Audio Engineering Society

Convention Paper 9830

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Identification of perceived sound quality attributes of 360° audiovisual recordings in VR using a Free Verbalization Method

Marta Olko[1], Dennis Dembeck[1], Yun-Han Wu[1], Andrea F. Genovese[1], and Agnieszka Roginska[1]

[1] *New York University*

Correspondence should be addressed to Marta Olko (olko.marta@gmail.com)

ABSTRACT

Recent advances in Virtual Reality (VR) technology have led to fast development of 3D binaural sound rendering methods that work in conjunction with head-tracking technology. As the production of 360° media grows, new subjective experiments that can appropriately evaluate and compare the sound quality of VR production tools are required. In this preliminary study, a Free Verbalization Method is employed to uncover auditory features within 360° audio-video experiences when paired with a 3-degrees-of-freedom head-tracking VR device and binaural sound over headphones. Subjects were first asked to identify perceived differences and similarities between different versions of audiovisual stimuli. In a second stage, subjects developed bipolar scales based on their verbal descriptions obtained previously. The verbal constructs created during the experiment, were then combined by the authors and experts into parent attributes by means of semantical analysis, similar to previous research on sound quality attributes. Analysis of the results indicated that there were three main groups of the sound quality attributes: attributes of sound quality describing the general impression of the 360° sound environment, attributes describing sound in relation to the head movement, and attributes describing audio and video congruency. Overall, the consistency of sound between different positions in 360° environment seems to create the new fundamental aspect of sound evaluation for VR and AR multimedia content.

1 Introduction

The need for spatial audio reproduction in novel contexts like VR applications or 360° degree video has been growing along with the recent developments in the gaming and multimedia industry. Delivering a truly immersive experience in VR systems requires high visual quality, intuitive user interaction, and authenticity of the perceived sound. New tools for 360° audio recording, post-production, rendering and playback in VR are facilitating the production pipeline available for artists, engineers, and customers. To appropriately evaluate and compare the quality of different VR

audio productions, comprehensive subjective assessment tests need to be employed.
Compared to static spatial audio experiences, such as binaural audio and surround sound systems, sound for head-tracked 360° experiences (as in VR) involves a different order of perceptual dimensions related to the possibility of shifting the point of listening perspective. The experience of sound in 360° is closer to a natural way of listening; thus, the list of factors that influence naturalness of the auditory sensation is assumed to be larger than in common playback systems. The conceptual differences between static channel-based audio and dynamic object audio may significantly influence how listeners evaluate the sound quality of traditional multichannel sound compared to the upcoming 360° audio formats. As a result, it may not be appropriate or sufficient to employ the same evaluation attributes used to rate static spatial experiences when judging dynamic audio presentations.
This paper illustrates a preliminary experiment aimed towards the investigation of appropriate attributes which comprehensively describe auditory perception in VR and are able to highlight its specific characteristics. Specifically, the focus is to study subjects' verbal elicitations and identifications of relevant auditory attributes within a dynamic binaural audio reproduction of a 3-degrees-of-freedom VR system. Discovered attributes can facilitate the future creation of judgment scales and assessment methods. Results and methods are compared with previous literature concerning elicitation of sound attributes.

2 Background

2.1 Elicitation methods for sound quality evaluation

In usual perceptual studies, before asking listeners to evaluate the spatial features of an audio signal, the attributes of sound quality need to be defined first by an experimenter. When a field becomes increasingly established, there is a higher possibility for the attributes to be validated, well-developed, and accurate in describing certain features. The experience gained from conducting experiments provides information to improve and refine the scales used, while listeners can sometimes be trained to focus on desired attributes of a given stimulus [1]. Unlike some well-established fields that are more consistent with their terminology, the words, and concepts used to describe sound are more likely to vary from individual to individual (Shaw and Gaines [2]). As a result, differences between verbal constructs provided by an experimenter and elicited constructs provided directly by the subjects may occur, especially with non-trained subjects who account for the majority of the population.
In several instances of studies on reproduced sound quality evaluation, subjects are asked to rate relatively vague pre-defined terms [3][4][5]. The major problem with provided attribute scales is that the subject is limited to respond in the ways predefined by the experimenter. In addition, some listeners might not be able to accurately map and connect their complex auditory perception using separable attributes. It is also hard for researchers to clarify which exact isolated attribute they want the listener to rate unless they provide extreme stimuli as an example. In the paper published by Colomes et al. in 2010 [6], the issue of unclear definitions in traditional single axis test methodologies, such as BS.1116 [7] and MUSHRA [8], is demonstrated. The paper aimed to validate the idea of sound families by comparing the results of a free categorization method and a multidimensional scaling method. The authors concluded that the use of sound families helps to minimize the bias created by the vague definition of sound attributes. Verbal elicitation tasks are designed to minimize the experimenter bias [9]. By encouraging the expression of personal sensations towards the stimulus under evaluation, the differences between the way each subject defines certain attributes can be put into context. In the paper published by Guastavino in 2004 [10], 26 subjects were presented with live recording materials in 1-D, 2-D (added speakers behind the listener) and 3-D (added speakers at height) configurations and were allowed to describe the perceptual impact of each stimulus freely. A semantical analysis, conducted by the researchers to all the phrasings generated by the free verbalization, served to group synonyms into several semantic themes. This method permits to gather information about how listeners subjectively perceive certain phenomena and describe them as spatial attributes using their own mental and verbal constructs and associations.

2.2 Spatial attributes in literature

Over the years, different approaches have been employed to identify the spatial attributes of sound in different reproduction systems. The attributes elicited were then used in subjective tests on the quality of various reproduction systems like surround, stereo headphones or Wave Field Synthesis. Although in the past there were several attempts to create a common lexicon of spatial sound attributes, in literature the terms used to describe spatial sound attributes are open to different kinds of interpretation. In general descriptive terms, Berg and Rumsey [11] indicated that spatial attributes stand for "the three-dimensional nature of sound sources and their environments". In order to satisfy two of the important requirements for psychological research, validity ("the test measures what it claims to measure") and reliability ("the repeatability of the measurement"), previous literature should be put in relevant context when making decisions on which spatial attribute to apply for rating a given setting.
In practice, the choice and definition of relevant attributes for judging spatial perception present a certain degree of variance according to the system being tested. In the paper written by Zacharov and Koivuniemi [12], source width and spatial impression are said to be the two spatial terms that repeatedly appeared in several spatial quality evaluation experiments done on mono, stereo, 5-channel and periphonic speaker systems. However, sometimes they were brought up in slightly different forms [13] [14]. In another paper published in 2010, Kamekawa and Marui [15] pointed out that the typical spatial attributes used in some of the multichannel surround sound system evaluation are localization (the seeming location of the sound sources), depth (the seeming spatial distance between the listener and the sound source), width (the width of the whole sound image), envelopment (the surround feeling from the side of and behind the listener) and presence (the feeling of "being there"). In the case of a stereo headphone system, Lorho indicated that five clusters of sound attributes were found after examining the dissimilarity between individual attributes elicited by subjects. The first category consists of spatial-related attributes such as width, reverb, and room size. The second cluster contains attributes concerning the timbral aspect of sound, e.g. clarity, brightness, and treble. The third cluster includes attributes related to various kinds of perceptual experiences, with three occurrences of the term noise. Moving on, the low-frequency emphasis is the core concept of the fourth cluster, which includes nine occurrences of the attribute bass. Finally, the fifth cluster is relatively close to the previous category and contains attributes of different sound natures [16]. In another paper, based on auditory virtual environment playback system, Silzle [17] stated that sound attributes elicited by listeners, which can also be called quality features, corresponded to quality elements on the service provider side. In addition, the evaluation results on quality features represent the quality of the listener's experience.
Differently, well-established standards for sound quality evaluation, such as IEC 60268 [18] and EBU 562-3 [19], defined three spatial attributes for sound quality evaluation. These are spaciousness (closed vs spacious), distance (distant vs near) and location of sources (unstable vs stable). Later versions of this standard also suggested three factors relating to spatial attributes: 1) image localization, which stands for how well-defined the spatial location of the reproduced sound sources is; 2) image stability, which depends on several factors - including pitch and loudness - and is also a function of the listener's position and head movement; 3) width homogeneity, which indicates if the stereophonic image is distributed uniformly between loudspeakers.
Previous research on elicitation of spatial sound attributes was performed using surround, binaural reproduction systems or virtual acoustic environments. This paper describes an experiment which is the first attempt to elicit attributes of spatial sound in the 360° audio format played back binaurally with head-tracking. The 360° format introduces new dimensions to the perception of the sound. The listener is provided with a full sphere in which object audio elements can be positioned and then delivered through speaker matrixes or binaurally through headphones. The signal delivered is commonly reproduced either within a spherical sound-field representation (Ambisonic) or as a speaker-independent sound object (Object-based audio). That is to say, any direction around the listener should be treated equally within an experimental investigation, as opposed to traditional

multichannel surround sound which is tied to discrete channel outputs and possesses the concept of a main "front" image [20].

2.3 Techniques used for audio production in VR application

Currently, there are two major flexible audio representations used for VR application — sound-field representations, also known as scene-based, and object-based representations. Susal et al. [21] described sound-field representations as "physically-based approaches that encode the incident wavefront at the listener location". Ambisonics is the common method for representing all the wavefronts in the spherical space around the listener [22]. In fact, it is relatively more similar to traditional channel-based technique compared to object-based representations, since the spatial information is directly encoded in the audio signal rather than stored as separated metadata. Scene-based audio is ideal for VR applications because of a more convenient process for acoustic capture, offline content creation, and post-production [23]. An ambisonic microphone is a tool that provides the ease of direct capturing of a spherical sound-field surrounding. New hybrid software tools combine the two capturing philosophies and allow artists and producers to design ambisonic scenes by encoding signals captured with spot microphones into ambisonic sound-fields. Those possibilities introduce new dimensions of modification of the sound scene and, as a result, might introduce new aspects of the perception of the sound quality.

3 Methodology

The purpose of this experiment was to extract a vocabulary of auditory differences and similarities in the stimuli presented to the subjects. Subjects composed their own attributes that were later gathered and reviewed by the researchers. In a previous study of related research [24], Berg and Rumsey generated spatial attributes by asking subjects to describe how one out of three stimuli was different from the other two, and how those two stimuli are similar to each other. Each subject was allowed to listen to every stimulus as many times as they wanted. The process was repeated until no more new attributes could be generated.

There are two major advantages of the triadic method. First, it prevents the researchers from asking the subjects for opposite expression directly. In other words, this method aims to guide the subjects to come up with phrases opposite in meaning naturally, by instructing them to describe the similarities and differences between the three stimuli [25]. However, an obvious disadvantage of grouping stimuli in triads is that the relatively small differences between two of the stimuli will be neglected if they are always presented with a distinct counterpart. Therefore, an alternative method of comparing the stimuli in pairs, which allows subjects to focus on small differences, is suggested.
An elicitation process was conducted where subjects generated their own bipolar constructs based on a triad of A/B pair comparisons of the recorded stimuli. In order to analyze this data, the verbal descriptors were grouped together in categories based on the Verbal Protocol Analysis and the semantical analysis. These groupings were then inspected for repeated or common verbal attributes used to identify the stimuli. Finding these common attributes was the desired goal of this study.

3.1 Subjects

Eighteen subjects with normal hearing, aged between 23 and 42 with a median age of 25, participated in the experiment. All subjects were expert listeners and students of New York University's Music Technology program. All of them listen to music actively several times a week. 11 subjects were native English speakers, 7 subjects were fluent in written and oral English as their second language.

3.2 Stimuli generation

Four individual musical performances were prepared for playback on a Samsung S7 smartphone and GearVR device. There were three versions/mixes of each video, with each version composed of a different audio mix while using the same visual. Each subject was presented with two out of the four video stimuli chosen by randomization. The stimuli were presented in three separate pairs to elicit differences and similarities between each version. Stimuli generation for the subjects to reflect upon was divided into three separate stages: recording, mixing, and encoding.

	Stimulus	Ensemble
1.	Choir	16 Vocalist
2.	Rock Band	2 Vocalist, Guitar, Bass, Drum Kit
3.	Solo Cello	Cello
4.	Percussion Ensemble	Marimba, Vibraphone, Udu

Table 1: Performance recordings for each of the stimuli

3.2.1 Recording

The recording process took place in the Dolan Studio at New York University. The 360° visuals were captured using a Giroptic 360° camera. The audio was recorded using both soundfield and object-based capturing techniques. To capture the soundfield recordings, the Sennheiser AMBEO VR microphone was used for all of the stimuli recordings, except for the percussion trio recording. In this case, the double MSZ technique was used (see [26]). All soundfield devices were placed in the center of the room, surrounded by the performance ensembles. The 360° camera was also positioned in the perspective of the soundfield recording devices. Various spot microphones (object-audio elements later encoded in Ambisonics by the renderer) were placed on individual musicians to capture the performance from a close perspective. The recordings are listed in Table 1.

3.2.2 Mixing

The three audio mixes for each video stimulus was rendered in ProTools HD using the Facebook Spatial Workstation-OSX v2.0 Beta2 plugin and were as follows:

- soundfield microphone only
- spot microphones and artificial reverb
- soundfield microphone and spot microphones

Two different reverberations were applied to the stimuli audio mixes by randomization. The first one utilized the Facebook Spatial Workstation plugin by activating the "Room" parameter. Through this parameter, room acoustic modeling is available to synthesize artificial reverberation in three-dimensional space with the ability to adjust the reverberation mix level and reflection order. The second reverberation method utilized was a stereo convolution reverberation, which was applied during the encoding stage.
The loudness of each stimulus was measured using the Facebook 360 Loudness meter. All stimuli were normalized to an integrated measurement of -15 LUFS.

3.2.3 Encoding

The 360° videos and eight channel spatial audio mixes were rendered and synchronized using the Facebook 360 Spatial Workstation Encoder. In order for the subjects to compare mixes in an A/B format, the three different mixes for each stimulus were rendered in pairs (ab, bc, ac). Subjects were then able to compare two different mixes within one video file.

3.3 Reproduction

The video stimuli were uploaded to the Facebook 360 application and played back on the Samsung GearVR using Sennheiser HD 650 headphones. The Facebook 360 application allowed for 360° visual playback and auditory binaural rendering of the eight-channel encoded mixes. The subjective testing took place in an acoustically treated research lab at New York University. Subjects were equipped with the GearVR while seated in a chair that allowed full 360° rotation. The playback of the video stimuli was streamed from a saved library within the Facebook 360° application. The loudness level of the playback was adjusted on the GearVR by the subjects at the beginning of each test to suit their loudness preferences and kept consistent throughout the experiment session.

3.4 Elicitation process

The goal of the elicitation process was to acquire verbal descriptors from the subjects personal vocabulary. The four video stimuli, each having three different mix versions presented in pairs, were randomly assigned to the subject. The stimuli versions, labeled A, B, C, were then uploaded to the

Facebook 360° application on the Samsung Gear VR. Subjects were first allowed to navigate the stimuli to experience all of the given A/B pairs. The duration of each stimulus averaged 30 seconds. Subjects viewed the pairings in order and were allowed to review and repeat the playback of each mix pair as desired. Participants were then instructed to listen for similarities and differences of the auditory experience in each version and subsequently instructed to write down the perceived experiential similarities and differences in their own format.

Once the subjects had finished viewing the video stimuli, they began dissecting verbal descriptors from their own documentation. They were asked to read all of their notes and create bipolar scales from each of the descriptive words they used. Subjects were encouraged to search for the words which are opposite in meaning and the most precise in the description of their perception. This created a list of bipolar constructs that were then gathered and processed by the researcher.

4 Analysis and discussion

4.1 Constructs elicited

The total number of constructs elicited by all subjects was 231. The minimum number of constructs generated by a single subject was 7, while the maximum number of constructs generated by a single subject was 20. The median value of the number of constructs elicited by subjects was 12.5.

4.2 Verbal Protocol Analysis

The first step in the analysis of results was to reduce redundancy of the obtained verbal descriptors when the same identical words were used by several subjects. After removing repeated instances of grading scales, 166 bipolar constructs were left. Verbal Protocol Analysis (VPA), proposed in the paper of Samoylenko [27], was employed in the analysis of results. In that paper, verbal descriptors describing timbre were analyzed on three levels: logical sense, stimulus relatedness, and semantic aspects. A similar analysis was used in this experiment to divide obtained descriptors into more general classes. The third level of analysis, which focuses on the semantical aspects of verbal units, was employed in this study. Verbal descriptors were categorized into attitudinal and descriptive. Attitudinal descriptors express the emotional relation to the sound (emv) and naturalness (ntl). Descriptive constructs were divided into those describing auditory modality only (UMD) or multiple sensory modalities (PMD). From all of the scales obtained during the experiment, 9% was attitudinal, and 91% was descriptive. Attitudinal descriptors were related to the preference, overall evaluation of the stimuli, and naturalness of the sound. Noticeably, there were several constructs describing naturalness of the sound change during head movement. From the descriptive features, 82% were unimodal and 18% were polymodal. Unimodal verbal descriptors were describing characteristics of auditory modality only. These constructs, which were a majority of all the obtained descriptors, were related to the general perception of the sound in the 360° scene.

It should be noted that grouping of the descriptors is a difficult task. Categorization based on semantical analysis is largely biased by the interpretation of the researcher. In order to reduce the bias, the categorization of the descriptors was conducted by researchers and a panel of experts. A panel of five experts, including some of the authors, was formed to read each of the scales carefully and to group them based on similar words usage, meaning, and comments of the subjects. The created groups of attributes were compared with the attribute definitions from previous studies effectuated on the spatial sound.

During the test, subjects were encouraged to comment on each of the scales to allow more precise interpretation of them. The attributes that defined the grouping of the descriptors during the analysis were as follows (the reference source for each attribute is shown in brackets): *Clarity* [28], *Externalization* [29], *Spatial impression* [28], *Depth perspective* [15], *Timbre* [28], *Sound image width* [15], *Location accuracy* [28], *Sound balance* [28], *Punch* [30], *Immersion/Presence* [10], and *Freedom from noise* [28]. The rate of appearance of the verbal descriptors assigned to each attribute is shown in Figure 1. There were no differences in the distribution of verbal descriptors elicited between native and non-native English speaker subjects.

Two categories of verbal descriptors related to polymodal sensations were found: audio-video congru-

ency and perception of sound during head movement. Figure 2 shows the number of unimodal and polymodal descriptors elicited by subjects. The number of polymodal descriptors is relatively small in comparison to unimodal. Audio-video congruency was described by subjects in four different aspects: sense of space (if the sense of space in sound was matching the space in the image), localization (if the localization of the sound sources was matching the image), distance (if the distance of the sound sources from listener was matching the video), and time synchronization between sound and image.

The other category of polymodal descriptors was related to the sound change during head movement. This category relates to the initial motivations behind the paper, to find new descriptive attributes for subjective perception of dynamic audio/video experiences in VR. The groups of scales identified during the analysis are reported in Table 2.

Verbal descriptors indicate that changes in the sound during head movements are perceived separately to the overall sound impression and might be a crucial element in the evaluation of the quality of sound in 360°. The results of the experiment are not robust enough to provide definitions to the new attributes with clear confidence. More research is required to validate the perception of sound during head movement.

5 Conclusion and future work

This preliminary study was the first attempt to investigate sound quality attributes in 360°. Verbal descriptors elicited by subjects and analyzed using the Verbal Protocol Analysis, and were divided into three main groups: attributes of sound quality describing the general impression of the sound environment, attributes describing sound in relation to the head movement, and attributes describing audio and video congruency.

Verbal descriptors identifying attributes of sound quality, relating to the general impression of the sound environment, were found to be the same as in the similar research on static spatial sound reproduction. Head-tracking allowed listeners to compare the change of sound from different positional

Attribute	**Scale**
Change of sound during head movement	How noticeable is the horizontal and frontal change in response to head movement
Sound balance during head movement	The signal is attenuated/not attenuated during head movement The amplitude change during head movement is/is not expected
Localization during head movement	Sound sources are easy/hard to localize during head movement Localization seems correct/incorrect during head movement
Width during head-movement	Width of the sound image is steady/changing during head movement
Depth during head movement	Depth or distance of the sources from the listener is steady/changing during head movement
Externalization during head movement	The changes in sound during head movement are happening inside/outside of the head
Clarity during head movement	Sound sources are present/absent when turning head toward the source Sound sources are focused/unfocused when turning head toward the source

Table 2: Attributes elicited during experiment describing sound in relation to the head movement

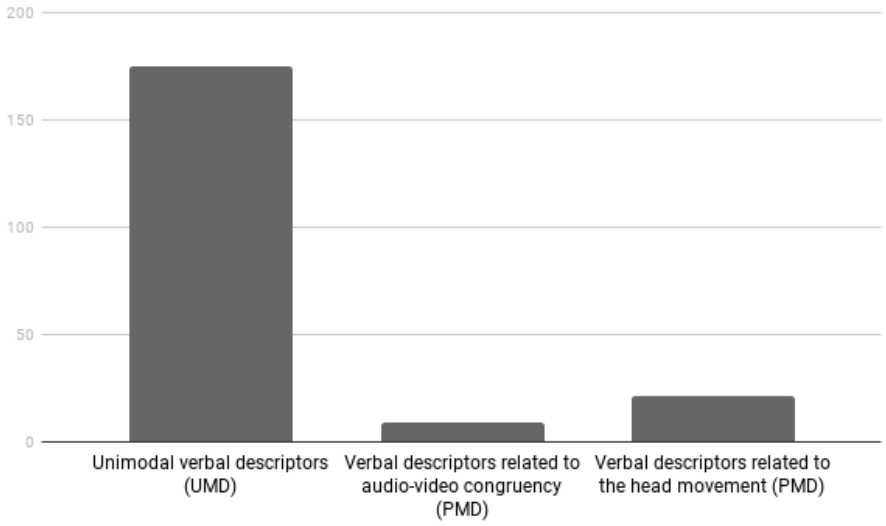

Fig. 1: Rate of appearance of spatial sound attributes

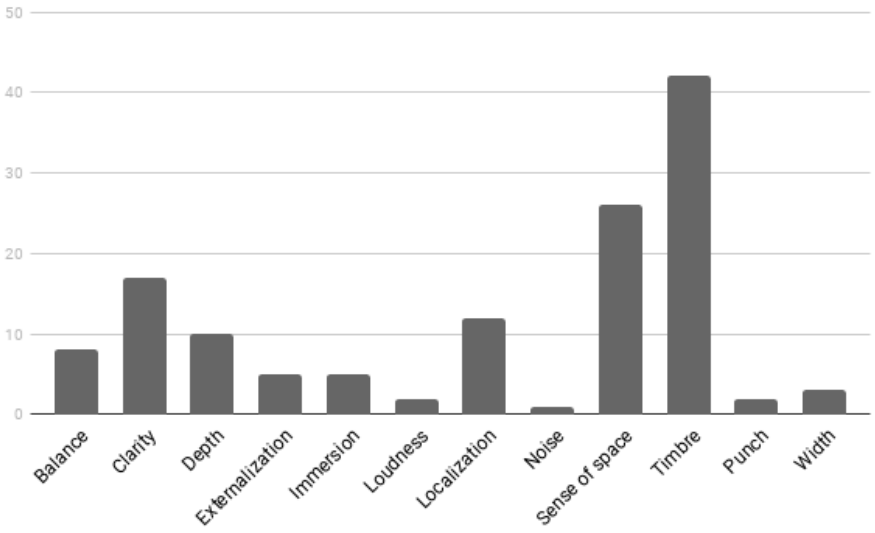

Fig. 2: Number of verbal descriptors elicited during experiment

perspectives. As a results, inconsistencies between head perspectives were noted by subjects. The study highlighted a number of verbal descriptors, describing the relation between sound and head movement in various aspects. The elicited scales were related to attributes stability and change during head movement. Overall, the consistency of sound between different positions in 360° environment seems to create the new fundamental aspect of sound evaluation for these type of experiences, relevant for upcoming VR and AR multimedia content.

The main limitation of this study is that the conducted experiment only comprised an elicitation stage. Due to constraints, subjects were not asked to use the elicited scales for numerical qualitative rating of the stimuli, which would allow a more robust statistical analysis of verbal descriptors and more precise identification of the attributes. Next studies aimed towards defining attributes of 360° sound should involve methods which allow statistical validation of obtained attributes, such as the Repertory Grid Technique. Other constraints including hardware limitations, low quality of videos, same recording space used in experiments, might have limited the number of attributes elicited in this study. More diversified stimuli might facilitate obtaining a bigger variety of verbal descriptors.

Nevertheless, this exploratory study should be regarded as a first attempt to explore the issue and to propose an experimental strategy to be applied to the new multimedia VR/AR devices that employ spatial audio. The experiment revealed also that evaluation of 360° sound format is much more time-consuming than evaluation of stereo or surround formats because of the infinite number of listener positions inside the scene. That should be taken into consideration in future test designs.

6 Acknowledgments

The authors would like to thank all of the subjects who participated in the experiment and the external members of the panel.

References

[1] Bech, S., "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *J. Audio Eng. Soc*, 40(7/8), pp. 590–610, 1992.

[2] Shaw, M. L. and Gaines, B. R., "Comparing conceptual structures: consensus, conflict, correspondence and contrast," *Knowledge Acquisition*, 1(4), pp. 341 – 363, 1989.

[3] Toole, F. E., "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc*, 33(1/2), pp. 2–32, 1985.

[4] Woszczyk, W., Bech, S., and Hansen, V., "Interaction Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes," in *Audio Engineering Society Convention 99*, 1995.

[5] Rumsey, F., "Controlled Subjective Assessments of 2-to-5-Channel Surround Sound Processing Algorithms," in *Audio Engineering Society Convention 104*, 1998.

[6] Colomes, C., Le Bagousse, S., and Paquier, M., "Families of Sound Attributes for Assessment of Spatial Audio," in *Audio Engineering Society Convention 129*, 2010.

[7] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, ITU-R Recommendation BS.1116-1, 1997.

[8] *Method for the subjective assessment of intermediate quality level of coding systems*, ITU-R Recommendation BS.1534-1, 2003.

[9] Kelly, G., *The Psychology of Personal Constructs*, Routledge, 1991.

[10] Guastavino, C. and Katz, B. F., "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, 116(2), pp. 1105–1115, 2004.

[11] Berg, J. and Rumsey, F., "In Search of the Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis of Scaled Verbal Descriptors," in *Audio Engineering Society Convention 108*, 2000.

[12] Zacharov, N. and Koivuniemi, K., "Unravelling the Perception of Spatial Sound Reproduction: Analysis; External Preference Mapping," in *Audio Engineering Society Convention 111*, 2001.

[13] Berg, J., *Systematic evaluation of perceived spatial quality in surround sound systems*, Ph.D. thesis, Luleå tekniska universitet, 2002.

[14] Mason, R. and Rumsey, F., "An Assessment of the Spatial Performance of Virtual Home Theatre Algorithms by Subjective and Objective Methods," in *Audio Engineering Society Convention 108*, 2000.

[15] Kamekawa, T. and Marui, A., "Developing Common Attributes to Evaluate Spatial Impression of Surround Sound Recording," in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, 2010.

[16] Lorho, G., "Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction," in *Audio Engineering Society Convention 119*, 2005.

[17] Silzle, A., "Quality Taxonomies for Auditory Virtual Environments," in *Audio Engineering Society Convention 122*, 2007.

[18] *Sound System Equipment–– Part 13: Listening Tests on Loudspeakers*, IEC 60268, 1997.

[19] *Subjective Assessment of Sound Quality*, EBU Rec. 562-3, 1990.

[20] Horsburgh, A. J., McAlpine, K. B., and Clark, D. F., "A Perspective on the Adoption of Ambisonics for Games," in *Audio Engineering Society Conference: 41st International Conference: Audio for Games*, 2011.

[21] Susal, J., Krauss, K., Tsingos, N., and Altman, M., "Immersive Audio for VR," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[22] Furness, R. K., "Ambisonics-An Overview," in *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*, 1990.

[23] Shivappa, S., Morrell, M., Sen, D., Peters, N., and Salehin, S. M. A., "Efficient, Compelling, and Immersive VR Audio Experience Using Scene Based Audio/Higher Order Ambisonics," in *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[24] Berg, J. and Rumsey, F., "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique," *J. Audio Eng. Soc*, 54(5), pp. 365–379, 2006.

[25] Choisel, S. and Wickelmaier, F., "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound," *J. Audio Eng. Soc*, 54(9), pp. 815–826, 2006.

[26] Geluso, P., "Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays," in *Audio Engineering Society Convention 132*, 2012.

[27] Samoylenko, E., "Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres," *International Journal of Psychology*, 31(6), pp. 255–278, 1996.

[28] *Assessment methods for the subjective evaluation of the quality of sound programme material – Music*, EBU Tech. 3286–E, 1997.

[29] Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M., "On the Externalization of Auditory Images," *Presence: Teleoper. Virtual Environ.*, 1(2), pp. 251–257, 1992.

[30] Fenton, S. and Wakefield, J., "Objective Profiling of Perceived Punch and Clarity in Produced Music," in *Audio Engineering Society Convention 132*, 2012.