S. Doma, C. A. Ermert and J. Fels, "A Magnitude-Based Parametric Model Predicting the Audibility of HRTF Variation"
J. Audio Eng. Soc., vol. 71, no. 4, pp. 155–172, (2023 April).
DOI: https://doi.org/10.17743/jaes.2022.0080

A Magnitude-Based Parametric Model Predicting the Audibility of HRTF Variation

SHAIMAA DOMA, *AES Student Member*, **COSIMA A. ERMERT, AND JANINA FELS**
(sdo@akustik.rwth-aachen.de) (cer@akustik.rwth-aachen.de) (jfe@akustik.rwth-aachen.de)

Institute for Hearing Technology and Acoustics, RWTH Aachen University, Aachen, Germany

This work proposes a parametric model for just noticeable differences of unilateral differences in head-related transfer functions (HRTFs). For seven generic magnitude-based distance metrics, common trends in their response to inter-individual and intra-individual HRTF differences are analyzed, identifying metric subgroups with pseudo-orthogonal behavior. On the basis of three representative metrics, a three-alternative forced-choice experiment is conducted, and the acquired discrimination probabilities are set in relation with distance metrics via different modeling approaches. A linear model, with coefficients based on principal component analysis and three distance metrics as input, yields the best performance, compared to a simple multi-linear regression approach or to principal component analysis–based models of higher complexity.

0 INTRODUCTION

Recent advances in audio playback technology have made high-quality sound accessible to the masses. A variety of applications, such as video games or home theater systems, benefit from the proper design and equalization of loudspeaker systems and headphones, allowing for better immersion in the spatial acoustic environment [1]. With this development—and with the increasing awareness and expectations of users—spatial reproduction is further optimized and tailored to the individual listener. Individualized head-related transfer functions (HRTFs) approximate the listener-specific sound directivity, which provides listeners with their learned cues for auditory localization of spatial sources. However, the available approximation methods are accompanied by a potential deterioration in perceived quality.

For a better understanding of the impact on quality, a holistic approach is required to relate auditory perception to a wide range of possible spectral errors, e.g., a loss of spectral or spatial detail, the presence of non-individual spatial cues, changes in sound level and/or coloration, or a combination of these (or other) factors. Several studies have examined the detectability of manually applied spectral changes, e.g., these changes included peaks and notches of different form and depth introduced to the spectra [2], a stepwise variation of inter-aural time differences (ITDs) [3, 4], or different degrees of notch smoothing applied to HRTFs [5].

However, to properly predict the audibility and perceptual effects of not only such controlled HRTF differences, but also arbitrary spectral errors, a generalized metric is needed. For example, in the context of developing HRTF individualization methods and spatial interpolation techniques, or of direct comparison between HRTF acquisitions methods and setups, the nature of underlying spectral deviations is not specifically known a priori. Studies performing comparisons of this kind generally follow one of the following approaches:

In the first approach, different levels of "pseudo"-arbitrary variations are defined in relation to a specific parameter. This can, for example, be a property of a filter used for HRTF reconstruction. In that case, the detection thresholds acquired in a listening experiment describe a required resolution of the parameter [6]. Although this information is relevant for the specific application, it is not possible to generalize for arbitrary HRTF differences, unless these differences can be approximately described as a filter degradation. Therefore, the second—more flexible—approach draws conclusions based on a selection of distance metrics from literature. These metrics may, on their part, be more or less suitable for capturing certain differences.

In the context of localization errors, metrics have been analyzed regarding how they scale with a given angular error [7]. Numerous binaural models have further integrated machine learning [8] and concepts of auditory cognition, quantifying the likelihood of misinterpreting a stimulus as coming from a deviant sound incidence direction [9–11].

Whereas spatial infidelities largely impact perceived playback quality, timbral infidelities have been reported to be even more detrimental [12], and these timbral differences would not be accurately represented by models targeting localization cues. Moreover, the spectral properties affecting the different perceptual attributes are closely linked. For example, the incidence angle directly influences sound color (as demonstrated in the directional bands theory [13]), entailing differences of up to 10 dB in narrow-band loudness sensitivity [14]. Given this complexity, it becomes evident that HRTF accuracy cannot be described on a unidimensional scale. An overarching model that scales with HRTF differences should incorporate multiple models and retain a multi-dimensional nature in its output.

The perceptual validation of such a model proves to be an intricate task. Especially higher levels of HRTF errors would need to be correctly mapped by the listeners to specific perceptual attributes. This requires them to have a proper (and common) understanding of appropriate vocabulary [15]. For example, in [16], the authors make do with the general term of "discoloration" of stimuli without addressing more detailed timbral descriptors. In the experiment, participants grade different degrees of discoloration. A model then receives two-channel spatial signals and predicts the degree of binaural discoloration as the mean of two separately calculated monaural values. Although the model integrates both ear channels, it omits a representation of binaural interaction and, more specifically, does not identify separate contributions of monaural and binaural cue differences. In another study on the degree of timbral and localization changes [17], ITDs are included, in addition to simple summation of the two separately calculated monaural errors.

The mentioned studies focus on error grading, while not explicitly identifying just noticeable differences (JNDs) for the perceptual descriptors. In fact, the question of mere detectability is rather detached from such definitions. Although it can be assumed that the different aspects of perceptual dissimilarity contribute to discrimination of stimuli, they may only be consciously identifiable for suprathreshold stimuli. Thus, they need not be considered in the error range close to the discrimination threshold.

Instead, a "generalized" JND would serve as a "worst-case" limit for permissible spectral deviations. Below this minimum required resolution, further HRTF optimization would no longer be of value. This is the goal of the present study: to examine the JND for HRTF dissimilarity given generic differences, i.e., no artificial degradation of the filters. A previous attempt for predicting discrimination thresholds for generic stimuli was done in [18] in the context of piano signals played on different instruments. Experimental results from a three-alternative forced choice (3AFC) discrimination task in noise provided thresholds in the form of SNR values. A single-channel (monaural) and physiologically motivated dissimilarity model was developed; its output achieved "moderate to high correlation" with the experimental data.

In the present work, the authors decided against applying a similar approach. The binaural nature of HRTF sets brings about an inherent complexity to the design of a suitable listening test paradigm. Possible interactions between the background noise and the already complex spatial and timbral percepts may affect the experimental results. Moreover, binaural unmasking effects [19] would need to be accounted for, based on the decision to use either inter-aurally identical or uncorrelated noise signals.

On this account, the present study targets a direct comparison of stimuli convolved with free-field HRTFs. In the range of small errors, spectral dissimilarity is regarded as binary information with a probability of occurrence for the two states: perceptually "similar" and "dissimilar." These probabilities are to be modeled, as a first step, for unilateral HRTF variation, i.e., only one of the two ear signals being varied. This limitation to single-channel JND components serves a later extension to a binaural JND model, in which monaural and binaural contributions are identified separately.

In the error range where the JND is to be expected, a modeling of higher-level cognitive representation of auditory percepts is not essential. Instead, seven signal-near metrics are used, with the goal of predicting the probability of detection for different kinds of spectral deviations. As shown in previous work by the authors [20], a preliminary assessment of distance metric interrelation indicates a variation in correlation patterns depending on the type of compared datasets. Here, a selection of seven metrics (not identical to the previous work) is analyzed numerically through both correlation and factor analysis, identifying common trends in their reactions to spectral errors. A thus-selected subset of metrics is employed in the design of a JND experiment for unilateral spectral deviations, which provides perceptual data for modeling the audibility of errors. Different modeling approaches and their performance are contrasted, upon which an optimal model is identified.

The paper is structured as follows: SEC. 1 introduces the magnitude-based distance metrics and the HRTF datasets employed for metric analysis. SEC. 2 follows an objective evaluation of distance metric interrelation, with the goal of defining pseudo-orthogonal subsets of metrics. The latter observations find application in SEC. 3, in which the paradigm for a JND listening experiment is designed and first experimental results are presented. These results form the basis for subsequent modeling of the discrimination probabilities in SEC. 4. Finally, SEC. 5 contextualizes the findings, followed by a brief summary and an outlook in SEC. 6.

1 MATERIALS

1.1 Distance Metrics

HRTFs provide information on incident sound arriving at both ears for a broad range of frequencies and directions. Due to high dimensionality, the task of identifying differences between HRTF sets is not straight-forward. Various distance metrics enable a direct comparison of two HRTF sets, summarizing the differences by means of a scalar

quantity or vector. In the present work, metrics with the following key aspects were selected:

- They quantify unilateral differences, i.e., they contrast data of a single ear in two HRTF sets through single-channel calculations.
- They focus on magnitude deviations, disregarding phase information. All calculations are therefore performed in the frequency domain, rather than on impulse responses.
- They provide scalar dissimilarity values in the directional domain, i.e., for each direction k defined by elevation angle θ and azimuth angle φ.

Four of the presented metrics incorporate psychoacoustic concepts to provide a closer relation to perception.

A well-established method for digital filter comparison is the Mean Squared Error (MSE). In the context of HRTFs [7], it is computed for each direction k by averaging the squared spectral difference over frequency bins $i \in [1, \mathrm{N}]$ as

$$\mathrm{MSE}(k) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} [\mathrm{HRTF}_1(k,i) - \mathrm{HRTF}_2(k,i)]^2. \quad (1)$$

Here, each frequency is equally weighted in the difference measure. Because of the non-linear nature of the frequency resolution of the human ear, errors in high frequencies may be overvalued in this metric, compared to low-frequency differences. The critical band model provides a mathematical description of the frequency processing in the auditory system [21]. As proposed in [7], this property can be incorporated into the metric by means of a frequency dependent weighting factor $\alpha(i)$, leading to the Critical-Band (CB) MSE:

$$\mathrm{CB}(k) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} (\alpha(i)[\mathrm{HRTF}_1(k,i) - \mathrm{HRTF}_2(k,i)])^2. \quad (2)$$

Analogous to the auditory frequency resolution, the weighting factor decreases for higher frequencies and is computed as

$$\alpha(i) = \frac{1}{\alpha_0 \Delta f_{\mathrm{CB}}(i)}, \quad (3)$$

with $\Delta f_{\mathrm{CB}}(i)$ being the critical bandwidth for each frequency bin i and the normalization value α_0 defined as

$$\alpha_0 = \sum_{i=1}^{N} \frac{1}{f_{\mathrm{CB}}(i)}. \quad (4)$$

Because both the MSE and CB rely on absolute differences, they will naturally yield larger dissimilarity values for directions of larger magnitude, e.g., for ipsilateral incidence. However, human perception of magnitude differences is not linear but approximately logarithmic [21]. As a consequence, differences at low magnitudes, e.g., at notches, can have a higher perceptual impact than would be captured on a linear scale. The MSE and CB may offer no adequate representation of such perceptual subtleties. Therefore, to counteract the influence of absolute magnitudes, a variation of the metrics is introduced: the logarithmic squared errors

$$\mathrm{MSE}_{\log}(k) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \log_{10}\left[\frac{\mathrm{HRTF}_1(k,i)}{\mathrm{HRTF}_2(k,i)}\right]^2 \quad (5)$$

and

$$\mathrm{CB}_{\log}(k) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \alpha(i) \log_{10}\left(\left[\frac{\mathrm{HRTF}_1(k,i)}{\mathrm{HRTF}_2(k,i)}\right]\right)^2. \quad (6)$$

Another metric performing squared error calculations is the Mel Frequency Cepstral Distortion (MFCD) [22]. Here, the HRTF spectrum is divided into $N_B = 24$ Mel bands by means of a gammatone filter-bank. Mel bands offer a linear scale that links physical frequencies to perceived pitch [21]. So-called Mel-frequency cepstral coefficients (MFCCs) are obtained by performing a discrete cosine transform on the energy within each Mel band [23]. Similarly to the MSE, the MFCD is defined as

$$\mathrm{MFCD}(k) = \frac{1}{\mathrm{N}_B} \sum_{n=1}^{\mathrm{N}_B} [\mathrm{MFCC}_1(\mathrm{k,n}) - \mathrm{MFCC}_2(\mathrm{k,n})]^2. \quad (7)$$

The variance of logarithmic magnitude differences is evaluated in the Inter-Subject Spectral Difference (ISSD), as introduced in [24]. This distance metric is defined for frequency bins up to 13 kHz. Originally, the metric uses directional transfer functions (DTFs), which are computed by omitting the information common to all directions of an HRTF set. This, however, makes the spectrum DTF(k) dependent on the HRTF spectra of all directions. As the current work targets a direct comparison of individual spectra, the original HRTF spectra are here employed instead (cf. [25]), yielding

$$\mathrm{ISSD}(k) = \sigma_i^2 \left(20 \cdot \log_{10}\left[\frac{\mathrm{HRTF}_1(k,i)}{\mathrm{HRTF}_2(k,i)}\right]\right), \quad (8)$$

with σ_i^2 being the variance over frequency and $f(i) < 13$ kHz. Finally, the Loudness Level Spectral Error (LLSE) is designed to capture coloration differences between HRTF. Based on [17], a pink noise signal is convolved with the two HRTF spectra. Subsequently, the loudness levels L_L of the resulting signals are computed in phon for each of the $m = 1 \ldots \mathrm{N}_E$ equivalent rectangular bandwidth filters [21], respectively. An error value for each direction is then obtained by evaluating the variance σ_m^2 of the difference in loudness levels:

$$\mathrm{LLSE}(k) = \sigma_m^2(\mathrm{L}_{L1}(m,k) - \mathrm{L}_{L2}(m,k)). \quad (9)$$

Because the computational procedures of the distance measures differ substantially from each other, it can be expected that different "types" of dissimilarity will be evaluated depending on the metric in use. For example, incorporating the frequency selectivity of the human auditory system (e.g., CB, $\mathrm{CB}_{\log}$, MFCD, and LLSE) avoids over-representation of errors in high frequencies. A brief overview of metric properties is summarized in Table 1.

Table 1. Overview of the selected distance metrics (in alphabetical order) and their main features.

Abbrev.	Arithmetic operation	Psychoacoustic model	Scale
CB	Squared mean	Critical band weights	Linear
CB_{log}	Squared mean	Critical band weights	Log
ISSD	Variance	···	Log
LLSE	Variance	ERB filters	Log
MFCD	Squared mean	Mel filters	Log
MSE	Squared mean	···	Linear
MSE_{log}	Squared mean	···	Log

ERB, equivalent rectangular bandwidth.

1.2 HRTF Data Sets

The analysis of the distance metrics is performed on the basis of three HRTF datasets, in the following termed: the "measured," "idealPCA," and "anthroPCA" datasets.

The "measured" set was directly extracted from the ITA HRTF database [26], in which acoustically acquired HRTFs at a resolution of $5^\circ \times 5^\circ$ as well as anthropometric features are provided for 47 individuals. For the "idealPCA" dataset, the "measured" HRTFs were approximated using Principal Component Analysis (PCA) reconstruction [27]. In this individualization approach, HRTF spectra are depicted as a weighted sum of Principal Components (PCs), with "ideal" weights provided for HRTF spectra originally used as input for the analysis. With increasing number of PCs, the discrepancy between reconstructed and original HRTF spectra in terms of ISSD error has been reported to reach a saturation level at 15 PCs, with no further improvement achieved by higher complexity [28]. In this work, however, the reconstruction is performed with 23 PCs, as the sufficiency of 15 PCs has not been validated with other distance metrics. The "anthroPCA" dataset was generated by approximating the ideal weights using a linear combination of six anthropometric features: "h," "w," "du," "df," "d6," and "d8," according to definitions in [26].

Note that these approximated HRTF spectra initially do not include phase information. All distance metrics received only magnitude spectra as input. For use in the perceptual experiment, additional phase spectra were reconstructed (cf. SEC. 3.2.3).

The simulated datasets ("ideal-" and "anthroPCA") were mainly chosen for this paper to enable the generation of both small and large error values. When comparing measured HRTFs from different individuals, distance metrics oftentimes yield large error values. This is not desirable here, given the aim of examining JND thresholds. Small spectral deviations would allow for providing both sub-threshold and supra-threshold stimuli in an experimental setting and would further lead to a more precise JND. In fact, the inclusion of less detailed spectra with, accordingly, less prominent differences, facilitated the stimulus selection for the subjective evaluation of distance metrics (see SEC. 3). Another motivation behind this selection of HRTF sets was to avoid the controlled manipulation or "degradation" [29] of spectra.

This should allow for more degrees of freedom and, thus, presumably more realistic variations in the HRTF spectra. An alternative to generating small errors by approximation would be to compare HRTF sets obtained from repeated measurements [30] of the same individual. This approach was, however, dismissed due to the extensive measurement effort required.

In the following, a distinction is made between inter-individual and intra-individual comparisons. In literature, intra-individual differences typically refer to the comparison of left and right HRTFs from a single dataset. In contrast, the term is here used to describe the comparison of left-ear HRTF spectra of different datasets ("measured," "idealPCA," and "anthroPCA") belonging to the same individual. For inter-individual comparison, equilateral HRTF spectra of different individuals are compared. When performing inter-individual comparisons, the left-ear HRTF of ID1 from the ITA HRTF database was used as a reference. Because of excessive data (2,304 directions per member), data from half of the database (24 subjects) were deemed sufficient for the present work.

2 NUMERICAL EVALUATION

The following sections observe the reaction of distance metrics to inter-individual and intra-individual differences between HRTFs. A special emphasis is laid on datasets eligible for an evaluation of audibility, see SEC. 3. Early informal listening led to the conclusion that comparisons involving the "measured" dataset were for the most part clearly distinguishable and therefore not suited for the planned JND experiment. The following analysis therefore focuses mainly on inter-individual comparisons within the "idealPCA" and "anthroPCA" datasets, respectively, and intra-individual comparisons between the two. To provide an anchor point for metric values, inter-individual comparisons within the "measured" dataset are also included.

2.1 Relative Value Ranges

Fig. 1 (top) visualizes the metric values for the four presented cases, i.e., inter-individual comparisons within the "measured," "idealPCA," and "anthroPCA" datasets as well as intra-individual comparison between the reconstructed HRTFs of ID 1 in the "idealPCA" and "anthroPCA" datasets, respectively. For better visibility of trends, each distance metric is normalized by its 97th percentile, with normalization factors calculated collectively over all four cases.

Inter-individual comparisons, i.e., cases in Figs. 1(a)–1(c), show most prominent reactions for metrics MSE_{log}, MFCD, and LLSE. The ISSD metric is furthermore strongly affected by the detailed non-individual cues between "measured" HRTFs of different individuals. Especially cases in Figs. 1(a) and 1(b) dominate the normalization factors for these metrics, leading to very small values in the inter-individual comparison case of Fig. 1(d). On the other hand, the case in Fig. 1(d), which contrasts the "anthroPCA" and "idealPCA" datasets, dominates the value range for metrics

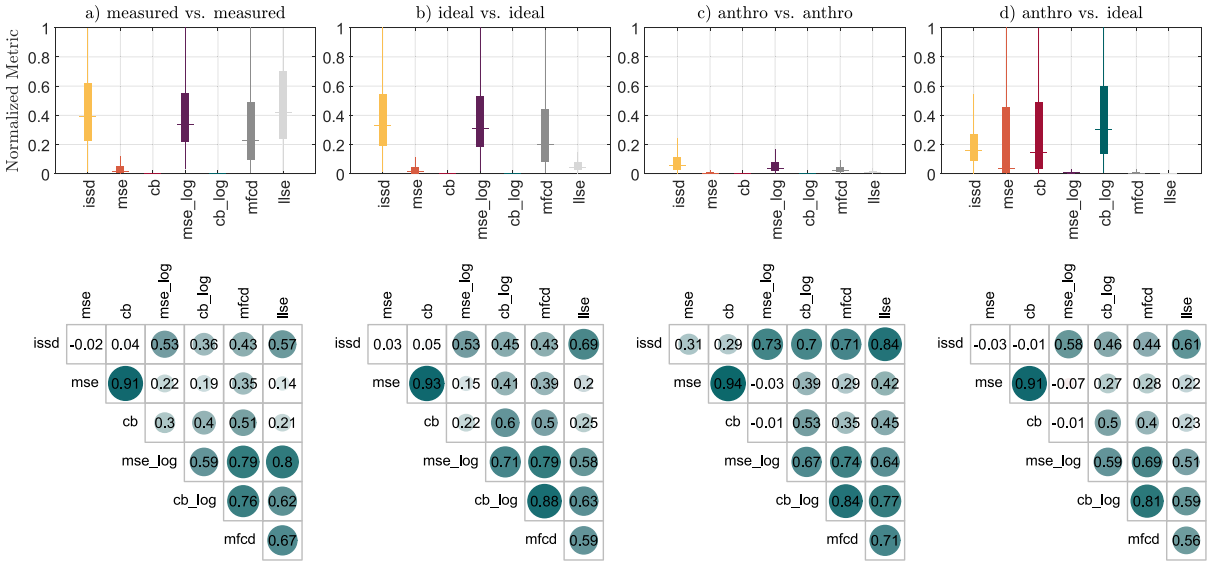

Fig. 1. Top: Relative value ranges of the distance metrics, normalized to their respective 97th percentile. Bottom: Bivariate Pearson correlation of log-transformed data points. The three cases (a)–(c) represent inter-individual comparisons between HRTF spectra of ID 1 and IDs 2–20 within the same dataset ("measured," "idealPCA," and "anthroPCA"), respectively. Case (d) represents the intra-individual comparison between the "anthroPCA" and "idealPCA" datasets, limited to HRTF spectra of ID 1.

MSE, CB, and CB_{log}. Smallest metric reactions are found for the case in Fig. 1(c), in which substantial loss of detail in reconstruction attenuates the differences between the non-individual cues being approximated.

2.2 Interrelation Analysis

Besides examining common trends in value ranges, mutual information and possible pseudo-orthogonal behavior of the distance metrics was investigated based on both intra-individual and inter-individual comparisons. The goal was to reduce the available metrics to a subset that would likely capture a variety of HRTF differences with little redundancy. The analysis focused on directional data, with each value contrasting a pair of HRTF spectra in a certain direction, rather than a whole set of spatial transfer functions. Calculations were performed using R (version 4.0.2) and RStudio (version 1.3.1073).

2.2.1 Correlation Analysis

Bivariate Pearson correlation [31] was first calculated to determine strong pairwise correlations and trends common to the different comparison cases. Given the lower zero boundary of all metrics and the presence of outliers towards high metric values, the data initially exhibited rather strong skewness, requiring a logarithmic transform for its reversal before conducting the correlation analysis.

The acquired correlation coefficients between distance metrics are displayed in Fig. 1 (bottom) for the three inter-individual and one intra-individual comparison set. Due to the large amount of data points, all correlations are significant ($p < 0.001$). The strongest and most consistent correlation is found between the metrics MSE and CB ($r > 0.90$), which otherwise (especially MSE) show the least correlation to the remaining metrics. Second highest is the correlation between MFCD and CB_{log}. Further moderate to strong correlations show a slight case dependency. Most pronounced is the contrast between the inter-individual cases and the one intra-individual case in Fig. 1(d). Variance in the correlation pattern becomes more prominent when observing ipsilateral and contralateral sound incidence directions separately, yet this differentiation is not further pursued at this point.

2.2.2 Factor Analysis

To gain further insight on how the metrics could be grouped, an Exploratory Factor Analysis (EFA) [32] was conducted as follows: The metric values were first log-transformed, since the analysis again relies on correlation matrices. Then, the datasets were tested on eligibility for the EFA by applying Bartlett's Test of Sphericity [33] and the Kaiser-Meyer-Olkin (KMO) criterion [34] as a measure of sampling adequacy. For subsets of data complying with these requirements, EFA could then be performed using a suitable number of factors as determined by the parallel analysis method [35]. Principal axis factoring was chosen, as it does not require strict multi-variate normality [36] and is preferred for exploratory analysis. Varimax rotation was applied to maximize "extreme" factor loadings.

With the seven distance metrics as a starting point, Bartlett's test of sphericity confirmed the presence of redundancy in the data ($p < 0.01$). The KMO criterion yielded "mediocre" to good values [34], ranging between 0.64 and 0.88 for most metrics. The sole exceptions lay in the MSE and CB metrics, with values as low as 0.41 to 0.51. The KMO is known to penalize pairwise grouping of variables in favor of larger groupings. According to common practice, it should lead to omission of the two concerned variables from the analysis, since they would "lead to erroneous in-

terpretation" [37]. However, as the study does not intend to create an ideal FA model but to detect tendencies for pseudo-orthogonal behavior, FA is run both before and after elimination of MSE and CB.

The analysis attempting to use $n_{\mathrm{Fac}} = 3$ factors (as suggested by parallel analysis) results in invalid models with negative specific variances for CB in all three datasets ("ultra-Heywood case"). Using $n_{\mathrm{Fac}} = 2$, two models remain invalid, whereas the case "anthroPCA" vs. "anthroPCA" successfully maps MSE and CB to Factor 1 (F1), and the remaining metrics to Factor 2 (F2). Indeed, the previously presented correlation results support the supposition that these two metrics would likely dominate the loadings for one factor and that they present a meaningful contrast to the other metrics.

After elimination of metrics MSE and CB from the FA, improved KMO values can be observed (range: [0.72, 0.88]). Again, parallel analysis suggests $n_{\mathrm{Fac}} = 3$ for the three datasets, and all analyses lead to valid models. The following common trends are found: MFCD and $\mathrm{CB}_{\mathrm{log}}$ have highest loadings for F1, ISSD and LLSE for F2, and $\mathrm{MSE}_{\mathrm{log}}$ for F3. For all three models, variable complexity is maximal for $\mathrm{MSE}_{\mathrm{log}}$. This metric is explained by 2.1 to 2.5 factors, indicating that it is not completely detached from F1 and F2 and thereby not totally separate from the other metrics.

2.3 Pseudo-Orthogonal Metric Subsets

Combining the results from the previous sections, insights on similarities in metric behavior can be gained. Firstly, the relative value ranges showed complementary behavior between groups of metrics: $\mathrm{MSE}_{\mathrm{log}}$, MFCD, and LLSE seem to react most strongly to inter-individual differences, whereas MSE, CB, and $\mathrm{CB}_{\mathrm{log}}$ are affected by intra-individual errors. The latter case is also reacted to by MFCD, if not as strongly as to the first.

Marked correlations could be found, e.g., between MSE and CB. This was to be expected, since both rely on bin-wise squared mean calculations on a linear scale. Correlation between ISSD and LLSE can be similarly attributed to the frequency domain variance calculations performed in both. The logarithmic Mel band energy in the calculation of MFCC coefficients may explain the strong correlation between the MFCD and the two metrics $\mathrm{MSE}_{\mathrm{log}}$ and $\mathrm{CB}_{\mathrm{log}}$, in contrast to their linear counterparts.

Combining these observations with the metric allocations in EFA, three groups can be identified:

- I: MSE and CB.
- II: MFCD and $\mathrm{CB}_{\mathrm{log}}$.
- III: ISSD and LLSE.

Evidently, these groups do not claim total *similarity* of the metrics included, but a tendency for a common response to spectral differences. A group thereby cannot be reduced to a single metric that explains all of its variance. For the current application, however, choosing a representative metric per group is of benefit, as a smaller number of metrics can be more easily considered for stimulus selection (see SEC. 3.2.2). In search of a generalized metric model that can handle different kinds of spectral alterations, a variety of these alterations should be included in a perceptual evaluation of the model. Stimuli to which predominantly one subgroups responds—and not the others—are likely to represent a specific category of errors.

On this account, the three groups are reduced to the following representative subset: MSE, MFCD, and ISSD. These three metrics are integrated in the experiment design, which will be described in the next section.

3 PERCEPTUAL EVALUATION

In contrast to previous studies that link distance metrics or binaural model output to physical quantities (e.g., angular displacement of HRTF incidence directions [7]), the present study aims to derive a direct connection between distance metrics and auditory distinguishability. The following sections describe a listening experiment paradigm developed for this purpose. The approach used for handling experimental output data as well as first findings on the relation between individual metrics and probabilities of stimulus discrimination are presented. These observations will serve as a foundation for modeling JNDs in the subsequent section.

3.1 Approach for Signal Presentation

The sought JND needs to be captured in an appropriately designed experimental setting. Multiple possibilities can be considered regarding the way the signals are presented and varied:

- Monaural presentation (with unilateral signal variation).
- Diotic presentation with bilateral signal variation.
- Binaural presentation with bilateral signal variation.
- Binaural presentation with unilateral signal variation.

Requirements for the present study as well as resulting potential drawbacks of the listed approaches are discussed in the following.

For single-channel filters, audibility assessment for spectral changes is a comparatively straightforward task. Unlike these simple filters, the information present in HRTFs and the auditory percepts evoked by them require special considerations. Not only are changes in coloration and/or level possible, but also a spatial difference may be perceived between two pairs of HRTFs, i.e., a relocated virtual source, an altered source width and/or degree of externalization.

It can be assumed that conscious recognition of the aforementioned percepts is mainly possible for supra-threshold (i.e., more prominent) differences. Close to the threshold, the affected perceptual aspects may not be identifiable independently. On this account, the authors do not seek separate JNDs for each perceptual attribute, but instead, a global JND, to which the perceptual attributes contribute collectively.

In this context, retaining bilateral stimulation (with plausible left and right ear HRTFs) is essential to account for the spatial component to the JND. A simple monaural playback, in which the HRTF spectra to be compared are successively presented to only one ear, with no playback at the averted ear, would therefore not be sufficient.

In the diotic case, this issue is not present. In spite of unnaturally symmetric monaural cues and the absence of binaural cue information, a spatial percept may be evoked, with a virtual source being perceived in the median plane. This scenario has the advantage that the signal variation (i.e., the switch between HRTF spectra) is identical at both ears. Similarly to the monaural case, it can therefore be attempted to describe the difference between the two HRTF spectra by means of a single set of distance metrics. Nonetheless, previous research about the perception of diotic playback leads to question its suitability in the present case:

Numerous studies have investigated a so-called "diotic advantage" [38] compared to monaural presentation, identifying a lower audibility threshold [39], lower speech reception thresholds in speech-shaped noise [40], lower detection thresholds for both amplitude and frequency modulation [19], and better distinguishability of speech processing algorithms in the context of quality assessment [41], among others. To explain this advantage, several hypotheses were put forward, that likely contribute to this effect. One theory justifies the improvement by the presence of "two independent observations" [19], which intrinsically leads to a higher probability of detection. Another theory attributes the effect to a different neural representation in absence of contralateral stimulation [42, 43]. A heightened "general attention" was further theorized in presence of bilateral stimulation [38].

Applying these theories leads to the following conclusions: The latter hypotheses relating to overall lowered sensitivity upon unilateral stimulation give further reasons against monaural playback and in favor of diotic (or binaural) presentation. However, the first theory implies, from a signal-theoretical perspective, that diotic playback would exaggerate the audibility of differences, since the (same) variation would be presented at both ears.

This leads to the third proposed approach: binaural presentation with both ear signals being varied individually. Because it is not intended to artificially "degrade" spectra according to specific parameters, this approach implies swapping two generic pairs of HRTFs within each experimental trial. Evidently, this is the most realistic scenario, which, however, introduces further complexity to the problem. The subjective percept (i.e., the ability to tell filters apart) would be influenced by two filters changing simultaneously (and, most importantly, in a non-identical way). These changes would need to be described via two different sets of distance metric values, which would flow into the selection of stimulus pairs for the experiment, as well as into later perceptual modeling.

In addition, changes in both ear signals would (unless closely monitored) introduce variations to binaural cues (i.e., to the direct difference values between the left and right ear). Certainly, the essential contribution of these binaural effects to audibility is not to be disputed. However, due to the described complexity, a separate evaluation of monaural and binaural cue differences would be of benefit as a first step.

Studying the contribution of monaural cue differences is enabled by the final suggested approach, which is applied in the present work: HRTF variation is introduced at only one ear, while an unchanged signal is maintained at the opposing ear, thus ensuring plausible binaural presentation. Simultaneously, inter-aural differences are controlled for. This is done, on the one hand, by applying identical ITD phases to the two pairs of magnitude spectra. On the other hand, the transitions in inter-aural cross-correlation (IACC) between HRTF pairs is minimized to be below the known JND for inter-aural noise cross-correlation [44], cf. SEC. 3.2.3.

Clearly, the selected approach is atypical. However, considering the drawbacks and limitations of the other approaches, it can be considered a first step towards deriving a more complex model for bilateral HRTF variations. Altogether, the arguments show this novel approach to be better suited for the given purpose.

3.2 Experimental Design

The goal of the present study is to derive a parametric model predicting audibility of unilateral HRTF differences. The input to such a model is a pair of slightly different HRTF spectra meant for one ear—which are to be varied during an experiment—while the other ear spectrum is kept constant. (See SEC. 3.2.3 for a detailed description of the stimuli.)

Accordingly, the database of feasible HRTFs needed to be limited to pairs in which the error detection threshold was to be expected. On this account, the "measured" HRTFs were discarded from the experiment, as they were, for the most part, too clearly distinguishable in both inter-individual and intra-individual comparisons. This lead to three cases, cf. Figs. 1(b)–1(d), which were integrated into the experiment design. For simplicity, the pool of HRTFs was further tightened to inter-individual comparisons between database IDs 1 and 2, and intra-individual comparisons for ID 1. Because only audibility was of relevance—and not the accuracy of source representation as, e.g., in the context of a localization task—the use of generic HRTFs from the HRTF database was deemed sufficient, irrespective of their similarity to the participants' own HRTF sets.

3.2.1 Paradigm

Audibility of a difference between two stimuli was examined in a classic three-alternative forced choice (3AFC) task. The positions of a target stimulus and two reference stimuli were balanced among the three buttons. Participants were asked to choose the one stimulus different from the other two.

As previously motivated, binaural playback was maintained in the experiment and the signal for only one ear was varied per stimulus pair, respectively. The sound incidence directions for the experiment were selected for signal

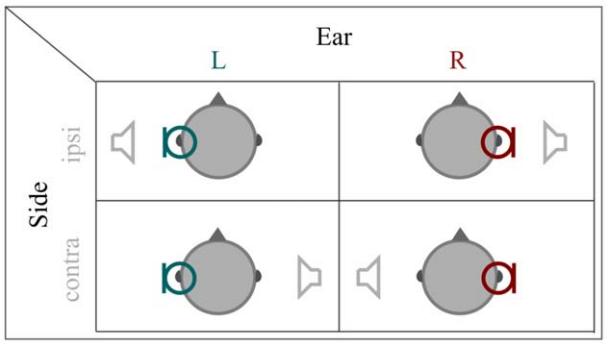

Fig. 2. Experimental conditions "Ear" × "Side." The microphones indicate the ear channel at which the signal is varied within a trial. The loudspeakers indicate the hemisphere of sound incidence relative to the ear at which the signal is varied.

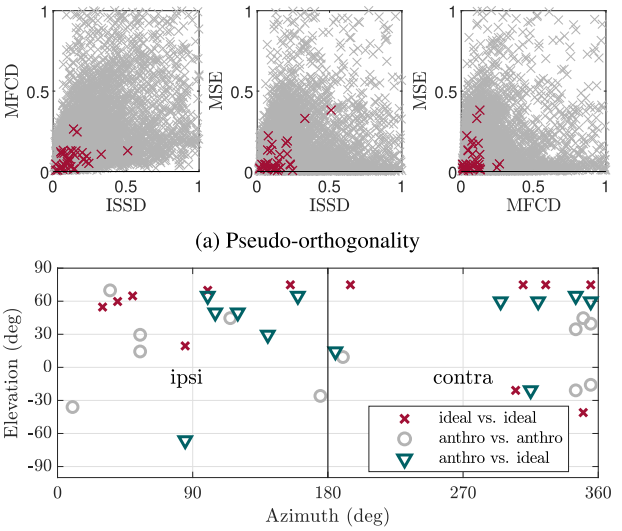

Fig. 3. Selection of stimulus pairs from HRTFs of ID1 and ID2 on the basis of small values of the three pre-selected metrics ISSD, MFCD, and MSE (a) and trying to achieve the best possible distribution in azimuth (b). The latter condition could be only partly satisfied while fulfilling the first.

variation at the left ear. These directions were additionally mirrored along the median plane, thereby allowing for a complementary set of trials with identical signal variations offered at the right ear. This guaranteed equal representation of both ears in the study. Supposing a potential difference in metric behavior for ipsilateral and contralateral incidence directions, this distinction was further considered by including virtual sound sources in both hemispheres, see Fig. 2.

In total, the study was composed of the following independent variables:

- Modality: cases in Figs. 1(b)–1(d).
- Side: ipsilateral/contralateral.
- Ear: L(eft) (optimized direction)/R(ight) (mirrored direction).

For each of the 12 conditions (modality × side × ear), six incidence directions were selected (see next section), leading to 72 stimulus pairs per block. Using a within-subject design, each participant was presented with three blocks (i.e., three repetitions in total), containing identical yet differently ordered (Latin-square balanced) stimuli.

3.2.2 Stimulus Selection

Within the selected conditions, stimuli close to the threshold of error detection needed to be chosen. As previously discussed, the individual metrics may not be sufficient to describe the audibility of differences. Yet, from a numerical perspective, they are indicators for the presence of a spectral deviation and can be assumed to scale with the strength of said deviation. Accordingly, the stimuli were selected to have rather low metric values.

Since the stimuli should additionally exploit the pseudo-orthogonal behavior of metric subgroups (as derived in Sec. 2.2), the three representative metrics served as a basis: MSE, MFCD, and ISSD. Metric values for the limited dataset (ID 1/ID 2) are visualized in Fig. 3(a) collectively for cases Figs. 3(b)–3(d). As before, the values are normalized by their respective 97th percentiles. In accordance with Fig. 1, visibly little to no common trends are present between the three metrics. The red crosses mark the directions used for the experiment. These points were statistically selected within empirically defined value ranges for the three metrics.

Besides the small metric values, it was aimed to provide a proper distribution of incidence directions within the respective hemispheres, particularly regarding the azimuth angle. Fig. 3(b) displays the HRTF incidence angles for the selected stimuli. Despite multiple iterations and some manual adaptations to the empirical value range, not all angular ranges were compatible with the constraint of small metric values. Particularly on the rear contralateral side (between 200° and 290°), no suitable stimuli could be identified for either of the comparison cases. Similarly, for frontal ipsilateral directions (azimuth angles ≤80°), the intra-individual comparison case (triangular marker) also came with increased metric values, leading to exclusion of these directions. The shown directions were used as they are for trials with signal variation on the left ear, and were mirrored along the median plane for trials with right ear signal variation.

3.2.3 Stimulus Preparation

The binaural signals for each experimental trial consisted of a triple pulse train of pink noise convolved with two pairs of HRTF spectra. The pulses each had a duration of 200 ms, with raised cosine ramps of 25 ms for fade-in and fade-out, and were separated by 150 ms of silence, producing to a total stimulus duration of 900 ms.

The preparation of HRTFs comprised multiple steps, as visualized in Fig. 4. First, HRTF magnitude spectra were reconstructed from PCs, as described in Sec. 1.2. As they only contained absolute values, a suitable phase was computed: A minimum-phase spectrum [45] accounted for phase con-

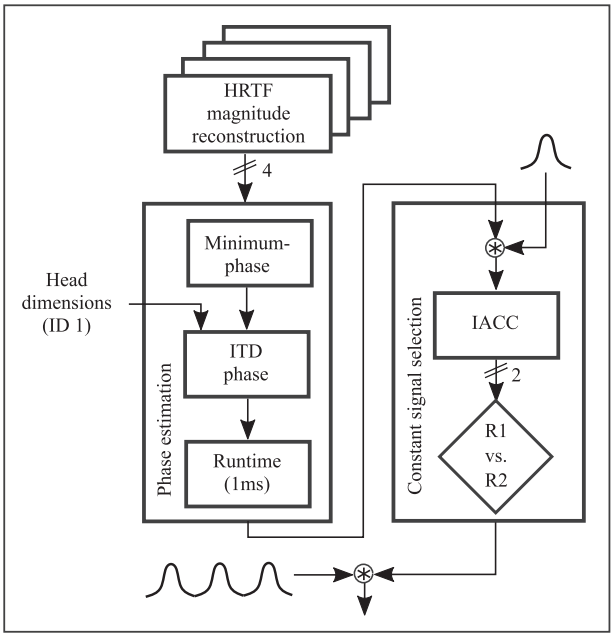

Fig. 4. Stimulus preparation for the JND experiment. Suitable phase spectra were calculated for the HRTF magnitude spectra reconstructed from PCs. The signal for the averted ear, which would not be varied between a pair of stimuli, was selected based on a minimization of the transition in IACC.

tributions of monaural cues, as contained in the magnitude spectra; An ITD phase component was subsequently calculated according to the analytical ellipsoidal model [46], with head dimension of ID 1 of the database as input; Finally, and to ensure causality of the filters, a time offset of 1ms was applied. The now complex-valued spectra could then be convolved with the noise pulses.

Up to this point, stimuli were considered from a monaural perspective. As left and right ear signals were combined for spatial audio playback, it was important to ensure that binaural effects would not interfere with the solely unilateral differences captured by the metrics in use. As previously noted, the signal of only one ear was varied between the HRTF pairs. This implicates that the other ear signal was kept constant and had to be selected among two available stimuli. Here, IACC should be considered. A transition in IACC may produce an audible difference between binaural signals, even when separate monaural signals might not be distinguishable. In [44], the JND for inter-aural noise cross-correlation was reported as $\Delta r^2 = 0.4$ for a reference stimulus of $r^2 = 0$. A worst-case (i.e., easiest discernibility) JND of $\Delta r^2 = 0.04$ was found for a reference of $r^2 = 1$.

For the HRTF spectra selected in the present study, mean and standard deviations of reference IACCs amounted to $\mu \pm \sigma = 0.827 \pm 0.135$. As the values were rather close to 1, a sensitivity to transitions in IACC close to (though not as bad as) the reported worst-case JND could be expected. For the selected HRTFs, the absolute transitions $|\Delta r^2|$ when replacing a left-ear spectrum by that of another HRTF set amounted to 0.049 ± 0.035, which, in fact, was dangerously close to the JNDs. Therefore, a minimization of the difference in IACC values between the two binaural signal pairs was applied. This served as a selection criterion for the right-ear signal, which would be employed as the fixed ear stimulus (either for right ear playback, or (after mirroring) as a left ear signal). Thus optimized IACC transitions amounted to $\mu \pm \sigma = 0.032 \pm 0.025$, with all but five stimuli lying below the JND curve reported in [44] (Fig. 4, curve denoting 75% correct in their 2AFC paradigm). Thereby, it was verified that the IACC transitions should, for the most part, play no part in distinguishing between presented stimuli.

3.2.4 Playback

The experiment was conducted in a custom-made hearing booth (length × width × height = 2.1 m × 2.1 m × 2 m). For playback, Sennheiser HD650 headphones were used, applying individual headphone equalization [47]. Level calibration was performed using an artificial head with IEC711 coupler (HMS III, HEAD Acoustics, Herzogenrath, Germany), a conditioning amplifier (Type 2690-A, NEXUS, Hottinger Brüel & Kjær GmbH, Darmstadt, Germany), and a soundcard (RME Fireface UC, Audio AG, Haimhausen, Germany), setting the level to ≈60 dB for frontal and a maximum of 66 dB for lateral incidence.

3.2.5 Participants

A total of 19 participants (six female, 13 male), aged 23−35 years ($\mu \pm \sigma = 27.3 \pm 2.8$), took part in the experiment. All, except for four, had prior experience with spatial audio and similar experiments. Pure-tone audiometry (up to 16 kHz) ensured normal hearing and sensitivity to changes in HRTF cues in the higher frequency range. All participants provided informed consent and received no compensation for their participation.

3.3 Experimental Results

As mentioned above, the participants received each stimulus once per block, leading to three repetitions in total. Thus, calculating separate probability values for each participant would produce values discretized to $p_{i,s} \in \{1/3\ 2/3, 1\}$ for individual i and stimulus $s \in [1,72]$, respectively. In a 3AFC paradigm, the guessing rate is $1/3$. Accordingly, the range between $1/3$ and 1 would therefore be very sparsely sampled, making it particularly difficult to fit the values to the sigmoid shape of a psychometric function. On this account, answers from all participants instead flow into a single data point per stimulus. Each point is defined by the overall probability of recognition in the 3AFC task (calculated based on $19 \times 3 = 57$ trials) and the seven metric values corresponding to a stimulus pair.

3.3.1 Distribution of Discrimination Probabilities

The percentage of correct responses for each stimulus, respectively, is displayed in Fig. 5. The tested conditions can be differentiated as follows: the vertical panels indicate the three different comparison sets (modalities); color coding denotes ipsilateral (petrol) and contralateral (gray) sound incidence; and the marker symbols correspond to the ear

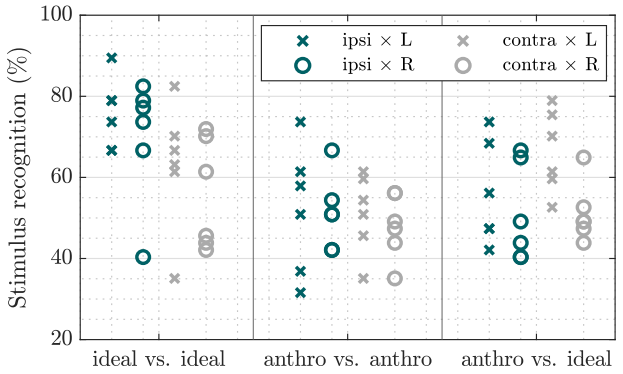

Fig. 5. Percentage of correct answers (over all participant responses) for each of the 72 stimuli in the 3AFC experiment. Data points are split after the different conditions ("Modality" × "Side" × "Ear"), see Fig. 2.

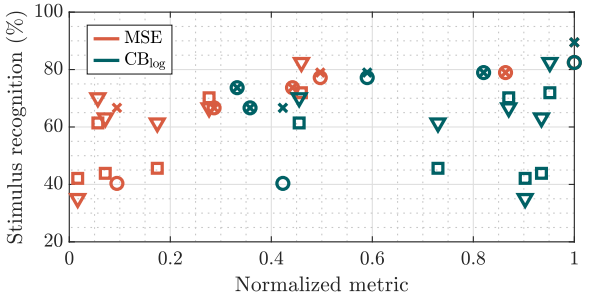

Fig. 6. Relation between stimulus distinguishability in the 3AFC task and normalized metrics for inter-individual comparisons of HRTF spectra from the "idealPC" dataset. The maximum of each metric over the 72 stimuli is used as a normalization factor. Marker shapes correspond to the four conditions ("Ear" × "Side"). Here, CB_{log} exemplifies cases less suitable for fitting to a psychometric function, compared to MSE.

signal that was varied in the trial, with "x" for the left and "o" for the right ear.

Collectively, the values range between 31.6% and 89.5% ($\mu \pm \sigma = 56.1 \pm 13.9\%$). For few individual conditions, especially the inter-individual comparison within the "anthroPCA" dataset (middle panel), the maximum probability barely reaches the 66.2% threshold point, which corresponds to the 50% chance of audibility outside of the experimental setting. (Note that the target point along the psychometric function is debated in literature. Here, the middle point is selected, as suggested in [48], while assuming a guessing rate of $\gamma = 1/3$ and a lapse rate of $\lambda = 0.01$.)

Most of the shown stimuli are independent from each other, i.e., a direct link between the stimuli of different conditions does not exist. The only exception lies in the differentiation between left and right ear playback, in which the binaural stimuli were presented once in their original form and once after mirroring along the median plane. On this account, it is feasible to point out a slight tendency towards lower detectability of spectral alterations in the right ear signal (circles), compared to the left ear (crosses).

In contrast, a possible impact of the other independent variables cannot be evaluated directly, since the observations are based on different stimuli per condition. The (partly empirical) choice of these stimuli, however, was of varying difficulty: for some conditions, finding inaudible stimulus pairs was a challenge, whereas for others, it was more demanding to find audible ones. This variation, although not leading to analytical conclusions, is in accordance with findings on the value range of distance metrics, which varied greatly for the different comparison cases, cf. Fig. 1.

3.3.2 Psychometric Representation

Given the partly limited ranges of the acquired audibility values, cf. Fig. 5, modeling a separate psychometric function for each condition, respectively, was not always feasible. Moreover, only few distance metrics showed a monotonic rise with increasing detectability rate of the corresponding stimuli. A "good" and "bad" example can be observed in Fig. 6 for inter-individual comparison within the "idealPCA" dataset. Normalized metric values for MSE and CB_{log} are displayed on the x axis, the percentage of correct answers in the 3AFC task on the y axis. The marker shapes represent different conditions ("Side" × "Ear"). (Note that the cases of left and right ear signal variation for the same "Side" variable are only mirrored stimuli and correspond to the same metric values. Therefore, squares and triangle (or circles and crosses) are always vertically aligned in pairs.)

Collectively, no general trend of increasing audibility with rising CB_{log} can be observed, whereas MSE data points seem to be a much better fit for reconstructing a psychometric function. The latter case was only true for very few conditions and metrics. This made most cases unsuited for a modeling approach based on gathered psychometric function fitting parameters, such as the slope or spread [49].

4 MODELING

In this paper, audibility of spectral deviations is to be modeled as a function of a set of numerical distance metrics. The choice of a suitable model type must be based on observations of the interaction behavior of the metrics and the collected perceptual information. The experimental results demonstrated that, e.g., an univariate regression model, relying on a single metric, would be insufficient to describe audibility. More complex model types are therefore suggested, embedding information from a varying number of distance metrics. The following sections present two different approaches and contrast their performance.

4.1 Multi-Linear Regression

A multi-linear regression (MLR) approach models the probabilities of recognition as a linear combination of a selection of N distance metrics.

$$p_{\text{correct}} = c_0 + \sum_{i=1}^{N} c_i \cdot X_i, \tag{10}$$

with c_i denoting linear coefficients and X_i the metric values. To maintain a consistent range for coefficients, metric

values can be normalized prior to summation, replacing X_i by

$$\hat{X}_i = \frac{X_i - \mu_{x,i}}{\sigma_{x,i}}. \tag{11}$$

The mean μ_i and standard deviation σ_i are calculated for metric i of dataset X, which is available for model training. For an arbitrary input dataset Y, vector $\hat{X}$ in Eq. (11) can be substituted by the normalized dataset $\hat{Y}$, while retaining the values μ_x and σ_x of the training data.

4.2 Principal Component Regression

A linear model requires the variables to be uncorrelated, thus avoiding collinearity issues. This prerequisite is only partially satisfied by the distance metrics. Therefore, as a second model type, a PCA-based approach is evaluated. Here, multi-dimensional data are projected onto a set of orthogonal coordinate axes (PCs) that capture the variance of input variables. In matrix representation, the normalized input data $\hat{\mathbf{X}}$, cf. Eq. (11), is expressed as a weighted sum of PCs:

$$\hat{\mathbf{X}} \approx \mathbf{W}_x \cdot \mathbf{V}_x. \tag{12}$$

Weight matrix $\mathbf{W}_x$ (also termed "score") and coordinates $\mathbf{V}_x$ are estimated based on a model training dataset $\mathbf{X}$. For an arbitrary input dataset $\mathbf{Y}$, corresponding scores can be calculated as

$$\mathbf{W}_y \approx \hat{\mathbf{Y}} \cdot \mathbf{V}_x^{-1} = \hat{\mathbf{Y}} \cdot \mathbf{V}_x^T, \tag{13}$$

with $\hat{\mathbf{Y}}$ denoting the input data after normalization with μ_x and σ_x. (Note that the equality of the inverse and transpose of $\mathbf{V}_x$ holds due to the orthonormality property of the matrix.) Applied to the present issue, a PCA model can provide a set of virtual metrics that are orthogonal by definition. Corresponding weights describe the reaction of these "metrics" for each data point, i.e., for each stimulus pair, and can therefore be treated similarly to the normalized distance metric values in SEC. 4.1. Recognition probabilities are then expressed as

$$p_{\text{correct}} = c_0 + \sum_{i=1}^{\text{N}_{\text{PC}}} c_i \cdot W_{y,i}, \tag{14}$$

where $W_{\text{y, i}}$ denotes a column vector i from the estimated score matrix.

The PC regression can be considered as a hierarchical linear modeling approach. Although it is more complex than the MLR model, it allows for integrating information from many metrics, and benefits—rather than suffers—from the information common to them.

4.3 Psychometric Adjustment

The presented models define the slope of a psychometric function, predicting the discrimination rate (or percent correct) in a 3AFC paradigm relative to a few selected distance metrics or to the score of "virtual" distance metrics (PCs). The probabilities p_{correct}, as reconstructed in Eq. (10) or (14), initially have no lower and upper limits, as opposed to typical data acquired in a 3AFC experiment. The common psychometric function possesses, besides the quasi-linear slope, a lower asymptote due to guessing rate $\gamma = {}^1\!/_3$. It further possesses an upper asymptote due to lapse rate λ, often assumed around 0.01 to minimize slope bias [50]. A hard cut is therefore introduced to the model output at 33.3% and 99%, respectively, leading to

$$p_{\text{limited}} = \min\{99\,\%, \max\{33.3\,\%, p_{\text{correct}}\}\}. \tag{15}$$

If p_{detect} is introduced as the general sensitivity, i.e., the probability of detection outside of the experimental setting, the percentage of correct answers given in the experiment can be expressed as

$$p_{\text{limited}} = \gamma + (1 - \lambda - \gamma) \cdot p_{\text{detect}}. \tag{16}$$

Accordingly, the sensitivity values are calculated as

$$p_{\text{detect}} = \frac{p_{\text{limited}} - \gamma}{(1 - \lambda - \gamma)}. \tag{17}$$

After this transformation, the probability p_{detect} is assumed to be no longer related to the experimental design. The 50% value corresponds to the sought audibility threshold. The chance level of $p_{\text{correct}} = 33.3\,\%$ is indicated by $p_{\text{detect}} = 0\,\%$.

4.4 Model Quality Measures

A common measure for quality of approximation is the root MSE (RMSE). In the present case, true and estimated discrimination probabilities for a given set of N_{St} stimuli are contrasted as

$$\text{RMSE} = \sqrt{\frac{1}{\text{N}_{\text{St}}} \sum_{i=1}^{\text{N}_{\text{St}}} [p_{\text{true}} - p_{\text{est}}]^2}. \tag{18}$$

Here, a distinction is made between a training and a test dataset. $\text{RMSE}_{\text{train}}$ is used to measure how well the model approximates the input data, i.e., the probabilities of discrimination acquired in the listening experiment. $\text{RMSE}_{\text{test}}$ is used to examine the suitability of the model for data that were not included when the model was created.

Moreover, the chosen metrics should explain as much as possible of the variance of the audibility values. The goodness-of-fit of the produced model is captured by the adjusted coefficient of determination (R^2), which describes the ratio of explained variance relative to the total variance of the dependent variable. Additionally, the p values of each explanatory variable indicate their respective significance for the model.

Finally, Pearson correlation coefficients r_{test} allow for better comparability to related studies that do not utilize linear regression. Here, they assess a linear relationship between true and estimated discrimination probabilities for the test data stimuli.

4.5 Model Selection

The following subsections present different variants of the two modeling approaches (MLR and PCA) and an analysis of their performance. On the basis of the discussion, it is possible to narrow down the variants to a set of best-

performing parameters in terms of model quality and complexity.

4.5.1 Parameter Variation

The output of the models depends on several factors:

- The number of distance metrics: All model types are examined for $n_{\text{Dist}} = 1$–7 input metrics. In the MLR models, the metrics flow directly into the linear combination. In PCA-based models, only n_{Dist} metrics flow into the calculation of PCs.
- The number of linear modeling coefficients: For the MLR approach, this number equals n_{Dist}, because the model directly creates a linear combination of the (centered and scaled) distance metric values. In contrast, for the PCA approach, the number of linear coefficients corresponds to the number of PCs, which, on their part, are calculated based on n_{Dist} different metrics. (Note that these two variables are varied independently.) Because the first few PCs account for a majority of the explained variance, only up to two linear modeling coefficients (i.e., two PCs) are considered sufficient for examining the performance of PCA-based models.
- The specific selection of metrics: Depending on the number of metrics to be selected, a large variety of combinations is possible. Instead of examining the model for discrete subsets of metrics, model quality is first assessed on the basis of a statistical selection of n_{Dist} metrics and 500 iterations per model type. The following two cases are considered: first, in the "random" case, n_{Dist} arbitrary variables are selected from the seven available metrics; second, in the "fixated" case, the three pseudo-orthogonal metrics MSE, ISSD, and MFCD (as employed for stimulus selection) are prioritized. This means that for $n_{\text{Dist}} \geq 3$, these three metrics are always included, with an additional $(n_{\text{Dist}} - 3)$ random selections from the remaining metrics. For $n_{\text{Dist}} < 3$, only a subset from MSE, ISSD, and MFCD is selected in the "fixated" case.

Furthermore, the choice of training data influences the resulting model. Furthermore, the (dis-)similarity between test and training data may affect the apparent model quality. For this reason, the selection of the training and test stimuli was randomized in a first step, which allowed for a proper assessment of the other factors. Iteratively, 54 stimuli (3/4) were selected for model creation, and the remaining 18 stimuli were used for subsequent testing. This approach is similar to that of k-fold cross-validation [51] with $k = 4$, yet with multiple runs and omitting the final averaging step over all folds.

4.5.2 Evaluation of Performance

Different combinations of model parameters and their effect on model quality measures are visualized in Fig. 7. The spread depicted in the different box plots results from

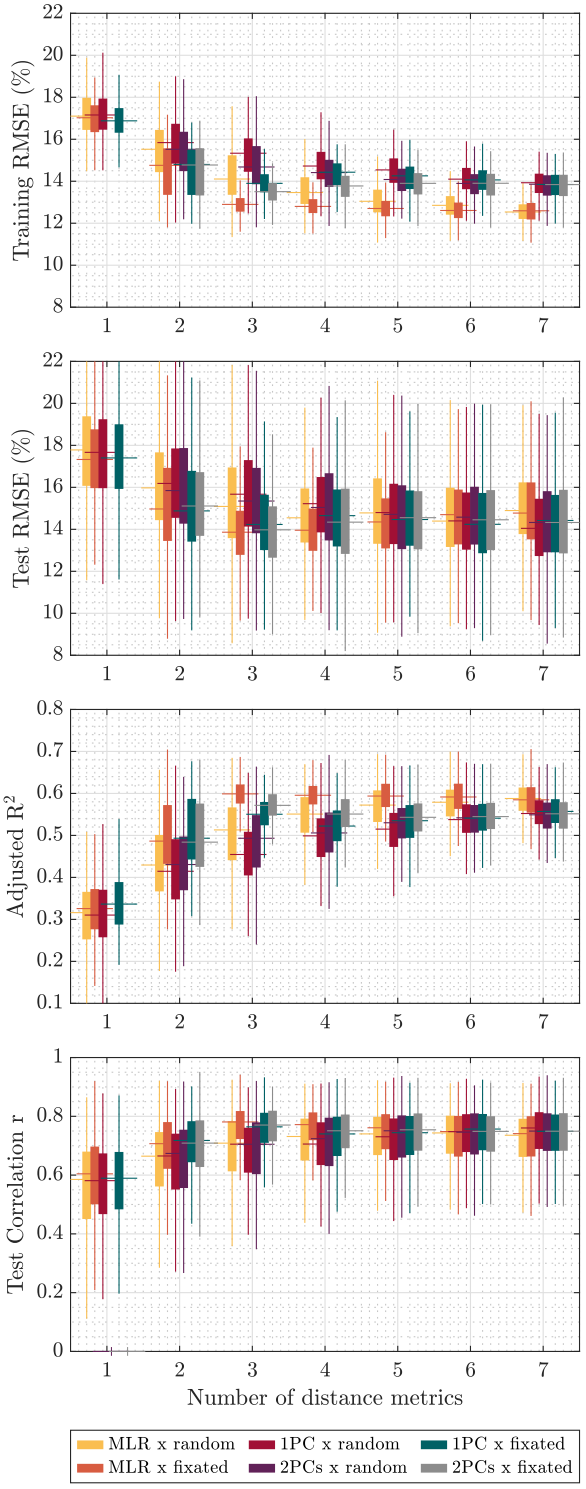

Fig. 7. Performance evaluation for the different model types as a function of n_{Dist} and for 500 iterations. A fully randomized selection of metrics shows deteriorated performance, compared to a prioritization (fixation) of the three metrics ISSD, MSE, and MFCD. Saturation is reached around three to five input metrics to the model.

500 iterations, both in terms of randomized metric subset and training/ test stimuli selection. Only for a "fixated" metric choice and $n_{\text{Dist}} = 3$, the data spread solely reflects variability due to stimuli selection.

On a descriptive level, with increasing number of distance metrics flowing into the model, the quality measures show improvement, i.e., lower RMSE values for training and test data, higher adjusted R^2 and higher Pearson correlation coefficients r_{test}. The quality measures further show asymptotic behavior and, depending on the model type, already reach saturation for as low as $n_{\text{Dist}} \approx 3-5$ distance metrics. For MLR and PCA models, respectively, asymptotes are located at 12.71% and 13.97% for $\text{RMSE}_{\text{train}}$, at 14.75% and 14.44% for $\text{RMSE}_{\text{test}}$, at 0.58 and 0.54 for the adjusted R^2, and at 0.72 and 0.73 for r_{test}.

Upon closer inspection, a marked influence of the model parameters is noticeable. A prioritization of the ISSD, MSE, and MFCD metrics in the "fixated" case enhances performance, compared to a fully randomized choice of metric subsets. Best overall performance is achieved by the MLR model applying the three pre-selected metrics (orange box plots), with $\text{RMSE}_{\text{train}}$ and $\text{RMSE}_{\text{test}}$ as small as $(\mu \pm \sigma) = (12.85 \pm 0.5)\%$ and $(13.83 \pm 1.6)\%$, respectively. It further shows the highest adjusted R^2 of (0.6 ± 0.04) and highest Pearson correlation coefficients r_{test} of (0.76 ± 0.08) (with p values as low as (0.002 ± 0.007) for r_{test}).

As can be derived from the figure, it is not beneficial to include more than three metrics; The performance measures even show a tendency for model deterioration with higher model complexity. A similar stagnation and deterioration is observed for the other two cases with "fixated" metric selection [PCA models with one PC (petrol) or two PCs (gray)].

Each additionally included distance metric necessitates more computations. For the ensuing complexity to be justified, each metric contribution needs to be significant. In the sense of purely linear models, each coefficient (or explanatory variable) should contribute significantly to the model ($p < \alpha$ with a significance level of, e.g., $\alpha = 0.01$). Fig. 8 displays the maximum p values over all explanatory variables. For linear models (yellow and orange), an increase in maximum p values with rising number of distance metrics can be observed. With as low as $n_{\text{Dist}} = 3$, the significance level is by far exceeded. These cases are therefore eliminated, including the (until now) best-performing MLR model with three distance metrics as input.

In the context of PCA models, n_{Dist} only affects the composition of PCs. The depicted p values refer to significance of the "virtual" metrics, which are linearly combined to estimate audibility. A rise from one to two PCs leads to exceeding the significance level, irrespective of metric selection (see purple and gray boxplots). Thus, models based on two PCs are also disqualified.

After these exclusions, and based on the model quality measures and significance values of the remaining model types, the PCA model with one PC and the three preselected input metrics (ISSD, MSE, MFCD) can be identified as most promising, with $\mu \pm \sigma = (13.88 \pm 0.65)\%$ for $\text{RMSE}_{\text{train}}$, $(14.31 \pm 1.91)\%$ for $\text{RMSE}_{\text{test}}$, (0.54 ± 0.04) for the adjusted R^2, and (0.75 ± 0.09) for r_{test} (with p values (0.003 ± 0.015) for r_{test}).

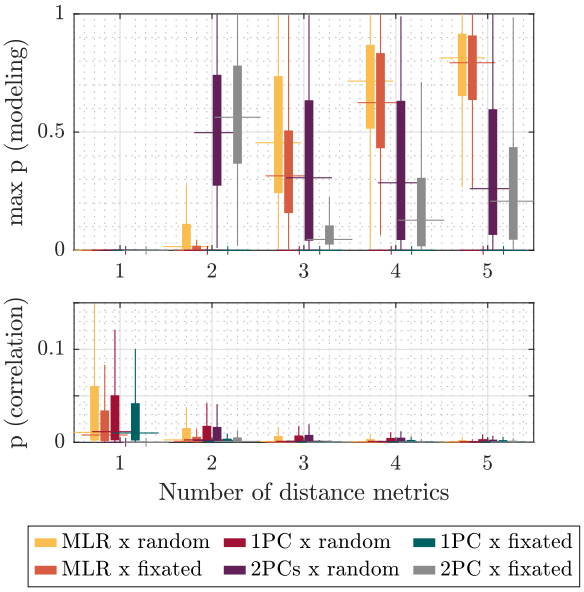

Fig. 8. Maximum p values among the explanatory variables in the linear models (top) and p values for test data correlation coefficients r_{test} (bottom). Exceeding the level $\alpha = 0.01$ in the top plot indicates an insignificant contribution of the metrics to the model output. Thus unnecessarily complex models are eliminated, including the best-performing three-metric MLR model according to Fig. 7. In the bottom plot, correlation values are increasingly significant with rising number of metrics.

It should be noted that the adjusted R^2 here only serves a relative comparison between the model types. Because of lack of a reference for the expected value range (to our knowledge), it is refrained from drawing conclusions regarding overall model effectiveness based on this value. Similarly, the acquired RMSE values are difficult to compare to other JND studies. In most cases, the 50% thresholds are evaluated, defining a minimum discriminable resolution of a specific variable, e.g., loudness. Here, contrarily, the unit of the model output is in percent. A more useful measure for effectiveness is therefore provided by the correlation values. A mean of 0.75 can be considered moderate to strong and further lies in a similar range to that achieved for some conditions in [18].

4.5.3 Specific Model Implementation

The cross-validation in the previous section already proved the superiority of a model type with one PC, that explicitly dictates the three metrics to use (ISSD, MSE, and MFCD). Now, the model can be re-trained using the whole data sample to obtain a final numerical model. For the finalized model, the now scalar quality measures yield 14.02% for $\text{RMSE}_{\text{train}}$ and 0.55 for the adjusted R^2. Both values are very close to the mean data obtained in cross-validation for this model type. Pearson correlation between the experimental and predicted discrimination probabilities yields 0.75 ($p < 0.001$). (Note that this value is not directly comparable to r_{test}, because all stimuli flow into training the final model, with no stimuli left for testing.)

For a pair of stimuli, the probability of a correct answer in the 3AFC task ($\hat{p}_{\text{correct}}$ in percent) is reconstructed as

$$\hat{p}_{\text{correct}} = c_0 + c_1 \cdot w_1, \tag{19}$$

with linear coefficients $c_0 = 57.6023$ (intercept) and $c_1 = 7.7673$. Weight (score) w_1 corresponds to PC_1 and is acquired by running PCA on the complete training dataset X (i.e., three metric values for each of the 72 stimuli).

For a given pair of stimuli, w_1 is estimated as the inner product

$$w_1 = \langle \text{PC}_1, \hat{d} \rangle = \langle \begin{pmatrix} 0.6671 \\ 0.6314 \\ 0.378 \end{pmatrix}, \begin{pmatrix} \hat{\text{issd}} \\ \hat{\text{mse}} \\ \hat{\text{mfcd}} \end{pmatrix} \rangle \tag{20}$$

with the centered and scaled metric vector $\hat{d}$ calculated as

$$\hat{d} = \left[\begin{pmatrix} \text{issd} \\ \text{mse} \\ \text{mfcd} \end{pmatrix} - \underbrace{\begin{pmatrix} 7.2916 \\ 0.1615 \\ 0.9294 \end{pmatrix}}_{\mu_X} \right] \odot \underbrace{\begin{pmatrix} 0.189 \\ 5.2127 \\ 1.4271 \end{pmatrix}}_{\frac{1}{\sigma_X}}. \tag{21}$$

Here, μ_X and σ_X represent the mean and standard deviation of the training data for each of the metrics. The Hadamard product operator $\odot$ indicates element-wise multiplication. Finally, the predicted values $\hat{p}_{\text{correct}}$ are transformed using Eqs. (15) and (17), yielding values independent from the 3AFC paradigm.

5 DISCUSSION

5.1 Integration and Reproduction of Variance

The numerical evaluation of the distance metrics and their interrelation behavior demonstrated the presence of mutual information. Although the correlation patterns featured common trends, especially a contrast between metric behavior for inter-individual and intra-individual HRTF comparisons could be observed. The pseudo-orthogonal metric subsets, created on the basis of correlation and factor analysis results, were used for the choice of stimuli for the listening experiment, and were later shown to represent an efficient choice of input metrics for the parametric model of error audibility.

The evaluation of different modeling approaches revealed that a model based on three metrics (ISSD, MSE, and MFCD) and a single PC provides a good balance between model complexity and performance. Clearly, a more complex approach could provide a more detailed description of the stimuli present. Eliminating all but three metrics from the model implicitly neglects some aspects of spectral deviation. Furthermore, a limitation to a single PC captures only 59.6% of the explained variance of metric behavior.

As shown, however, the potentially added variance information of the second PC (29.8%) does not help represent the variance of perceptual data in the model output. Similar observations hold true for an increased number of metrics, which only led to insignificant contributions in both the PCA and MLR model types. The variance of perceptual data can thereby not be fully explained by the given metrics. This could indicate the need for further metrics that capture other aspects of spectral dissimilarity, possibly including also phase information. (Note that, although minimum-phase spectra were reconstructed for the HRTF magnitudes before convolution with the noise pulse train, phase differences were not directly evaluated by the employed metrics.)

5.2 Model Applicability

In the following sections, the generalized validity of the model is discussed, reviewing the meaningfulness of a unilateral model in practice, as well as the applicability on arbitrary HRTF data.

5.2.1 Relevance of Unilateral Modeling

The concept and motivation for choosing an approach of binaural presentation with unilateral signal variation were outlined in SEC. 3.1. Still, the question of usability and informative value of the model must be addressed.

In the conducted experiment, only unilateral signal variations were presented. In practice, however, a direct comparison of HRTFs very likely entails changes to both ear signals. In order to apply the model in practice, it needs to be extended to a bilateral variation model. For the development of such a model, the presented unilateral JND approach would be of value in two ways: On the one hand, it could serve a selection of binaural stimuli covering combinations of different degrees of similarity. As indicated in SEC. 3.2, the task of finding near-threshold stimuli already proved quite hard for the single-channel variation. Attempting a similar approach for binaural variation, with the pure distance metric values as a starting point, would be very restrictive. The present model would facilitate stimulus selection and thereby allow for the inclusion of more conditions.

On the other hand, the current model output could serve as direct input for the new, more complex model. The latter would then predict distinguishability using appropriate binaural weighting, in addition to integrating binaural effects, e.g., the IACC transitions that have been controlled for in the present work. It was decided against the inclusion of a bilateral extension in the present study, as it would have exceeded the intended scope of this paper.

5.2.2 Validity for Generic Datasets

As previously noted, only a subset of available HRTFs were used for the listening experiment. This included the elimination of "measured" HRTF set, for which an empirical selection of near-threshold stimuli proved to be difficult. This raises the question whether the derived model is applicable to HRTF datasets other than those used for the stimuli.

To examine this, the model was applied to the four HRTF comparison cases, as introduced in SEC. 2. The different stages of the model output are visualized in Fig. 9. The direct output of the linear model is represented by p_{correct}. Introducing a cut-off to the model output yields p_{limited}, in which the unrealistic probabilities above 100% are elimi-

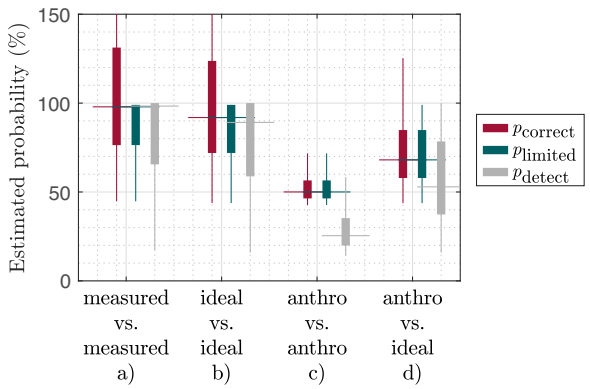

Fig. 9. Modeled probability of discrimination for the four comparison cases. A distinction is made between the percentage of correct answers in the 3AFC paradigm as a direct output of the PCA model (p_{correct}), with introduced cutoff for guessing and lapse rate (p_{limited}), and transformed to a general probability of detection (p_{detect}).

nated, whereas the median values remain unchanged. The gray boxplots (p_{detect}) show the actual predicted probabilities of detection (for a non-experimental setting).

The inter-individual comparison Figs. 9(a) and 9(b) show the highest probabilities, with a median and interquartile range of median (IQR) = 98.3 ([65.5, 100])% and 89.2 ([58.8, 100])%, respectively. This observation is in accordance with the difficulties found in selecting suitable stimuli for the experiment, which lead to the total exclusion of the "measured" dataset (a) and to an unequal azimuthal distribution of stimulus incidence directions in Fig. 9(b), cf. Fig. 6. Approximately half of the stimulus pairs in the intra-individual comparison dataset in Fig. 9(d) lies below the threshold for audibility, with median (IQR) = 52.9 ([37.4, 78.4]), The same goes for most of the dataset in Fig. 9(c), with median (IQR) = 25.4 ([19.9, 35.3]). A comparison of p_{correct} of Figs. 9(b)–9(d) with the experimental data in Fig. 5 indicates that the rather empirical selection of stimuli is representative of the datasets.

Informal listening by the authors showed that the model is successful in finding HRTF pairs below the JND for the "measured" dataset—a task that had been particularly difficult when relying on separate minimization of the three metrics. It can therefore be assumed that the application of the model is not strictly limited to the reconstructed HRTFs used for the experiment. Though the "measured" data played a part in metric selection, it could be argued that the analysis already covered a variety of HRTF variations. Still, a validation based on other datasets (e.g., from other HRTF databases) would be useful.

5.3 Limitations

The derived model is based on the available perceptual data. Here, the selection of stimuli close to the JND for the listening experiment had a major influence on modeling possibilities, restricting the model to the slope of the psychometric function. By including stimuli that are further above or below the threshold, it would have been possible to model the full sigmoid shape. An increased number of stimuli would further enable a distinction between conditions, with the cost of a more time-consuming listening experiment. Separate models created for, e.g., ipsilateral and contralateral incidence directions, would take into account the influence of HRTF magnitude and hemisphere-specific cues on distance metric behavior.

As in former JND studies for psychoacoustic properties, the loudness level of stimuli might affect the sensitivity to spectral changes. Especially in low-level spectral components, changes (e.g., in notch quality of HRTFs) may go undetected if below the auditory threshold. Although the choice of 60–65 dB was deemed sufficient to assess the concept of the model in a first step, a potential level dependence should still be examined in future work.

Further perceptual validation is required to ensure the applicability of the model to other types of spectral alterations, possibly not represented in the used datasets. However, it should be noted that, e.g., HRTFs acquired in different measurement or simulation setups likely possess large differences [52]. They might feature too clearly audible differences and thus have to be excluded from the JND assessment. Supra-threshold stimulus pairs of this kind would derive more benefit from a model predicting the type or the markedness of a perceptual difference, rather than its mere presence.

6 CONCLUSION

In this study, seven magnitude-based distance metrics for HRTFs were used on measured HRTFs from the ITA HRTF database and on approximations thereof, representing two levels of spectral detail loss. Metric behavior for the different inter-individual and intra-individual comparison cases was analyzed using correlation and factor analysis. Although the interrelation patterns of the metrics could be partly attributed to the arithmetic operations involved in their calculation, metric variance was not fully explained by the common factors. For the purpose of the study, the metrics ISSD, MSE, and MFCD were selected, representing three subgroups with a tendency for related metric responses within each.

A listening test paradigm was designed for the assessment of perceptibility of unilateral differences of HRTFs. The choice of test stimuli was optimized, selecting HRTF pairs to which the three metrics did not respond equally. Furthermore, a minimization of the influence of inter-aural cross-correlation on the detection of dissimilarities between stimuli was considered. The experiment provided detection probabilities around the perception threshold, allowing for modeling the slope of a psychometric function to describe these probabilities.

A multi-linear regression approach and a linear model based on the score of principal components were examined regarding their performance. Different complexities of the model in terms of the number of linear coefficients and the number of integrated metrics were contrasted. A trade-off was made between model accuracy and complexity, choosing to model the variance of the three pre-selected models metrics ISSD, MSE, and MFCD using one principal com-

ponent. The weighting score of this PC served as a "virtual" metric, approximating detection probabilities in a univariate linear model. It was shown that more complex approaches, e.g., integrating more than three distance metrics, did not improve the goodness-of-fit of the model and partly even caused a deterioration of performance.

Observing the residual variance of the perceptual data, which could not be modeled by increasing model complexity, the question arises whether other types of metrics should be considered. This could include, e.g., phase-based calculations or variants of the employed metrics, evaluating spectral deviations within specific frequency ranges. Nonetheless, the model will remain subject to the reservation of inter-individual variation, as responses can (and will) certainly deviate from the modeled mean threshold.

For better applicability, future work should extend the results to a binaural model with simultaneous variation of both ear signals. This bilateral approach could either make use of the direct output of the present single-channel model or follow similar methodology in deriving a model from a subset of metrics. For the acquisition of perceptual data on detection of bilateral variation, the unilateral model could further help in the selection of near-threshold stimulus pairs.

It should be noted that the employed pulsed noise signals, combined with free-field conditions, emphasize the audibility of spectral differences, leading to comparatively strict JND values. For more natural situations, involving in-door playback of more common sounds, the sensitivity to changes is expected to be substantially lower. Further validation with binaural room impulse responses and, e.g., speech or music signals could provide insight on the accuracy level of HRTF representation required in practice.

7 ACKNOWLEDGMENT

This research was funded by the German Research Foundation (DFG, project no. 402811912, Individual Binaural Synthesis of Virtual Acoustic Scenes). The authors would like to thank Natálie Brožová for implementing and conducting the listening experiment. Many thanks to Hark Braren, Lukas Vollmer, Dr. Manuj Yadav, and Dr. Ming Yang for the fruitful discussions on modeling and evaluation, as well as to the experiment participants for their time. The authors would furthermore like to express their gratitude to the four anonymous reviewers, whose remarks have significantly improved the quality of this paper.

Declaration of Conflicting Interests

Part of this work has been previously presented at the 48th Annual Conference on Acoustics, DAGA 2022, Stuttgart, Germany. The authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

8 REFERENCES

[1] H. Lee, "A Conceptual Model of Immersive Experience in Extended Reality," PsyArXiv preprint (2020 Sep.). https://doi.org/10.31234/osf.io/sefkh.

[2] R. Bücklein, "The Audibility of Frequency Response Irregularities," *J. Audio Eng. Soc.*, vol. 29, no. 3, pp. 126–131 (1981 Mar.).

[3] L. Simon, A. Andreopoulou, and B. Katz, "Investigation of Perceptual Interaural Time Difference Evaluation Protocols in a Binaural Context," *Acta Acust. united Acust.*, vol. 102, no. 1, pp. 129–140 (2016 Jan.).

[4] R. Bomhardt, I. C. P. Mejía, A. Zell, and J. Fels, "Required Measurement Accuracy of Head Dimensions for Modeling the Interaural Time Difference," *J. Audio Eng. Soc*, vol. 66, no. 3, pp. 114–126 (2018 Mar.). https://doi.org/10.17743/jaes.2018.0005.

[5] M. Kohnen, R. Bomhardt, J. Fels, and M. Vorländer, "Just Noticeable Notch Smoothing of Head-Related Transfer Functions," presented at the *Fortschritte der Akustik (DAGA)*, pp. 333–335 (Munich, Germany) (2018 Mar.).

[6] J. Breebaart, F. Nater, and A. Kohlrausch, "Spectral and Spatial Parameter Resolution Requirements for Parametric, Filter-Bank-Based HRTF Processing," *J. Audio Eng. Soc.*, vol. 58, no. 3, pp. 126–140 (2010 Mar.).

[7] R. Nicol, V. Lemaire, A. Bondu, and S. Busson, "Looking for a Relevant Similarity Criterion for HRTF Clustering: A Comparative Study," presented at the *120th Convention of the Audio Engineering Society* (2006 May), paper 6653.

[8] I. Ananthabhotla, V. K. Ithapu, and W. O. Brimijoin, "A Framework for Designing Head-Related Transfer Function Distance Metrics That Capture Localization Perception," *JASA Express Letters*, vol. 1, no. 4, paper 044401 (2021 Apr.). https://doi.org/10.1121/10.0003983.

[9] E. H. Langendijk and A. W. Bronkhorst, "Contribution of Spectral Cues to Human Sound Localization," *J. Acoust. Soc. Am.*, vol. 112, no. 4, pp. 1583–1596 (2002 Oct.).

[10] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory Model Based Direction Estimation of Concurrent Speakers From Binaural Signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605 (2011 May/Jun.). https://doi.org/10.1016/j.specom.2010.05.006.

[11] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014 Aug.). https://doi.org/10.1121/1.4887447.

[12] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976 (2005 Aug.). https://doi.org/10.1121/1.1945368.

[13] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).

[14] V. P. Sivonen and W. Ellermeier, "Directional Loudness in an Anechoic Sound Field, Head-Related Transfer Functions, and Binaural Summation," *J. Acoust. Soc. Am.*, vol. 119, no. 5, pp. 2965–2980 (2006 May). https://doi.org/10.1121/1.2184268.

[15] A. Lindau, V. Erbes, S. Lepa, et al., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 984–994 (2014 Sep.).

[16] T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney, "Predicting the Colouration Between Binaural Signals," *Appl. Sci.*, vol. 12, no. 5, paper 2441 (2022 Feb.). https://doi.org/10.3390/app12052441.

[17] J. Huopaniemi, N. Zacharov, and M. Karjalainen, "Objective and Subjective Evaluation of Head-Related Transfer Function Filter Design," *J. Audio Eng. Soc.*, vol. 47, no. 4, pp. 218–239 (1999 Apr.).

[18] A. Osses and A. Kohlrausch, "Perceptual Similarity Between Piano Notes: Simulations With a Template-Based Perception Model," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3534–3552 (2021 May). https://doi.org/10.1121/10.0004818.

[19] E. Zwicker and G. B. Henning, "The Four Factors Leading to Binaural Masking-Level Differences," *Hear. Res.*, vol. 19, no. 1, pp. 29–47 (1985 Jun.).

[20] S. Doma, N. Brožová, and J. Fels, "Interrelation Analysis of Distance Metrics for Head-Related Transfer Functions," presented at the *Fortschritte der Akustik (DAGA)*, pp. 290–291 (Stuttgart, Germany) (2022 May).

[21] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models* (Springer, Berlin, Germany, 1999), 2nd ed. https://doi.org/10.1007/978-3-662-09562-1.

[22] S. Shimada, N. Hayashi, and S. Hayashi, "A Clustering Method for Sound Localization Transfer Functions," *J. Audio Eng. Soc.*, vol. 42, no. 7/8, pp. 577–584 (1994 Jul.).

[23] K.-S. Lee and S.-P. Lee, "A Relevant Distance Criterion for Interpolation of Head-Related Transfer Functions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1780–1790 (2011 Sep.). https://doi.org/10.1109/TASL.2010.2101590.

[24] J. C. Middlebrooks, "Individual Differences in External-Ear Transfer Functions Reduced by Scaling in Frequency," *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1480–1492 (1999 Sep.). https://doi.org/10.1121/1.427176.

[25] J.-G. Richter, G. Behler, and J. Fels, "Evaluation of a Fast HRTF Measurement System," presented at the *140th Convention of the Audio Engineering Society* (2016 May), paper 9498.

[26] R. Bomhardt, M. de la Fuente Klein, and J. Fels, "A High-Resolution Head-Related Transfer Function and Three-Dimensional Ear Model Database," *Proc. Mtgs. Acoust*, vol. 29, paper 050002 (2016 Nov.). https://doi.org/10.1121/2.0000467.

[27] S. Hwang and Y. Park, "Interpretations on Principal Components Analysis of Head-Related Impulse Responses in the Median Plane," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. EL65–EL71 (2008 Apr.). https://doi.org/10.1121/1.2884094.

[28] R. Bomhardt, H. Braren, and J. Fels, "Individualization of Head-Related Transfer functions Using Principal Component Analysis and Anthropometric Dimensions," *Proc. Mtgs. Acoust*, vol. 29, paper 050007 (2016 Nov.). https://doi.org/10.1121/2.0000562.

[29] T. McKenzie, C. Armstrong, L. Ward, D. T. Murphy, and G. Kearney, "Predicting the Colouration Between Binaural Signals," *Appl. Sci.*, vol. 12, no. 5, paper 2441 (2022 Feb.). https://doi.org/10.3390/app12052441.

[30] K. A. J. Riederer, "Repeatability Analysis of Head-Related Transfer Function Measurements," presented at the *105th Convention of the Audio Engineering Society* (1998 Sep.), paper 4846.

[31] K. Pearson, "Notes on Regression and Inheritance in the Case of Two Parents," *P. Roy. Soc. Lond.*, vol. 58, pp. 240–242 (1895 Jun.).

[32] L. L. Thurstone, "Multiple Factor Analysis," *Psychol. Rev.*, vol. 38, no. 5, pp. 406–427 (Sep. 1931).

[33] M. S. Bartlett, "Tests of Significance in Factor Analysis," *Brit. J. Psychol.*, vol. 3, no. 2, pp. 77–85 (1950 Jun.).

[34] H. F. Kaiser, "An Index of Factorial Simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31–36 (1974 Mar.).

[35] J. L. Horn, "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, vol. 30, no. 2, pp. 179–185 (Jun. 1965).

[36] I. Izquierdo Alfaro, J. Olea Díaz, and F. J. Abad García, "Exploratory Factor Analysis in Validation Studies: Uses and Recommendations," *Psicothema*, vol. 26, no. 3, pp. 395–400 (2014). https://doi.org/10.7334/psicothema2013.349.

[37] C. D. Dziuban and E. C. Shirkey, "When Is a Correlation Matrix Appropriate for Factor Analysis? Some Decision Rules," *Psychol. Bull.*, vol. 81, no. 6, paper 358 (Jun. 1974). https://doi.org/10.1037/h0036316.

[38] A. Langhans and A. Kohlrausch, "Differences in Auditory Performance between Monaural and Diotic Conditions. I: Masked Thresholds in Frozen Noise," *J. Acoust. Soc. Am.*, vol. 91, no. 6, pp. 3456–3470 (1992 Jun.).

[39] B. C. Moore, B. R. Glasberg, and T. Baer, "A Model for the Prediction of Thresholds, Loudness, and Partial Loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240 (1997 Apr.).

[40] A. Bronkhorst and R. Plomp, "Binaural Speech Intelligibility in Noise for Hearing-Impaired Listeners," *J. Acoust. Soc. Am.*, vol. 86, no. 4, pp. 1374–1383 (1989 Oct.).

[41] A. Nagle, C. Quinquis, A. Sollaud, A. Battistello, and D. Slock, "Quality Impact of Diotic Versus Monaural Hearing on Processed Speech," presented at the *123rd Convention of the Audio Engineering Society* (2007 Oct.), paper 7220.

[42] L. Collet, D. T. Kemp, E. Veuillet, et al., "Effect of Contralateral Auditory Stimuli on Active Cochlear Micro-Mechanical Properties in Human Subjects," *Hear. Res.*, vol. 43, no. 2–3, pp. 251–261 (1990 Jan.). https://doi.org/10.1016/0378-5955(90)90232-E.

[43] W. Buño, Jr., "Auditory Nerve Fiber Activity Influenced by Contralateral Ear Sound Stimulation," *Exper. Neurol.*, vol. 59, no. 1, pp. 62–74 (1978 Mar.).

[44] I. Pollack and W. Trittipoe, "Binaural Listening and Interaural Noise Cross Correlation," *J. Acoust. Soc. Am.*, vol. 31, no. 9, pp. 1250–1252 (1959 Sep.).

[45] A. Kulkarni, S. Isabelle, and H. Colburn, "On the Minimum-Phase Approximation of Head-Related

Transfer Functions," in *Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 84–87 (New Paltz, NY) (1995 Oct.). https://doi.org/10.1109/ASPAA.1995.482964.

[46] R. Bomhardt, M. Lins, and J. Fels, "Analytical Ellipsoidal Model of Interaural Time Differences for the Individualization of Head-Related Impulse Responses," *J. Audio Eng. Soc.*, vol. 64, no. 11, pp. 882–894 (2016 Nov.). https://doi.org/10.17743/jaes.2016.0041.

[47] B. Masiero and J. Fels, "Perceptually Robust Headphone Equalization for Binaural Reproduction," presented at the *130th Convention of the Audio Engineering Society* (2011 May), paper 8388.

[48] B. Treutwein, "Adaptive Psychophysical Procedures," *Vis. Res.*, vol. 35, no. 17, pp. 2503–2522 (1995 Sep.).

[49] B. Treutwein and H. Strasburger, "Fitting the Psychometric Function," *Percept. Psychophys.*, vol. 61, no. 1, pp. 87–106 (1999 Jan.).

[50] S. A. Klein, "Measuring, Estimating, and Understanding the Psychometric Function: A Commentary," *Percept. Psychophys.*, vol. 63, no. 8, pp. 1421–1455 (2001 Nov.).

[51] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *J. R. Stat. Soc. B*, vol. 36, no. 2, pp. 111–133 (1974 Jan.).

[52] A. Andreopoulou, D. R. Begault, and B. F. Katz, "Inter-Laboratory Round Robin HRTF Measurement Comparison," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 895–906 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2400417.

THE AUTHORS


Shaimaa Doma


Cosima A. Ermert


Janina Fels

Shaimaa Doma is a research assistant at the Institute for Hearing Technology and Acoustics (IHTA), RWTH Aachen University, where she started her Ph.D. in 2019. She received her M.Sc. in Electrical Engineering and Information Technology from RWTH Aachen University in 2018, with a specialization in Biomedical Engineering. In 2017, she was an intern at the Advanced Bionics European Research Center, Hanover, Germany.

•

Cosima A. Ermert received her M.Sc. in Electrical Engineering and Information Technology from RWTH Aachen University in 2020. After an internship in the clinical research group at Oticon A/S in Denmark, she started working as a research assistant at the Institute for Hearing Technology and Acoustics (IHTA) at the RWTH Aachen University. In her research, she focuses on distance metrics for HRTF comparison and cognitive research in virtual environments.

•

Janina Fels is a full professor and director of the Chair and Institute for Hearing Technology and Acoustics at RWTH Aachen University, Germany, since 2020. In 2012–2020, she was Professor for Medical Acoustics at RWTH Aachen University, Germany. She studied electrical engineering (diploma 2002) at RWTH Aachen University, Germany, where she also received her Ph.D. from the Institute of Technical Acoustics in 2008. In 2009, she was a post-doc at the Center for Applied Hearing Research (CAHR) at the Technical University of Denmark (DTU) and Widex, Denmark. From 2012–2015, she was a visiting scientist at the Institute of Neuroscience and Medicine, Structural and Functional Organization of the Brain (INM-1) at Forschungszentrum Jülich, Germany. In 2013, she was awarded the Lothar Cremer Prize by the German Acoustics Society for her innovative and pioneering work in the field of binaural technology and medical acoustics. In 2014, she was appointed to the Young College of the North Rhine-Westphalian Academy of Sciences and Arts. In 2020, she was elected as a Review Board Member for Acoustics for the German Research Foundation (DFG).