Bibliographical reference

A Spatial Audio Quality Inventory (SAQI), Alexander Lindau, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkman and Stefan Weinzierl, Acta Acustica **vol. 100** (Number 5), 2014, pp. 984-994

DOI

https://doi.org/10.3813/AAA.918778

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 100 (2014) 984 – 994

DOI 10.3813/AAA.918778

A Spatial Audio Quality Inventory (SAQI)

Alexander Lindau, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkman, Stefan Weinzierl
Audio Communication Group, Technische Universität Berlin, Germany.
alexander.lindau@tu-berlin.de

Summary
The perceptual evaluation of spatial audio systems may be based on singular auditory qualities such as the localization accuracy or the perception of coloration, on overall criteria of perceptual accuracy such as plausibility and authenticity or on detailed catalogues of auditory qualities. However, only the latter will be suited for the perceptual characterization of a simulation's technical shortcomings and allow for its focused improvement. Therefore, a common vocabulary containing all perceptual attributes which are relevant in this context appears desirable. Existing vocabularies for the evaluation of sound field synthesis, spatialization technologies and virtual acoustic environments were often generated *ad hoc* by authors or were focused on specific perceptual aspects or on specific spatialization techniques only. To overcome limitations with respect to the relevance and completeness of these vocabularies we have developed a Spatial Audio Quality Inventory (SAQI) for the perceptual evaluation of all spatial audio technologies used for the (re)synthesis of acoustic environments. It is a consensus vocabulary comprising 48 verbal descriptors of auditive qualities assumed to be of practical relevance when comparing (re)synthesized sound fields to real or imagined references or amongst each other. The vocabulary was generated by a Focus Group of 21 German speaking experts for virtual acoustics. Five additional experts helped verifying the unambiguity of all descriptors and the related explanations. Moreover, an English translation was generated and verified by eight bilingual experts. This article describes the applied methodology and presents the English version of the final vocabulary.

PACS no. 55.09.Ka, 66.10.Lj

1. Introduction

1.1. Objective

Today, 'Spatial Audio' is an umbrella term related to different technical approaches to spatial sound field (re)synthesis. These sound fields may be delivered as ear signals via headphones, or synthesized over an extended listening area using either few or up to several hundreds of loudspeakers [1]. Technologies include static and dynamic binaural synthesis, multichannel loudspeaker reproduction as well as sound field synthesis with approaches such as Wave Field Synthesis (WFS) or Higher Order Amibsonics (HOA).

Spatial Audio systems have often been evaluated with respect to singular auditive qualities such as, e.g., localization accuracy, the perception of coloration or distance [2, 3]. However, it remains unclear in how far these qualities are relevant for the sensory accuracy of such a system as a whole. To overcome this uncertainty, measures of overall perceptual accuracy such as plausibility [4] or authenticity [5, p. 373], have been suggested with operational definitions and experimental designs given in [6] and [7]. These measures differentiate between assessments with respect to an inner reference resulting from experience and expectation (*plausibility*) or to an external, explicitly given reference (*authenticity*). While being of value for, e.g., the benchmarking of systems and approaches, these overall measures will give no insight into specific perceptual deficiencies and the related technical shortcomings which would be required for a further improvement of the system under test. Hence, in the current study we developed a descriptive sensory vocabulary for spatial audio technologies which may conveniently be applied for purposes such as the directed technical improvement, the perceptually motivated cost reduction, the detailed documentation of the perceptual state of the art, or the qualitatively differentiated benchmarking.

Received 16 April 2014,
accepted 14 July 2014.

1.2. State of the art

Descriptive sensory vocabularies have been developed for various fields of interest in audio, as e.g. room acoustics, loudspeakers, multichannel recording and reproduction systems [8, 9, 10, 11], as well as for room correction systems, audio codecs, algorithms for headphone spatialization [12] and also for more complex Virtual Acoustic Environments (VAEs) [13, 14, 15]. Previous studies often applied vocabularies which were initially generated *ad hoc* from experience/knowledge of the authors and then

reduced by applying factor analysis to listeners' ratings of a set of exemplary stimuli. Recently, empirically more substantiated procedures have been applied for the development of consensual sensory vocabularies in the field of audio, such as Quantitative Descriptive Analysis – QDA [11], Free Choice Profiling – FCP [12], Flash Profiling – FP [8] or the Repertory Grid Technique – RGT [10]. One study [15] was concerned with the development of a descriptive vocabulary ('quality features') for dynamic binaural auralizations from geometrical room acoustic simulations. To this end, the author did not use a stimulus- but an expert-based approach (Delphi method, [16]). However, the scope of this study was limited to the specific implementation described. Furthermore, while initially in [15] the central evaluation criterion 'quality' was understood quite broadly as the degree of agreement with a desired characteristic, in the course of the study it was narrowed by referring to three specific use cases (localization test, chat room, edutainment scenario) raising again doubts about the universality of the presented vocabulary.

By taking into account auditive qualities which were mentioned by at least two of the five studies [11, 12, 13, 14, 15] most closely related to our research objective, eleven attributes could be identified as potentially relevant for assessing spatial audio technology. Ordered for frequency of appearance (in brackets) those were: spectral coloration (5), spaciousness (5), localizability (5), steadiness of movements (5), source width (4), loudness (2), loudness balance (2), distance (2), internalization vs. externalization (2), impulse-like artifacts (2), and dynamic responsiveness (2). However, for a comprehensive perceptual evaluation of virtual environments the existing vocabularies did not appear sufficiently complete, nor do they cover all aspects of particular importance. For instance, while there is usually a descriptor for the perceived width of individual sources, there is none for the dimensions of complex ensembles of sources or for the spatial environment, such as the height or depth of rooms. Furthermore, elementary problems of the spatial rendering are not covered, such as offsets in perceived location. A reason for these gaps might be that none of the authors explicitly targeted *comparative* assessments, either with regard to reality or between different approaches to sound field (re)synthesis.

Existing standards dedicated to terminological issues in the field of sensory evaluation [17], electro-acoustics [18], and acoustics [19] either neglect the auditory domain [17] or they are limited to only a few relevant terms such as loudness, timbre, and noisiness [18]. The German standard for acoustic terminology [19] includes a comparatively large number of potentially relevant terms such as tone color, (auditory) spaciousness, localization, lateralization, loudness, roughness, pitch, intracranial locatedness, (flutter) echo, reverberance, duration of reverberation, tone impulse, (speech) intelligibility, or transparency. However, doubts remain with regard to completeness and relevancy for the intended subject of spatial audio technologies.

Hence, in order to allow for an exhaustive differential diagnosis of perceptual shortcomings of virtual acoustic environments of any type, covering all aspects of particular relevance for the perceived quality of the different technologies involved, the development of a new sensory consensus vocabulary (CV) appeared mandatory. Furthermore, this aim was supposed to be approached by utilizing an empirically substantiated approach preferably based on an agreement within a larger group of experts.

2. Methods

2.1. General considerations

Techniques for the spoken elicitation of descriptive consensus vocabularies as summarized in [20, pp. 43], may be divided into individual (incl. Repertory Grid Technique, Free Choice Profiling, Flash Profiling) and group-based approaches (incl. Quantitative Descriptive Analysis, Flavour Profile, Texture Profile, Sensory Spectrum). Whereas individual elicitation methods are confronted with the still unsolved problem of merging individual vocabularies into a valid group vocabulary, group methods directly aim at deriving a consensual language. In the latter case, panels of naïve subjects are often instructed to develop a descriptive language by discussing sensory impressions of selected stimuli under guidance of a moderator. Such procedures are time-consuming [11] and the chosen set of stimuli is critical with respect to the representativeness of results. To obtain such a set of stimuli representative for all approaches to sound field spatialization is difficult, if not at all impossible. Therefore, as an alternative, three established approaches for non-stimulus based CV generation were taken into consideration: the Nominal Group method [21], the Delphi method [16] and the Focus Group method [22]. All three procedures rely on the assumption that the superior practical and theoretical experience of an expert compensates for the lack of adequate stimuli. The Nominal Group and the Delphi method are interview or survey-based procedures, respectively, for finding an agreement in a course of single or repeated interrogation sessions, i.e. without direct contact between experts. In contrast, the Focus Group may be applied, if experts can be accessed for face-to-face moderated roundtable discussions. While the two first approaches are considered as reducing group bias, the latter is expected to lead to more vivid and potentially more effective discussions.

2.2. The Focus Group approach

As partners in a larger research consortium for virtual acoustics (SEACEN, Simulation and Evaluation of Acoustical Environments[1]) the authors had a comparably easy access to experts in the field in order to arrange face-to-face meetings, making a Focus Group approach feasible.

Methodologically, a Focus Group may be regarded as a combination of a guided interview and a group discussion. This combination is particularly well-suited for the

[1] http://www.seacen.tu-berlin.de

elicitation of expert-knowledge, as experts are routinely used to discourse-based revelation of consensual knowledge [23]. The incompleteness of results and irrelevancy during discussions can be reduced by using the so-called *dual-moderator setting*, where one moderator guides the discussion, while a co-moderator keeps track of the predefined agenda and discussion guidelines. The moderator is supposed to control for unwanted group effects, e.g. by restraining 'leading' and motivating 'hiding' discussants, and being sensitive to non-verbal communication. The co-moderator is supposed to monitor the moderator's behavior and the general compliance with the discussion guidelines. Experimenter bias and group effects may be further addressed by extending the scheme to a so-called *two-way Focus Group*. There, during the first part of each discussion round the panel is split up in two groups, one group discussing (in dual-moderator scheme) and the other group observing the discussion from a remote room without interfering directly (for instance via one-way AV-monitoring). The observer group acts as a control mechanism: Being less exposed to group effects, the observers are supposed to follow the discussion more objectively and rationally. In the second part, observers and discussants are brought together, discussing the observers' comments on the preceding discussion. If the targeted objective cannot be reached within a single discussion a *serial* Focus Group scheme allows for repeated discussion rounds.

2.3. Discussion panel

As participants several German speaking experts were invited (PhD candidates, post-docs and professors), representing a wide professional experience regarding recording, simulation, reproduction and evaluation of spatial sound fields (see acknowledgements for the composition of the group). While there were some changes over the different meetings regarding group size and composition, the panel size of 10–15 participants (21 experts in total, aged 25 to 67 yrs., 1–2 females per meeting) may be considered as optimal [24]. According to [24] the panel may further be regarded as a 'homogenous real group', i.e. with discussants coming from similar educational background and being known to each other before. In contrast to groups of differing backgrounds, homogenous groups are expected to lead more effective discussions. Moreover, real groups may (a) be more influenced by given hierarchies and role models, and (b) be more prone to 'private' conversations than randomly assigned groups, effects that have to be made aware of and controlled for by the moderator.

Discussions were held at four meetings in Berlin and Aachen over a period of 6 months. During those meetings repeated discussion rounds were scheduled for between one and four days. Discussions were conducted in the two-way dual-moderator scheme: After separating the experts into panel and observer group, the results were continuously updated on a projection screen, with the discussion audiovisually transmitted to the observer group along with the current state of the discussion. The AV-transmission was recorded for documentation purposes.

2.4. Main discussion objectives

The vocabulary to be developed was defined as aiming at a consensual psychological measurement instrument, i.e. a questionnaire comprising a list of auditory qualities and respective rating scales to be used in listening tests. As primary objects of assessment we considered Virtual Acoustic Environments of all kinds. The sensory scope was defined to comprise all perceivable differences of VAEs when comparing them to reality (be it imagined or explicitly given), or when comparing between different technical implementations thereof. At last, some typical intended applications of the future vocabulary were given such as technical improvement, effort reduction, and benchmarking. These main objectives were aggregated into a mission statement:

> *Creation of a consensus vocabulary for evaluating apparatus-related perceptual differences between technically generated acoustic environments or with respect to a presented or imagined acoustic reality.*

The experts were instructed that terms should relate to auditive perceptions, not physical quantities or measures. Furthermore, they were asked to aim at completeness of the overall vocabulary, while, at the same time, consider the assumed practical relevance. Descriptors should be formulated as semantically unidimensional and mutually exclusive as possible. Since the terms should be self-explaining to other experts in the field, the use of established names was to be preferred over inventing new ones. If the experts found it difficult to give a self-explaining term, a short supplementary statement could be added to the descriptor. In particularly difficult cases – i.e. when defining a quality's name and giving a supplementary statement was not fully satisfying – it could also be agreed upon creating some illustrative audio examples later on. Hence, demanding such an audio example was understood as an exception, not a rule. Finally, the group was instructed to propose suitable comparative rating scale label only if they were assumed to be necessary for clarity, leaving details of the operationalization mainly to the authors of the study.

2.5. Discussion rules and agenda

All major decisions were to be agreed upon by simple majority, with the moderator being excluded from voting. However, after thorough and sometimes long discussions, a full consensus could be reached in most cases. After each discussion round, feedback rounds were conducted protocolling group comments on the moderator's performance, the progress of the discussion, setting and organization. Comments of the observer group were kept for the record until discussed to the satisfaction of the group.

After defining these rules, an initial agenda was created by means of 20-minute brainstorming sessions conducted at the beginning of the first two discussion rounds and resulting in a list of 62 initial descriptors serving as a basis for the subsequent discussion. Each discussion round lasted about 3.5 hours, with 2 rounds conducted at

maximum per day. A discussion round began with a random separation into discussants and observers. After 90 minutes of observed discussion a 20-minute consolidation break was given to the observers. Finally, discussants and observers joined for a 90-minute roundtable discussion. If the discussion round was the last of a meeting, it was concluded by a feedback round. In total, 16 such rounds were completed summing up to 56 hours of discussions.

Following recommendations in [23] the group was involved in the development of objectives, agenda and rules, thus increasing the involvement of participants and maintaining a thematically focused discussion. Thus, for instance, the group repeatedly specified the mission statement more precisely, added a rotation rule for panel and observer group, motivated the moderator to – in case of lively discussions – apply a speaker list or demanded from the observers to present their statements as an organized and consolidated list of pleas.

2.6. Finalization

Discussions resulted in a preliminary vocabulary which was finalized in a four-stage procedure: First, a postcheck of semantic consensus was carried out by creating short written circumscriptions for all descriptors. Those experts that took part in the majority of discussion rounds (12 participants) were invited again to comment on the proposed circumscriptions in a written on-line discussion moderated by the authors. During the Focus Group discussions three descriptors ('roughness', 'comb filter coloration', and 'dynamic compression effects') had been perceived as not being sufficiently self-explaining. Therefore, illustrative audio stimuli demonstrating the respective qualities were created by use of proprietary or self-implemented signal processing algorithms. Finally, consensual circumscriptions for all descriptors and a representative selection of illustrative audio stimuli were agreed upon.

Second, the vocabulary was subjected to an external evaluation of understandability. For this purpose, five additional experts for spatial audio technologies were asked to individually explain what descriptors meant to them while being given only the descriptor names, the (optional) short supplementary statements and audio examples, as well as objective and method of the vocabulary development. They were also asked to state, whether audio examples were adequate.

Third, after receiving all written statements (ca. 250) those were analyzed and checked for semantic equivalence with the group's circumscriptions. For about a third of the descriptors minor semantic problems were identified. In a fourth step, corrections derived from the analysis of additional experts' comments were agreed upon during final face-to-face discussion of a core group of five experts. Thereby, one attribute was confirmed to be obsolete. Moreover, labels for rating scales for each descriptor were agreed upon, considering, where available, earlier proposals of the group. Further, it was agreed upon including the final circumscriptions into the vocabulary as in most cases this was supposed to resolve remaining uncertainties identified by external experts, this way resulting in the final German version of the SAQI [25].

2.7. Translation to English

As descriptive vocabularies are known to be sensitive to language, care has to be taken in translation to conserve the original meaning [20, p. 46]. Translators should hence be sensitive to both the obviously meant (denotation), and the, potentially interculturally differing, ascribed meaning (connotation). According to guidelines related to the translation of psychological tests [26], it is recommended to invite as translators at least two experts in the field which are fluent in both languages. To ensure validity of the translation it is further proposed to back-translate the questionnaire and to consider some more experts for a final review.

Regarding the target language we assumed a 'technical community language' to exist in the field of acoustics, which is neither a real US, UK nor any other native English. Furthermore, we would consider any scientist in the field a 'native' speaker of this 'community language'. Accordingly, as translators, we invited one native US, one native UK, one Greek and two Dutch acousticians. Three of them were researchers in virtual acoustics and all had good knowledge of German. They were provided with the German descriptors, the pre-translated circumscriptions, and the audio examples and were asked to produce adequate English terms. Translations were finally discussed in a teleconference including the translators and the authors.

Additionally, three more German experts for virtual acoustics living for a longer period in English-speaking countries produced back-translations which were in turn semantically analyzed by the authors. It has to be emphasized that besides being a test of the semantic compatibility of the English and German version, the back-translation can also be regarded a test of the assumed 'bilingualism' of (here: German) experts in the 'community English'. Thus, finding the back-translated versions only slightly different in meaning from the original German SAQI was considered as incidental empirical evidence in support of our above hypothesis.

3. Results

The final vocabulary is termed *Spatial Audio Quality Inventory* (SAQI, cf. Table I). It consists of 48 descriptors of auditive qualities which can be roughly sorted into eight categories (timbre, tonalness, geometry, room, time behavior, dynamics, artifacts, and general impressions) and are to be considered as describing 'perceived differences with respect to [descriptor name]'.

While some attributes of the SAQI reflect a 'bottom-up' perspective of perception, being closely related to temporal or spectral properties of the audio signal ('Loudness', 'High/Mid/Low frequency tone color', 'Horizontal/Vertical direction'), other attributes reflect a more 'top-down' perspective representing higher-order psychological constructs, supra-modal, affective, aesthetic or attitudinal aspects ('Clarity', 'Naturalness', 'Presence').

Table I. Spatial Audio Quality Inventory (SAQI) - English version.
*sound examples may be downloaded from http://dx.doi.org/10.14279/depositonce-1

Quality	Circumscription	Scale End Label
Difference	Existence of a noticeable difference.	none – very large
Timbre		
Tone color bright-dark	Timbral impression which is determined by the ratio of high to low frequency components.	darker – brighter
High-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
Mid-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
Low-frequency tone color	Timbral change in a limited frequency range.	attenuated – emphasized
Sharpness	Timbral impression which e.g., is indicative for the force with which a sound source is excited. Example: Hard/soft beating of percussion instruments, hard/soft plucking of string instruments (class. guitar, harp). Emphasized high frequencies may promote a 'sharp' sound impression.	less sharp – sharper
Roughness*	Timbral impression of fierce or aggressive modulation/vibration, whereas individual oscillations are hardly distinguishable. Often rated as unpleasant.	less rough – more rough
Comb filter coloration*	Often perceived as tonal coloration. 'Hollow' sound. Example: speaking through a tube.	less pronounced – more pronounced
Metallic tone color	Coloration with pronounced narrow-band resonances, often as a result of low density of natural frequencies. Often heard when exciting metallic objects such as gongs, bells, tin cans. Applicable to room simulations, plate reverb, spring reverb, too.	less pronounced – more pronounced
Tonalness		
Tonalness	Perceptibility of a pitch in a sound. Example for tonal sounds: voiced speech, beeps.	more unpitched – more pitched
Pitch	The perception of pitch allows arranging tonal signals along a scale "higher – lower".	lower – higher
Doppler effect	Continuous change of pitch (see above). Often perceived as a 'continuous detuning'. Example: 'Detuned' sound of the siren of a fast-moving ambulance.	less pronounced – more pronounced
Geometry		
Horizontal direction	Direction of a sound source in the horizontal plane.	shifted anticlockwise – shifted clockwise (up to 180°)
Vertical direction	Direction of a sound source in the vertical plane.	shifted down – shifted up (up to 180°)
Front-back position	Refers to the position of a sound source before or behind the listener only. Impression of a position difference of a sound source caused by 'reflecting' its position on the frontal plane going through the listener.	dichotomous scale: not confused/confused
Distance	Perceived distance of a sound source.	closer – more distant
Depth	Perceived extent of a sound source in radial direction.	less deep – deeper
Width	Perceived extent of a sound source in horizontal direction.	less wide – wider
Height	Perceived extent of a sound source in vertical direction.	less high – higher
Externalization	Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance. Terminologically often enclosed between the phenomena of in-head localization and out-of-head localization. Examples: Poorly/not externalized = perceived position of sound sources at diotic sound presentation via headphones, good/strongly externalized = perceived position of a natural source in reverberant environment and when allowing for movements of the listener.	more internalized – more externalized

Table I. Continuation

Localizability	If localizability is low, spatial extent and location of a sound source are difficult to estimate, or appear diffuse, resp. If localizability is high, a sound source is clearly delimited. Low/high localizability is often associated with high/low perceived extent of a sound source. Examples: sound sources in highly diffuse sound field are poorly localizable.	more difficult – easier
Spatial disintegration	Sound sources, which - by experience - should have a united spatial shape, appear spatially separated. Possible cause: Parts of the sound source have been synthesized/simulated using separated algorithms/simulation methods and between those exists an unwanted offset in spatial parameters. Examples: fingering noise and playing tones of an instrument appear at different positions; spirant and voiced phonemes of speech are synthesized separately and then reproduced with an unwanted spatial separation.	more coherent – more disjointed
Room		
Level of reverberation	Perception of a strong reverberant sound field, caused by a high ratio of reflected to direct sound energy. Leads to the impression of high diffusivity in case of stationary excitation (in the sense of a low D/R-ratio). Example: The perceived intensity of reverberation differs significantly between rather small and very large spaces, such as living rooms and churches.	less – more
Duration of reverberation	Duration of the reverberant decay. Well audible at the end of signals.	shorter – longer
Envelopment (by reverberation)	Sensation of being spatially surrounded by the reverberation. With more pronounced envelopment of reverberation, it is increasingly difficult to assign a specific position, a limited extension or a preferred direction to the reverberation. Impressions of either low or high reverberation envelopment arise with either diotic or dichotic (i.e., uncorrelated) presentation of reverberant audio material.	less pronounced – more pronounced
Time behaviour		
Pre-echoes	Copies of a sound with mostly lower loudness prior to the actually intended the starting point of a sound.	less intense – more intense
Post-echoes	Copies of a sound with mostly decreasing loudness after the actually intended the starting point of a sound. Example: repetition of one's own voice through reflection on mountain walls.	less intense – more intense
Temporal disintegration	Sound sources, which - by experience - should have a united temporal shape, appear temporally separated. Causes similar to "Spatial disintegration", however, here: due to timing-offsets in synthesis. Example: fingering noise and playing tones of an instrument appear at different points in time.	more coherent – more disjointed
Crispness	Characteristic which is affected by the impulse fidelity of systems. Perception of the reproduction of transients. Transients can either be more soft/more smoothed/less precise, or - as opposed - be quicker/more precise/ more exact. Example for 'smoothed' transients: A transmission system that exhibits strong group delay distortions. Counter-example: Result of an equalization aiming at phase linearization.	less pronounced – more pronounced
Speed	A scene is identical in content and sound, but evolves faster or slower. Does not have to be accompanied by a change in pitch. Examples of technical reasons: rotation speed, sample rate conversion, time stretching, changed duration of pauses between signal starting points; movements proceed at a different speed.	reduced – increased
Sequence of events	Order or occurrence of scene components. Example: A dog suddenly barks at the end, instead - and as opposed to the reference - at the beginning.	unchanged – changed
Responsiveness	Characteristic that is affected by latencies in the reproduction system. Distinguishes between more or less delayed reactions of a reproduction system with respect to user interactions.	lower – higher
Dynamics		
Loudness	Perceived loudness of a sound source. Disappearance of a sound source can be stated by a loudness equaling zero. Example of a loudness contrast: whispering vs. screaming.	quieter – louder

Table I. Continuation

Dynamic range	Amount of loudness differences between loud and soft passages. In signals with a smaller dynamic range loud and soft passages differ less from the average loudness. Signals with a larger dynamic range contain both very loud and very soft passages.	smaller – larger
Dynamic compression effects*	Sound changes beyond the long-term loudness. Collective category for a variety of percepts caused by dynamic compression. Examples: More compact sound of sum-compressed music tracks in comparison to the unedited original. 'Compressor pumping': Energy peaks in audio signals (bass drums, speech plosives) lead to a sudden drop in signal loudness which needs a susceptible period of time to recover.	less pronounced – more pronounced
Artifacts		
Pitched artifact	Perception of a clearly unintended sound event. For example, a disturbing tone which is clearly not associated with the presented scene, such as an unexpected beep.	less intense – more intense
Impulsive artifact	Perception of a clearly unintended sound event. For example, a short disturbing sound which is clearly not associated with the presented scene, such as an unexpected click.	less intense – more intense
Noise-like artifact	Perception of a clearly unintended sound event. For example, a noise which is clearly not associated with the presented scene, such as a background noise from of a fan.	less intense – more intense
Alien source	Perception of a clearly unintended sound event. Examples: an interfering radio signal, a wrongly unmuted mixing desk channel.	less intense – more intense
Ghost source	Spatially separated, nearly simultaneous and not necessarily identical image of a sound source. A kind of a spatial copy of a signal: a sound source appears at one or more additional positions in the scene. Examples: two sound sources which are erroneously playing back the same audio content; double images when downmixing main and spot microphone recordings; spatial aliasing in wave field synthesis (WFS): sound sources are perceived as ambivalent in direction.	less intense – more intense
Distortion	Percept as a result of non-linear distortions as caused e.g. by clipping. Scratchy or 'broken' sound. Often dependent on signal amplitude. Perceptual quality can vary widely depending on the type of distortion. Example: clipping of digital input stages.	less intense – more intense
Tactile vibration	Perception at the border between auditory and tactile modality. Vibration caused by a sound source can be felt through mechanical coupling to supporting surfaces. Examples: Live Concert: bass can be 'felt in the stomach'; headphone cushions vibrate noticeably on the ear/head.	less intense – more intense
General		
Clarity	Clarity/clearness with respect to any characteristic of elements of a sound scene. Impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected. The term is thus to be understood much broader than the in realm of room acoustics, where Clarity is used to predict the impression of declining transparency with increasing reverberation.	less pronounced – more pronounced
Speech intelligibility	Impression of how well the words of a speaker can be understood. Typical of low speech intelligibility: train station announcements. Typical for high speech intelligibility: Newscaster.	lower – higher
Naturalness	Impression that a signal is in accordance with the expectation/former experience of an equivalent signal.	lower – higher
Presence	Perception of 'being-in-the-scene', or 'spatial presence'. Impression of being inside a presented scene or to be spatially integrated into the scene.	lower – higher
Degree-of-Liking	Difference with respect to pleasantness/unpleasantness. Evaluation of the perceived overall difference with respect to the degree of enjoyment or displeasure. Note that 'preference' might not be used synonymously, as, e.g., there may be situations where something is preferred that is - at the same time - not liked most.	lower – higher
Other	Another, previously unrecognized difference.	less pronounced – more pronounced

Table II. Assessments entities: Events or objects a perceived difference may be addressed to.

All audible events				
Intended audible events (elements of the presented virtual scene)			Unintended audible events	
Foreground sources	Background sources	Room acoustic environment	Reproduction system	Laboratory environment

Table III. Modifications: Temporal and causal variations that may help defining a perceived difference in more closely.

The perceived difference is ...		
... constant	... varying periodically or otherwise rule-based with time	... varying non-regularly with time
	... in a continuous / discontinuous manner	
... and depending on scene events / user interaction / independent.		

Each descriptor is complemented by a short written clarifying circumscription and suitable dichotomous, uni- or bipolar scale label, respectively. As mentioned above, for three of the descriptors, illustrative audio examples were created ('Roughness', 'Comb filter coloration', 'Dynamic compression effects'). For handling possibly overlooked or newly emerging aspects, an open category ('Other') was included in the vocabulary, to be named by subjects of the listening tests. Readers are cordially invited to share their experience with this category with the corresponding author.

Furthermore, it appeared reasonable to be able to address a perceived difference to certain reference objects of an acoustical environment. Hence, five basic assessment entities were defined, providing an ideal-type ontology for a virtual acoustic scene: *foreground sources, background sources*, the simulated *room acoustical environment*, the *reproduction system* itself (e.g., as source for loudspeaker artifacts, amplifier noise) and the *laboratory environment* (e.g., as source for HVAC noise or exterior/environmental sounds). In combination, these five entities are thought to incorporate all possible objects of interest (Table II). Additionally, it was perceived as valuable to be able to further differentiate observed perceptual differences with respect to their time-variance. Thus, in the first stage, perceived differences might be either *constant* or *time-varying*. Second, an observed time-variance might be *periodical or otherwise rule-based* or *non-regular* and at the same time be perceived as *continuous* or *discontinuous* (Table III). Finally, all perceived differences may be defined more closely with respect to their cause, i.e. whether they *depend on user interaction, depend on scene events* or on none of them (*independent*, Table III).

4. Application

In order to illustrate the practical application of the SAQI, we shortly describe an exemplary listening test procedure (cf. Figure 1). Further references as, e.g., to the SAQI Manual, to audio examples and a free listening test software implementing the SAQI can be found in the outlook section.

Before conducting a SAQI test, it should to be understood, that the vocabulary in its complete form, i.e. comprising quality names, short descriptions, scale labels and audio examples is intended to be self-explaining to any expert in the field. Furthermore, these experts are assumed to be able to train laymen to the proper understanding of the SAQI, e.g., by giving more detailed explanations, discussing suitable real-life examples or by providing more illustrative sound examples. Hence, for the remainder of this section it is assumed that the test subject has been trained by the researcher to the proper understanding of the vocabulary.

As can be seen from Figure 1, a SAQI test starts with the auditive comparison of a test stimulus and a given or imagined reference. Then the subject should first be asked whether it perceived any difference at all, because, if the subjects denies, the test could be stopped at this point. Otherwise, the perceived overall difference may be rated using a unipolar intensity scale. As visible from Table 1 scale end labels are proposed in a comparative style, i.e. "darker/brighter" instead of "dark/bright". This comes rather naturally, as any SAQI assessment is intrinsically a comparison task, independent from the used references being imagined or explicitly given. After rating, the researcher may optionally ask the subject to indicate the temporal behavior, user- or scene-related dependencies and/or to assign reference objects to the perceived difference which can be done by using multiple-choice questions. These options can be selected with regard to the research interest or with respect to the used stimuli. The procedure is repeated for all selected attributes contained in the SAQI, potentially in randomized presentation order while the test stimuli are accessible for continued comparison. More exploratory research questions might ask for the application of the full SAQI, whereas for very specific hypotheses, some attributes of the SAQI (e.g., the 'artifacts' section) might be sufficient. However, applying the full SAQI will (a) increase comparability across studies, and (b) will not only reveal perceptual issues but also show in which respect a simulation is unproblematic. Finally, subjects should be asked to specify and to rate other differences that have potentially been overlooked. Rating the

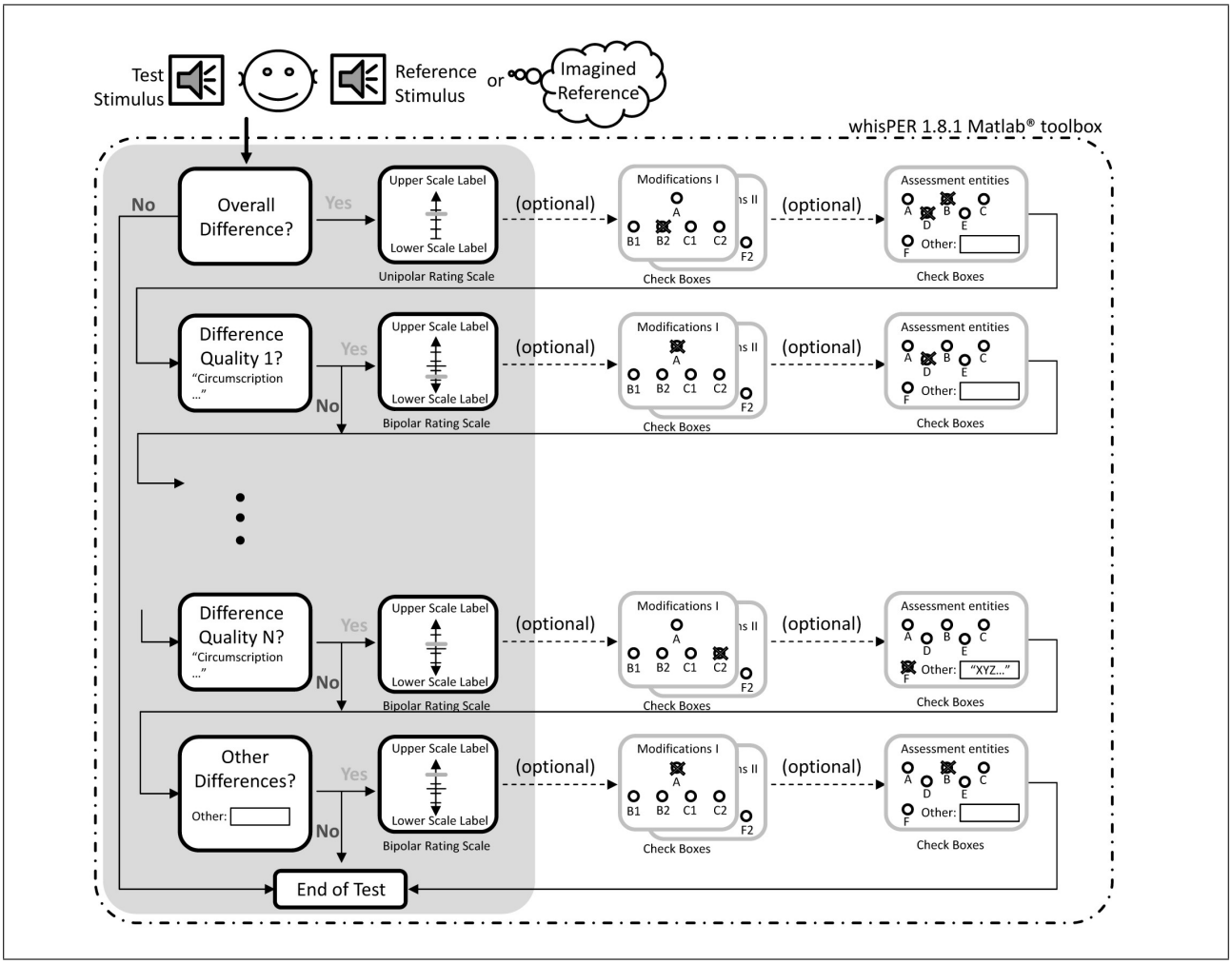

Figure 1. Illustration of an exemplary listening test applying the Spatial Audio Quality Inventory (SAQI). The depicted procedure corresponds to the implementation in the free listening test software WhisPER [27].

complete SAQI including decisions about assessment entities takes about 10-15 minutes for a subject who is familiar with the procedure and meaning of all qualities.

5. Discussion

Throughout the current investigation, the expert Focus Group approach proved to be an effective method for deriving a consensual vocabulary for the assessment of acoustic environments generated by spatial audio systems. It yielded not only a comprehensive semantic differential but also some valuable extensions such as a systematics for reference entities, for the temporal behavior of auditory attributes, illustrative audio examples, and a glossary of terms.

Different mechanisms of self-control, used to ensure the objectivity of the approach, turned out to be important, e.g. in cases, where the moderator tended to influence the discussion or where the 'power of persuasion' was not fully balanced within the group. In these cases, both the comments of the observer group and the external evaluations were perceived as valuable. Construct validity of the discussions benefited from recent debates on quality measures for virtual acoustic environments [6, 7, 28]. Thus, – e.g. from discussing the mission statement – disputants were sensitized regarding a clear separation between perceptual impressions and physical measures and were aware of the relevance of assessments regarding both inner and external references. Furthermore, the content validity of the SAQI is assumed to be high, since the panel covered a substantial and diverse range of expertise regarding the topic under discussion. To obtain an impression of the completeness of the derived vocabulary, it can be confronted with results from previous studies as summed up in the first paragraph of section 1.2. Thus, it can be verified that all previously identified aspects are also covered by the SAQI. Although the Focus Group discussions took more time than expected, the overall duration of the vocabulary development was comparable to those reported for other approaches (i.e. for QDA [11], RGT [20], or the Delphi method [15]). Finally, when accepting the presented English translation of the SAQI as a 'community language version' the effort for further translations might be greatly reduced: If most experts in the field are considered 'bilingual' in their native and the community language, they should be able to produce valid translations by

themselves. Hence, the English version is also intended as a 'bridge' to the international community allowing a convenient creation of national versions of the SAQI.

6. Outlook

A database of audio examples covering most aspects of the SAQI is currently being developed to be used (a) for the training of subjects and test panels, and (b) as anchor stimuli to increase reliability and absolute comparability of SAQI test results. The retest reliability of SAQI items shall be assessed, across German and English, by providing a sample of identical stimuli to be included in different listening tests. Based on these results, we will also be able to analyze the interdependency of items.

The German and English versions of the SAQI have been incorporated in the Matlab® listening test environment WhisPER [27] which is freely available online. Procedures for the convenient modification of the questionnaire, e.g., for selecting language, test paradigm (paired comparison or direct assessment), descriptors, or assessment entities as well as their time variance are explained in the software's documentation. A French version of the SAQI is currently prepared in cooperation with the *Institut de Recherche et Coordination Acoustique/Musique* (IRCAM, Paris). Practical guidelines (e.g., for test subject training, test customization), audio examples, and Matlab® routines for direct visualization and importing of test results into statistical analysis software can be obtained from the online repository hosting the SAQI Test Manual [29].

Acknowledgement

This investigation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG FOR 1557, DFG WE 4057/3-1). The authors – all of them members of the Focus Group themselves – would like to thank for participation (in alphabetical order): Benjamin Bernschütz, Clemens Büttner, Diemer de Vries, Alexander Fuß, Matthias Geier, Martin Guski, Michael Horn, Stefan Klockgether, Johannes Nowak, Rob Opdam, Zora Schärer, Frank Schultz, Sascha Spors, Michael Vorländer, and Hagen Wierstorf. We also would like to thank the five external experts for providing their semantic circumscriptions of the pre-final qualitative descriptors: Matthias Frank, Markus Noisternig, Sönke Pelzer, Andreas Silze, and Franz Zotter. Moreover we would like to thank the five translators: Jude Brereton, Kees de Visser, Brian Gygi, Charalampos Saitis, and Steven van de Par while our final regards go to the three back-translators Frank Melchior, Nils Peters and Ulrich Reiter.

References

[1] 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. Proc. of the IEEE, 101, 2013, 1920–1938.

[2] H. Wierstorf, A. Raake, S. Spors: Localization in wave field synthesis and higher order ambisonics at different positions within the listening area. Proc. of the AIA-DAGA 2013 Conference on Acoustics, Meran, 2013, 2376–2379.

[3] H. Wittek, S. Kerber, F. Rumsey, G. Theile: Spatial perception in wave field synthesis rendered sound fields: Distance of real and virtual nearby sources. Proc. of the 116th AES Convention, Berlin, 2004, preprint no. 6000.

[4] R. S. Pellegrini: Quality assessment of auditory virtual environments. Proc. of ICAD 2001 - Seventh Meeting of the International Conference on Auditory Display, Espoo, 2001, 161–168.

[5] J. Blauert: Spatial hearing. The psychophysics of human sound localization. 2nd edition. MIT Press, Massachusetts, USA, 1997.

[6] A. Lindau, S. Weinzierl: Assessing the plausibility of virtual acoustic environments. Acta Acustica united with Acustica **98** (2012) 804–810.

[7] F. Brinkmann, A. Lindau, M. Vrhovnik, S. Weinzierl: Assessing the authenticity of individual dynamic binaural synthesis. Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, 2014, 62–68. http://dx.doi.org/10.14279/depositonce-11.

[8] T. Lokki, J. Pätynen, A. Kuusinen, H. Vertanen, S. Tervo: Concert hall acoustics assessment with individually elicited attributes. J. Acoust. Soc. Am. **130** (2011) 835–849.

[9] T. Lokki, J. Pätynen, A. Kuusinen, S. Tervo: Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. J. Acoust. Soc. Am. **132** (2012) 3148–3161.

[10] J. Berg, F. Rumsey: Spatial attribute identification and scaling by repertory grid technique and other methods. Proc. of the 16th International AES Conference: On Spatial Sound Reproduction, Rovaniemi, 1999, 51–66.

[11] K. Koivuniemi, N. Zacharov: Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. Proc. of the 111th AES Convention, New York, 2001, preprint no. 5424.

[12] G. Lorho: Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction. Proc. of the 119th AES Convention, New York, 2005, preprint no. 6629.

[13] E. M. Wenzel, S. H. Foster: Real-time digital synthesis of virtual acoustic environments. Proc. of the ACM Symposium on Interactive 3D Computer Graphics, 1990, 139–140.

[14] T. Lokki, H. Järveläinen: Subjective evaluation of auralization of physics-based room acoustics modeling. Proc. of ICAD, Espoo, 2001, 26–31.

[15] A. Silzle: Quality taxonomies for auditory virtual environments. Proc. of the 122nd AES Convention, Vienna, 2007, preprint no. 6993.

[16] H. Dichanz: Delphi-Befragung. – In: Qualitative Medienforschung. L. Mikos, C. Wegener (eds.). UVK, Konstanz, 2005, 297–303.

[17] ISO 5492: Sensory analysis – Vocabulary. Multilingual version. International Standardization Organization, Geneva, 2009.

[18] IEC 60050-801: International electrotechnical vocabulary. Chapter 801: Acoustics and electroacoustics. http://www.electropedia.org/iev/iev.nsf/index?openform&part=801.

[19] DIN 1320: Akustik. Begriffe. Beuth, Berlin, 2009.

[20] S. Bech, N. Zacharov: Perceptual audio evaluation: Theory, method and application. Wiley, Chichester, 2006.

[21] M. Gallagher, T. Hares, J. Spencer, C. Bradshaw, I. Webb: The nominal group technique: A research tool for general practice. Family Practice **10** (1993) 76–81.

[22] D. W. Stewart, P. N. Shamdasani, D. W. Rook: Focus groups: Theory and practice. Sage, Newbury Park, CA, 1990.

[23] A. Bogner, M. Leuthold: Was ich dazu noch sagen wollte... Die Moderation von Experten-Fokusgruppen. – In: Das Experteninterview. A. Bogner, B. Littig, W. Menz (eds.). VS Verlag für Sozalwissenschaften, Wiesbaden, 2005.

[24] G. Dürrenberger, J. Behringer: Die Fokusgruppe in Theorie und Anwendung. Rudolph-Sophien-Stift gGmbH, Stuttgart, 1999.

[25] A. Lindau et al.: Ein Fokusgruppenverfahren für die Entwicklung eines Vokabulars zur sensorischen Beurteilung virtueller akustischer Umgebungen. Fortschritte der Akustik: Proc. of the 40th DAGA, Oldenburg, 2014, 553–554.

[26] R. K. Hambleton: The next generation of the ITC test translation and adaptation guidelines. Europ. J. Psych. Ass., **17** (2001) 164–172.

[27] S. Ciba, A. Wlodarski, H.-J. Maempel: WhisPER – A new tool for performing listening tests. Proc. of the 126th AES Convention, Munich, 2009, preprint 7749, http://www. ak.tu–berlin.de/whisper, http://dx.doi.org/ 10.14279/depositonce–31.

[28] H.-J. Maempel, S. Weinzierl: Demands on measurement models for the perceptual qualities of virtual acoustic environments. Proc. of the 59th Open Seminar on Acoustics (OSA), Boszkowo (PL), 2012.

[29] A. Lindau: SAQI. Test manual. http://dx.doi.org/10.14279/depositonce-1, 2014.