Assessing the plausibility of virtual acoustic environments

Alexander Lindau
Audio Communication Group, Technical University of Berlin, Germany,
alexander.lindau@tu-berlin.de

Stefan Weinzierl
Audio Communication Group, Technical University of Berlin, Germany

Summary

Aiming at the perceptual evaluation of virtual acoustic environments (VAEs), 'plausibility' is introduced as a quality criterion that can be of value for many applications of virtual realities. We suggest a definition as well as an experimental operationalization for plausibility, referring to the perceived agreement with the listener's expectation towards an equivalent real acoustic event. A listening test methodology for the criterion-free assessment of the deviation from this non-explicit, inner reference is proposed. It requires the rating of corresponding real and simulated stimuli in a Yes/No test paradigm, and the analysis of the results according to signal detection theory. The specification of minimum effect hypotheses allows the testing of plausibility with any desired strictness. The approach is demonstrated with the perceptual evaluation of a system for dynamic binaural synthesis in two different development stages.

PACS no. 43.60.+d, 43.66.+y

1 Introduction

Dynamic binaural synthesis has reached a high degree of realism in the simulation of acoustic environments. The actual quality of binaural simulations is, however, typically assessed by observing only singular aspects of spatial hearing, e.g. by comparing the localization accuracy in real and simulated sound fields [1] - [6] or by comparing to references that are simulations themselves [7]. The results from such studies seem though inappropriate as measures in how far a dynamic binaural synthesis as a whole is able to provide substitutes for real sound fields.

More holistic measures are provided by analysing the 'immersion' or a 'sense of presence' of subjects in virtual environments. Widely used in the evaluation of visual virtual environments (and up to now only incidentally assessed in acoustics), these features are generally interpreted as multidimensional constructs including aspects such as a spatial presence experience ('being there'), a sense of 'involvement' and a judgement of 'realness' [8], [9]. Some of these underlying facets are, however, strongly related to the quality of the presented content, the provided modes of interaction, the usability of the interface applied, and the personality of the users addressed. Therefore, for the purpose of system development and evaluation, these constructs seem less appropriate.

As a system-oriented criterion, 'authenticity' was suggested [10:373], referring to the perceived identity between simulation and reality. This would make maximum demands on the performance of virtual environments. The necessary immediate comparison between simulation and reality can, however, not always be realised experimentally, and it is, at the same time, not always required in applications where users don't have any 'reality' as an external reference.

A more appropriate criterion for most applications could be the 'plausibility' of a virtual environment, defined as

a simulation in agreement with the listener's expectation towards an equivalent real acoustic event.

Referring to an inner reference as the result of each listener's personal experience and expectations [11] rather than to the exact perceptual identity of both environments, plausibility corresponds well to the situation in which most users will evaluate the quality of a simulation. In the following, we will present an approach for the criterion-free assessment of plausibility.

2 Plausibility: An experimental approach

The plausibility of virtual environments could theoretically be rated directly on a linear scale with values between '0' and '1' without any given reference. However, a strong and inter-individually different response bias can be expected due to personal theories about the credibility of virtual realities and the performance of media systems in general. Therefore, a criterion-free assessment of plausibility would be preferable. Following the definition given above, requiring the evaluation of a simulation with regard to an inner reference, any forced choice paradigm is precluded, because a direct comparison with an external, given reference would be necessary. Decisions with regard to an inner reference can, however, be collected by using a Yes/No test paradigm and by removing the response bias ex post with an analysis according to signal detection theory [12].

2.1 Plausibility as a signal detection problem

Signal detection theory (SDT) provides a model for the perceptual process when detecting weak signals in the presence of internal noise. The SDT concept can be easily adapted to the discrimination task involved in evaluating the plausibility of a virtual environment. In our case, the reality takes the role of the 'no signal' condition, while the simulation represents the 'signal' condition, assuming that the latter contains small artefacts (the 'signal') compared to the original.

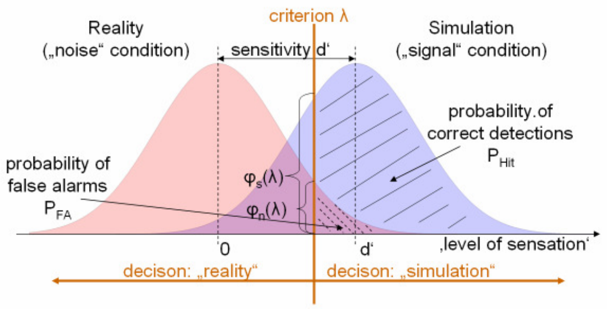

Figure 1. Parameters of the equal-variance Gaussian signal detection model, adapted to an evaluation of the 'plausibility' of virtual environments

As in standard SDT approaches, we use a simple observer model which assumes Gaussian probability density distributions of equal variance for the 'reality' ('noise alone') and the 'simulation' ('signal plus noise') conditions (cf. Figure 1). The sensory difference of the two stimuli is expressed by the distance of the two distributions' maxima (sensitivity d' in Figure 1). Individual response bias, i.e. the tendency to regard the simulation as reality or vice versa, is reflected by individually differing response criteria λ_i. If the sensation level is perceived to be above that individual criterion – which is assumed to be stable over time – the observer will give a positive response. Hence, observers with criteria $\lambda_i > d'_i/2$ show a conservative answering behaviour, i.e. a tendency to believe in the realness of the stimulus, whereas subjects with smaller criteria will respond more progressively, i.e. even referring to real stimuli as being simulated. A criterion of $\lambda_i = d'_i/2$ would indicate a perfectly 'fair' observer. Applying the inverse cumulative normal distribution Z(p) the individual criteria λ_i, and the sensitivities d'_i can be estimated as (with ^ indicating estimated variables) from empirically observed proportions of false alarms (p_{FA}) and correct detections (p_{Hit}) as

$$\hat{\lambda}_i = Z(1-p_{FA_i}) \,, \tag{1}$$

and

$$\hat{d}'_i = Z(p_{Hiti}) - Z(p_{FAi}) \,, \tag{2}$$

respectively. An alternative measure of bias is the ratio of the values of normalized noise and signal probability density at the position of the criterion:

$$\hat{\beta}_i = \varphi_s(\hat{\lambda}_i) \big/ \varphi_n(\hat{\lambda}_i) = \varphi(\hat{\lambda}_i - \hat{d}_i') \big/ \varphi(\hat{\lambda}_i) \tag{3}$$

In contrast to λ_i it allows a more direct interpretation of bias (without relating to d'_i): Subjects exhibiting $\beta_i < 1$ have the tendency to report "Yes (i.e. 'simulation')", whereas a $\beta_i > 1$ indicates a "No (i.e. 'reality')"-tendency.

2.2 Minimum effect hypothesis and sample size

In terms of the SDT observer model, proving 'perfect' plausibility would require proving a sensitivity d' of zero. From the view of inferential statistics, however, a direct proof of the null hypothesis is impossible. Thus, one has to draw back to rejecting a directional and specific alternative hypothesis by negating an effect that is small enough to be regarded as perceptively irrelevant (a minimum-effect hypothesis, [13]).

Values of d' can not easily be interpreted for the formulation of a meaningful minimum effect hypothesis. Besides, they can be related directly to detection rates of nAFC test paradigms, as both Yes/No and nAFC paradigms can be described using the same probabilistic representation [14].

For the equal variance Gaussian signal detection model the probability of correct responses P_c in the 2AFC paradigm and sensitivity d' in Yes/No tasks are related by

$$P_c = \Phi(d'/\sqrt{2}) = \Phi(.707 \cdot d'), \qquad (4)$$

and

$$d' = \sqrt{2} \cdot Z(P_c) = .707 \cdot Z(P_c), \qquad (5)$$

respectively, with Φ(z) as the cumulative Gaussian normal distribution[2]. These relations allow formulating hypotheses more intuitively in terms of P_c. As an example, for our listening test we assumed plausibility to be reached, if the probabilities of correct responses in an equivalent 2AFC test design were less than $P_c = 0.55$, i.e. exceeding the pure guessing rate by less than 5%. Please note that this resembles a far stricter criterion than determination of the inflection point of the psychometrical function (at $P_c = 0.75$) commonly targeted as a population's difference or detection threshold.

By converting values of d' to equivalent values of P_c and vice versa one can make use of the well developed theory of designing nAFC tests with given type I and II error levels. Applying the binomial distribution for the analysis of 2AFC tests results yields the resulting type I and II error levels for different numbers of trials and different effect sizes P_c [16]. Approximating the binomial by the normal distribution, analytical formulae for calculating optimum sample size, implied effect size, or critical detection rate P_{c_crit} for arbitrary levels of confidence have been derived [17].

Accordingly, a test of the above mentioned proportion of correct responses $P_c = 0.55$ at 0.05 type 1 and 2 error level (i.e. with 95% test power) requires 1077 singular decisions. Type 1 and 2 errors are set to the same level to avoid favouring either of the two hypotheses in the inferential test. The critical number of correct responses can be calculated to be 566, hence $P_{c\_crit} = .5255$, and $d'_{crit} = .084$ respectively.

2.3 Aggregation of SDT indices

SDT tests are typically conducted using only few, well trained subjects. If similar effects can be shown, results are thought to be generalizable. In the context of our study, instead of testing only two or three trained experts, it appeared more adequate to evaluate the plausibility of VAEs using a larger and therefore more representative sample of subjects.

As the calculation of d' involves nonlinear transformations the calculation of an average sensitivity d'_{avg} from the pooled Yes/No decisions of all observers does not give the same result as when averaging over the observers' d'_i. [14]. To obtain a measure of the average sensitivity of the group, individual values for d'_i were calculated first and then averaged to get d'_{avg}. Considering a figure of 100 decisions necessary for obtaining stable individual SDT parameters [18], the plausibility test did thus require a sample of at least 11 subjects, each giving 100 decisions.

3 Listening test setup

3.1 Realizing the binaural simulation

The presented approach towards assessing the plausibility of virtual environments requires a test setup where real and simulated stimuli can be presented in the same acoustic setting. To guarantee constant test conditions, both real and simulated stimuli were generated from pre-recorded audio material and electro acoustical sound sources placed in a large lecture hall (auditorium maximum of TU Berlin, V = 8500 m³, RT = 2.0 s, r_{crit} = 3.6 m). Five mid-size loudspeakers (Meyersound UPL-1) were placed at different positions on the stage, floor, and balcony areas. The head and torso simulator (HATS) FABIAN [19] with freely moveable artificial head above torso was placed at a central seat in the frontal half of the audience area. Datasets of binaural room impulse responses (BRIRs) were measured individually for each of the five loudspeakers and for horizontal head movements in a range of ±80° with an angular resolution of 1°. Distances between loudspeakers and the central listener seat varied between 3–5 times r_{crit}.

For the listening test, subjects were placed at the same seat as the HATS. Dynamic auralization was realized using a fast convolution algorithm with head-tracking [19]. To hide the presentation mode, subjects kept their headphones on throughout the test. This was enabled by letting the dummy head wear acoustically relatively transparent headphones (STAX SR-2050II) throughout the BRIR measurements.

[2] Obviously, with the 2AFC paradigm, sensitivity is expected to be increased by factor $\sqrt{2}$ as compared to the Yes/No task, an illustrative derivation can be found in [15].

3.2 Testing two development stages of a VAE

A system for dynamic binaural synthesis was perceptually evaluated in 2007 [19] with encouraging but not fully satisfying results. Following qualitative reports of perceived deficiencies such as spectral coloration, latency, instability of localization, and cross fade artefacts, several technical improvements were implemented. These include a perceptually optimized headphone compensation [20], an inaudible system latency [21], a reduction of cross fade artefacts and localization instability by individualizing the interaural time delay by means of ITD extraction and manipulation and by replacing BRIRs with minimum phase representations [22], and perceptually validated thresholds for the transition of dynamic and static parts of the room impulse response [23]. For an exemplary and comparative realisation of the test design introduced above, plausibility was tested for both simulator stages (i.e. from years 2007 and 2010, cf. Table I) in two independent listening tests. Average system latency was measured according to the procedure described in [21]. Individualization of the ITD was realized using the procedure described in [22], i.e. by measuring individual head diameters (intertragus distances) of subjects prior to the listening test.

Table I. Reported perceptual artefacts of the 2007 stage of the simulation and resp. treatment in 2010.

Perceptual artefacts	**Technical treatment**	
	in year 2007	*in year 2010*
Spectral coloration	Linear phase, regularized LMS headphone filter	Minimum phase, regularized LMS headphone filter
Latency	avg. TSL*: 112 ms (minimally)	avg. TSL*: 65 ms
Unstable localization	-	individual ITD manipulation
Cross fade artefacts	BRIRs not time aligned	BRIRs time aligned

*TSL = total system latency, cf. [21]

3.3 Listening test procedure

According to sample size calculations above, each of the two tests was conducted with eleven subjects, while subjects were not the same across tests. 100 real and simulated stimuli were presented to each subject in individually randomized order. The actual sequence of the presentation mode (real vs. simulated) was – again individually for each subject – drawn from a uniform dichotomous random distribution, implying that real/simulated proportions varied slightly among subjects. From binomial distribution one can though calculate that in more than 95% of cases, the proportion of real vs. simulated stimuli was within 0.5 ± 0.1. Thus, slightly unequal proportions were tolerated to minimize interdependence of succeeding answers [16] and to prevent subjects from making assumptions about the absolute amount of "Yes" and "No" answers in the test. After each presentation subjects had to decide whether, in their opinion, the presentation was real or simulated. As real and simulated presentations were rather similar, before beginning the test, subjects were allowed to take headphones off once while playing either stimulus. This was considered necessary as it was observed in pre-tests that stimuli were so similar that people actually thought they never heard a simulation. In order to suppress memory effects of minor auditive differences, which, while not interfering with our definition of plausibility, could bias individual results in one direction, stimuli were randomly varied in content (20) and source location (5). Hence, a particular combination of 'content and source' was presented only once in each test, either as real or simulated stimulus. Contents varied from artificial signals such as steady state or burst noises, over male and female speech in foreign and native language, to recordings of single instruments and monophonic down mixes of pop songs. Loudness differences between stimuli were compensated beforehand. Equal loudness between real and simulated presentation mode was established through adjustment by two expert listeners. The duration of the stimuli was set to approximately 6 seconds (including reverberation), which sufficed to move one's head in a reasonable range. Each stimulus was presented only once. To help maintaining a constant level of concentration, subjects could decide when to proceed to the next stimulus.

3.4 Subjects

Across both listening tests subjects were of an average age of 29 years (86.5% male). Subjects had an average of 6 years of musical education and more than half of them had already taken part in listening tests using dynamic binaural technology. Subjects could thus be regarded as an experienced sample of a typical, untrained population.

4 Results

As stimuli were rather short and stimulus replay was not allowed, none of the subjects needed more than 15 minutes to complete the 100 trials. From decisions, individual hit and false alarm rates were calculated. Using equations 2 and 3 individual sensitivities d'_i, and criteria β_i were estimated. Using Lilliefors and t-tests for independent samples, distributions of criteria β_i were – independently for both simulator stages – tested for deviation from normality, for deviation from each other and from unity at p = .05 type 1 error level. All tests were negative except for the 2007 stage test where a significant "No"-tendency ('no simulation') was observed. Extreme values of β were 1.24 ("No"-tendency), and 0.97 (a slight "Yes"-tendency), respectively. Mean values of β were 1.08, and 1.02 for the 2007 and 2010 stage respectively, hence, shifted systematically into the "No" direction.

Individual sensitivities d'_i were averaged independently for both tests to determine the group's discriminability d'_{avg} for either simulator stage. Assuming validity of the equal variance observer model, these averaged sensitivity values d'_{avg} were used to model the group's response behaviour (cf. Figure 2).

To test our minimum effect hypothesis H_1 stating a residual 2AFC detection rate of the 'simulation' condition P_c of 55%, the group's sensitivities were transformed to the proportion of correct responses P_c in an equivalent 2AFC task and compared to the critical proportion P_{c_crit} (cf. Table II).

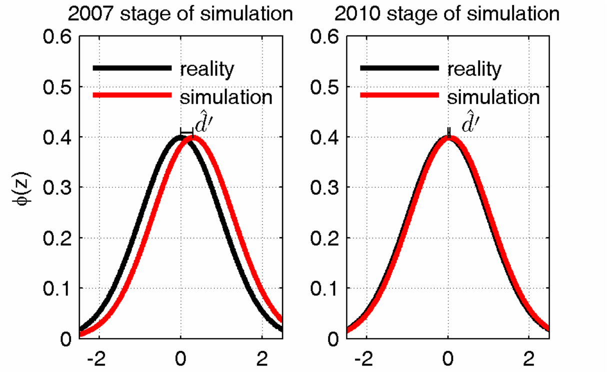

Figure 2. Probability density distributions modelling the group's performance in detecting the 'simulation' condition in both simulator stages.

Table II. Group's sensitivity values, proportions of correct responses in an equivalent 2AFC test, and critical values to fall below for negating our minimum effect hypothesis H_1.

	2007 stage	**2010 stage**	**critical values**
d'	0.2956	0.0512	0.084
P_c	0.5828	0.5144	0.5236

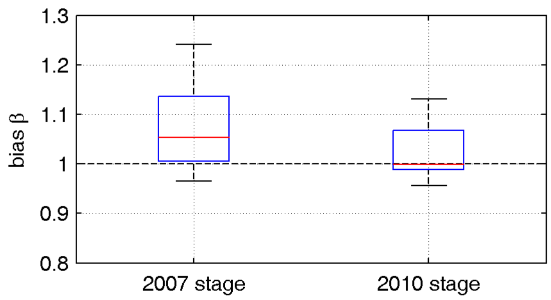

Figure 3. Distribution of individual biases β_i in both simulator stages. Box plots show medians, interquartile ranges and extreme values (as whiskers)

5 Discussion

Comparing the average sensitivity measures with critical values according to our minimum effect hypothesis (Table II), it can be seen that with the improvements summarized in Table I the current implementation of a data based dynamic binaural simulation succeeds to satisfy a strict test of plausibility, whereas the simulator in the legacy stage exceeds the critical sensory difference d_{crit}'. With the current simulator implementation, subjects were almost perfectly guessing, whereas with the preceding stage a noticeable amount of residual discriminability is observable.

Results suggest that spectral differences resulting from the non-individual morphology of the HATS, while limiting the perceived *authenticity* (perceptual identity) of non-individual acoustic simulations, do not interfere with perceived plausibility. On the other hand, plausibility is shown to be sensitive to excessive latency, cross fade artefacts and instable localization. These findings could be explained by the absolute memory of sound colouration being comparatively weak, whereas latency, cross fading artefacts and instable localization are perceived as permanent sources of disturbance. The distribution of individual biases β_i (Figure 3) illustrates the importance of a criterion-free listening test design. The systematic shift into the "No" direction suggests that subjects were on average biased towards deciding for a 'real' stimulus, an observation which could be attributed to the considerable and unexpected naturalism of the simulation. Interestingly our results also imply that – when using state of the art auralization in a consistent natural environment – frequently claimed 'systematic artefacts' of binaural simulation such as a perception of elevation or a reduction of externalisation did apparently not occur, at least not to a degree where the simulation could be recognized as such.

6 Conclusions

Starting from a definition of plausibility a test method was consequently developed. Plausibility was explained to require indistinguishability from an inner reference. Yes/No tasks were found a suitable test paradigm. Defining an appropriate minimum effect alternative hypothesis and applying signal detection theory analysis it was shown how discriminability from an inner reference can be tested criterion-free and with high confidence. The approach was applied to evaluate an environment for data based dynamic binaural synthesis. With recent technical improvements our simulator could be shown to satisfy a strict test of plausibility.

7 References

[1] Bronkhorst, A. W. (1995): "Localization of real and virtual sound sources." In: *J. Acoust. Soc. Am.*, Vol. **98**(5), pp. 2542-2553

[2] Møller, H. et al. (1996): "Binaural Technique: Do We Need Individual Recordings?" In: *J. Audio Eng. Soc.,* **44**(6), pp. 451-469

[3] Møller, H.et al.(1997): "Evaluation of Artificial Heads in Listening Tests." In: *Proc. of the 102nd AES Conv.*, München, preprint no. 4404

[4] Djelani, T. et al. (2000): "An Interactive Virtual-Environment Generator for Psychoacoustic Research II: Collection of Head-Related Impulse Responses and Evaluation of Auditory Localization." In: *Acta Acustica united with Acustica*, **86**, pp. 1046-1053

[5] Minnaar, P. et al. (2001): "Localization with Binaural Recordings from Artificial and Human Heads." In: *J. Audio Eng. Soc*., **49**(5), pp. 323-336

[6] Liebetrau, J. et al. (2007): "Localization in Spatial Audio - from Wave Field Synthesis to 22.2." In: *Proc. of the 123rd AES Conv.*, New York, preprint no. 7164

[7] Pulkki, V.; Merimaa, J.: "Spatial Impulse Response Rendering: Listening tests and applications to continuous sound." In: *Proc. of the 118th AES Conv.*. Barcelona, peprint no. 6371

[8] Schubert, T.; Friedmann, F.; Regenbrecht, H. (2001): "The Experience of Presence: Factor Analytic Insights.", In: *Presence: Teleoperators and Virtual Environments*, **10**(3), pp. 266-281

[9] Lessiter, J. et al. (2001): "A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory.", In: *Presence: Teleoperators and Virtual Environments*, **10**(3), pp. 282-297

[10] Blauert, J. (1997): *Spatial Hearing. The Psychophysics of Human Sound Localization*. 2nd ed., Cambridge, MA.: MIT Press.

[11] Kuhn-Rahloff, C. (2011): *Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen*. doct diss, TU Berlin

[12] Green, D. M.; Swets, J. A. (1974): *Signal Detection Theory and Psychophysics*, Huntington: Krieger

[13] Murphy, K. R.; Myors, B. (1999): "Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model." In: *J. Appl. Psychol.*, **84**(2), pp. 234-248

[14] Wickens, T. D. (2002): *Elementary Signal Detection Theory*, New York: Oxford University Press

[15] Macmillan, N. A.; Creelman, C. D. (2005): *Detection Theory. A user's Guide*. 2nd ed., Mahwah, NJ: Lawrence Erlbaum Ass. Inc.

[16] Leventhal, L. (1986): "Type I and Type 2 Errors in the Statistical Analysis of Listening Tests." In: *J. Audio Eng. Soc.,* **34**(6), pp. 437-453

[17] Burstein, H. (1988): "Approximation Formulas for Error Risk and Sample Size in ABX Testing." In: *J. Audio Eng. Soc*., **36**(11), pp. 879-883

[18] Kadlec, H. (1999): "Statistical Properties of d' and ß Estimates of Signal Detection Theory." In: *Psychological Methods*, **4**(1), pp. 22-43

[19] Lindau, A.; Hohn, T., Weinzierl, S. (2007): "Binaural resynthesis for comparative studies of acoustical environments.", *Proc. of the 122nd AES Conv.*, prepr. no. 7032

[20] Lindau, A.; Brinkmann, F. (2010): "Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings." In: *Proc. of the 3rd Int. Workshop on Perceptual Quality of Systems*. Dresden, pp. 137-142

[21] Lindau, A. (2009): "The Perception of System Latency in Dynamic Binaural Synthesis." In: *Proc. of 35th DAGA*. Rotterdam, pp. 1063-1066

[22] Lindau, A.; Estrella, J. and Weinzierl, S. (2010): "Individualization of dynamic binaural synthesis by real time manipulation of the ITD." In: *Proc. of the 128th AES Conv.,* London, prepr. no. 8088

[23] Lindau, A.; Kosanke, L.; Weinzierl, S. (2010): "Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses." In: *Proc. of the 128th AES Conv.,* London, prepr. no. 8089

(c) European Acoustics Association, ISBN: 978-84-694-1520-7, ISSN: 221-3767