Demands on measurement models for the perceptual qualities of virtual acoustic environments

H.-J. Maempel[1], S. Weinzierl[1]

[1]Audio communication group, Berlin Institute of Technology, Einsteinufer 17c, DE-10587 Berlin, Germany, e-mail: {hans-joachim.maempel, stefan.weinzierl}@tu-berlin.de

Summary

Virtual acoustic environments (VAEs) are frequently intended for the reproduction of and the interaction with acoustic scenes. While different technical approaches such as sound field synthesis and binaural synthesis vary in their capability of accurately reproducing certain features of the physical sound field, there is no agreement as to the perceptual criteria for the evaluation of VAEs. Frequently suggested global attributes such as *presence*, *authenticity*, *plausibility* and *naturalness* are reviewed. We propose a systematization of different properties, discuss the suitability for different research objectives, and consider demands for their measurement. In this context, methodological issues regarding operationalization, experimental references, and criterion-free test procedures are discussed.

1. Introduction

Increasingly sophisticated simulation techniques have been developed, aiming at the reproduction and sometimes also the interaction with acoustic scenes: dynamic binaural synthesis, and different types of sound field synthesis such as wave field synthesis or higher order ambisonics. Thereby, the approaches for sound field reproduction have shifted from the exploitation of psychoacoustic effects (as in classical two-channel or multichannel stereophony) towards the generation of physically correct sound fields and ear signals. A couple of parameters have been suggested to measure the physical error of the synthesized sound fields such as a difference in sound pressure level or a 'reproduced aliasing-to-signal ratio' (RASR) [1].

The question arises, however, which qualities are suitable for the evaluation of virtual acoustic environments (VAEs) as a whole and how they are to be measured. In the following we review some qualities that have been suggested for the evaluation of VAEs, propose a systematization of potential qualities, and discuss the suitability of different classes of qualities for certain research objectives. Finally, methodological demands for the direct and the indirect measurement of these properties are discussed including operationalization, internal and external references, and criterion-free test procedures.

2. Evaluating virtual environments

VAEs are often intended for simulating the reality. Thus it is not surprising that many criteria for the evaluation of VAEs implicitly or explicitly refer to a real reference. Throughout the literature there are numerous properties in this vein, e.g. *immersion*, *presence*, *authenticity*, *plausibility*, *naturalness*, *realness*, *fidelity*. These concepts are not always easy to distinguish. A global concept that does not refer to reality is *sound quality*. In order to more differentially evaluate VAEs, also characteristics are applied that are related to the description of auditory scenes or sound images in general, such as *timbre*, *loudness*, *localization*, and *envelopment*. Furthermore, VAEs are expected not to produce artifacts and to have imperceptible latencies. Based on the concepts of *immersion*, *presence*, *authenticity, plausibility* and *naturalness* we discuss to what extent these global criteria are appropriate for the evaluation of VAEs and how they can be measured in order to meet the classical criteria for test quality.

Prominent criteria for the evaluation of virtual environments are *immersion* and the *sense of presence*. While the first is often applied to denote system features (technical description) which are conducive to the latter (perceptual phenomenon) [2], there is no consistent definition of *presence* except the consensus that it is a multidimensional concept. Lombard and Ditton define *presence* as the "perceptual illusion of nonmediation" and stress the illusion of physical presence ('being there') as well as of social presence ('being together and communicating') [3]. According to Slater et al. [4] the *sense of being there*, the *simulation being the dominant reality*, and *experiencing the simulation as a place* contribute to *presence*. The Presence Questionnaire (PQ) developed by Witmer and Singer [5] covers three aspects, revealed by cluster analysis: *involvement/control*, *naturalness*, and *interface quality*. As noted by Schuemie et al., the PQ measures presence by measuring its causes [6] (p. 189). Schubert et al. [7] developed the Igroup Presence Questionnaire (IPQ). Its items form the factors *spatial presence*, *involvement* and *realness*, which load on a 2nd order factor that is regarded to represent the general *presence*. Even more aspects are covered by the ITC Sense Of Presence Inventory (ITC-SOPI) [8], the Cross-Media Presence Questionnaire [9], and the Swedish viewer-user presence questionnaire (SVUP) [10]. The Continuous Presence Assessment [11] allows for the collection of time responses of *presence* as a global measure, the Presence Counter [12] is to reveal the time points at which participants transit from the real world to the virtual world (and vice versa), whereas the presence scale ("completely unreal" – "just like in real life") is intended for a retrospective measurement [10] (p. 4). For a detailed review of presence concepts cf. [6].

The concept of *authenticity*, proposed by Blauert [13], refers to the sensory identity between a stimulus and an external reference, i.e. a second presented stimulus. According to Pellegrini [14] *authenticity* "describes the property of two entities to be indistinguishable to a human observer." (p. 18). Applied to VAEs, *authenticity* implies the reality serving as the external reference, yielding the most demanding perceptual criterion conceivable. However, Blauert argues that in many applications perceptual *plausibility* was more important than *authenticity* [13].

The definitions of *plausibility* appear to be less consistent. According to Pellegrini [14] a *plausible* VAE provides "a suitable reproduction of all required quality features for a given specific application" rather than a copy of "an existing environment in all its physical aspects" (p. 19). Kuhn-Rahloff defines *plausibility* as "the result of a process, that determines to what extent a perceptual object corresponds to an inner reference resulting from individual experiences. The reference of the plausibility judgment is the result of all imaginable perceptual events within the respective system of relational and classifying inference" [15] (p. 99, translated by the authors). Lindau and Weinzierl define a plausible simulation as "a simulation in agreement with the listener's expectation towards a corresponding real event" [16] and provide an elaborated measurement model applying signal detection theory [17] on plausibility decisions [18]. More specifically, Reiter regards a simulation to be plausible, "as long as there is no obvious contradiction between the visual and the acoustic representation of a virtual scene", allowing the human senses to "merge auditory and visual impressions" [19] (p. 159). Apparently, there is an agreement in terms of regarding the reference as an internal reference, no matter if events are imaginable within a class of experienced realities, an expected corresponding reality, or a visual percept. However, Pulkki and Merimaa measure *plausibility* by comparing the simulation in question with another simulation rather than referring it to reality [20].

The concept of *naturalness* is one of the most unclear. Already in 1964 Queen puts *naturalness* on a level with the "listener's illusion of no reinforcement" [21] (p. 1). According to Theile a "natural stereophonic sound image […] should satisfy aesthetically and it should match the tonal and spatial properties of the original sound at the same time" [22] (p. 761). Letowski and Dreisbach just list *naturalness* as one aspect of the *overall sound quality* [23]; however, they state that a familiarity with a sound is a prerequisite for judging its *naturalness*. Fredriksson and Zacharov regard a sound reproduction "close to the real-life experience" as being natural [24] (p. 2). So definitions or explanations refer to an original sound being not further defined, to the absence of an electroacoustic transmission system, and to the listening experience. Particularly the latter aspect overlaps the concept of *plausibility*. Berg and Rumsey [25] even empirically found *naturalness* to be connected with *plausibility*, with a sense of *being-in-the-room* and with *matching of the listener's references*.

The synopsis of the reviewed concepts indicates that (a) their definitions often remain unclear or are missing at all which suggests that one relies on a colloquial denotation, (b) there is rarely an agreement with respect to provided definitions, (c) definitions of several concepts overlap each other, (d) for many definitions provided neither theoretical nor empirical reasons are given, (e) some definitions are operational, i.e. they are no real definitions.

This appears to hold true also for the other criteria mentioned at the beginning. E.g., the definition of *plausibility* by Pellegrini [14] and the definition of *sound quality* by Blauert (the adequacy of a sound in the context of a specific technical goal and/or task) [26] can hardly be disentangled. And Berg found that there are different denominations and partly conflicting definitions of *envelopment* [27]. Obviously, many perceptual criteria are intuitively understood and colloquially denoted, however they lack a concurrent translation into a sufficiently differentiated professional terminology – a problem known as the problem of correspondence [28] (p. 2. et sqq.). As a consequence, there is also no self-evident solution for transferring them into numerical data (operationalization).

3. Systematization of properties

In order to allow for a more general discussion of the measurement of evaluation properties, a systematization might be helpful. Firstly, a difference can be made between physical and psychological properties or between technical and perceptual measurements, respectively. These types of properties often are referred to as 'objective' and 'subjective' properties. However, we consider that the indication of technical measures as 'objective' is misleading and is due to a confusion of the scientist's perspective with the perspective of a test person being either part of a measuring instrument or an object of the investigation himself [29]. Provided that the intention of acoustic communication is always the perception by human beings, perception is usually the ultimate test criterion for audio transmission systems and audio content, whereas physical properties are often only predictors for psychological properties. In these cases, technical measurements are less valid than perceptual measurements. Thus, we will not discuss the measurement of purely physical properties (e.g. *immersion* in terms of technical system features) in the following.

Looking at the psychological properties, it makes sense to distinguish unimodal (e.g. auditory and visual) from multimodal (e.g. audio-visual) properties (cf. [30]). The latter may be further differentiated into intermodal and supramodal properties. Intermodal properties are the result of a directed comparison of several modalities, e.g. the perception of synchrony [31]; though supramodal properties also derive from unimodal properties, they are the result of a higher stage of processing and therefore more general and abstract. *Presence*, emotional responses and structural representations (e.g. the imagined shape and size of a room [32]) may be regarded as supramodal properties. Thus, an intermodal property such as the audio-visual matching proposed by Reiter [19] would be too specific for the general evaluation of virtual environments. In order to evaluate virtual *acoustic* environments in particular, purely auditory and audio-visual (supramodal) properties come into consideration.

Different stages of perceptual processing such as feature analysis and feature integration generate different levels of perceptual properties often referred to as lower and higher attributes. This distinction is constitutive not only with respect to concepts and theories of *sound quality*: In this context Letowski uses the terms "global" and "parametric" assessment of sound [33] (p. 91). Pellegrini links the different levels to two ways of listening (analytically and globally) and calls the results of analytical listening "elementary attributes" [34] (p. 4). In the following, we refer to properties on a high perceptual level as holistic properties, and to features on lower levels as specific properties. Holistic properties describe the perceptual quality of a simulation rather comprehensively. There are obviously further perceptual levels in between holistic properties and basal sensations, i.e. multiple specific properties may be integrated by more unspecific meta-properties (e.g. different aspects of *presence*). Basal specific properties (e.g. *loudness*) are generally unidimensional, whereas higher specific properties are frequently, however not necessarily, multidimensional (e.g. *timbre*). The different perceptual levels and their interrelations entail different empirical accessibilities: A couple of specific and holistic properties (e.g. *localization*) are accessible by direct measurement (manifest variables), other properties (e.g. *presence*) have to be indirectly observed by measuring lower-level properties (i.e. the aspects of *presence*) that indicate the higher.

Furthermore, Berg and Rumsey distinguish between properties related to the transmission system or to the recipient (descriptive versus attitudinal features) [30]. More generally speaking, properties of evaluation differ with respect to the object they are referred to: audio content, acoustic transmission systems, electroacoustic transmission systems, the listener, and relations between these entities.

4. Research objectives and suitability of properties

Which types of properties are appropriate for the evaluation of VAEs depends on the objective of the simulation. The measurement of holistic properties such as *authenticity* and *plausibility* makes sense when a VAE is evaluated in comparison to another system or to reality. The knowledge of the values of these properties is required either in order to apply the VAE as a research tool in the field of psychology of perception, e.g. in order to present experimental stimuli making up for acoustic scenes that would not be feasible or variable in reality, or in order to apply the VAE as an entertainment medium for untrained listeners. These objectives require only few degrees of freedom of parameters. On the other hand, if the VAE is used as an exploratory research tool in order to provide experts with experiences of combinations of conditions that have been inaccessible so far, e.g. in case of auralization for room acoustical design or for simulating cultural environments, many degrees of freedom of parameters are required.

The measurement of specific properties is indicated when the VAE is the object of investigation itself, typically in the development process when shortcomings of the system are to be identified by expert listeners. In this case, expert knowledge can help to derive technical characteristics from the specific perceptual properties.

Holistic and specific properties have to be measured within the same investigation in order to reveal correlations between these perceptual levels. This conduces to fundamental research in the field of psychology of perception in terms of revealing perceptual and cognitive mechanisms as well as constructing valid questionnaires.

Since different perceptual levels turn out to be adequate depending on the objective of a simulation, the question remains how these properties can be measured. Since VAEs have been evaluated applying predominantly specific properties (particularly localization accuracy, cf. e.g. [35] [36]) rather than holistic properties so far, we focus on the latter.

5. Measurement models

We use the term *measurement model* in a broader sense: to denominate both direct measurements, i.e. applying special paradigms and methods for observing a property, and indirect measurements, i.e. observing indicators that represent a construct. For the properties in question, both approaches are practicable. E.g., one might attempt to measure *plausibility* as a holistic property and at the same time ask by which specific properties *plausibility* is represented.

When regarding holistic properties as constructs that are operationalized by indicator variables (typically lower-level properties) the question arises, how valid indicators are to be determined and to be measured. Indicators may be determined by a theoretical definition, which is typically based on a respective discourse within the scientific community, as in case of *presence*. Alternatively, indicators may be determined empirically. To this end, qualitative preliminary studies, mostly using experts as test subjects, are carried out yielding the relevant properties. The elicitation of properties can rely on stimulus-based procedures such as the Repertory Grid Technique (RGT) [37] and the Verbal Protocol Analysis [38] [30]. The properties may also be elicited without presenting acoustic stimuli, as it is usually done in focus groups. From non-auditory sensory research, procedures generating individual vocabularies such as the Free Choice Profiling (FCP) [39] are known as well as procedures generating consensual vocabularies such as Flavor Profile, Quantitative Descriptive Analysis (QDA), Spectrum Descriptive Analysis, and Texture Profile [40]. Based on the properties found empirically a preliminary questionnaire is constructed, followed by an item analysis revealing redundancies of the items, indicating their reliability and their contribution to the construct to be indirectly measured, and showing the dimensionality of the scale as a whole. We consider that a theoretical or empirical definition of constructs and their revision by item analysis is indispensable in order to ensure validity, reliability, and intersubjectivity (quasi-objectivity). In case a construct is not a directly measurable concept at the same time but rather a latent variable, the relation between the indicators and the latent variable as well as the mutual relations between different latent variables may be examined by means of structural equation modeling [41].

Regarding the holistic properties *authenticity*, *plausibility* and *naturalness* as constructs, adequate indicators for them obviously remain to be determined. Considering *presence*, several elaborated questionnaires already exist as described above. They typically apply metric rating scales (of sub-aspects, where appropriate) or statement batteries for the collection of subjects' self reports. However, no consistent definition has become accepted to date. Hence, methods for a direct measurement of *presence* in the aggregate that have been proposed [10] [11] [12] lack validity as well. Whether the presence concept does cover all relevant aspects of VAEs, particularly of simulations referring to reality, is questionable. "Presence, when defined as a subjective sensation, can be a goal in itself for certain applications such as games and movies." However, "[…][b]ased on the current status of presence research, much uncertainty remains as to the usefulness of presence." [6] (p. 188).

Regarding *authenticity*, we agree to the consistent definition of Blauert and Pellegrini [13] [14], thereby proposing to gear it to the *auditory* indistinguishability (rather than the physical) of two acoustic stimuli, because *authenticity* is to be regarded as a psychological property. Thus, the only measurement strategy that comes into consideration, is a simple test for detection. However, when applying a yes/no paradigm, the subject's answer will contain both the detection performance (sensory component) and a response bias (psychological component), i.e. the subject's internal criterion leading to a 'yes' decision. Generally, the response bias is persistent in terms of a personality trait. The fact that the two components are confounded is known as the criterion problem. Normally, only the sensory component is of interest. The separation of the sensory and the psychological component is practicable by applying criterion-free answering paradigms: Forced-choice paradigms preclude the psychological component to affect the answer by consecutive presentations of the test stimulus and the reference stimulus in random order (hidden reference). The test subject is asked not to decide *whether* she/he has detected a signal but *at what position* ('interval') he/she has detected a signal. In case of uncertainty she/he has to guess – inevitably independently of the decision criterion. Thus, the psychological component that cannot be quantified is replaced by the guessing rate. Naturally, the guessing rate has to be accounted for in the analysis. However, there is a difference between forced-choice paradigms based on two intervals/alternatives (2AFC) and paradigms based on more intervals/alternatives (nAFC, $n>2$): 2AFC paradigms demand the definition of a property (e.g. *louder*) in order to allow the test subject to specify the correct interval/alternative. In contrast, nAFC paradigms ($n>2$) allow for the test on the unspecific property *different*. Thus, one does not have to rely on the correct understanding and applying of a property by the subject, and criterion-free judgments on the identity of two stimuli can be collected. That is why nAFC paradigms ($n>2$) are optimal in order to test the subject's ability to discriminate between a test stimulus and an external reference, i.e. to test on *authenticity*. The ABX test also applies a criterion-free paradigm quite similar to the 3AFC paradigm; moreover, it is a well-elaborated procedure with respect to statistical values [42].

A method allowing for an ex post elimination of the response bias is the signal detection theory (SDT) [17]

[43]. Therefore it has been recently applied to the measurement of *plausibility* according to the definition of Kuhn-Rahloff [15] whereupon *plausibility* may be considered as the perceptual compatibility of an acoustic stimulus with an internal reference based on experience (see above). By applying the yes/no paradigm, Lindau and Weinzierl ask their test subjects whether the test stimulus is in conflict with her/his mental representation of a corresponding real event, i.e. an internal reference [16]. This is a recognition task. The answers of a test subject pass a verification filter in order to check their correctness regarding the physical reality, i.e. the factual stimulus condition. This verification filter resulting hits, misses, correct rejections, and false alarms is a prerequisite for both forced-choice paradigms and the analysis in accordance to the SDT allowing for separating the discriminability from the response bias ex post. Insofar, this plausibility measure compares the relation between the test subject and the simulation with the relation between the test subject and the reality. Though the suggested measurement model for plausibility appears to be optimal with respect to test quality criteria and practicability so far, one should be aware of its limitations. At first, the investigator does not have access to the subject's recognition criterion. In case a subject applies a different criterion, e.g. 'like vs. dislike', his/her sensitivity decreases because the investigator still applies the verification filter 'real vs. simulated'. In order to match the subject's recognition criterion and the investigator's verification filter a clear definition and understanding of the recognition criterion and a sufficient expertise of the test subjects in this regard is essential. The second problem affects the empirical approach to all properties referring to reality: It is not feasible to repeatedly realize a physically identical real (in terms of naturally originated) acoustic event, e.g. a melody played on the violin. Thus, in compliance with the criteria of test quality it is only possible to test second order simulations for its relation to the reality as long as this reality is defined as a first order simulation (generally realized using sound converters).

Agreeing to the most evident definition of *naturalness* by Fredriksson and Zacharov ("close to the real-life experience") [24] (p. 2) that also refers to reality, this holistic property had to be measured in exactly the same way as described here for *plausibility*. Since the two properties were identical in terms of an operational definition, we suggest to abandon *naturalness* as a criterion for the perceptual evaluation of VAEs or to alternatively define *naturalness* as a construct.

Evaluation properties that are not based on predominantly operational definitions such as *authenticity* and *plausibility* require real definitions which often need to be more precise than colloquial denotations and meanings as well as elaborated measurement models in order to expand the contribution of listening tests to theory formation and hypothesis testing.

References

[1] Ahrens, J., & Spors, S. (2008). "An Analytical Approach to Sound Field Reproduction Using Circular and Spherical Loudspeaker Distributions". In: *Acta Acustica united with Acustica* 94:988-999.

[2] Slater, M. & Wilbur, S. (1997). "A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments". In: *Presence* 6:603-616.

[3] Lombard, M., & Ditton, T. (1997). "At the heart of it all: The concept of presence". In: *Journal of Computer Mediated Communication* 3(2): w/o paging.

[4] Slater, M., Usoh, M., & Steed, A. (1994). "Depth of presence in virtual environments". In: *Presence* 3:130-144.

[5] Witmer, B. G., & Singer, M. J. (1998). "Measuring presence in virtual environments: A presence questionnaire". In: *Presence* 7:225-240.

[6] Schuemie, M. J., Van der Straaten, P., Krijn, M., & Van der Mast, C. A. P. G. (2001). "Research on Presence in Virtual Reality: A Survey". In: *Cyberpsychology & Behaviour* 4(2):183-200.

[7] Schubert, T. W., Friedmann, F., & Regenbrecht, H.T. (1999). "Decomposing the sense of presence: Factor analytic insights". In: *2nd International Workshop on Presence, Univ. of Essex, UK, 6-7 April 1999*.

[8] Lessiter, J., Freeman, J., Keogh, E., & Davidoff, J. (2000). "Development of a new cross-media presence questionnaire: The ITC-Sense of presence". In: *Presence 2000 Workshop, March 27-28, Delft.*

[9] Lombard, M., & Ditton, T. (2000). "Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument". In: *Presence 2000 Workshop, March 27-28, Delft.*

[10] Larsson, P., Västfjäll, D., Olsson, P., & Kleiner, M. (2007). "When What You Hear is What You See: Presence and Auditory-Visual Integration in Virtual Environments". In: *Presence 2007, 10th Annual International Workshop on Presence, October 25-27, 2007, Barcelona, Spain*:11-18.

[11] IJsselsteijn, W. A., Freeman, J., Avons, S. E., Davidoff, J., de Ridder, H., & Hamberg, R. (1997). "Continuous assessment of presence". In: *Perception* 26 (suppl):42-43.

[12] Slater, M., & Steed, A. (2000). "A virtual presence counter". In: *Presence* 9:413-434.

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

[14] Pellegrini, R. (2001). *A virtual reference listening room as an application of auditory virtual environments*. Diss. Bochum: Ruhr-Univ., Inst. f. Kommunikationsakustik.

[15] Kuhn-Rahloff, C. (2012). *Realitätstreue, Natürlichkeit, Plausibilität: Perzeptive Beurteilungen in der Elektroakustik*, Berlin: Springer.

[16] Lindau, A., & Weinzierl, S. (in press). "Assessing the plausibility of virtual acoustic environments". In: *Acta Acustica united with Acustica*.

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

[18] Ashby, Gregory F. (2000). "A Stochastic Version of General Recognition Theory." In: *J. Math. Psych.* 44:310-329.

[19] Reiter, U. (2011). "Perceived Quality in Game Audio." In: Grimshaw, M. (ed.): *Game Sound Technology and Player Interaction: Concepts and Developments.* Hershey, New York: IGI Global:153-174.

[20] Pulkki, V., & Merimaa, J. (2005). "Spatial Impulse Response Rendering: Listening tests and applications to continuous sound." In: *118th AES Convention, Barcelona*. Peprint no. 6371.

[21] Queen, D. (1964). "Considerations for Naturalness in Portable Sound Reinforcement Systems". In: *16th AES annual Meeting, October 1994*.

[22] Theile, G. (1991). "On the Naturalness of Two-Channel Stereo Sound". *In: JAES* 39(10):761-767.

[23] Letowski, T., & Dreisbach, L. (1992). "Pleasantness and Unpleasantness of Speech". In: *11th International AES Conference*:159-165.

[24] Fredriksson, M., & Zacharov, N. (2002). "Natural reproduction of music and environmental sounds". In: *112th AES Convention, May 2002, Munich, Germany*.

[25] Berg, J., & Rumsey, F. (2002). "Validity of Selected Spatial Attributes in the evaluation of 5 channel microphone techniques". In: *112th AES Convention, May 2002, Munich, Germany*.

[26] Blauert J. (1994). "Product-sound Design and Assessment: An Enigmatic Issue from the point of view of engineering?". In: *inter-noise 1994, Yokohama Japan***,** *August 29-31*.

[27] Berg, J. (2009). "The contrasting and conflicting definitions of envelopment". In:*126th AES Convention 2009, Munich, Germany.*

[28] Steyer, R., & Eid. M. (2001). *Messen und Testen*. 2., korr. Aufl. Berlin et al.: Springer.

[29] Weinzierl, S., & Maempel, H.-J. (2012). "Sind Hörversuche subjektiv? Zur Objektivität akustischer Maße". In: *DAGA, 38. Jahrestagung für Akustik, Darmstadt*:315-316.

[30] Berg, J., & Rumsey, F. (2000). "In search of the spatial dimensions of reproduced sound: Verbal Protocol Analysis and Cluster Analysis of scaled verbal descriptors". In: *108th AES Convention, Paris*. Preprint no. 5139.

[31] Heide, B. L., & Maempel, H.-J. (2011). "Die Wahrnehmung audiovisueller Synchronität in elektronischen Medien". In: *26th VDT International Convention, 25.-28.11.2010 Congress Center Leipzig*:525-537.

[32] Maempel, H.-J. (2012). "Experiments on audio-visual room perception: a methodological discussion". In: *DAGA, 38. Jahrestagung für Akustik, Darmstadt*:783-784.

[33] Letowski, T. (1995). "Sound Quality Scales and Systems". In: *VI Sympozjum Rezyserii i Inzynierii Dzwieku*:90-101.

[34] Pellegrini, R. S. (1999). "Comparison of Data and Model Based Simulation Algorithms for Auditory Virtual Environments". In: *106th AES Convention, May 1999, Munich, Germany*.

[35] Møller, H., Hammershøi, D., Johnson, C. B., & Sørensen, M. F. (1997). "Evaluation of Artificial Heads in Listening Tests." In: *102nd AES Convention*, *München*. Preprint no. 4404.

[36] Liebetrau, J. et al. (2007). "Localization in Spatial Audio - from Wave Field Synthesis to 22.2." In: *123rd AES Convention, New York*. Preprint no. 7164.

[37] Berg, J., & Rumsey, F. (2006). "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique". In: *JAES* 54(5):365-379.

[38] Samoylenko, E., McAdams, S., & Nosulenko, V. (1996). "Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres". In: *Intern. J. of Psychology* 31:255-278.

[39] Carpenter, R. P., Lyon, D. H., & Hasdell, T. A. (Eds.) (2000). *Guidelines for Sensory Analysis in food product development and quality control*. Gaithersburg/MD: Aspen Publ.

[40] Hootman, R. C. (Ed.) (1992). *Manual on descriptive analysis testing for sensory evaluation*. Philadelphia/PA: ASTM.

[41] Bowen, N. K. (2012). *Structural equation modeling*. Oxford, New York: Oxford Univ. Press.

[42] Leventhal, L. (1986). "Type I and Type II Errors in the Statistical Analysis of Listening Tests". In: *JAES* 34(6):437-453.

[43] Gelfand, S. A. (2004). *Hearing. An Introduction to psychological and physiological acoustics.* 4th rev. and exp. ed., New York: Dekker.