Analyzing the Degree of Immersion of Music Reproduction by means of Acoustic Fingerprinting

Jakob Bergner[1], Daphne Schössow[1], Stephan Preihs[1], Jürgen Peissig[1],
Yves Wycisk[2], Kilian Sander[2], Reinhard Kopiez[2], Friedrich Platz[3]
[1] *Leibniz Universität Hannover, Email: jakob.bergner@ikt.uni-hannover.de*
[2] *HMTM Hannover*
[3] *HMDK Stuttgart*

Introduction

The term *Immersion* in virtual environments is widely defined as "the effect that users lose the awareness that they experience an illusory reality but perceive their environment as real" [1]. Obviously immersion refers to a perceptual, cognitive and possible psychological construct that is formed during a user experience. At the same time physical properties of any reproduction can be assessed to investigate their influence on immersion. The term *Immersive Audio* usually refers to specific audio technologies that promise to create enhanced auditive experiences which allows the user to not only listen to music but dive into a sort of concert situation, i.e. to have the feeling to witness an actual music performance [2, 3]. With home entertainment systems, the promise is that a larger number of loudspeakers will create a higher level of immersion. However, it is not definitively explained, what alteration in the soundfield leads to this perception or in other words, which acoustic parameter actually changes measurably with the use of immersive audio technologies. Thus, this work tries to contribute to an identification of relevant acoustic dimensions for the assessment of immersive audio reproduction as realization of (virtual) acoustic environments.

Samples of Immersive Music Reproduction

The samples of music reproductions analyzed in this work cover a specific aspect of immersive audio. The stimulus set comprises 8 excerpts of musical pieces of varying genre, ensemble size and recording/production technique. Each musical piece is available in four versions of different channel-based loudspeaker reproduction formats: *mono* (center loudspeaker), *stereo* (left + right lsp.), *2D* (5.1) and *3D* (5.1.4). For the production of the stimuli two audio engineers with experience in multi-channel mixing were engaged to produce three well-sounding mixes (stereo, 2D and 3D) from provided multi-track recordings without any other restrictions. The loudness of the stimuli within the four playback formats was calibrated to minimise the median deviation of the short-term LUFS according to ITU-R BS.1770-4 [4]. Between the musical pieces a subjective loudness adaptation was applied aiming for plausibility in the reproduction of music with different ensemble sizes and types.

All 32 Stimuli were played back through a loudspeaker system with positioning according to ITU-R BS.2051-2 [5] and equalization based on ITU-R BS.1116-3 [6] with applied room gain [7] preserving an expected low frequency behaviour for loudspeaker reproduction in rooms. The stimuli were then re-recorded at the listening position with a spherical microphone array (mhacoustics Eigenmike) for capturing the three dimensional soundfield. From the microphone array signals three signal representations were deduced: 4th-order Ambisonics (HOA), a binaural representation with an appropriate decoder [8] and a monophonic pressure representation based on the 0th-order Ambisonics channel.

Acoustic Fingerprinting

The method of acoustic fingerprinting describes a process to obtain fundamental acoustic dimensions that are suitable to characterize and compare acoustic environments [9]. It is inspired by the approach to assess soundscapes defined in ISO 12913-1/2/3 [10, 11, 12] and associated approaches for the perceptual and emotional assessment of soundscapes [13]. The method assumes that a range of acoustic indicators with potential relevance for human auditory perception can be taken into account to identify underlying acoustic dimensions. A manageable number of dimensions then forms a fingerprint with characterizing and comparable properties. The approach to identify the dimensions that is followed in this work is data-driven. A large number of observations of a large number of indicators are fed into multivariate analysis. The indicators are separated into three a-priori categories *loudness*, *quality* and *spaciousness* as listed in the following:

Quality: MFCC, Spectral Brightness, ~ Centroid, ~ Crest Factor, ~ Decrease, ~ Entropy, ~ Flatness, ~ Flux, ~ Irregularity, ~ Kurtosis, ~ Roll-Off, ~ Skewness, ~ Spread, Timbral Booming, Roughness, Sharpness, Fluctuation Strength

Loudness: SPL (A-/Z-weighting), Octave Bands, Loudness (ISO 532-1/2), LUFS (ITU-R BS.1770-4)

Spaciousness: ILD, ITD, IACC, IC, Direction of Arrival (hor., vert.) [14], Diffuseness [14], Directivity Index (hor., vert., sph.), Ambisonics Energy Ratio

The indicators of the categories loudness and spaciousness were calculated both for a broadband frequency range as well as for individual frequency bands listed in Table 1. All indicators were calculated as time series with window length of $l_w = 0.1\,\text{s}$ and hop size of $l_h = 0.05\,\text{s}$. Each indicator time series was then scaled to an expected

Table 1: Frequency limits in Hz of analysis bands.

ID	0	1	2	3	4	5	6	7
f_{lo}	65	242	527	986	1724	2910	4816	7880
f_{hi}	238	521	975	1707	2882	4772	7809	12691

value range and logarithmic sampling was applied where necessary. The preprocessing stage was completed by applying z-standardization (removal of mean; normalization to unit variance) to all indicators individually. In total the input data consists of 227 indicators and 33235 observations each.

The indicator's time series were then subject to multivariate analysis methods, precisely to *Factor Analysis* (FA) [15]. FA assumes that underlying *latent* factors become manifest in observed indicators as shown in Figure 1. FA can be used to transform data from the original space into an optimized space of latent dimensions. The operation itself to obtain the factor scores $\mathbf{Y}$ is realized by matrix multiplication as shown in Eq. 1

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{L} \quad . \tag{1}$$

where $\mathbf{X}$ is a $[N_{\text{obersavtions}} \times N_{\text{indicators}}]$ matrix of the original data and $\mathbf{L}$ a specific loading matrix of dimension $[N_{\text{indicators}} \times N_{\text{factors}}]$. The loading matrix comprises the individual weights of each indicator into each factor. The sum over columns, i.e. among indicators yields the sum of square loadings or explained variance of a certain factor

$$s_j^2 = \sum_{i=1}^{N_i} l_{ij}^2 \quad . \tag{2}$$

The relative loading represents the direction of the transformation and can be described as

$$\mathbf{L_{rel}} = \mathbf{L} \cdot \text{diag}\{s\}^{-1} \tag{3}$$

In FA it is an important decision how many factors to keep, i.e. in this case underlying acoustic dimensions. The Kaiser Criterion assumes factors with $s_j^2 \leq 1$ to be relevant since they inhibit more variance than a single indicator. However the parallel analysis according to Horn is a more convenient method since it compares the explained variance with a random sample of the same size by means of a Monte-Carlo simulation. The results for both criteria can be found in Table 2 for pure and *varimax* rotated FA. In this work we follow the parallel analysis suggestion for *varimax* rotated FA and keep the 8 most prominent factors. Their explained variance portion can be taken as scree plot from Figure 2 (top) and the associated relative loading matrix $\mathbf{L_{rel}}$ is visualized in Figure 2 (bottom).

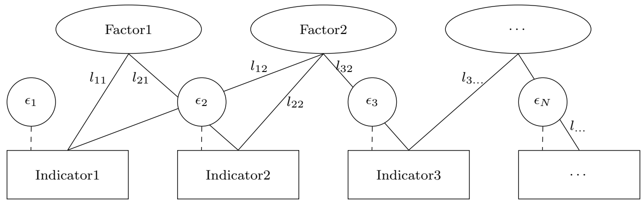


Figure 1: Concept of Factor Analysis with loadings l_{ij} and unique variances ϵ_i.

Table 2: Relevant number of factors accoring to Kaiser's criterion and parallel analysis for pure and rotated FA.

	Kaiser Criterion $s_j^2 \leq 1$		Parallel Analysis	
	$N_{relevant}$	cum. s_j^2	$N_{relevant}$	cum. s_j^2
FA	25	160.31 (70.65%)	9	129.27 (56.95%)
FA *varimax*	26	164.56 (72.49%)	8	113.90 (50.18%)

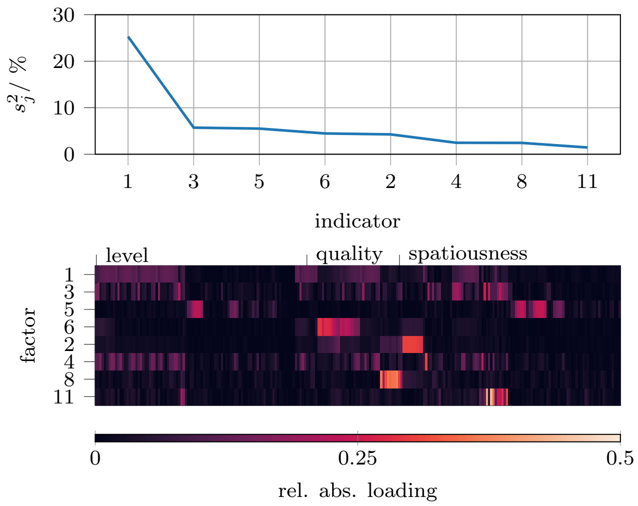


Figure 2: Top: explained variance portion in % scree plot), bottom: relative loading matrix $\mathbf{L_{rel}}$.

Results

The distribution of the resulting factor scores $\mathbf{Y}$ can be found in Figure 3. Each of the 8 most relevant factors is shown individually where factor scores (ordinate) for the individual loudspeaker setups (color coded) are grouped for each music piece (abscissa). Outliers exist but are omitted in the visualization for clarity. We can observe different patterns between the factors. For example factor 1 (top) shows differences between the musical pieces but seems to be stable between the loudspeaker setups. Other factors like factor 5 (third from top) show distinct differences between loudspeaker setups but not that much between musical pieces. In order to find underlying acoustic dimensions that actually change with the number and arrangement of loudspeakers, appropriate statistics were applied. A test for normality for each subgroup (distribution of factor scores for a specific musical piece and a specific loudspeaker setup) failed in many cases. For this purpose Shapiro-Wilk tests were conducted and additionally validated with Q-Q plots to compensate their weakness for large sample sizes. Subsequently non-parametric methods were applied, namely a *Friedman Rank Sum test* (as alternative for one-way repeated measure ANOVA) for testing the null hypothesis H_0: *"There is no difference in scores of a*

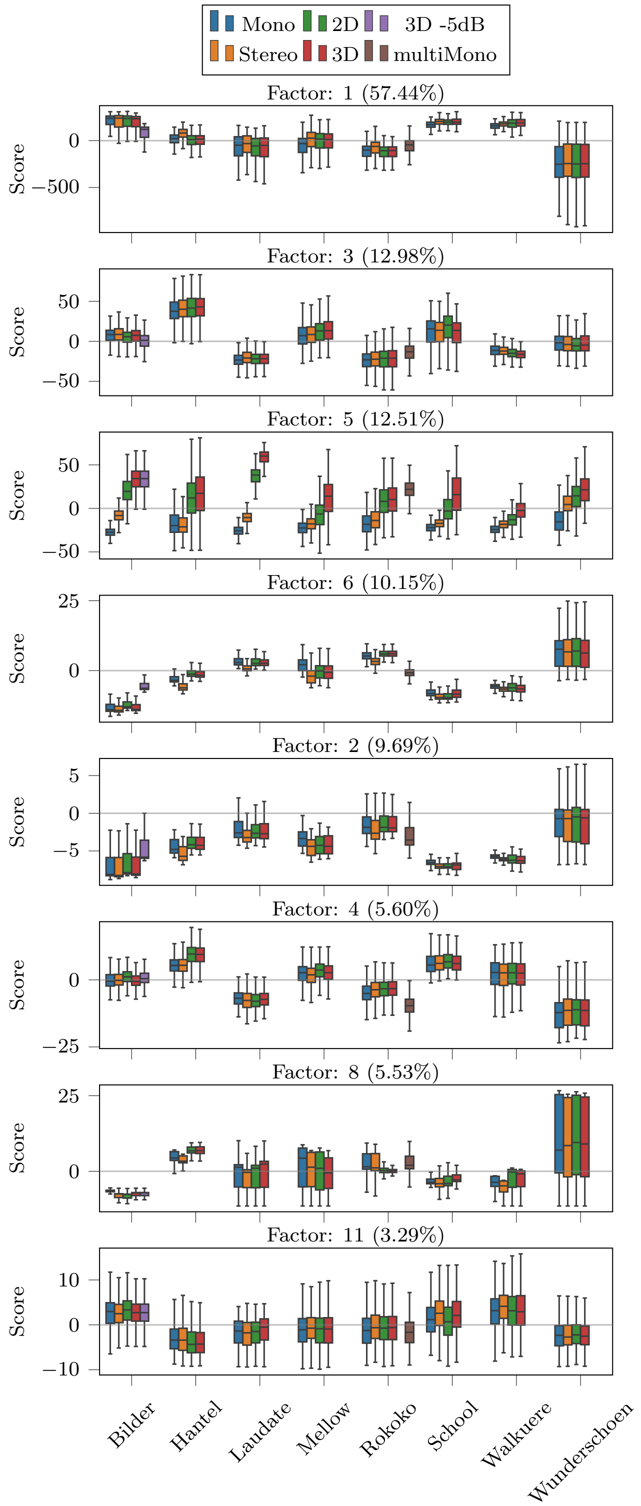


Figure 3: Boxplots of the factor score distributions for the first 8 most relevant factors.

specific factor between mono, stereo, 2D and 3D loudspeaker setups." with a level of significance of $\alpha = 0.05$. The results of the Friedman test can be found in Table 3. It shows a highly significant factor 5 ($p < 0.001$) and three moderately significant factors (2, 4, 8) ($p < 0.05$). Since Kendall's coefficient of concordance is moderate to low for these factors ($W < 0.5$) only factor 5 with excellent effect size $W > 0.9$ is taken into account for further analysis. Posthoc *paired Wilcoxon tests* were conducted to examine if the individual loudspeaker setups differ from each other, which could be confirmed for all musical pieces with low p-values ($p < 0.001$).

Table 3: Friedman test statistics: p-values and Kendall's W.

	1	3	5	6	2	4	8	11
W	.28125	.01250	.95625	.38125	.37500	.38125	.23125	.05625
p	.08031	.96003	.00004	.02736	.02929	.02736	.13568	.71730

Obviously factor 5 reliably reflects the differences of loudspeaker setups while the investigation of the indicator composition of this factor confirms this assumptions. Table 4 shows the indicators with highest absolute loadings l_{i5} (up to 51% of the total factor's explained variance). It can be seen that this factor is mainly composed by three indicators. The indicator *sphDIAz* refers to a spherical directivity index with respect to the horizontal or azimuthal plane based on plane wave decomposition of the higher-order Ambisonics signals. It is present for the frequency bands 4, 5, 6 and 7 which covers a frequency range between 1.7 and 12.7 kHz (cf. Table 1). The indicator *diff* refers to diffuseness according to [14] based on the magnitude of the three-dimensional intensity vector. Lastly, *sphDI* refers again to a spherical directivity index, this time respecting the full sphere. These three indicator groups mainly form factor 5, which is why the semantic description of this factor as "Diffusity" might be reasonable.

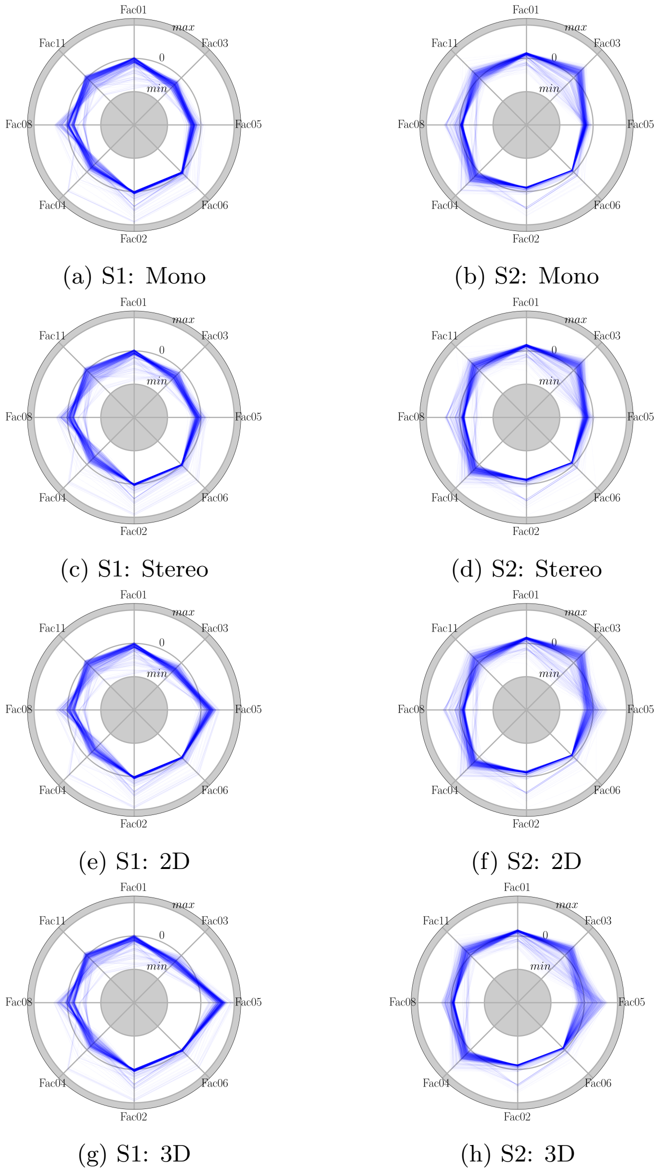
Table 4: Indicator composition of factor 5 with relative loadings $l_{\mathrm{rel},ij}$ in parentheses. Trailing numbers of the indicators denote the frequency band according to Table 1.

Descr.	s_j^2	Indicator composition
"Diffusity"	12.51 (5.51%)	sphDIAz5(-0.242), sphDIAz6(-0.242), diff6(0.235), sphDIAz4(-0.233), sphDI6(-0.233), diff4(0.233), diff5(0.232), sphDI5(-0.231), sphDI4(-0.216), sphDIAz7(-0.213)

An overall comparison of the identified underlying acoustic dimensions can be found in Figure 4. It shows the acoustic fingerprint of two exemplary musical pieces for all four loudspeaker setups. The polar axes of each fingerprint represents the respective factor or acoustic dimension. Each time window of 0.1 s is represented by a faint blue polar line which also supports a general understanding of the temporal distribution of factor scores. This visualization allows to compare general characteristics of the musical piece as well as the progression of the dimensions between the loudspeaker setups. It can be seen that factor 5 (axes to the right; 3 o'clock) increases with the number of loudspeakers in use as it was assumed.

Discussion and Outlook

This work investigated which acoustic properties actually change when music reproduction setups with increasing



(a) S1: Mono (b) S2: Mono

(c) S1: Stereo (d) S2: Stereo

(e) S1: 2D (f) S2: 2D

(g) S1: 3D (h) S2: 3D

Figure 4: Acoustic fingerprints ob music pieces Laudate (S1; left column) and School (S2; right column) for the four loudspeaker setups.

number of loudspeakers are used that promise enhanced immersive experience. By means of factor analysis (FA) 8 relevant underlying acoustic dimensions could be identified of which solely the dimension “Diffusity” shows significant differences between loudspeaker setups. The underlying dimensions can be taken into account to form an acoustic fingerprint that can be used to characterize and compare general acoustic environments.

The next steps include investigations according to generalize the results. The questions would be if a single loading matrix can be utilized for all kind of acoustic environments or if this should be adapted dependent on the application. Further important future work includes an in-depth temporal analysis since the current processing chain only asses score distributions. The question if and how indicators and factors are modulated is expected to be relevant for the temporal characteristis of human auditory perception. Methods to assess these aspects inlcude cross-correlations, derivatives and Fourier Series as well as statistical models such as autoregressive integrated moving average (ARIMA) or functional principle component analysis (FPCA).

References

[1] Wikipedia. Immersion (virtuelle realität), 2022. URL `https://de.wikipedia.org/wiki/Immersion_(virtuelle_Realit%C3%A4t)`.

[2] C. Eaton and H. Lee. Subjective evaluations of three-dimensional, surround and stereo loudspeaker reproductions using classical music recordings. *Acoustical Science and Technology*, 43(2):149–161, 2022.

[3] E. Hahn. Musical emotions evoked by 3D audio. *AES International Conference on Spatial Reproduction*, pages 380–387, 2018.

[4] ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level, 2015.

[5] ITU-R BS.2051-2: Advanced sound system for programme production, 2018.

[6] ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems, 2015.

[7] J. A. Pedersen and H. G. Mortensen. Natural timbre in room correction systems (Part II). In *AES 32nd International Conference*, volume 2, pages 916–926, 2007.

[8] F. Zotter and M. Frank. All-Round Ambisonic Panning and Decoding. *J. Audio Eng. Soc.*, 60(10):807–820, 2012.

[9] J. Bergner, S. Preihs, and J. Peissig. Soundscape Fingerprinting - Methods and Parameters for Acoustic Assessment. In *Fortschritte der Akustik - DAGA*, 2021.

[10] ISO 12913-1: Acoustics. Soundscape. Part 1. Definition and conceptual framework, 2018.

[11] ISO 12913-2: Acoustics. Soundscape. Part 2. Data Collection and reporting requirements, 2019.

[12] ISO 12913-3: Acoustics. Soundscape. Part 3. Data Analysis, 2020.

[13] A. Fiebig, P. Jordan, and C. C. Moshona. Assessments of Acoustic Environments by Emotions - The Application of Emotion Theory in Soundscape. *Frontiers in Psychology*, 11, 2020.

[14] V. Pulkki. Spatial sound reproduction with directional audio coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[15] J. Bortz and C. Schuster. *Statistik für Human- und Sozialwissenschaftler*. Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-12770-0_8.