Acta Acustica 2023, **7**, 45

https://doi.org/10.1051/aacus/2023040

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

Towards predicting immersion in surround sound music reproduction from sound field features

Roman Kiyan[1,*], Jakob Bergner[1], Stephan Preihs[1], Yves Wycisk[2], Daphne Schössow[1], Kilian Sander[2], Jürgen Peissig[1], and Reinhard Kopiez[2]

[1] Institute of Communications Technology, Leibniz University Hannover, Hannover, Germany
[2] Institute for Musicology, Hannover University of Music, Drama, and Media, Hannover, Germany

Received 13 February 2023, Accepted 2 August 2023

Abstract – When evaluating surround sound loudspeaker reproduction, perceptual effects are commonly analyzed in relationship to different loudspeaker configurations. The presented work contributes to this by modeling perceptual effects based on acoustic properties of various reproduction formats. A model of immersion in music listening is derived from the results of an experimental study analyzing the psychological construct of immersive music experience. The proposed approach is evaluated with respect to the relationship between immersion ratings and sound field features obtained from re-recordings of the stimuli using a spherical microphone array at the listening position. Spatial sound field parameters such as inter-aural cross-correlation (IACC), diffuseness and directivity are found to be of particular relevance. Further, immersion is observed to reach a point of saturation with greater numbers of loudspeakers, which is confirmed to be predictable from the physical properties of the sound field. Although effects related to participants and musical pieces outweigh the impact of sound field features, the proposed approach is found to be suitable for predicting population-average ratings, i.e. immersion experienced by an average listener for unknown content. The proposed method could complement existing research on multichannel loudspeaker reproduction by establishing a more generalizable framework independent of particular speaker setups.

Keywords: Spatial audio, Multichannel loudspeaker reproduction, Auditory perception modeling, Feature selection, Sound field analysis

1 Introduction

In spatial audio reproduction over loudspeakers, the trend is towards greater numbers of speakers, with possible configurations including speakers in the listening plane in systems such as 5.1 surround as well as speakers at different heights in 5.1.4 or 22.2 setups. These configurations are employed in the hope of achieving better reproduction of spatial cues and – as a consequence – more immersive listening experience [1, 2].

Although increasing the number of loudspeakers and developing new techniques for generating signals to feed them is a superficially obvious course of action, the relationship between technological elaborateness and subjective listening experience remains somewhat ambiguous. This has been investigated in a number of studies [3–6] by evaluating various perceptual attributes or listener preference in listening tests. The attributes used to assess spatial audio reproduction are mostly rooted in audio quality evaluation [7]. In the absence of glaring deficiencies and artifacts in modern multichannel audio systems, however, overall listening experience (OLE) is becoming increasingly important over basic audio quality (BAQ) [8]. This view on spatial audio evaluation requires different assessment paradigms. Notions of OLE, emotional arousal and immersion have also been evaluated for various loudspeaker setups and program material [9–11].

The cited studies have in common that they evaluate perceptual attributes against reproduction formats. Instigated by the idea of determining differences in acoustic properties of music reproduced with different loudspeaker configurations [12], this work presents an analysis of an experimental study through a model based on sound field features of the stimuli. Musical stimuli in different loudspeaker reproduction formats have been assessed by participants using the Immersive Music Experience Inventory (IMEI) [13], in addition to the collection of psycho-physiological feedback. This paper, however, focuses only on the relationship of immersion to acoustic features – exploring both descriptive and predictive modeling.

This paper is structured as follows: The rest of Section 1 is concerned with delineating the concept of immersion

*Corresponding author: roman.kiyan@ikt.uni-hannover.de

underpinning the current study as well as reviewing related studies with respect to key observations that the model to be developed should account for. Section 2 presents details of the experimental study and Section 3 introduces the proposed modeling approach. Since this approach follows a general formulation, the selection process of which specific sound field features are suitable for modeling is described in Section 4. Results concerning descriptive and predictive modeling are reported and discussed in Section 5 and a summary and conclusion are given in Section 6.

1.1 Immersion in spatial audio

Various conceptualizations and definitions of *immersion* have been proposed with varying degrees of discrimination towards related concepts such as *presence* or *envelopment*. In acoustics and audio, these terms often serve as *attributes* to evaluate spatial sound reproduction [14–16]. Unless specifically prompted, judgments are generally meant to be independent of preference or emotion [17]. In audiovisual media, virtual reality (VR) and computer games, immersion is commonly understood in a broader sense, also encompassing aspects of user emotion, attention and other psychological and cognitive factors. For instance, Zhang et al. [18] make a distinction between immersion in video content in a *spatial* and an *emotional* sense. Agrawal et al. [19] ascribe a cognitive character to immersion while deeming *envelopment* to be of lower-level perceptual nature. Jennett et al. [20] define *presence* in computer games to be related to the fidelity of a virtual environment while reserving immersion for the effect of being absorbed in the playing activity. There is no single consensus on terminology, however, with some authors effectively assigning reversed roles to immersion and presence [21, 22].

In spatial audio evaluation, the provision for higher-level psychological effects may be viewed in the light of evaluating OLE in contrast to BAQ [8, 9], with such notions of immersion increasingly gaining a foothold in the spatial audio community [10, 11, 23]. Using IMEI to study immersive experience, we follow the definition of immersion by Wycisk et al. [13], which is based on the definition by Witmer and Singer as a "psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences" [22], while simultaneously emphasizing emotional and mental aspects of the experience highlighted by Georgiou and Kyza [24]. The IMEI questionnaire asks for ratings of ten items that have been selected in accordance with a many-facet Rasch model of immersion. Each item is rated on a four-point scale and an overall pseudo-continuous immersion rating is obtained by averaging the ten item ratings.

1.2 Perceptual and emotional impact of spatial audio formats

Several studies have dealt with perceptual and emotional effects of spatial sound. In particular, recent attention has been focused on assessing differences between stereo, 2D (loudspeakers surrounding the listener in one horizontal plane) and 3D (including loudspeakers at different heights) sound reproduction. A brief review of these studies shall identify common observations that should ideally be captured by a model of immersion based on sound field features.

1.2.1 Review of related studies

Considering the definition of immersion adopted in the current study, the work of Schoeffler et al. [9] investigating the differences in ratings of BAQ and OLE for spatial audio stimuli is noteworthy. With BAQ being rated against a reference and OLE being assessed without reference, the latter was found to reach a point of saturation with an increasing number of loudspeakers, yielding a limited benefit of 3D reproduction formats over 2D. Similarly, Agrawal et al. [11] found no significant differences in the perceptual attribute of *envelopment* between 5.1 and 7.1.4 sound reproduction in their study examining the impact of audio formats on immersion in movie scenes – while both 5.1 and 7.1.4 reproduction have been rated higher in envelopment than stereo. Additionally, a dependency on the specific movie excerpt presented to the participants was found. Hahn [10] has reported a difference in emotional response to classical music reproduced on an Auro3D 9.1 loudspeaker setup as opposed to a 5.1 arrangement, on the other hand, observing overall emotional arousal and pleasantness to increase with the spatiality of loudspeaker setups as well as being dependent on the program material. Other studies have dealt with ratings of listener preference. Francombe et al. [5] reported increased preference for 3D loudspeaker setups over 2D in music and other multimedia content, but they found the benefit of 3D setups with extended channel counts (particularly 22.2) to be limited as compared to 5.1.4. Silzle et al. [3] found 3D formats to be preferred over 2D and stereo, also observing increased preference for 22.2 over 5.1.4 reproduction. They observed differences in preference to be content-dependent.

Ratings of listener envelopment, overall tonal quality, presence and overall listening experience have been evaluated by Eaton and Lee [4] for stereo, 5.1, 5.1.4 and 22.2 setups in reproduction of classical music. Differences between reproduction formats were found to be content-dependent, which was hypothesized to be due to different instrumentation as well as microphone and recording techniques. Guastavino and Katz [6], studying soundscape reproduction and evaluating a number of attributes, found marked differences between stereo and 2D reproduction, with 3D being rated in between on most attributes.

1.2.2 Key observations

Although the works cited above analyze widely different attributes or notions of immersion, two common themes emerge from the review. Firstly, many of the attributes were found to depend on program material. Secondly, a point of *saturation* in the respective attributes is observed in many studies. The number of loudspeakers at which this effect occurs – and whether it occurs at all – appears to

differ between experiments. It may be assumed that the dependency on the reproduction format can be modeled using sound field features. At the same time, acoustic properties alone may be insufficient for adequate modeling of content-dependency.

1.3 Modeling based on sound field parameters

The aforementioned studies have focused on the speaker setup as the independent variable. Analyses of how the physical sound field influences perception have mostly been carried out in the field of concert hall acoustics – where there is no speaker setup to serve as a proxy variable for sound field parameters. The use of loudspeakers in listening experiments, however, leads to experimental procedures similar to studies on multichannel loudspeaker reproduction [25–27]. Acoustic metrics for modeling attributes such as *listener envelopment* in concert halls were also applied to multichannel loudspeaker reproduction independent of the original context [28]. Interaural parameters and properties of reverberation are typically evaluated as acoustic predictors of perception, in accordance with general knowledge on the psychoacoustics of spatial sound [29].

In soundscape research, a greater variety of psychoacoustic metrics is commonly employed for modeling human perception [30], including loudness and spectral properties. Adapting the sound field parameters used by Bergner and Peissig [31] and Bergner et al. [12], these additional classes of sound field features are included in the initial set of parameters utilized in the modeling approach presented here.

A notable study concerning the methodology of selecting features for perceptual modeling is that of Sarroff and Bello [32]. They took a machine-learning approach to the prediction of perceived spaciousness in stereo recordings. Good prediction performance was achieved using support vector regression, but the features obtained by an automatic feature selection procedure were found to be lacking in interpretability to some degree. As a limitation of the approach, they addressed that only two spatial features were used among mostly temporal and spectral properties of the music recordings – highlighting that the selection of features fed into an automatic model building method is crucial to the final results.

2 Experimental study

This section shall introduce the technical setup and the music program material used in the experimental study.

2.1 Technical setup

The experimental study has been carried out in the Immersive Media Laboratory (IML) [33] at the Institute of Communications Technology (IKT). The listening room is largely compliant with ITU-R BS.1116-3 [34] in terms of background noise, reverberation time and frequency responses of the loudspeakers – except for incorporating *room gain* at low frequencies [35], which was deemed more appropriate for music listening. The setup used in the study is made up of nine Neumann KH 120 A full-range speakers and two Neumann KH 810 G subwoofers. Full-range speaker positions are according to ITU-R BS.2051-2 [36] setup D. Further details on the listening room, loudspeaker setup and equalization are given by Hupke et al. [37] and Bergner et al. [12].

A graphical user interface (GUI) for the IMEI questionnaire has been implemented using the `QUEST` software [38]. The questionnaire GUI was presented on a tablet computer positioned on a stand next to the listening position. In addition to questionnaire responses by the participants, psychophysiological parameters were recorded using electromyography (EMG), electrodermal activity (EDA), breath monitoring and pupillometry equipment. This paper is concerned with the IMEI questionnaire results only.

2.2 Stimuli

Eight musical pieces have been utilized in the experimental study. The pieces represent a variety of genres and production techniques, with versions for different loudspeaker setups prepared for the study. In this paper, the term *stimulus* is used to refer to a particular piece in a particular reproduction format. The formats mono and stereo refer to a single center loudspeaker and two loudspeakers at $\pm 30°$ in the listening plane, respectively. 2D and 3D refer to loudspeaker arrangements 5.1 and 5.1.4 according to ITU-R BS.2051-2 [36]. In view of the tradeoff between the number of program material items and reproduction formats that may reasonably be evaluated in a single-session listening experiment, the selection of formats has been limited to the minimal configurations introducing rear (2D) and height loudspeakers (3D) among common setups.

Additional stimuli have been generated for the pieces Bilder and Rokoko by lowering the playback level of the 3D version by 5dB and playing back the mono signal over all speakers of the 3D setup (referred to as multi Mono in Tab. 1), respectively. The 3D variant of the piece Hantel has been presented twice over the course of the experiment. Note that not all combinations of pieces and formats have been used in order to limit the experiment to an appropriate duration. Combinations with all pieces have been tested for the stereo, 2D and 3D setups, whereas mono, multiMono and 3D (-5dB) are considered as anchor-like or control conditions, respectively. An overview of the stimuli is given in Table 1.

Mixes for stereo, 5.1 and 5.1.4 were prepared by two sound engineers, aiming to produce stimuli that are aesthetically similar across reproduction formats and vary only in their spatial characteristics. The mono version has been generated by averaging the stereo channels. The level of the stimuli has been normalized within each piece with respect to median deviation of the short term LUFS time series between each format and the respective stereo version. The overall level for each piece has been adjusted by an audio engineer taking into consideration the genre and ensemble for each piece. For reference, the average

Table 1. Musical pieces and reproduction formats used in the experimental study. Short names of the pieces used throughout this paper are highlighted in bold.

Piece	Composer	Genre, Ensemble	Production	Dur. in s	Stereo avg. lvl. in dB(A)	Mono	Stereo	2D	3D	3D (−5dB)	multi Mono	3D Retest
Walkürenritt	R. Wagner	Opera (orchestra, fem. voices)	Manual upmix from commercial 5.1 [40]	62.5	80.4		X	X	X			
School's Out	A. Cooper, M. Bruce, G. Buxton, D. Dunaway, N. Smith	Rock (band with male voice)	spot mics + 3D ambience (live)	57.5	79.0		X	X	X			
In a **Mellow** Tone	D. Ellington, performed by J. Berger	Jazz (band with fem. voice)	3D mic. setup + support mics	35.4	73.4		X	X	X			
Im **Wunderschön**en Monat Mai	R. Schumann	Art song (piano, male voice)	3D mic. setup + support mics	38.6	67.9		X	X	X			
Laudate Dominum	J. Vila	Choir (12 singers)	3D mic. setup + support mics	33.4	71.6		X	X	X			
Bilder einer Ausstellung – Das große Tor von Kiew	M. Mussorgsky	Classical (large orchestra)	3D mic. setup + support mics	37.3	76.3		X	X	X	X		
Die **Hantel**	F. Thiesen	Electropop (synthesizers, voices)	multitrack studio production	61.8	75.5	X	X	X	X			X*
Rokoko variations – Finale	P. Tchaikovsky	Classical (woodwind quintet, cello)	manual upmix from commercial 5.1 [41]	68.1	71.1	X	X	X	X		X	

* Not used in the statistical evaluations.

A-weighted level at the listening position is given for the stereo version of each piece in Table 1. Stimulus duration has been chosen to provide sufficient time for an emotional response to form [39], and to represent a musically coherent excerpt of the respective piece with consideration of the genre and tempo, resulting in stimuli of varying duration. For further details on the stimuli, the reader is referred to Bergner et al. [12]. In similar studies, stimuli have been prepared using automatic downmixing algorithms [3, 4, 9], decoding from Ambisonics [6], by muting channels [3] as well as individual mixes being produced for different reproduction formats [5, 10]. It is important to note that the current modeling approach is based entirely on sound field features of each stimulus which are expected to capture any remaining inconsistencies between mixes.

2.3 Subjects and study design

The experimental study has been carried out in a within-subject design with each participant listening to all stimuli marked with a "X" in Table 1. In total, 57 subjects participated in the study (31 female and 26 male, avg. age 26.1 years, std. dev. 6.5 years). The order of the stimuli has been randomized for each participant under the constraint that the same piece (in different formats) would not occur more than twice in a row. The order of the IMEI items has also been randomized for each participant but kept constant across each participant's session. In addition to IMEI item ratings, each participant's reported personal liking of the musical pieces has been collected on the same four-point scale as the IMEI questionnaire. The duration of the laboratory experiment was approximately 70min. After setting up the psychophysiological measurement equipment (EDA, EMG, pupillometry), an experimenter instructed the participants in the procedure and then left the room. Participants independently started the playback of a stimulus via the GUI. The IMEI items were displayed after each stimulus had finished playing.

3 Modeling based on sound field features

This section details the definition and measurement of sound field features as well as their application in the proposed modeling framework.

3.1 Sound field parameters

In order to analyze sound field properties of the stimuli, re-recordings at the listening position using an EM32 *Eigenmike*® [42] have been carried out. Sound field parameters have been computed based on a fourth order spherical harmonic representation of the sound field, a zeroth-order pressure representation and a magnitude least-squares binaural rendering [43] generated using head-related transfer functions (HRTFs) of a Neumann KU100 head simulator [44] as implemented in the IEM BinauralDecoder plugin [45]. Rendering from Ambisonics has been chosen over re-recording with an actual dummy head in order to exclude uncertainties arising from the use of different equipment and multiple re-recording runs. Since the chosen rendering method has been shown to perform well perceptually for music reproduction [46], it is considered to reproduce binaural cues with adequate precision. Because of the participants' limited ability to move out of the listening position due to the presence of the psycho-physiological measurement equipment, the current study focuses on the *sweet spot* only. The variation of sound field parameters around the listening position is dependent on loudspeaker setups and rendering techniques used [47]. A re-recording at the *sweet spot* is assumed to be sufficiently representative of the sound field experienced by the human subjects for the channel-based reproduction formats considered here. The parameter computation pipeline based on the MATLAB Audio Toolbox [48], the Spherical Array Processing Toolbox [49], the Auditory Modeling Toolbox [50], and AudioCommons [51] has been adapted from Bergner and Peissig [31] and Bergner et al. [12]. The quantities used as sound field parameters are listed in Table 2 and can be grouped roughly into four categories: timbre/quality, temporal variation, loudness, and spatial/binaural properties. Sound field parameters have been computed as time series with a 0.1 s sliding window at a stride of 0.05 s. These values have been selected to be reasonably certain that variability due to sound events in the musical excerpts may be captured while limiting computational effort. Parameters have been calculated from the broadband audio signals (B) and, where applicable, in four frequency bands: (0) bass below 100 Hz, (1) low mids from 100 Hz to 500 Hz, (2) high mids up to 2.5 kHz, and (3) high frequencies above 2.5 kHz. Barring the division into low and high mids at 500 Hz, the crossover frequencies are inspired by Olive et al. [52]. The four broad frequency bands have been utilized for two reasons. Firstly, since the current study deals with music reproduction, it can be assumed that the given bands roughly match the partitioning of the auditory frequency spectrum typically undertaken by sound engineers when making mixing decisions. Secondly, results obtained for such broad frequency subdivisions are considered to be more readily interpretable than finer frequency resolutions such as octave or Gammatone bands.

3.2 Derived features

Being a questionnaire result, the IMEI score can only capture an integrative or aggregated measure of the impression made by a particular stimulus on a participant [53]. Therefore summary statistics of the sound field parameter time series have been used as features for the modeling approach.

With sound field parameters being on hand in the form $\psi_i(n, k)$ with time window index n and frequency band index $k \in \{B, 0, 1, 2, 3\}$, the following features have been defined:

$$\psi_{i,\mathrm{mean}}(k) = \underset{n}{\mathrm{mean}}\,[\psi_i(n, k)], \tag{1a}$$

which indicates the average value of feature ψ_i over time in each band k across one stimulus, and

$$\psi_{i,\mathrm{var}}(k) = \underset{n}{\mathrm{var}}\,[\psi_i(n, k)], \tag{1b}$$

Table 2. Sound field parameters used in the evaluation, sorted by category. Full names are given for parameters whose short names may be ambiguous. For full definitions, see Bergner and Peissig [31] and Bergner et al. [12]. Parameters are marked according to the pressure (^{P}), Ambisonics (^{A}) or binaural (^{B}) stimulus representations they depend upon. Underlined parameters are computed from broadband signals and in four frequency bands.

Category	Parameters
Timbre, quality	rough^{P} (roughness), sharp^{P} (sharpness), spectralCentroid^{P}, spectralCrest^{P}, spectralDecrease^{P}, spectralEntropy^{P}, spectralFlux^{P}, spectralKurtosis^{P}, spectralRolloffPoint^{P}, spectralSkewness^{P}, spectralSlope^{P}, spectralSpread^{P}, booming^{P} (timbral booming), MFCC00^{P}, …, MFCC12^{P}
Temporal variation	fluct^{P} (fluctuation strength), modDepthP1^{P}, modDepthP2^{P}, modDepthP3^{P}, (periodic modulation depth), modF1^{P}, modF2^{P}, modF3^{P} (periodic modulation frequency), modDepthS (stochastic modulation depth)
Loudness	LA^{P}, LAeq^{P}, LAmax^{P}, LApeak^{P}, loudnessZwicker^{P}, lufsInt^{P}, lufsMom^{P}, lufsPeak^{P}, lufsRange^{P}, lufsShort^{P}, oct00^{P}, …, oct10^{P} (octave band energy)
Spatial, binaural	diff^{A} (diffuseness), doaAz^{A}, doaEl^{A} (horizontal and vertical direction of arrival), niacc^{B} (normalized inter-aural cross correlation), ild^{B}, itd^{B} (inter-aural level and time difference), sphDI^{A}, sphDIAz^{A}, sphDIEl^{A} (overall, horizontal and vertical spherical directivity index), sphGPRatio^{A} (spherical gradient/pressure ratio), sphGRatio^{A} (spherical gradient ratio), sphPRatio^{A} (spherical pressure ratio)

which refers to the temporal variance of feature ψ_i in each band k. Mean and variance have been used as features due to the parameter distributions being reasonably unimodal and symmetric. Alternative statistics such as median and inter-quartile range are also possible choices within the current approach.

To distinguish between the "raw" sound field *parameters* referred to as ψ_i and the resulting sound field *features,* the latter shall be termed ϕ_j. The sound field features derived from the same parameter in different frequency bands will be considered to be distinct features (e.g. $\phi_1 = \psi_{1,\mathrm{mean}}(0)$, $\phi_2 = \psi_{1,\mathrm{mean}}(1)$, … $\phi_5 = \psi_{1,\mathrm{mean}}(0)$, …). Independent of physical units, each feature ϕ_j has been normalized to the interval [0, 1] across the entire data set for descriptive modeling and across subsets used to cross-validate the predictive model as described in Section 5.

3.3 Modeling approach

A linear mixed modeling approach has been selected due to the repeated-measures study design. In a linear mixed model, the response variable is considered to be composed of a linear combination of fixed effects, a linear combination of random effects and Gaussian noise according to

$$\vec{y} = \Phi \cdot \vec{c}_{\mathrm{f}} + P \cdot \vec{c}_{\mathrm{r}} + \vec{\epsilon}, \tag{2}$$

where $\vec{y}$ is the vector of outcome observations, the matrix Φ contains the fixed-effect predictor values and $\vec{c}_{\mathrm{f}}$ includes the model coefficients for those predictors – this part is identical to a "regular" linear model. P contains the random effects specification while $\vec{c}_{\mathrm{r}} \sim \mathcal{N}(\vec{0}, \Sigma_{\mathrm{r}})$ are the random effects coefficients assumed to be normally distributed with zero mean and covariance matrix Σ_{r}. Finally, the error $\vec{\epsilon} \sim \mathcal{N}(\vec{0}, \Sigma_{\epsilon})$ is assumed to be normally distributed as well with covariance matrix $\Sigma_{\epsilon} = I\sigma_{\epsilon}{}^{2}$, i.e. errors are assumed to be uncorrelated [54].

3.3.1 Model specification

In the model proposed here, Φ contains an intercept K as well as the values $\phi_{j,\zeta\xi}$ of the sound field features. The indexes ζ and ξ refer to the musical piece and the reproduction method, respectively (thus uniquely identifying each stimulus). Concerning the random effects, P's entries encode to which participant η and which piece ζ each observation belongs. Based on the experimental design, a random offset per participant η has been used to account for systematic tendencies in individuals' ratings. An additional random interaction between participants and pieces η, ζ incorporates the effect that a particular piece of music may have on a particular participant's ratings irrespective of the reproduction methods – e.g. a participant particularly liking the genre, interpretation or piece of music. Overall, each IMEI rating $y_{\eta\zeta\xi}$ by participant η for piece ζ in reproduction version ξ is modeled as

$$y_{\eta\zeta\xi} = K + c_{\mathrm{f},1} \cdot \phi_{1,\zeta\xi} + c_{\mathrm{f},2} \cdot \phi_{2,\zeta\xi} + \ldots + c_{\mathrm{r},\eta} + c_{\mathrm{r},\eta\zeta} + \epsilon_{\eta\zeta\xi}. \tag{3}$$

Importantly, any explicit information on the reproduction format (e.g. in the form of format labels) is withheld from the model, the modeling approach is thus based on the sound field features ϕ_j only.

Given the IMEI scores $\vec{y}$, the coefficient vectors $\vec{c}_{\mathrm{f}}$ and $\vec{c}_{\mathrm{r}}$ are to be estimated under the distributional assumptions stated above in order to fit the given data. Unlike in linear regression, no closed-form solution to this problem exists [54], necessitating treatment as an optimization problem in $\vec{c}_{\mathrm{f}}$ and $\vec{c}_{\mathrm{r}}$ (or, more precisely, $\vec{c}_{\mathrm{f}}$ and Σ_{r}). This optimization may be performed in a maximum likelihood framework using a maximum likelihood (ML) or restricted maximum likelihood (REML) metric. When comparing mixed models with different fixed effects, model fitting using ML is recommended [55]. All models have been fitted using ML as implemented in the R package `lmerTest` [56] based on `lme4` [57].

3.3.2 Prediction

Equation (3) defines a descriptive modeling approach for given data, but it may not be used *as-is* for prediction in new instances. This is because the random effects are

Table 3. Results of five runs of automatic feature selection using the genetic algorithm optimizing for AIC (lower is better). A model with randomly chosen features is shown in the last row for comparison.

Selected features	AIC
`modF3Bands3_mean`, `niaccBands3_mean`, `modDepthP3Bands3_var`, `diffBands3_mean`	2753
`sphDIAzBands3_mean`, `modDepthP3Bands3_var`, `sphDIBands3_mean`, `niaccBands3_mean`	2758
`LAmaxBandsl_mean`, `diffBands3_var`, `sphGPRatioBands3_mean`, `niaccBands0_var`	2765
`ildBands2_mean`, `sphGPRatioBands3_mean`, `lufsPeakBands2_mean`, `modDepthSBands0_var`	2762
`niaccBands3_mean`, `modDepthSBands0_var`, `diffBands3_mean`, `lufsIntBands0_var`	2756
`lufsIntBands2_mean`, `modDepthP3Bands3_var`, `lufsRange_mean`, `ildBandsl_mean`	2864

specific to the participants η and the pieces ζ used in model fitting and cannot be known for a new observation. If random effects are used to model disturbances to the actual effects of interest, only the fixed effects part $\mathbf{\Phi} \cdot \vec{c}_{\mathrm{f}}$ of a mixed model may be used for prediction [58]. However, *training* a model including the random effects has been considered necessary to not confound effects related to participants and pieces with the effects of the sound field features. In other words, the random effects are considered to account for participant and piece effects on IMEI in order to be able to deduce effects of the sound field parameters that are valid for the overall population.

4 Feature selection

The procedures described in Section 3 result in a total of 386 features[1]. However, a model based on a few impactful features is considered to be more expedient than a highly complex model. What is more, including all available features at once inevitably leads to numerical issues in mixed model estimation.

In order to arrive at a parsimonious but expressive model, two main strategies may be employed: dimensionality reduction and feature selection. The former is concerned with aggregating relevant information in a set of variables of lower dimensionality. Feature selection, on the other hand, aims to identify a subset of relevant features among those available. Dimensionality reduction by methods such as principal component analysis (PCA) presents a level of abstraction that lessens interpretability, which is the primary reason for adopting a feature selection approach in the current work. Feature selection may be carried out using *wrapper* or *embedded* methods [59]. Wrapper methods perform the search for an optimal feature vector by re-fitting the desired model architecture for different feature subsets, whereas embedded methods incorporate feature selection into the fitting procedure itself. Although a generalization of L_1 regularization of linear models – also known as the least absolute shrinkage and selection operator (LASSO) – has been proposed as an embedded method for mixed models [60], a wrapper approach has been selected here for simplicity.

A wrapper requires a metric for comparing candidate models. Nakagawa's pseudo-R^2 [61] is a metric commonly used to evaluate goodness of fit in mixed models. Generalizing the common R^2 in linear models, two definitions of the pseudo-R^2 exist: The *conditional* $R^2_{cond.}$ assesses variance explained by both fixed and random effects whereas the *marginal* $R^2_{marg.}$ considers the fixed effects only. Given the model specification Equation (3), $R^2_{marg.}$ is of primary interest here. Another common metric is the Akaike information criterion (AIC) [62] which is based on information theoretical reasoning on model fit, with a lower AIC value indicating a better model. Of these metrics, AIC is considered to be more appropriate for model selection whereas R^2-type metrics assess variance explained by a particular model. Unlike R^2, AIC will not automatically favor a more complex model in an attempt to maximize explained variance [55].

To summarize, wrapper feature selection shall implement the search for an optimal feature vector $\vec{\phi}_{\mathrm{opt.}}$ such that a model $\mathcal{M}(\vec{\phi}_{\mathrm{opt.}})$ fitted using those features is the optimal model subject to metric $\mathcal{G}(\mathcal{M})$:

$$\vec{\phi}_{\mathrm{opt.}} = \underset{\vec{\phi}}{\mathrm{argmax}} \left[\mathcal{G}\left(\mathcal{M}\left(\vec{\phi} \right) \right) \right]. \tag{4}$$

4.1 Genetic feature selection

The most generic of feature selection methods is the evaluation of all feature combinations. For the current data set this is intractable, necessitating a heuristic to selectively probe the space of feature vectors. The genetic algorithm has been proposed for feature selection by defining a vector coding for the inclusion and exclusion of features to be optimized [63]. Here, the coding scheme consists of a vector containing $n_{\mathrm{feat.}}$ unique integer indexes, each representing one feature occupying the respective slot [64]. The results of multiple optimization runs with $n_{\mathrm{feat.}} = 4$ are shown in Table 3. The algorithm has been configured with a population size of 32 with two elitist individuals, a mutation chance of 0.1, and has been run for 64 generations in each instance. Five runs have been carried out optimizing for minimum AIC, each starting from a random feature vector.

[1] The total number emerges from $\underbrace{2}_{\text{mean and variance}} \cdot \left(\underbrace{31 \cdot 5}_{\text{broadband and in bands}} + \underbrace{11}_{\text{broadband only}} + \underbrace{11 + 13 + 3}_{\texttt{oct}, \texttt{MFCC}, \texttt{booming}} \right) = 386.$

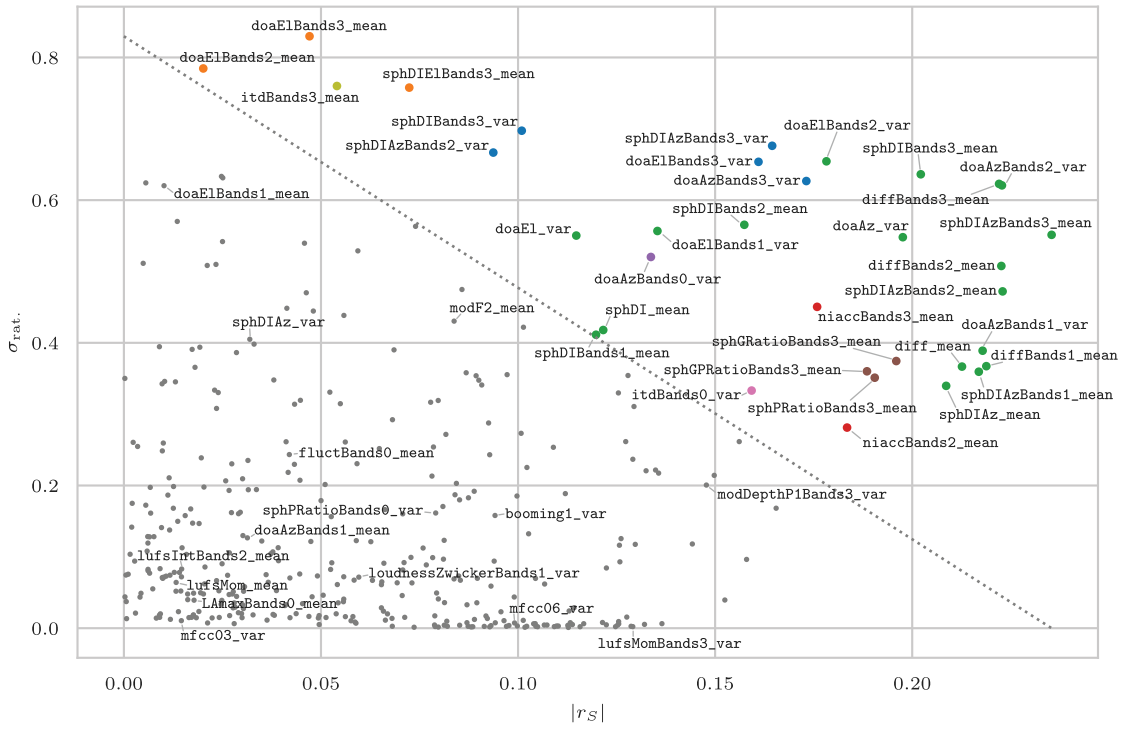


Figure 1. Scatter plot of the within-piece to overall variance ratio $\sigma_{\text{rat.}}$ for each sound field feature against the Spearman rank correlation $|r_S|$ between each feature and the observed IMEI scores. The dotted line indicates the cutoff for the combined rating $q_j > 1.0$. Colors represent the clusters resulting from the procedure described in Section 4.2.2 (cf. Tab. 4). For features with $q_j \leq 1.0$, most labels are hidden for visual clarity.

The emerging models score better than a model based on features selected at random, with features from the spatial and loudness categories occurring most often in the selected subsets. However, optimization converges towards different feature vectors in each run, returning similar AIC values. This suggests that the dependency of model quality on the feature vector is highly non-convex, which in turn implies that a *ranking* of features might be more appropriate than a search for *the* optimal vector $\vec{\phi}_{\text{opt.}}$. On top of that, the interpretability of the selected feature sets is questionable. For these reasons, a pre-selection among the available features (known as a *filter* approach to feature selection [65]) has been implemented.

4.2 Feature pre-selection

4.2.1 Feature screening

The results obtained from genetic feature selection are somewhat reminiscent of the results obtained through automatic feature selection by Sarroff and Bello [32] in that the optimality and interpretability of the resulting feature vector are to be doubted. They highlight as a limitation of their approach that features may be selected because they are confounded with the choice of program material as a consequence of limited data being available. Thus – as a first step in the current feature pre-selection procedure – features that vary strongly between the musical pieces but not between reproduction methods have been identified.

The quantity

$$\sigma_{\text{rat.},j} = \underset{\zeta}{\text{mean}} \left[\left\{ \frac{\underset{\gamma \in \Gamma_\zeta}{\text{var}} \left[\phi_j\right]}{\underset{\gamma}{\text{var}} \left[\phi_j\right]} \forall \zeta \right\} \right], \tag{5}$$

has been defined, where pieces are indexed by ζ and Γ_ζ refers to the subset of all observations belonging to piece ζ (with γ denoting a particular single observation). The metric $\sigma_{\text{rat.},j} < 1$ thus describes the ratio of feature ϕ_j's average variance within each of the pieces to the feature's overall variance. Features that are strongly confounded with the pieces (ζ) have $\sigma_{\text{rat.},j} \to 0$ while features that vary across reproduction methods within the pieces yield $\sigma_{\text{rat.},j} > 0$. As a second metric in the pre-screening method, the absolute Spearman rank correlation $|r_S|_j := |r_S[\phi_j, y]|$ of each feature to the experimental IMEI scores has been used. With the quantities $\sigma_{\text{rat.},j}$ and $|r_S|_j$, only features ϕ_j with

$$q_j = \frac{\sigma_{\text{rat.},j}}{\max\limits_j \sigma_{\text{rat.},j}} + \frac{|r_S|_j}{\max\limits_j |r_S|_j} > 1.0, \tag{6}$$

have been considered for further evaluation. This threshold has been selected by inspection of the distribution of $\sigma_{\text{rat.},j}$ and $|r_S|_j$ shown in Figure 1 to keep approximately 10% of features. It is apparent that a large number of features – particularly ones relating to loudness and spectral properties of the stimuli – have low values of $\sigma_{\text{rat.},j} < 0.25$, indicating that they are highly confounded

Table 4. Features remaining after pre-screening according to criterion q_j. Each row represents equivalent features in the frequency bands stated in Section 3.1, with an entry in the table indicating that a particular feature in a particular band passes the pre-selection criterion. Letters and colors represent the clusters resulting from the procedure described in Section 4.2.2. Cluster colors correspond to Figure 1.

Feature name	Bands (cluster)					Description
	B	0	1	2	3	
`diff_mean`	(C)		(C)	(C)	(C)	Mean diffuseness
`doaEl_mean`				(B)	(B)	Mean vertical elevation
`niacc_mean`				(D)	(D)	Mean normalized inter-aural cross-correlation (IACC)
`itd_mean`					(H)	Mean inter-aural time difference (ITD)
`sphDI_mean`	(C)		(C)	(C)	(C)	Mean spherical directivity index
`sphDIAz_mean`	(C)		(C)	(C)	(C)	Mean horizontal spherical directivity index
`sphDIEl_mean`					(B)	Mean vertical spherical directivity index
`sphGPRatio_mean`					(F)	Mean spherical gradient/pressure ratio
`sphGRatio_mean`					(F)	Mean spherical gradient ratio
`sphPRatio_mean`					(F)	Mean spherical pressure ratio
`doaAz_var`	(C)	(E)	(C)	(C)	(A)	Variance of horizontal direction of arrival (DOA)
`doaEl_var`	(C)		(C)	(C)	(A)	Variance of vertical DOA
`itd_var`		(G)				Variance of ITD
`sphDI_var`					(A)	Variance of spherical directivity index
`sphDIAz_var`				(A)	(A)	Variance of horizontal spherical directivity index

with the musical pieces. This is unsurprising given that the stimuli have been designed with a variation of spatial parameters in mind. While loudness and timbre features could have been discarded simply based on this knowledge, the pre-selection procedure underpins this decision by confirming that the statistics of the sound field features indeed follow the expected patterns. The 34 features remaining after pre-screening – corresponding to spatial sound field properties – are listed in Table 4. The mid to high frequency bands are primarily represented among the remaining features. Only two of the 34 features are based on the bass band and five are broadband features.

Selecting features by (rank) correlation to the target variable may be argued to foster overfitting the given data. This risk is minimized by including considerable variation in terms of genre, ensemble and production techniques. The broad repertoire should reasonably ensure that a feature discarded in the pre-screening process due to low q_j would not play a major predictive role even if a bigger data set were available.

4.2.2 Feature clustering

Some of the features remaining after pre-selection have been found to exhibit high degrees of collinearity. Since collinear features may lead to unreliable model parameter estimates [66], the features in the reduced set have been clustered by mutual absolute Pearson correlation $|r_P[\phi_{j_1}, \phi_{j_2}]|$ between all combinations of features ϕ_{j_1}, ϕ_{j_2} and models in further evaluation have been prevented from including multiple features belonging to the same cluster. Agglomerative hierarchical clustering has been employed with distance metric $1-|r_P[\phi_{j_1}, \phi_{j_2}]|$. By using complete linkage, a minimum correlation of 0.5 between features in each cluster has been ensured. The eight clusters obtained this way are given in Table 4 with the same color coding as in Figure 1. As expected, clusters are mostly made up of features based on the same sound field parameter in different frequency bands.

5 Results and discussion

In this section, results pertaining to the usefulness of the pre-selected sound field features for descriptive modeling of the experimental data will be presented first and an optimal model will be specified. Then, prediction performance of this optimal model specification will be evaluated using a cross-validation scheme, focusing on population mean effects. Finally, the effect of participants' personal liking of the musical pieces will be introduced in order to analyze the prediction accuracy of individual ratings.

5.1 Selected feature set and descriptive modeling

5.1.1 Optimal model specification

After feature pre-selection, all possible models with $n_{\text{feat.}} = 2, 3, 4$ features chosen from Table 4 under the constraint of not including features from the same cluster have been evaluated by AIC. In contrast to the full feature set, this yields a computationally manageable 9474 models. The optimal model is stated in Table 5. It includes mean diffuseness between 500 Hz and 2.5 kHz as well as mean normalized IACC and mean vertical spherical directivity index above 2.5 kHz. Finally, temporal variance of the spherical directivity index above 2.5 kHz serves as a predictor. Based on t-tests (using Satterwaithe's method for estimation of degrees of freedom in the mixed model as implemented in the R package `lmerTest` [56]), the three former effects are determined to be statistically significant whereas the effect of the last feature is not significant at

Table 5. Best model according to the feature selection procedure of Section 4. *t*-tests performed using Satterwaithe's method as implemented in R package `lmerTest` [56].

Fixed effects	Estimates	Standard error	*df*	*t*	p
Intercept	2.19	0.09	435.23	24.368	<0.001
`diffBands2_mean`	0.52	0.07	1524.52	7.173	<0.001
`niaccBands3_mean`	−0.43	0.06	1513.04	−7.667	<0.001
`sphDIElBands3_mean`	0.23	0.07	1473.81	3.496	<0.001
`sphDIBands3_var`	0.11	0.06	1410.96	1.780	0.075
Random effects	$\sigma^2_{\mathrm{r},\eta} = 0.131$ (participant)			AIC = 2764.1	
	$\sigma^2_{\mathrm{r},\eta\zeta} = 0.128$ (participant and piece)			$R^2_{\mathrm{marg.}} = 0.078$	
Residual variance	$\sigma^2_{\epsilon} = 0.225$			$R^2_{cond.} = 0.573$	

the $\alpha = 0.05$ level. Regarding model diagnostics (using the R package `performance` [67]), normality of residuals has been confirmed by a Shapiro-Wilk test ($p = 0.369$) and homogeneity of residual variance across predictor variable values has been found to be fulfilled by a Bartlett test ($p = 0.672$). Shapiro-Wilk tests analyzing the normality of random offsets per participant ($p = 0.030$) and per participant and piece ($p = 0.070$) have found the former to deviate from a normal distribution. Upon inspection of a quantile-quantile plot this deviation has been deemed acceptable.

5.1.2 Feature ranking

Although the optimal model of Table 5 yields the lowest AIC among the set of models permissible after feature pre-selection, it can be reasonably assumed that different model specifications performing similarly well in terms of AIC can be found. Therefore, a feature ranking has been established in parallel to the optimal model search by computing the difference in mean AIC for models including a particular feature versus models not including it. This is displayed in Figure 2. The observed AIC values are in the range from 2764.1 to 2965.2 with a median of 2827.6. Figure 2 shows high-frequency IACC and high-mid frequency diffuseness to be among the highest-ranking features. Inclusion of vertical spherical directivity index, however, even leads to an increase in AIC on average. This suggests that its usefulness as a predictor (being included in the optimal model) is linked to the particular combination with the other predictors instead of the feature's own merit. The inclusion of the variance of high-frequency spherical directivity index is neutral with respect to AIC on average. Indeed, AIC values of the optimal model (AIC = 2764.1) and a version with this feature removed (AIC = 2765.3) barely differ. Further, it is of note that the lowest AIC values resulting from the feature pre-selection strategy are comparable to those obtained from blind AIC optimization using the genetic algorithm as listed in Table 3. However, the models emerging from the proposed approach include more salient features than those chosen by the genetic algorithm, as shall be discussed below.

5.1.3 Discussion of selected features

Diffuseness and normalized IACC are likely feature choices. Diffuseness as defined by Pulkki [68] is a quantity designed to assess a sound field on a continuum from a single plane wave to a completely diffuse field. IACC is known to be linked to perceptual properties such as *auditory source width* (ASW) [69, 70] as well as *envelopment* [69, 71] and *spaciousness* [71], with a deliberate increase in IACC being able to reduce perceived ASW and envelopment in reproduction using binaural room impulse responses [69]. In line with these insights, the regression coefficients stated in Table 5 indicate that higher diffuseness and lower IACC lead to increased immersion. In concert halls, spatial perception is influenced by the azimuthal angular distribution of reflections [25]. Although not featured in the optimal model specification, multiple features based on the horizontal spherical directivity index are among the highest-ranking features in Figure 2. Perceived spaciousness has also been reported to depend on temporal variations of inter-aural parameters [72, 73], which are not prominently featured among the pre-selected features (Tab. 4) and thus are absent from the optimal model.

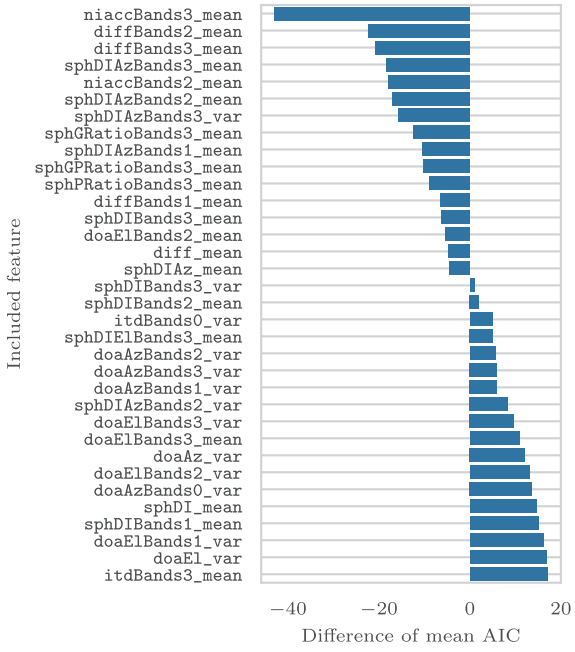
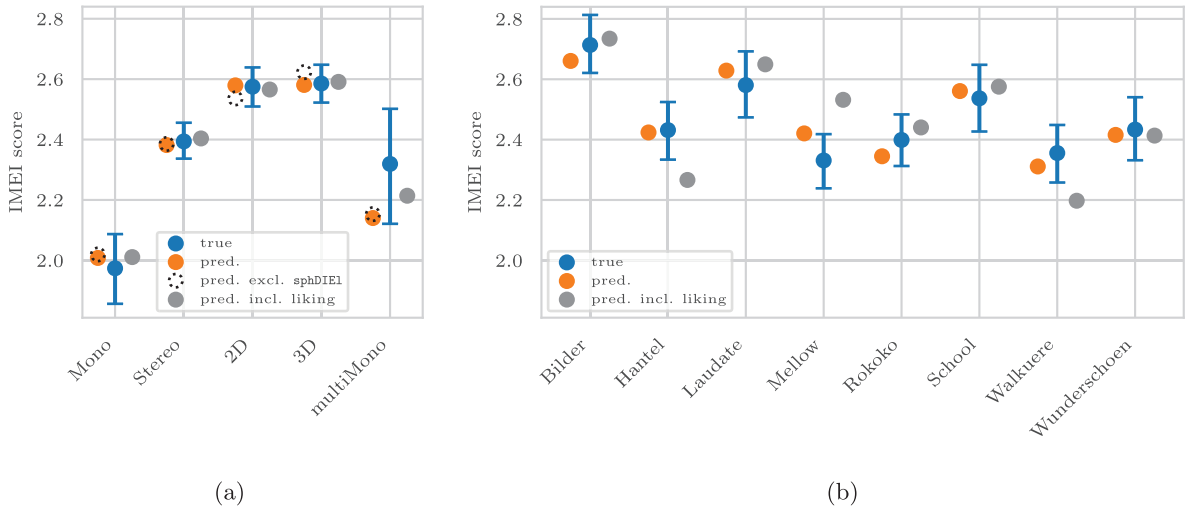


Figure 2. Difference in mean AIC (lower is better) of models including and excluding each feature after pre-selection. All possible combinations of features in Table 4 not assigned to the same cluster have been evaluated.



(a) (b)

Figure 3. Means and bootstrapped 95% confidence intervals of true IMEI ratings as well as mean model predictions in the cross-validation procedure. Predictions are according to different model specifications. As predictions (except for those including liking) contain no between-participant variance, no confidence intervals are shown. (a) By reproduction format. (b) By piece.

The selection of vertical spherical directivity index with a positive regression coefficient is somewhat unexpected because this implies that immersion *increases* if the sound field is *more directional* (narrower) in the vertical dimension. For most of the pieces in the repertoire, the feature `sphDIElBands3_mean` is *lowest* in 3D reproduction. Since this feature is *detrimental* to a model's AIC on average, it is hypothesized that its inclusion is only a good choice in conjunction with IACC and diffuseness. This is to be discussed in detail in Section 5.2.1.

It is important to be aware that the results obtained here may depend on the specific feature computation pipeline. For example, the implementation of ITD and IACC used in this work is based on the ear signals' energy envelopes (termed "IACCe" by Andreopoulou and Katz [74]). This is implemented and recommended as a default by the Auditory Modeling Toolbox [50] and has been considered appropriate in the light of the time resolution of 0.05 s chosen here. Different time resolutions and different sound field parameter definitions may, however, influence results. An analysis of different parameter definitions and implementations is considered to be beyond the scope of this paper.

5.2 Predictive modeling

In addition to descriptively modeling the experimental data, prediction of IMEI ratings from the sound field features has been investigated using the model specification of Table 5. The evaluation has been carried out by cross-validation. 64 instances of partitioning into training and test sets have been performed, with each of those instances setting aside the data for one of the 8 musical pieces and for 10 of the 57 participants for validation. Evaluation of each model with newly estimated coefficients is thus always performed on data for an unknown piece and an unknown participant. As mentioned in Section 3.3.2, random effects cannot be known for new data, and are considered as disturbances in the current modeling framework. Hence only the fixed effects part of a model is used for prediction. When fitting each of the 64 models, normalization of sound field features to [0, 1] is performed based on the respective training set only to prevent information on the test set from bleeding into training data. The normalization parameters are then stored and applied to the test data. This approach relies on a training set with all features spanning representative value ranges. Alternatively, normalization may be carried out on the basis of expected ranges the features may take. However, this requires the incorporation of prior assumptions on feature distributions. It shall also be remarked that the size of the collected data set has necessitated validation to be carried on the same set of data used for initial feature selection. For more thorough validation of the proposed approach, larger-scale data collection and evaluation is required.

5.2.1 Prediction of mean ratings

In order to evaluate whether the proposed modeling approach is able to represent the saturation of immersion and of content-dependency mentioned in Section 1.2 and observed in the current study, it is necessary to compare mean model predictions to the means of observed IMEI ratings by reproduction format and by piece. Figure 3 displays the means and bootstrapped 95% confidence intervals of observed IMEI scores by reproduction format and by piece along with prediction results averaged over the 64 cross-validation runs. A good match between true and predicted means is apparent for mono, stereo, 2D and 3D reproduction with an average absolute difference of experimental mean ratings and mean predictions of 0.01. Notably, the experimental results show a very small difference between the mean ratings for 2D and 3D, which is in line with the saturation effects often observed in similar studies

[6, 9, 11]. This behavior is well-captured by mean model predictions. The feature `sphDIElBands3_mean` plays a key role in modeling the relationship between the 2D and 3D formats as confirmed by repeating the cross-validation procedure with a model specification omitting this feature. As shown in Figure 3a, this leads to underestimation of mean 2D IMEI scores and overestimation of 3D scores, with means of predictions and observed ratings differing by ±0.04 in both cases. The proposed feature selection procedure can thus be deemed to have succeeded at identifying a suitable choice of predictors. More generally, the prediction results show that ratings of immersion – on a population-average level – may indeed be traced down to acoustic properties of the stimuli.

The mean prediction for format multiMono is imprecise, however, with mean predictions differing by 0.18 from the mean experimental rating (although the predicted mean is still within the true mean's confidence interval). This may be linked to the fact that only one piece (Rokoko) was presented in this format. Although all predictions were made for previously unseen participants and pieces, the training data did contain observations with the same *format* for mono, stereo, 2D, and 3D. Training on data omitting a format altogether has not been performed because the limited number of reproduction methods means that removing one format would strongly distort the distribution of sound field features learned by a model.

Observed and predicted mean IMEI scores by piece are displayed in Figure 3b. The average absolute difference between experimental mean IMEI scores and predictions is 0.04, with the greatest difference being observed for the piece Mellow at 0.09. Note that features being confounded with the musical pieces accounting for the prediction results can be ruled out by design of the feature selection procedure. Variation in IMEI ratings between pieces can thus be partially explained by their acoustic properties.

While the proposed modeling approach shows good prediction performance in the cross-validation scheme applied here, the reliability of the presented observations may benefit from further substantiation. In particular, more variation in sound field features (i.e. more different stimuli, even at the cost of fewer participant ratings per stimulus) would be desirable. Although 1596 observations of IMEI ratings have been used in the current evaluation, only 28 distinct values have been available for each sound field feature. This is due to the study being subject to various constraints resulting in the design stated in Section 2.

5.2.2 Prediction of individual ratings

Distinct from *mean* predictions is the evaluation of how precise *individual* IMEI predictions are with respect to the observed ratings. In order to predict individual ratings, subjective effects need to be re-introduced to the model specification. To this end, the specification of Table 5 has been extended by adding reported liking as an ordinal predictor variable with a fixed effect. Cross-validation has been repeated for this extended model specification with the same data partitioning as for the "sound field only" approach. Per-instance error observed in cross-validation is plotted against the respective predicted value in Figure 4 for both specifications.

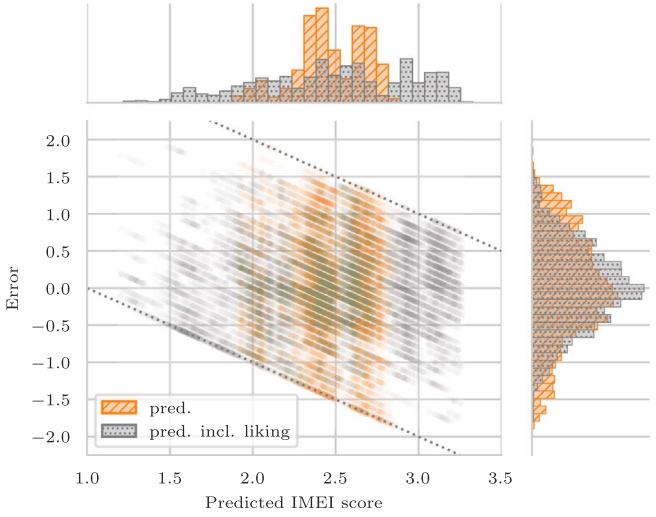


Figure 4. Errors plotted against predicted IMEI scores in the cross-validation procedure for a model specification according to Table 5 and one additionally including liking. Marginal distributions are displayed as histograms. Minimum and maximum possible errors due to boundedness of IMEI scores are indicated by dotted lines.

The models including liking predict more extremal values in the range of 1.2–3.26 whereas the predictions of the sound field only models are capped at 1.9 and 2.85. The error distribution reveals that the sound field only models yield greater absolute errors. Indeed, mean absolute error (MAE) for this specification is 0.584 across all observations in cross-validation whereas it is 0.446 for the models including liking. A model including liking is thus able to predict individual IMEI scores more accurately than a model only using population average effects of sound field features. The descriptive model of Table 5 – which expresses subjective impact on immersion as random effects – confirms this finding. This is evident from the metrics $R^2_{\text{marg.}} = 0.078$ and $R^2_{\text{cond.}} = 0.573$ of the descriptive model. Since $R^2_{\text{cond.}}$ includes variance explained by random effects whereas $R^2_{\text{marg.}}$ indicates variance explained by fixed effects only, is clear that a more substantial amount of variance in the ratings is actually explained by the effects of participants and pieces rather than the sound field features.

On the other hand, Figure 3b shows that mean predictions by piece are less accurate with respect to experimental mean IMEI ratings when liking is part of the model specification. In the mean predictions by format shown in Figure 3a, this effect appears to be averaged out across pieces whereas some mean predicted IMEI ratings by piece are biased. Although not the focus of this paper, these observations show that subjective effects are in need of further investigation. In particular, the current (sound field only) modeling approach regards effects of sound field features to be independent of participant and piece effects.

A more general approach may include individual differences in perception, for instance by incorporating per-person random slopes in the mixed model framework in addition to the random offsets used here.

The slopes at the bottom and top of the point clouds in Figure 4 are due to IMEI scores being bounded between 1 and 4 (e.g. a predicted score of 1.5 may not have an error lower than −0.5 as the true score cannot be below 1). This points to the perceptual response to the stimuli being somewhat nonlinear in the first place. Results obtained here support the conclusion that a linear relationship may be assumed within a certain interval, but modeling of scores at the bounds of the domain may benefit from a nonlinear approach such as generalized linear mixed effects models including variable transformations through nonlinear link functions. This is further substantiated by the fact that existing auditory models do utilize psychoacoustically motivated nonlinear dependencies on parameters such as IACC [69].

6 Summary and conclusion

This paper presents an approach for modeling immersion – as quantified by IMEI – based on acoustic properties of musical stimuli reproduced using different surround sound loudspeaker setups. Using immersion ratings collected in the experimental study, a linear mixed effects approach has been employed in order to estimate population-average effects of sound field features while accounting for subjective effects and effects of the musical pieces on individual subjects.

To arrive at sound field parameters serving as meaningful predictors of immersion, a feature selection procedure based on feature statistics and model fit has been implemented. Feature selection confirms quantities related to spatial sound perception such as IACC and diffuseness, particularly in the mid to high frequency range, to be relevant in modeling immersion. With the selected features, a descriptive model of the experimental data has been estimated. Application of the model specification in a predictive capacity has been explored by means of cross-validation. The modeling approach has been found to predict mean ratings for the various reproduction methods with high accuracy and mean ratings per musical piece with a somewhat lower accuracy. This permits to draw the tentative conclusion that the immersion *saturation* phenomenon observed in this and other studies – i.e. an increase in the number of loudspeakers yielding diminishing returns in terms of immersion and related attributes – can be traced down to acoustic features of sound reproduced over different loudspeaker setups. The content-dependency of immersion, which is also commonly observed, can partially be explained by acoustic features. A study design featuring greater variation in acoustic features is identified as a prerequisite for more general validation of the proposed modeling approach and an evaluation of the general validity of the features identified as predictors. Furthermore, an analysis of the model's sensitivity to the sound field parameters is a necessary next step towards a generalization of the approach. Specifically, the dependencies of the computed sound field features (and, by extension, model predictions) on the computation pipeline's implementation specifics require further evaluation. Additionally, an investigation of the sound field parameters' spatial variation and uncertainty – potentially leading to an incorporation of this variation into the model – could help to ensure the model's validity outside of the currently assumed *sweet spot* listening conditions.

Although this is not the focus of this paper, the role of subjective effects has been found to be highly relevant to the psychological construct of immersion. In descriptive modeling, the effects of participants and pieces have been found to explain substantially more variance than the sound field features. In predictive modeling, introducing reported liking of the musical piece as an additional predictor yields more accurate predictions of individual immersion ratings over a model based on sound field features only. Nonetheless, sound field features alone serve as good predictors of immersion on the population level.

Overall, the proposed approach may be regarded as a step towards a framework for modeling and predicting perceptual and psychological responses to surround sound loudspeaker reproduction independent of particular speaker arrangements, production techniques and content types and formats. Comparing acoustic properties could help to explain and contextualize varying results obtained in different studies on multichannel loudspeaker reproduction – which is otherwise relegated to hypothetical deliberations based on different experimental conditions and program material.

Conflict of interest

The authors declare that they have no conflicts of interest in relation to this article.

Acknowledgments

The authors are thankful for the research grant of the project Richard Wagner 3.0 funded by "Niedersachsisches Vorab", a joint program by the Volkswagen Foundation in conjunction with the Lower Saxony Ministry for Science and Culture (ZN3497).

Data availability statement

Questionnaire responses, sound field parameter data and derived feature data for the stimuli as well as the results of the statistical evaluations described in this paper are available at https://data.uni-hannover.de/dataset/immersive-music-experience-in-surround-sound-music-reproduction [75].

A Matlab toolset and exemplary code for computation of sound field features from spherical harmonic, binaural and pressure representations of stimuli is available at https://gitlab.com/janywhere/sosca-indicators/ [76]. Python and R code implementing the feature selection and modeling

methodology presented in this paper is available at https://gitlab.uni-hannover.de/roman.kiyan.jr/immersionmodeling/ [77].

References

1. F. Rumsey: Surround Sound. In: A. Roginska, P. Geluso, Eds., Immersive sound, Routledge, 2017, pp. 180–220. https://doi.org/10.4324/9781315707525.
2. S. Kim: Height Channels. In: A. Roginska, P. Geluso, Eds., Immersive Sound, Routledge, 2017, pp. 221–243. https://doi.org/10.4324/9781315707525.
3. A. Silzle, S. George, E.A. Habets, T. Bachmann: Investigation on the quality of 3D sound reproduction. In: Proceedings of the International Conference on Spatial Audio, Detmold, Germany, 10–13 November, Verband Deutscher Tonmeister, 2011, pp. 334–341.
4. C. Eaton, H. Lee: Subjective evaluations of three-dimensional, surround and stereo loudspeaker reproductions using classical music recordings. Acoustical Science and Technology 43, 2 (2022) 149–161. https://doi.org/10.1250/ast.43.149.
5. J. Francombe, T. Brookes, R. Mason, J. Woodcock: Evaluation of spatial audio reproduction methods (part 2): analysis of listener preference. Journal of the Audio Engineering Society 65, 3 (2017) 212–225. https://doi.org/10.17743/jaes.2016.0071.
6. C. Guastavino, B.F.G. Katz: Perceptual evaluation of multidimensional spatial audio reproduction. Journal of the Acoustical Society of America 116, 2 (2004) 1105–1115. https://doi.org/10.1121/1.1763973.
7. A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkman, S. Weinzierl: A Spatial Audio Quality Inventory (SAQI), Acta Acustica united with Acustica 100, 5 (2014) 984–994. https://doi.org/10.3813/aaa.918778.
8. M. Schoeffler, J. Herre: The relationship between basic audio quality and overall listening experience, The Journal of the Acoustical Society of America 140, 3 (2016) 2101–2112. https://doi.org/10.1121/1.4963078.
9. M. Schoeffler, A. Silzle, J. Herre: Evaluation of spatial/3D audio: basic audio quality versus quality of experience. IEEE Journal of Selected Topics in Signal Processing 11, 1 (2017) 75–88. https://doi.org/10.1109/jstsp.2016.2639325.
10. E. Hahn: Musical emotions evoked by 3D audio. In: Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science. 2018. Available at https://www.aes.org/e-lib/browse.cfm?elib=19640.
11. S. Agrawal, S. Bech, K.D. Moor, S. Forchhammer: Influence of changes in audio spatialization on immersion in audiovisual experiences, Journal of the Audio Engineering Society 70, 10 (2022) 810–823. https://doi.org/10.17743/jaes.2022.0034.
12. J. Bergner, D. Schössow, S. Preihs, J. Peissig: Identification of discriminative acoustic dimensions in stereo, surround and 3D music reproduction. Journal of the Audio Engineering Society 71, 7/8 (2023) 420–430. https://doi.org/10.17743/jaes.2022.0071.
13. Y. Wycisk, K. Sander, R. Kopiez, F. Platz, S. Preihs, J. Peissig: Wrapped into sound: development of the immersive music experience inventory (IMEI). Frontiers in Psychology 13 (2022) 951161. https://doi.org/10.3389/fpsyg.2022.951161.
14. J. Berg, F. Rumsey: Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In: Audio Engineering Society Conference: 19th International Conference: Surround Sound - Techniques, Technology, and Perception. 2001. Available at https://www.aes.org/e-lib/browse.cfm?elib=10057.
15. C. Colomes, S. Le Bagousse, M. Paquier: Families of sound attributes for assessment of spatial audio. In: 129th Audio Engineering Society Convention. Audio Engineering Society. 2010. Available at https://www.aes.org/e-lib/browse.cfm?elib=15728.
16. N. Zacharov, T.H. Pedersen: Spatial sound attributes – development of a common lexicon. In: 139th Audio Engineering Society Convention. 2015. Available at https://www.aes.org/e-lib/browse.cfm?elib=17992.
17. F. Rumsey: Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm. Journal of the Audio Engineering Society 50, 9 (2002) 651–666. Available at https://www.aes.org/e-lib/browse.cfm?elib=11067.
18. C. Zhang, A. Perkis, S. Arndt: Spatial immersion versus emotional immersion, which is more immersive? In: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2017. https://doi.org/10.1109/qomex.2017.7965655.
19. S. Agrawal, A. Simon, S. Bech, K. Bærentsen, S. Forchhammer: Defining immersion: literature review and implications for research on audiovisual experiences, Journal of the Audio Engineering Society 68, 6 (2020) 404–417. https://doi.org/10.17743/jaes.2020.0039.
20. C. Jennett, A.L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, A. Walton: Measuring and defining the experience of immersion in games. International Journal of Human-Computer Studies 66, 9 (2008) 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004.
21. N.C. Nilsson, R. Nordahl, S. Serafin: Immersion revisited: a review of existing definitions of immersion and their relation to different theories of presence. Human Technology 12, 2 (2016), 108–134. https://doi.org/10.17011/ht/urn.201611174652.
22. B.G. Witmer, M.J. Singer: Measuring presence in virtual environments: a presence questionnaire. Presence: Teleoperators and Virtual Environments 7, 3 (1998) 225–240. https://doi.org/10.1162/105474698565686.
23. J. Kelly, W. Woszczyk, R. King. Are you there? A literature review of presence for immersive music reproduction. In: 149th Audio Engineering Society Convention October 27–30, 2020. Online. Available at https://www.aes.org/e-lib/browse.cfm?elib=20926.
24. Y. Georgiou, E.A. Kyza: The development and validation of the ARI questionnaire: An instrument for measuring immersion in location-based augmented reality settings. International Journal of Human-Computer Studies 98 (2017) 24–37. https://doi.org/10.1016/j.ijhcs.2016.09.014.
25. J.S. Bradley, G.A. Soulodre: Objective measures of listener envelopment. Journal of the Acoustical Society of America 98, 5 (1995) 2590–2597. https://doi.org/10.1121/1.413225.
26. H. Furuya, K. Fujimoto, A. Wakuda, Y. Nakano: The influence of total and directional energy of late sound on listener envelopment. Acoustical Science and Technology 26, 2 (2005) 208–211. https://doi.org/10.1250/ast.26.208.
27. J. Pätynen, T. Lokki: Perception of music dynamics in concert hall acoustics. Journal of the Acoustical Society of America 140, 5 (2016) 3787–3798. https://doi.org/10.1121/1.4967157.
28. G.A. Soulodre, M.C. Lavoie, S.G. Norcross: Objective measures of listener envelopment in multichannel surround systems. Journal of the Audio Engineering Society 51, 9 (2003) 826–840. Available at https://www.aes.org/e-lib/browse.cfm?elib=12284.
29. J. Blauert: Spatial hearing: the psychophysics of human sound localization, MIT Press, 1997.
30. M.S. Engel, A. Fiebig, C. Pfaffenbach, J. Fels: A review of the use of psychoacoustic indicators on soundscape studies. Current Pollution Reports 7, 3 (2021) 359–378. https://doi.org/10.1007/s40726-021-00197-1.

31. J. Bergner, J. Peissig: On the identification and assessment of underlying acoustic dimensions of soundscapes, Acta Acustica 6 (2022) 46. https://doi.org/10.1051/aacus/2022042.
32. A.M. Sarroff, J.P. Bello: Toward a computational model of perceived spaciousness in recorded music. Journal of the Audio Engineering Society 59, 7/8 (2011) 498–513. Available at https://www.aes.org/e-lib/browse.cfm?elib=15975.
33. R. Hupke, M. Nophut, S. Li, R. Schlieper, S. Preihs, J. Peissig: The immersive media laboratory: Installation of a novel multichannel audio laboratory for immersive media applications. In: 144th Audio Engineering Society Convention. 2018. Available at https://www.aes.org/e-lib/browse.cfm?elib=19522.
34. ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems. International Telecommunications Union. 2015.
35. J. Abildgaard Pedersen, F. El-Azm: Natural timbre in room correction systems (Part II). In: The Proceeding of the Audio Engineering Society Conference: 32nd International Conference: DSP for Loudspeakers, Hillerød, Denmark, 21–23 September, Audio Engineering Society, 2007. Available at https://www.aes.org/e-lib/browse.cfm?elib=14201.
36. ITU-R BS.2051-2: Advanced sound system for programme production, International Telecommunications Union, 2018.
37. R. Hupke, J. Ordner, J. Bergner, M. Nophut, S. Preihs, J. Peissig: Towards a Virtual Audiovisual Environment for Interactive 3D Audio Productions. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27–29 March, Audio Engineering Society, 2019. Available at https://www.aes.org/e-lib/browse.cfm?elib=20438.
38. D. Schössow: QUEST - Questionnaire Editor SysTem, 2022. Available at https://doi.org/10.5281/ZENODO.7360198.
39. J.P. Bachorik, M. Bangert, P. Loui, K. Larke, J. Berger, R. Rowe, G. Schlaug: Emotion in motion: investigating the time-course of emotional judgments of musical stimuli. Music Perception 26, 4 (2009) 355–364. https://doi.org/10.1525/mp.2009.26.4.355.
40. Richard Wagner – Die Walküre. The Metropolitan Opera. Deutsche Grammophon 0734855. 2011.
41. Pyotr Tchaikovsky - Rococo Variations for Cello and Wind Quintet. Orchestra Academy of the Bayerisches Staatsorchester. Hänssler Classic. 2018.
42. J. Meyer, G. Elko: A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. In: IEEE International Conference on Acoustics Speech and Signal Processing, Orlando, FL, USA, 13–17 May, IEEE, 2002. https://doi.org/10.1109/icassp.2002.5744968.
43. C. Schörkhuber, M. Zaunschirm, R. Höldrich: Binaural rendering of ambisonic signals via magnitude least squares. In: Fortschritte der Akustik - DAGAMunich, Germany, 19–22 March, Deutsche Gesellschaft für Akustik e.V. (DEGA), 2018, pp. 339–342.
44. B. Bernschütz: A spherical far field HRIR/HRTF compilation of the Neumann KU 100. In: Fortschritte der Akustik – AIA-DAGA 2013, Merano, Italy, 18–21 March, German Acoustical Society (DEGA), Berlin, 2013, pp. 592–595.
45. IEM Plugin Suite: Institut für Elektronische Musik und Akustik, Universität für Musik und darstellende Kunst Graz, 2021. Available at https://plugins.iem.at.
46. H. Lee, M. Frank, F. Zotter: Spatial and timbral fidelities of binaural Ambisonics decoders for main microphone array recordings. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio, York, UK, 27–29 March, Audio Engineering Society, 2019. Available at https://www.aes.org/e-lib/browse.cfm?elib=20392.
47. S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, F. Zotter: Spatial sound with loudspeakers and its perception: a review of the current state. Proceedings of the IEEE 101, 9 (2013) 1920–1938. https://doi.org/10.1109/jproc.2013.2264784.
48. The Mathworks Inc: Audio Toolbox. 2022. Available at https://de.mathworks.com/products/audio.html.
49. A. Politis: Microphone array processing for parametric spatial audio techniques. PhD thesis, Department of Signal Processing and Acoustics, Aalto University, Finland, 2016.
50. P. Majdak, C. Hollomey, R. Baumgartner: AMT 1.x: A toolbox for reproducible research in auditory modeling. Acta Acustica 6 (2022) 19. https://doi.org/10.1051/aacus/2022011.
51. Institute of Sound Recording, University of Surrey: AudioCommons timbral models. 2019. Available at https://www.audiocommons.org/materials/.
52. S. Olive, T. Welti, E. McMullin, Listener preferences for in-room loudspeaker and headphone target responses. In: 135th Audio Engineering Society Convention, New York, NY, 17–20 October, 2013. Available at https://www.aes.org/e-lib/browse.cfm?elib=17042.
53. C. Zhang: The why, what, and how of immersive experience, IEEE Access 8 (2020) 90878–90888. https://doi.org/10.1109/access.2020.2993646.
54. J.O. Rawlings, S.G. Pantula, D.A. Dickey, Eds.: Applied regression analysis, Springer-Verlag, 1998. https://doi.org/10.1007/b98890.
55. E. Cantoni, N. Jacot, P. Ghisletta: Review and comparison of measures of explained variation and model selection in linear mixed-effects models. Econometrics and Statistics (2021). https://doi.org/10.1016/j.ecosta.2021.05.005.
56. A. Kuznetsova, P.B. Brockhoff, R.H.B. Christensen: lmerTest package: tests in linear mixed effects models. Journal of Statistical Software 82, 13 (2017) 1–26. https://doi.org/10.18637/jss.v082.i13.
57. D. Bates, M. Mächler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4, Journal of Statistical Software 67, 1 (2015) 1–48. https://doi.org/10.18637/jss.v067.i01.
58. S. Welham, B. Cullis, B. Gogel, A. Gilmour, R. Thompson: Prediction in linear mixed models. Australian & New Zealand Journal of Statistics 46, 3 (2004) 325–347. https://doi.org/10.1111/j.1467-842x.2004.00334.x.
59. I. Guyon, A. Elisseeff: An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182. Available at https://www.jmlr.org/papers/v3/guyon03a.html.
60. A. Groll, G. Tutz: Variable selection for generalized linear mixed models by L1-penalized estimation, Statistics and Computing 24, 2 (2012) 137–154. https://doi.org/10.1007/s11222-012-9359-z.
61. S. Nakagawa, H. Schielzeth: A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4, 2 (2012) 133–142. https://doi.org/10.1111/j.2041–210x.2012.00261.x.
62. H. Akaike; A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 6 (1974) 716–723. https://doi.org/10.1109/tac.1974.1100705.
63. H. Vafaie, K. De Jong: Genetic algorithms as a tool for feature selection in machine learning. In: Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92, Arlington, VA, USA, 10–13 November, IEEE. 1992, pp. 200–203. https://doi.org/10.1109/TAI.1992.246402.
64. M. Chiesa, G. Maioli, G.I. Colombo, L. Piacentini: GARS: Genetic Algorithm for the identification of a Robust Subset of features in high-dimensional datasets. BMC Bioinformatics 21 (2020) 1. https://doi.org/10.1186/s12859-020-3400-6.

65. R. Kohavi, G.H. John: Wrappers for feature subset selection, Artificial Intelligence 97, 1–2 (1997) 273–324. https://doi.org/10.1016/s0004-3702(97)00043-x.
66. N. Morrow-Howell: The M word: multicollinearity in multiple regression, Social Work Research 18, 4 (1994) 247–251. https://doi.org/10.1093/swr/18.4.247.
67. D. Ludecke, M. Ben-Shachar, I. Patil, P. Waggoner, D. Makowski: performance: an R package for assessment, comparison and testing of statistical models, Journal of Open Source Software 6, 60 (2021) 3139. https://doi.org/10.21105/joss.03139.
68. V. Pulkki: Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society 55 (2007) 503–516. Available at https://www.aes.org/e-lib/browse.cfm?elib=14170.
69. S. Klockgether, S. van de Par: A model for the prediction of room acoustical perception based on the just noticeable differences of spatial perception, Acta Acustica united with Acustica 100, 5 (2014) 964–971. https://doi.org/10.3813/aaa.918776.
70. M. Morimoto, K. Iida, Y. Furue: Relation between auditory source width in various sound fields and degree of interaural cross-correlation, Applied Acoustics 38, 2–4 (1993) 291–301. https://doi.org/10.1016/0003-682x(93)90057-d.
71. T. Hidaka, T. Okano, L. Beranek: Interaural cross correlation (IACC) as a measure of spaciousness and envelopment in concert halls, Journal of the Acoustical Society of America 92, 4 (1992) 2469–2469. https://doi.org/10.1121/1.404472.
72. J. Blauert, W. Lindemann: Auditory spaciousness: some further psychoacoustic analyses. Journal of the Acoustical Society of America 80, 2 (1986) 533–542. https://doi.org/10.1121/1.394048.
73. J. Catic, S. Santurette, J.M. Buchholz, F. Gran, T. Dau: The effect of interaural-level-difference fluctuations on the externalization of sound. Journal of the Acoustical Society of America 134, 2 (2013) 1232–1241. https://doi.org/10.1121/1.4812264.
74. A. Andreopoulou, B.F.G. Katz: Identification of perceptually relevant methods of inter-aural time difference estimation, Journal of the Acoustical Society of America 142, 2 (2017) 588–598. https://doi.org/10.1121/1.4996457.
75. R. Kiyan, J. Bergner, S. Preihs, Y. Wycisk, D. Schössow, K. Sander, J. Peissig, R. Kopiez: Immersive music experience in surround sound music reproduction [Data set]. Leibniz University Hannover Research Data Repository. 2023. https://doi.org/10.25835/3vx9ls5h.
76. J. Bergner: Soundscape Analysis – Indicators [Code]. GitLab. 2023. https://gitlab.com/janywhere/sosca-indicators/.
77. R. Kiyan: immersionmodeling [Code]. Leibniz University Hannover GitLab. 2023. https://gitlab.uni-hannover.de/roman.kiyan.jr/immersionmodeling/.

Cite this article as: Kiyan R. Bergner J. Preihs S. Wycisk Y. Schössow D, et al. 2023. Towards predicting immersion in surround sound music reproduction from sound field features. Acta Acustica, **7**, 45.