manner. Furthermore, the above-mentioned models for spectral evaluation typically address suprathreshold differences, relating them to consciously perceived properties of the auditory stimulus. Little research has been conducted on modeling near-threshold HRTF differences for arbitrary (i.e., not artificially modified) data.

The current work follows up on a JND model for unilateral HRTF differences, previously developed by the authors [6]. Here, the model is expanded, introducing near-threshold bi-lateral HRTF variation. This implicates monaural cue changes at both ears as well as binaural cue distortions. A selection of suitable descriptors for these distortions is introduced in Section 2. Different degrees of variation of the descriptors are incorporated into a listening experiment, as described in Section 3. The acquired perceptual data are subsequently presented in Section 4, where the relation between the detectability of differences and the given descriptor values is examined. A simplified linear model is derived based on these findings. Model performance and applicability are finally discussed in Section 5, followed by a short summary of findings.

2 Materials

Bilateral changes in HRTFs can be quantified using monaural descriptors (assessing the changes in left and right ear spectra individually) and binaural descriptors (assessing changes in binaural cue information). A selection from both categories is used in this work.

As an indicator for detectability of uni-lateral changes, a JND model [6] is used, that is based on the following three metrics:

The Mean Squared Error (`mse`) metric performs a bin-based calculation of the mean squared difference between two HRTF spectra, both belonging to the same incidence direction. This corresponds to a normalized integration over the difference spectrum. The Inter-Subject Spectral Difference (`issd`) [7] calculates the variance over the difference spectrum, thus being sensitive to changes in spectral shape, rather than in gain. Finally, the Mel-Frequency Cepstral Distortion (`mfcd`) [8] operates on the basis of cepstral coefficients of 24 mel bands. Using a mean squared error calculation, it quantifies the difference in spectral energy within these spectral bands.

The above-mentioned linear model exploits the common variance of the three metrics by using linear coefficients derived from Principal Component Analysis (PCA). The output of the model is an estimated probability of detecting a uni-lateral change in an HRTF spectrum, with a hard cut-off introduced at $p_{\text{mon}} = 0\%$ and 100 %. A value of $p_{\text{mon}} = 0\%$ corresponds to no detection (i.e., guessing in an experimental setting) and 50 % to the JND.

In addition to this model output, the three monaural input metrics are included as individual descriptors. Three binaural descriptors are furthermore considered. The first two describe changes in interaural level difference (ILD): Similarly to the `mse` metric, the $\texttt{mse}_{\text{ILD}}$ is introduced, calculating the mean squared error between the two ILD spectra (i.e., between the two difference spectra, on their part calculated from the left and right ear HRTFs). Analogously, the $\texttt{issd}_{\text{ILD}}$ expresses the variance of the difference spectrum between the two ILD spectra.

A third metric, p_{IACC}, predicts the detectability of a change in interaural cross-correlation (IACC), based on the findings of Pollack and Trittipoe [9]. Their work comprised a two-alternative forced choice (2-AFC) experiment, from the results of which they derived just noticeable differences of interaural noise cross-correlation, in relation to a reference IACC value. With a guessing rate of $\gamma = 50\%$ and assuming a lapse rate of $\lambda = 1\%$, their percentage of correct responses p_{2AFC} is transformed as

$$p_{\text{IACC}} = \frac{p_{\text{2AFC}} - \gamma}{(1 - \lambda - \gamma)}, \tag{1}$$

yielding a simplified paradigm-independent predictor for IACC transitions. Figure 1 depicts its dependence on both the amount of change and the actual (reference) IACC present in the signals. The values are reconstructed and interpolated from data in [9], Fig. 4, and transformed using eq. (1). The line corresponding to $p_{\text{IACC}} = 50\%$ is in the following referred to as the JND of IACC.

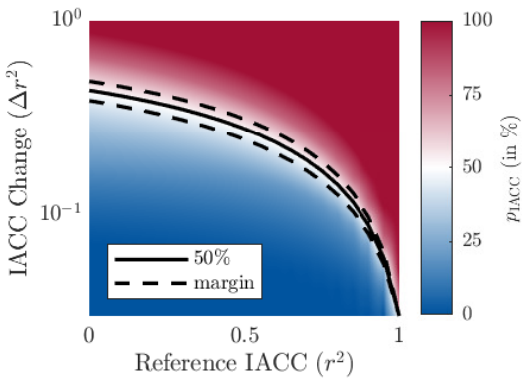

Fig. 1: Probabilities of detection of changes in interaural cross-correlation (IACC), as a function of the initial (reference) IACC between the two ear signals. The solid line indicates the JND (50% threshold). Dashed lines refer to the safety margins for stimulus selection above and below said threshold, see Section 3.1. (Information is based on data from [9], Fig.4.)

3 Experimental Design

The goal of the experiment is to provide perceptual data, to be used for modeling a binaural JND. This JND can be considered a minimum required resolution for HRTF similarity.

A three-alternative forced choice (3-AFC) paradigm, with the task of distinguishing one target from two reference signals, is chosen. A triple pink noise pulse, convolved with two pairs of HRTFs (in free-field), is presented for comparison. Each stimulus pair is repeated twice, with the total number of trials split into six experimental blocks. The order of blocks and of the trials within is latin square balanced. A mandatory break after the third block is introduced to minimize effects of exhaustion.

Playback is done using headphones (Sennheiser HD650) in an acoustically optimized hearing booth. The presentation level is calibrated to a mean of 60 dB SPL, using a Head Acoustics HMS III artificial head with IEC 711 ear simulator and a Nexus Type 2690-A conditioning amplifier. An individual headphone equalization [10] ensures that all participants receive approximately the same signal at the ear canal entrance. While the relative differences between the presented stimuli would still be present without this equalization, the detectability of differences would be affected by the absolute level in the respective frequency bands. E.g., a spectral notch caused by individual headphone-pinna interactions may lie below the audibility threshold, rendering superimposed differences in that frequency band undetectable, and thus hindering the comparability between the performance of different participants.

3.1 Conditions

The experiment is intended to be representative for arbitrary (or at least a wide variety of) data. To enable proper modeling using the introduced descriptors, it should further cover a large range of differences in both monaural and binaural cue information. The following features are therefore considered in the selection of stimuli.

3.1.1 HRTF databases and transitions

A total of three HRTF databases is integrated into the study. The first database comprises `measured` datasets from the ITA HRTF database [11]. The second database (`idealPCA`) is a reconstruction of these HRTF spectra using a linear combination of 23 Principal Components (in the spectral domain), weighted using the ideal score output of the PCA [12]. Finally, the third database (`anthroPCA`) is an approximation of the second. Instead of the ideal weighting score, however, an approximation of the PC score is attempted using multi-linear regression and six anthropometric dimensions of the respective database members [13].

The three databases provide different degrees of spectral detail loss for the HRTFs of the database members. This allows, on the one hand, for `intra`-individual transitions, i.e., a direct contrasting between the different levels of detail. On the other hand, `inter`-individual transitions represent the effect of non-individual cues, as HRTFs of different database members are compared. Combining the three databases and two transition cases yields a factor with six levels.

3.1.2 Hemispheres

In order to rule out directional biases (e.g., right- [14] or left-side [15] advantage effects), ipsi- and contralateral incidence directions are equally covered in the choice of HRTF spectra, which results in another factor with two levels.

3.1.3 Degrees of predicted monaural detectability

The monaural prediction model, as briefly outlined in Section 2, is used as an indicator for detectability of separate ear variations. Different combinations of left and right ear values are meant to allow for cases where the audible difference at one ear dominates perception, as well as cases where both ears have an equally indiscriminate or pronounced error. Three ranges are defined as follows:

`Low`:			p_{mon}	$<$	30 %
`Medium`:	30 %	$\leq$	p_{mon}	$<$	50 %
`High`:	50 %	$\leq$	p_{mon}	$<$	75 %

Evidently, these ranges are not symmetrical around the 50 % threshold. For modeling purposes, the goal is to gain perceptual data points centered around the 50 % threshold for the bi-lateral variation experiment. Therefore, since it is expected that the presence of a variation at the second ear would increase the probability of detection, a bias towards lower detectability is introduced to the uni-lateral value ranges. This choice of ranges was validated in pilot runs prior to the experiment. Combinations of the three value ranges for the left and right ear result in a factor with nine levels.

3.1.4 Transitions in IACC

The a priori known probabilities of detection of IACC transitions (p_{IACC}) are further considered. Half of the stimulus pairs are chosen to be below the known JND threshold of 50%. Ideally, for these stimuli, monaural predictors for the two ears may be sufficient to explain most of the perceived (or not perceived) differences. The other half of the stimuli is chosen above the JND, where (supra-threshold) binaural interactions are expected to superimpose monaural error behaviour. A safety margin, accounting for ± 0.05 at $r^2 = 0$, is introduced (see the dashed lines in Figure 1). All stimuli are selected outside this margin, providing a more distinct gap between the two conditions.

3.2 Stimulus selection and preparation

Given the available HRTF data, pairs of HRTF spectra are sought, with the constraint of representing combinations of the presented variables. Out of the given 216 condition combinations ($= 6 \cdot 2 \cdot 9 \cdot 2$), only 210 cases could be fulfilled, as a small number of conditions could not be met simultaneously.

With this approach for selection, multiple spatial directions are covered. It should be noted that only a direct comparison of HRTF spectra for the same direction, respectively, is offered. Thereby, the spatial cues corresponding to these direction are contrasted between the different datasets. These cue variations are restricted to changes in magnitude. All HRTF spectra are processed, replacing their phase by a minimum phase (calculated from the respective magnitude spectra), and adding a linear phase component for the runtime and an interaural phase difference (IPD) fixed for all participants. This direction-dependent IPD component is estimated for the respective incidence angles using an analytical ellipsoidal model [16], with mean head dimensions of the ITA HRTF database [11] as model input.

3.3 Participants

A total of 25 participants (15 female, 10 male), aged 19 - 28 years (median: 26), took part in the experiment. All possessed normal hearing, as verified by means of a high frequency audiogram. This ensured the perceptibility of spectral variations in the whole frequency range relevant for spatial cue detection.

4 Results

4.1 Experimental output

The listening experiment provides information on the distinguishability of various stimulus pairs. The percentage of correct answers for these data points is each calculated by averaging over the data of all participants (25×2 repetitions per stimulus pair). This information can be set in relation with the descriptors introduced in Section 2.

Figure 2 displays the relation between p_{correct} and the a-priori predictions of detectability for independent unilateral variation at the left and right ear, respectively. The color and size of the points indicate the percentage of correct answers achieved. The black grid marks the three ranges (`low`, `medium`, `high`) used for stimulus selection. When all 210 stimulus pairs are taken into account (top), the monaural model output does not seem to be (on its own) a good predictor for audibility, especially in the `medium` range. This is further examined in Sec. 4.6 through regression analysis. Omitting the

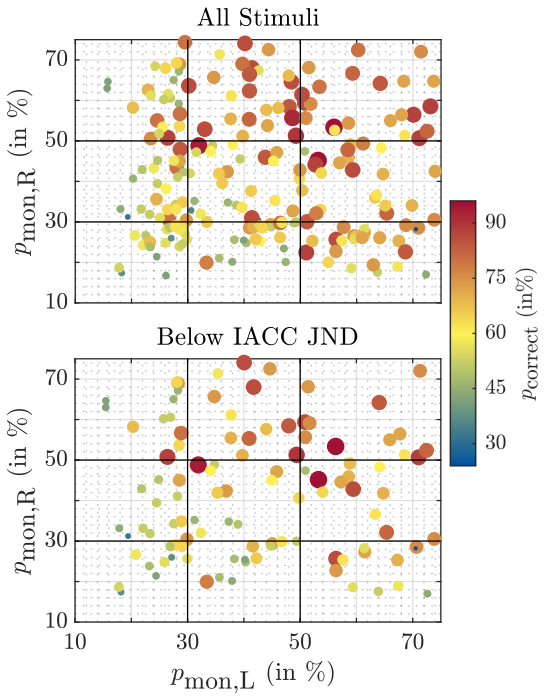

Fig. 2: Relation between the correct response rate and the monaural model predictions. Top: Stimuli with both sub- and supra-threshold IACC transitions are included. Bottom: The monaural model performs better as the superimposed supra-threshold IACC differences are excluded.

stimuli with supra-threshold IACC transitions (bottom) removes the superimposed contribution to audibility and leads to slightly better (though not ideal) correspondence between the correct response rate and the value ranges of p_{mon}.

4.2 Modeling approach

Multi-linear regression is chosen, approximating the percentage of correct answers p_{correct} by a linear combination of N descriptors:

$$p_{\mathrm{correct}} = c_0 + \sum_{i=1}^{N} c_i \cdot \hat{X}_i. \tag{2}$$

The centered and normalized descriptors $\hat{X}_i$ are calculated as

$$\hat{X}_i = \frac{X_i - \mu_{x,i}}{\sigma_{x,i}}. \tag{3}$$

The mean $\mu_{x,i}$ and standard deviation $\sigma_{x,i}$ are determined based on the available model training data X and are later used to center and normalize arbitrary model input. As only the slope of the psychometric function is modeled, a limitation to plausible values is required. Given the 3-AFC paradigm, a guessing rate of $\gamma = 33.3\%$ and a lapse rate of $\lambda = 1\%$ are assumed, leading to

$$p_{\mathrm{limited}} = \max\{33.3\,\%, \min\{99\,\%, p_{\mathrm{correct}}\}\}. \tag{4}$$

After this limitation, the data points are transformed to a paradigm-independent value range, where with the guessing rate is mapped to $p_{\mathrm{detect}} = 0\%$ and the sought threshold of $p_{\mathrm{correct}} = 66.2\%$ is mapped to 50%:

$$p_{\mathrm{detect}} = \frac{p_{\mathrm{limited}} - \gamma}{(1 - \lambda - \gamma)}. \tag{5}$$

4.3 Binaural weighting

The input to the model requires a single value per descriptor. This also holds for the monaural descriptors, for which initially a value for the left and right ear comparisons are available, respectively. Three options for combining these two values are considered.

`mean`: The ipsi- and contra-lateral ear descriptors are equally weighted in a simple averaging calculation.

`wMean`: A directional weighting is introduced, giving higher importance to differences occurring at the ipsi-lateral ear. Weighting curves, as constructed by Baumgartner et al. [3] based on data from [17] and [18], are used, see Fig. 3. Note that the lateral angle ϕ is only equal to the azimuth angle in the front half of the horizontal plane and is otherwise elevation- and azimuth-dependent.

`max`: The ear with the higher value for a given monaural descriptor is selected.

4.4 Correlation analysis

The selection of N potential descriptors to be used in eq. (2) is subject to restrictions of collinearity [19]. Subsets of descriptors with little to no correlation are therefore to be identified. Pearson correlation requires an approximately normal distribution of input data points. On this account, a logarithmic transform is applied to counteract the initial skewness of the metric data. The monaural prediction values p_{mon} and the results p_{correct}

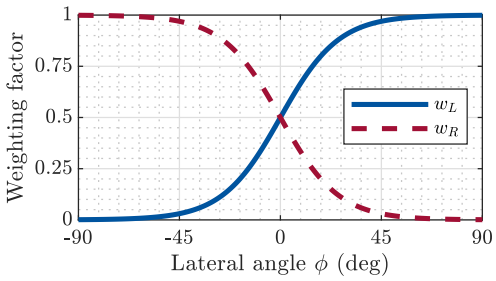

Fig. 3: Binaural weighting curves used in the `wMean` condition for averaging over monaural descriptors of the left and right ear (after [3].)

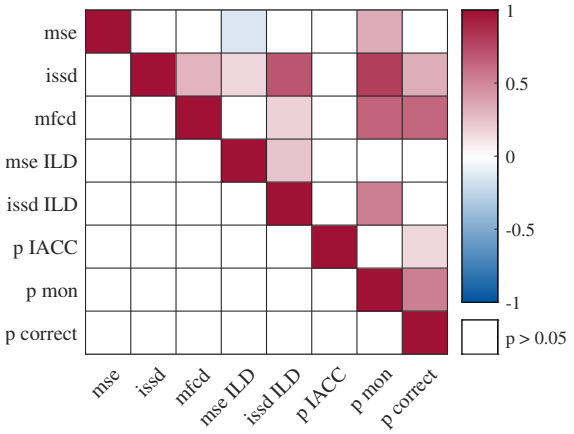

Fig. 4: Pearson correlation coefficients r for the descriptors and experimental output p_{correct}. Left and right ear values of monaural descriptors are equally weighted (`mean`). Coefficients of no statistical significance ($p > 0.05$) are omitted.

are exempted from this transform, as they already represent a good approximation of a normal distribution. p_{IACC} is further excluded from the transform due its bimodal distribution (a direct result of selecting the stimuli in two regions separated by a buffer zone around the pre-known JND). Accordingly, the correlation values for p_{IACC} may not be valid and should be treated with caution.

Mostly, similar trends are found in the correlation patterns for the three binaural weighting approaches. Results for the `mean` case are shown in Figure 4. Rather obvious is the moderate to strong correlation between the monaural model output p_{mon} and the monaural metrics from which it is calculated (especially `issd` with $r = 0.797$, and `mfcd` with $r = 0.641$). Another moderate to strong correlation of $r = 0.706$ between the `issd` and $\texttt{issd}_{\text{ILD}}$ is explicable in that both metrics are based on spectral variance calculation. Furthermore, the spectral progress of the ILD is implicitly affected by the monaural cues present at the two ears. Accordingly, the descriptors $\texttt{issd}_{\text{ILD}}$ and p_{mon} are also moderately correlated ($r = 0.534$).

4.5 Model variants

The presence of correlated descriptors results in three potential descriptor combinations, see `i-iii` in Table 1. Pairs of descriptors that are mutually exclusive to each other are accounted for in these sub-sets. E.g., p_{mon} is omitted in models `i` and `iii`, and the three monaural metrics are omitted from model `ii`. For direct assessment of the need for binaural descriptors, two additional control models (`iv` and `v`) are introduced. Another degree of freedom is given by the three binaural weighting approaches for monaural descriptors. Furthermore, in addition to global modeling based on all stimuli, two separate models, based only on data points below (`belowIACC`) and above the IACC JND (`aboveIACC`) are considered.

Applying the linear modeling approach to the N descriptors leads to $N+1$ linear coefficients. Using a backward step-wise linear regression approach, the least significant coefficients are successively excluded, until $p < 0.05$ is fulfilled for all model components. The resulting descriptors for the three cases are shown in Table 2.

As can be seen, the subsets are reduced to a maximum of four variables per model type. Neither of the binaural descriptors related to the ILD provide a significant contribution, whereas p_{IACC} is only consistently present when the training data considers all stimuli (both below and above the IACC JND). In contrast, the stepwise regression approach completely discards p_{IACC} in both the `belowIACC` and `aboveIACC` model variants, leading to purely monaural descriptors as input. While expected for data below the IACC JND, this behavior is rather surprising for `aboveIACC` models, but could be explained by the small variation in p_{IACC} values covered *within* the two ranges.

Generally, a strong dominance of the `mfcd` metric (and of the monaural predictions p_{mon} in absence of the `mfcd`) is observed in all cases.

Table 1: Feasible combinations of (uncorrelated) descriptors for linear modeling (`i`-`iii`) and purely monaural input for control models (`iv`,`v`). Note that the sub-sets will be further reduced using step-wise linear regression.

	mse	issd	mfcd	mse ILD	issd ILD	p IACC	p mon
i	x	x	x	x	-	x	-
ii	-	-	-	x	-	x	x
iii	x	-	x	x	x	x	-
iv	x	x	x	-	-	-	-
v	-	-	-	-	-	-	x

4.6 Model evaluation

The different model types are contrasted using two numerical measures of quality. The root mean squared error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} [p_{\mathrm{exp}} - p_{\mathrm{pred}}]^2}. \qquad (6)$$

and the linear correlation r_{pears} are used to assess the similarity between the detectability values p_{exp}, as obtained from the experiment (and transformed using eq. (4) and (5)), and their corresponding model predictions p_{pred}. In order to assess the generalizability of the model, these two quality measures are used on specified test data points, not overlapping with model training data. Within 500 iterations, $5/6$ of the available points are randomly selected for training the different model types, leaving $1/6$ as test data points for subsequent evaluation. The box plots in Figure 5 show the distribution of error values over the iterations, respectively.

Improved model performance is indicated by lower RMSE and higher Pearson correlation values. In this regard, best results are given when using a direction-dependent weighting while averaging left and right ear monaural descriptors (`wMean`), followed by simple averaging (`mean`) and maximum selection (`max`). More generally, models trained and tested on stimuli below the IACC JND threshold show superior values, followed by the global modeling (using all stimuli).

On the level of descriptors as model input, variants `ii` and `v`, in their reduced form down to p_{mon} (and p_{IACC}), show worst performance over all condition combinations. For the `belowIACC` case, two-sample t-tests on RMSE and r_{pears} values show the variants `i`, `iii` and `iv` to be significantly better than their 12 alternatives ($p < \alpha_{\mathrm{adj}} = {}^{0.05}/_{105}$ after Bonferroni correction). The quality measures of the three models, however, do not vary significantly from each other. Note that `i` and `iv` here use identical parameters (`issd` and `mfcd`) and the visible discrepancies are only due to the randomized selection of training and test stimuli. The merged datasets of these two cases result in an $\mathrm{RMSE}_{\mathrm{test}}$ as small as $(\mu \pm \sigma) = (15.72 \pm 2.1)\%$ and an r_{pears} of (0.71 ± 0.1), which is slightly (yet not significantly) better than for variant `iii`.

Though the `belowIACC` case shows best results, the test stimuli, here, all lie strictly below the IACC JND – a condition generally not given for arbitrary HRTF input. The more generalizable global modeling case is therefore considered. For the `wMean` condition, the step-wise reduction of `i`, `iii` and `iv` leads to retaining `mfcd`, in addition to `issd` and/or p_{IACC}. In terms of RMSE values, the three models do not differ significantly. Correlation values r_{pears}, however, show `i` to be superior to all other models except model variant `iv`, which on its part does not differ significantly from `iii`. This indicates a small tendency for superiority of model `i` in the `wMean` case, with an $\mathrm{RMSE}_{\mathrm{test}}$ of $(\mu \pm \sigma) = (15.54 \pm 1.53)\%$ and r_{pears} of (0.68 ± 0.08).

5 Discussion

The presented JND model for bi-lateral HRTF variation puts a special focus on spectral magnitude differences. All descriptors, except for p_{IACC}, process solely magnitude information. Evidently, the phase flows into the IACC calculation. However, since IPDs are fixated for all stimulus pairs, the only phase changes present are due to the minimum phase component – which stands in direct relation to the magnitude spectrum.

In practice, a switch between two generic HRTF sets would likely entail changes to the ITD as well, which would certainly contribute to distinguishability of stimuli. As the JND of ITDs has been widely examined [20, 21], however, it is purposefully not included within the covered descriptors. Future studies could examine possible interactions and the relative importance of phase- and magnitude-related factors.

The derived models are intended to be representative for various types of HRTF variations. The experimental

Table 2: Reduced descriptor combinations after backward step-wise linear regression. The three binaural weighting approaches are indicated by the symbols ○ (mean), ● (wMean) and ∧ (max).

	All Stimuli							Below IACC JND							Above IACC JND						
	mse	issd	mfcd	mse ILD	issd ILD	p IACC	p mon	mse	issd	mfcd	mse ILD	issd ILD	p IACC	p mon	mse	issd	mfcd	mse ILD	issd ILD	p IACC	p mon
i	○∧	○●∧	○●∧	-	-	○●∧	-	-	●	○●∧	-	-	-	-	∧	○●∧	○●∧	-	-	-	-
ii	-	-	-	-	-	○●∧	○●∧	-	-	-	-	-	-	○●∧	-	-	-	-	-	-	○●∧
iii	○∧	-	○●∧	-	-	○●∧	-	-	-	○●∧	-	-	-	-	○∧	-	○●∧	-	-	-	-
iv	○∧	○●	○●∧	-	-	-	-	-	●	○●∧	-	-	-	-	∧	○●∧	○●∧	-	-	-	-
v	-	-	-	-	-	-	○●∧	-	-	-	-	-	-	○●∧	-	-	-	-	-	-	○●∧

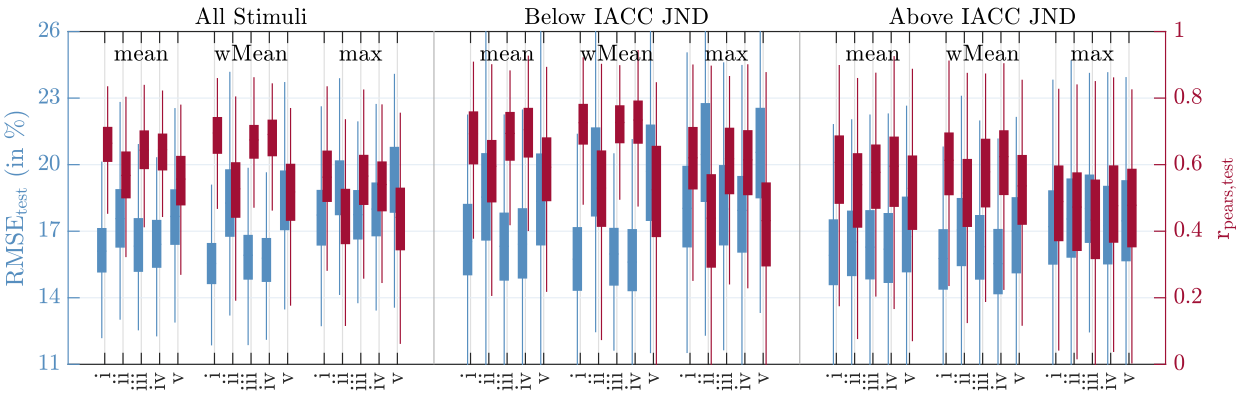

Fig. 5: Evaluation of model quality based on RMSE and Pearson correlation of experimental and predicted values (p_{detect}). Each box plot represents 500 iterations of randomly selecting test ($1/6$) and training data subsets ($5/6$ of total points.)

design, accordingly, integrates a comparison of HRTF datasets with different degrees of detail loss as well as typical inter-individual HRTF differences. Nonetheless, it can be observed in the correlation pattern (Fig. 4) that values of the `mse` metric are only slightly correlated ($r = 0.348$) with the monaural prediction values (p_{mon}). This indicates that the range of `mse` values is barely exploited by the choice of stimuli. The metric is furthermore not (significantly) correlated with the correct response rate (p_{correct}), which explains its elimination within the step-wise regression for more than half of the model variants.

In the previous study [6], the relevance of the `mse` metric for the monaural JND model was enforced by representing the full space spanned by all three monaural descriptors in the selection of stimuli. Here, however, a less constrained approach for stimulus selection, only considering p_{correct}, seems to support the relatively lower impact of the `mse` metric. In the context of the present study, a stimulus choice not covering a useful range of this metric should not be considered as an indicator for lack of variation in HRTF error types, since multiple conditions were set in the experimental design.

Further implications arise from the way IACC transitions were integrated in stimulus selection. The approach was based on the initial assumption that this property was to be considered as binary information. E.g., it was presumed that IACC transitions would be omitted from modeling in the sub-threshold case. This

assumption has been confirmed, as p_{IACC} exceeded the significance level in the step-wise linear regression in the `belowIACC` models. When training the model using all stimuli, however, p_{IACC} was retained as a descriptor in some model variants. Thus, the safety margin, introduced between the sub- and supra-threshold regions, see Figure 1, may in that case have negatively affected the achievable model quality. A representation of data points close to the IACC JND within the training data could provide more precise predictions.

The quality measures support a final model based direction-dependent weighting of metrics `mfcd` and `issd`, in addition to p_{IACC}. Since the latter only contributes to a "tendency" for enhancement in predictions, it is questionable, whether it is worth the increase in model complexity. Nonetheless, if the value turns out to fall below the JND threshold, a model specifically optimized for the `belowIACC` case may offer a more significant improvement.

Finally, the model performance reported above shows to be somewhat below that of the monaural model, where an RMSE value of $(\mu \pm \sigma) = (14.31 \pm 1.91)\%$ and a correlation coefficient r_{pears} of (0.75 ± 0.09) were achieved for the test data [6]. In both cases, one can assume that the predictions of detectibility indicate "tendencies", rather than safe predictions. Nonetheless, the monaural model was of great benefit in the context of stimulus selection in the present work. In the previous study, the fact that stimulus selection was based on three individual numerical descriptors had led to restrictions in the pool of HRTFs to be used. E.g., the `measured` database had to be excluded from the subjective evaluation, since sub-threshold stimulus pairs could not be identified easily. The tendencies resulting from the present model could be equally valuable, depending on the context of use. E.g., they could be applied in a comparable scenario, selecting stimuli for a JND experiment involving more complex acoustic scenes. In the direct evaluation of HRTF approximation methods, the model could still serve as an indicator for perceivable similarity – that is yet to be used with caution.

6 Summary

This work aimed at deriving JNDs of bi-laterally varied HRTF sets, with focus on changes in the magnitude spectrum. A 3-AFC experiment was conducted, assessing the detectibility of changes in the spectra, while retaining a fixed ITD. A uni-lateral JND model, also focusing on magnitude-related variations, was integrated into the stimulus selection process and allowed for identifying sub-threshold transitions, even for the intricate case of comparing measured HRTFs of different individuals. In the subsequent modeling steps, however, the monaural model output proved to be less effective in estimating a bi-lateral JND.

Instead, the metric `mfcd` showed to be dominant in the models and, accordingly, well-suited for predicting audible differences. The use of weighted means (according to [3]) of the left and right ear monaural descriptors was further shown to improve model performance.

In terms of binaural cue descriptors, p_{IACC} seems to be of higher importance for JND modeling, compared to the ILD – which however may be due to the choice of the two employed ILD-related descriptors. These descriptors may be perceptually less relevant for near-threshold variations.

The acquired perceptual data relates to free-field HRTFs convolved with pulsed noise signals. The experiment can therefore be considered as a worst-case situation, where differences between HRTF spectra are emphasized. A more lenient JND threshold can be expected for a different choice of raw signals (e.g., music or voice recordings) and for more complex acoustic scenes. The derived model could be of use in designing experimental settings to examine JNDs in such complex scenarios. E.g., it could be applied in stimulus selection, evaluating variations in both direct and reflected sound components – which would be partly masked given the temporal structure of the binaural room impulse response.

In overall, the model yields tendencies for detectability, rather than safe predictions. However, it comes with the advantage of being rather simple. A more complex approach for modeling may be able to provide a more accurate estimation – a trade-off that needs to be considered in future work.

7 Acknowledgment

This research was funded by the German Research Foundation (DFG), project no. 402811912. The authors would like to thank Liang Shang for conducting the experiment, and all the participants for their time.

References

[1] Andreopoulou, A., Begault, D. R., and Katz, B. F., "Inter-laboratory round robin HRTF measurement comparison," *IEEE Journal of Selected Topics in Signal Processing*, 9(5), pp. 895–906, 2015.

[2] Nicol, R., Lemaire, V., Bondu, A., and Busson, S., "Looking for a relevant similarity criterion for HRTF clustering: a comparative study," in *Audio Eng. Soc. Convention 120*, 2006.

[3] Baumgartner, R., Majdak, P., and Laback, B., "Modeling sound-source localization in sagittal planes for human listeners," *J. Acoust. Soc. Am.*, 136(2), pp. 791–802, 2014.

[4] McKenzie, T., Armstrong, C., Ward, L., Murphy, D. T., and Kearney, G., "Predicting the colouration between binaural signals," *Appl. Sci.*, 12(5), p. 2441, 2022.

[5] Kohnen, M., Bomhardt, R., Fels, J., and Vorländer, M., "Just noticeable notch smoothing of head-related transfer functions," in *Fortschritte der Akustik - DAGA 2018*, ISBN 978-3-939296-13-3, DEGA, Berlin, 2018.

[6] Doma, S., Ermert, C. A., and Fels, J., "A magnitude-based parametric model predicting the audibility of HRTF variation," *J. Audio Eng. Soc.*, 2023, doi:10.17743/jaes.2022.0080.

[7] Middlebrooks, J. C., "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Am.*, 106(3), pp. 1480–1492, 1999.

[8] Shimada, S., Hayashi, N., and Hayashi, S., "A Clustering Method for Sound Localization Transfer Functions," *J. Audio Eng. Soc.*, 42(7/8), pp. 577–584, 1994.

[9] Pollack, I. and Trittipoe, W., "Binaural listening and interaural noise cross correlation," *J. Acoust. Soc. Am.*, 31(9), pp. 1250–1252, 1959.

[10] Masiero, B. and Fels, J., "Perceptually robust headphone equalization for binaural reproduction," in *Audio Eng. Soc. Convention 130*, 2011.

[11] Bomhardt, R., de la Fuente Klein, M., and Fels, J., "A high-resolution head-related transfer function and three-dimensional ear model database," *Proc. Mtgs. Acoust*, 29(1), p. 050002, 2016, doi: 10.1121/2.0000467.

[12] Hwang, S. and Park, Y., "Interpretations on principal components analysis of head-related impulse responses in the median plane," *J. Acoust. Soc. Am.*, 123(4), pp. EL65–EL71, 2008.

[13] Bomhardt, R. and Fels, J., "Individualization of head-related transfer functions by the principle component analysis based on anthropometric measurements," *J. Acoust. Soc. Am.*, 140(4), pp. 3277–3277, 2016, doi:10.1121/1.4970411.

[14] Kimura, D., "Cerebral dominance and the perception of verbal stimuli," *Can. J. Psychol.*, 15(3), p. 166, 1961.

[15] King, F. L. and Kimura, D., "Left-ear superiority in dichotic perception of vocal nonverbal sounds." *Can. J. Psychol.*, 26(2), p. 111, 1972.

[16] Bomhardt, R., Lins, M., and Fels, J., "Analytical Ellipsoidal Model of Interaural Time Differences for the Individualization of Head-Related Impulse Responses," *J. Audio Eng. Soc.*, 64(11), pp. 882–894, 2016, doi:10.17743/jaes.2016.0041.

[17] Morimoto, M., "The contribution of two ears to the perception of vertical angle in sagittal planes," *J. Acoust. Soc. Am.*, 109(4), pp. 1596–1603, 2001.

[18] Macpherson, E. A. and Sabin, A. T., "Binaural weighting of monaural spectral cues for sound localization," *J. Acoust. Soc. Am.*, 121(6), pp. 3677–3688, 2007.

[19] Stewart, G. W., "Collinearity and least squares regression," *Stat. Sci.*, 2(1), pp. 68–84, 1987.

[20] Aussal, M., Alouges, F., and Katz, B., "HRTF interpolation and ITD personalization for binaural synthesis using spherical harmonics," in *Audio Eng. Soc. Conference: UK 25th Conference: Spatial Audio in Today's 3D World*, 2012.

[21] Bomhardt, R., Mejía, I. C. P., Zell, A., and Fels, J., "Required Measurement Accuracy of Head Dimensions for Modeling the Interaural Time Difference," *J. Audio Eng. Soc.*, 66(3), pp. 114–126, 2018, doi:10.17743/jaes.2018.0005.