Audio Engineering Society

Convention Paper 10594

Presented at the 152nd Convention
2022 May, In-Person and Online

Sensory evaluation of spatially dynamic audiovisual sound scenes: a review

Pia Nancy Porysek Moreta[1,2], Søren Bech[1,2], Jon Francombe[1], Jan Østergaard[2], Steven van de Par[3], and Neofytos Kaplanis[1]

[1]*Bang & Olufsen a/s, 7600 Struer, Denmark*
[2]*Aalborg University, Department of Electronic Systems, 9220 Aalborg, Denmark*
[2]*Carl von Ossietzky University, Department of Medical Physics and Acoustics, 26129 Oldenburg, Germany*

Correspondence should be addressed to Pia Nancy Porysek Moreta (pipo@bang-olufsen.dk)

ABSTRACT

Spatial audio systems are capable of rendering moving sound sources. This is becoming more common in commercial and domestic environments, driven by interest in spatial audio and virtual reality, and underpinned by object-based audio delivery formats. When consulting standard methods for sound quality evaluation in terms of spatially dynamic sound scenes (SDSS), challenges emerge indicating the methods are inadequate for the application. This article presents an overview of state-of-the-art sound quality evaluation methods, particularly focusing on their appropriateness for evaluation of SDSS. Limitations of current methodologies are discussed, and research of temporal evaluation methodologies used in other sensory sciences are reviewed for their potential applicability to audio quality assessment.

1 Introduction

Everyday life is full of movement. Whether passing the main street of a city on a busy day or taking a relaxed stroll in a secluded park, we always encounter a stream of sounds creating a dynamic, ever-changing experience. Triggered by physical cues perceived through our ears, our brain creates auditory 3D images of the world and movements around us. These auditory images are continuously updated and linked to visual cues to gain detailed spatial information about our surroundings. Advanced spatial audio technologies aim to reproduce this enveloping real-life sensation, subsequently referred to as spatially dynamic sound scenes (SDSS), while also allowing new creative and artistic ways to render the movement of sound in complex 3D audio scenes. In this context, it is not surprising that the reproduction of 3D sound plays a particularly important role in the entertainment sector, including cinemas, home cinemas, and gaming.

Reproduction of audio that involves perceivable movement and, therefore, a form of envelopment in the sound environment, is not a new invention. Due to human perception being more sensitive in the horizontal than the vertical plane [1, 2], earlier spatial audio reproduction systems (such as 5.1 surround sound) focused on optimization of sound in the horizontal plane. The

field has continued to develop, and immersive audio, 3D-Sound, and next-generation ("Next Gen") spatial audio technology have all become buzzwords in recent years. The primary development of these systems is the addition of an acoustical vertical dimension to provide a more immersive perception. They also allow mutlidimensional audio-visual experiences; for example, virtual reality with full-sphere video and six degrees of freedom movement.

There are two main types of 3D spatial audio reproduction system: binaural and soundfield. Binaural rendering is mostly used with headphones and applied in gaming and other personal entertainment (e.g. augmented and virtual reality in commonly known products such as Oculus Rift and PlayStation VR). In terms of sound reproduction quality of 3D audio, binaural rendering faces various challenges such as optimal reproduction of externalization, especially under dynamic conditions [3]. Due to reliance on accurate simulation of localization cues, influenced by the individual nature of head-related transfer functions [4] and binaural room responses [5], 3D audio experiences may vary for different users with headphones.

While binaural reproduction recreates natural 3D sound at the listener's ears, soundfield-based systems establish spatial audio effects by recreating soundfields corresponding to an acoustic scene with multiple loudspeakers. The sound experience is not restricted to a single listener or spatial point [6] and the spatial effect can be experienced by multiple listeners at a time. Therefore, such systems are useful for scenarios such as cinema, home theatre, and multi-player gaming.

Along with possible options like channel-based and model-based (Higher Order Ambisonics, Wave-Field Synthesis), the object-based audio workflow aims to create an immersive sound experience, while allowing new artistic, personalized and interactive options [7–9]. Gasull Ruiz et al. [10] describe object-based audio as means to overcome the drawbacks of the commonly used channel-based approach, where an audio mix is specially produced for a certain loudspeaker setup. In the object-based audio approach, the spatial audio scene is saved in form of audio objects instead of fixing the specific speaker signals in production. Audio objects link a sound source with a set of associated information (metadata) like position, gain, delay, or type of the virtual sound source, which can change as a function of time and therefore represent the movement of the audio object. A complex audio scene is established by combining several so-called virtual sound sources.

Despite the growing interest in spatial audio experiences, standardized procedures and metrics for determining the subjective quality of spatially dynamic sound scenes (SDSS) are still missing. The available standardized methods for evaluating sound quality face various challenges with this type of acoustic environment. Typically, methods for sound quality evaluation obtain a single value to describe the entire measured experience. Those methods, therefore, do not consider the changes in perception that occur with the dynamic movement of sound. Fortunately, there are potentially useful methods for considering temporal changes in perception from other fields of sensory science. Experimental studies [11, 12] show that descriptive methods from sensory food quality evaluation can be implemented in the sensory evaluation of sound. As the flavor palette and the intensity of perceived flavors in our mouth change over the process of chewing [13, p.179], methods for evaluating food quality over time have been developed. These include time-intensity methods [14] as well as more descriptive methods [15, 16]. Recent literature shows initial experiments combining temporal food quality evaluation with attributes of sound evaluation in SDSS [17, 18].

Moreover, SDSS often combine audio and visual aspects. Research around multisensory perception shows that there is a need to investigate the importance and interaction of visual aspects when evaluating audio quality [19, 20], especially under the consideration of key use-case scenarios of SDSS.

In this article, current audio quality evaluation methodologies are reviewed in order to determine whether they are sufficient for evaluating the perceptual quality of SDSS. In Section 2 the focus is on the definition of SDSS and the perception of movement. This is followed in Section 3 by an introduction to state-of-the-art perceptual sound quality evaluation methods, their limits regarding spatially dynamic stimuli, and a review of methods that might tackle those challenges. In Section 4 multi-sensory research in terms of audio-visual cross-influences is summarized. Finally, in Section 5, the main conclusions emerging from the literature are summarized and future work is suggested.

2 Spatially dynamic sound scenes and perception of moving audio sources

It is commonly known that there is a difference between a physically measurable acoustic event and how the same event is perceived by a listener [21, p.1 ff.]. A simple example of this concept is the following: the changing level of an approaching car is physically measurable. However, if the pedestrian completely blocks out the outside noise with headphones, their auditory perception does not correspond to the physical changes of sound in their environment. It seems, thus, logical, to differentiate between a physical "sound event" and a perceived "auditory event". Those different events are, of course, related to and associated with each other, while also being able to be classified in relation to each other [21, p.3]. Investigating the connection between sound and auditory events is one of the underlying questions of sensory evaluation. This process is complicated by the indirect nature of assessment of perceptual features, as the experimenter is not able to directly assess the perception of a listener and, therefore, relies on a form of description of their experience [21, p.6 ff.].

This underlying assumption is also of importance when perceptually evaluating SDSS. The term dynamic, by definition, implies something that is "characterized by constant change, activity, or progress" (Oxford Languages). A SDSS implies, consequently, an acoustic scene that is characterized by some spatial change. In terms of perceptual evaluation, the first question would, therefore, be: when and how do we perceive spatial changes, and, hence, auditory movement?

"Spatial hearing" describes the way sound sources are perceptually localized and the perception of spatial attributes at the listener's position. Spatial cues (e.g., interaural time difference and interaural level difference) change when the sound sources move in relation to the listener [22], the listener moves in relation to the sound source [23], or both move in relation to each other [24]. Carlile and Leung [25] identified location, trajectory, and velocity of motion relative to that of the listener's head as the main perceptual factors for auditory motion.

When reviewing research on the topic of perception of auditory motion Carlile and Leung [25] summarized scientific advances and concluded, that a lot of scientific findings were influenced by the limitation of stimulus paradigms. These limitations were, among other things, caused by restricted technical apparatus (e.g., limited ability to reproduce moving sound sources without physically moving a loudspeaker). Consequently, research has been primarily based on simple linear or rotational movement rather than more ecologically valid source movement [26].

3 Sensory evaluation of spatially dynamic sound scenes

In this section some basic definitions in terms of sound quality evaluation are given, before focusing on the limitations of modern quality evaluation methods regarding the evaluation of SDSS. Thereafter, temporal quality evaluation methods from other fields and their application in sound evaluation are reviewed.

3.1 An introduction in sound quality evaluation

There has been extensive research on the topic of sound quality assessment since the early days of sound creation, processing, and reproduction. In the past, quality assessment has proven to be a valuable and essential tool for the development and further improvement of new technology (e.g., enhancement algorithms, reproduction systems, etc.). As numerous as the areas of application for quality evaluation are, so is the number of definitions.

Letowski [27] describes the general term of quality as a reference to "a character of an object or a merit of its superiority". Assuming that a physically measurable sound creates an auditory image of the sound in the listener's mind, sound quality is expressed as satisfaction or dissatisfaction with that image. The auditory image is formed by a sum of auditory sensations and can be assessed by comparison to an auditory image of another stimulus or a concept in the listener's mind based on expectation and memory. Since this definition was proposed, there have been multiple specifications of this underlying concept. When considering sound quality, Raake and Wierstorf [28] similarly describe the consideration of human listeners who use received acoustic signals to extract features and assign meaning. Moreover, it is stated that quality assessment depends on the context it is used in. Raake and Wierstorf consequently divide quality into an audio-technology and an engineering context. The engineering context is characterized by instrumental measurements (i.e., pressure levels, frequency responses, decay times, etc.), while

the audio-technology context, which this article will focus on, covers "*any kind of processing between the generation of a sound by its initial source(s) and its recording via different audio-technology systems along the chain up to the listener*". Therefore, the audio-technology context does not only cover technical but also features of human perception. This characterization resembles the often used division into subjective and objective evaluation methods. While subjective evaluation results according to this definition are obtained by directly asking the listener about their perception, objective evaluation methods aim to predict listener perception by either using physical measurements or based on perception models [29, p.11]. As criticized by Zacharov [30, p.61,62] the term "subjective" in this context is misleading, as it implies listening tests being purely "*based on, or influenced by, personal feelings, tastes, or opinions*". It is further argued, that some types of listening experiments can be objective. This will be explained in detail with the help of the so-called filter model [29, 30].

The filter model, as described by Bech and Zacharov [29, p.39 ff.], shows how sound as a physical phenomenon creates auditory images in the listener's mind, on which they can form an opinion of the sound under evaluation. The model distinguishes the perception process into three domains. The physical domain is characterized by a measurable physical stimulus, thus, assessable by objective evaluation methods e.g., sound pressure levels or frequency analysis. The processing along the auditory pathway can be described with two filters. The first is a filter where the audio signal is transformed to the perceptual domain. Here, attributes of perception, based on sensory modalities, are assessed (e.g., loudness). These types of attributes can either be called perceptual or descriptive and form an objective quantification of sensory strength of the perceived stimulus [30]. Next, the perceived stimulus is influenced by the second filter made up of cognitive factors (e.g., mood, context, emotion, background, expectation), which forms the affective (or hedonic) domain and ultimately the preference. In this model, it is only this domain that is associated with personal opinion and, consequently, connected with subjectivity. Lastly, according to this model, it is possible to assess basic audio quality (BAQ), by adding attributes of the perceptual domain with a certain weighing according to the situation under evaluation, assessing BAQ without the influence of cognitive factors and, thus, assessing a form of comparable objective listening test evaluation. Therefore, the term "sensory" evaluation when speaking of listening tests, in general, is preferred.

There is a significant focus on the expansion of instrumental audio quality algorithms to predict human perception and reduce expenses in time and cost of the evaluation processes. Still, listening tests with human listeners are the gold standard. One particular downside of perceptually-motivated perceptual models is that their accuracy is not known when applied to novel, complex audio technology [29, p.11].

3.2 The limits of modern quality evaluation

There are various ITU standard recommendations regarding the assessment of quality and methodologies for assessing audio quality (e.g., ITU-R BS.1116 [31], ITU-R BS.1534 [32]). When using state-of-the-art methodologies of sound quality or experience assessment on spatially dynamic sound fields, different challenges can be identified.

Recommendations for state-of-the-art quality evaluation suggest that the stimulus should be of low complexity to enable an easy understanding of the attribute under evaluation and prevent confusion [33]. Stimuli are, thus, often selected on the basis of perceptual constancy; dynamic aspects are considered to be a distraction and, therefore, not included in stimulus sets. Hence, it is assumed that perceptual attributes under evaluation are stationary enough that the resulting mean value over time can be considered an accurate representation of the experience [29]. A SDSS implies an acoustic scene that is characterized by some spatial change. While some aspects of this kind of sound scene might remain constant over time (e.g., timbre), some other aspects are by definition dynamic (e.g., the spatial characteristics of the sound source when movement is rendered). Thus, by evaluating the dynamic nature of a sound (e.g., the movement of a sound object), some perceptual characteristic describing SDSS are assumed not to be constant over time and due to their changing nature, high in complexity. As a consequence, it can be concluded, that a single value result of quality quantification does not correctly describe the actual experience of the listener.

Another closely connected point of discussion comes into question when looking at the length of the stimulus. Due to the auditory working memory that is used

to distinguish perceptual differences having a range of a few seconds at most [34], it is generally advised to keep the stimulus under test as short as possible. This is to "*avoid fatiguing of listeners, increased robustness and stability of listener responses, and to reduce the total duration of the listening test*", as stated in the ITU recommendation BS.1534 for subjective assessments of intermediate quality level of audio systems [32]. Further, the recommendation points out that a signal exceeding the length of 12 seconds might lead to bias in the result, inducing primacy and recency effects. Both effects describe the likelihood of recalling information depending on their position of presentation, as the first and last item in a memory task can usually be memorized best. Particularly the recency effect means that quality impairments towards the end of a stimulus are weighted more strongly in the overall judgment, and duration neglect suggests that the peak intensity of a negative stimulus is a much greater factor than the duration of the negative stimulus in the rating [35]. The standard methods also allow participants to isolate a specific loop of the content enabling the evaluation of only isolated loops of the signal that differ greatly in spectral and temporal features across the signal. Moreover, there is no guarantee that different participants isolate the same loops.

In summary, a longer stimulus combined with an retrospective evaluation promotes bias in the results, which leads to the result not being representative for the experience. Depending on SDSS under evaluation (e.g., the evaluation of a looming motion) and considering the perception of movement depending on time, it is valid to say that the traditional evaluation methods are insufficient for evaluation of SDSS.

3.3 Temporal quality evaluation

In the previous section, some problems with using existing methods for evaluation of SDSS were outlined. Despite this, such methods are still commonly used for evaluation in situations where there are moving sound sources; for example, to evaluate algorithms for spatial audio systems [36] or compare different rendering systems for spatial sound [37]. Only very few perceptual methods have been developed or applied that try to account for SDSS evaluation and the challenges mentioned previously.

In a broader picture of assessing momentary quality of experience Weiss et al. [38] give two approaches for overcoming the drawbacks of retrospective evaluation: dividing the content into shorter, separately evaluated segments or a continuous evaluation. Besides being a time consuming procedure for long test items, segmentation of the content (e.g., as used in the double stimulus continuous quality scale [39] method), might not always be appropriate for evaluating some SDSS; for example, if a long item was required to allow ecological validity and/or narrative context, or if certain dynamic aspects of the scene (such as movement paths) would be affected by the segmentation.

Continuous evaluation has only rarely been performed for evaluation of audio stimuli, but has been used in other sensory sciences such as evaluation of picture and food quality. This type of evaluation has been shown to reduce the recency effect bias on retrospective judgements [35].

Single stimulus continuous quality evaluation (SSCQE) [39–41] was first introduced in the evaluation of television pictures then implemented in speech quality evaluation. It allows assessors to continuously rate quality using a slider mechanism with an associated interval scale. While SSCQE has proven to be a valuable method of rating quality variations over time [42] there are two notable criticisms[38]. Firstly, the task of constantly evaluating and operating the slider takes attention away from the main task of evaluating the stimulus [43]. Secondly, the results can lose accuracy when aggregated across experiment participants because of their different reaction times to quality changes [44]. Additionally, Kokotopoulos [45] showed that there are differences in reaction time depending on quality changes from good to bad or vise versa. In an attempt to mitigate against some of these drawbacks, other types of rating devices for continuous tracking of quality changes have been investigated [38]. For example, Jumisko-Pyykkö at al. [46] used a simplified continuous assessment (pressing a button whenever perceived quality degradation occurred) to investigate the perceived unacceptability of instantaneous audio, visual and audiovisual errors.

Borowiak et al. [47] proposed another method for continuous quality evaluation for long duration audiovisual content. Instead of setting scores on predefined rating scales, their approach was to let the assessors adjust the quality when degradation occurred by means of an adjustment device. According to the authors, low distraction was achieved by the intuitiveness of the methodology, which led to higher focus on the content.

In food quality evaluation continuous time intensity scaling has long been a subject of research, as temporal methods give more detailed information about aspects like flavor and texture changes [13]. In time-intensity (TI) evaluation, perceived sensational changes of one descriptive attribute (e.g., sweetness) are continuously monitored [14]. As the process of evaluating only one attribute at a time quickly gets time consuming, the method of temporal dominance of sensation (TDS) [15] was developed. Here, a list of up to ten attributes is presented with the task to select and score the dominant attribute. Moreover, participants are instructed to change the selected attribute if the perceived dominance changes. Each run results in a set of scored sensory attributes quoted at different times along the food tasting. Another method that was proposed to investigate the temporal sensory characteristics of food in more detail is temporal check-all-that-apply (TCATA) [16]. The TCATA method, as proposed by Castura et al. [16] is an extension of the check-all-that-apply (CATA) method [48] and facilitates dynamic assessment of the multidimensional sensory properties of a product as they evolving over time. In the more basic CATA method, rapid product profiles are obtained; consumers are presented with a list of attributes and asked to indicate which words or phrases appropriately describe their experiences. In the TCATA method, the selection and deselection of attributes continues over time. Changes are logged, so the assessor is able to evaluate the sensory change in products.

3.4 Temporal evaluation methods applied to audio quality

In an experimental study, Gil et al. [17, 18] combined the food quality assessment method TCATA [16] with attributes for evaluation of spatial audio assessment [49]. In a first pilot experiment, Gil et al. [17] used two previously selected attributes ("dynamic" and "enveloping") to describe two sound scenes, where one of the sound scenes contained a source that was moving around the listener and the other one was considered static. The result of this preliminary experiment indicated that both tested attributes were selected when movement was involved in the sound scene. This observation was made for both stereo and 40-loudspeaker layouts. In a second experiment, Gil at al. [18] used the same methodology on a larger scale (four attributes) and were, overall, able to measure perceptual responses varying over time that were consistent with physical changes, while using different loudspeaker systems and following different patterns of movement around a listener.

4 Visual cues in sensory evaluation of sound

SDSS are used most commonly in conjunction with moving images (for example, in object-based audio with cinema, or 6DOF rendering in virtual reality). Traditionally, sensory research has shied away from investigating multi-sensory integration [50]. This is at least in part due to the common belief that uni-sensory processing is first executed by dedicated neural pathways, assumed to be largely independent and hierarchically organized [51], and the process of sensory integration taking only place at later stages [52, 53]. However, recent findings show that the connectivity between early uni-sensory areas in the brain is closer than previously thought [54]. One example for auditory-visual interactions is the ventriloquist effect, where a listener couples a sound to a visual stimulus rather than the actual sound source (e.g., when watching a movie and sound source seems to be the actor's lips rather then the loudspeaker setup) [55]. Another auditory-visual effect is the McGurk effect [56] describing the influence of a visually formed syllable on the auditory perception of another syllable. Additionally, visual cues have been shown: (1) to be beneficial for speech intelligibility in noisy environments [57]; (2) to affect percepts such as annoyance [58]; and to be involved in localization perception [24]. Finally, there has also been evidence of visual cues influencing the perception of spatial audio. Woodcock et al. [19] investigated the influence of visual stimuli on a set of attributes relevant to the perception of spatial audio; they showed that the presence or absence of visuals has a significant effect on *realism, sense of space* and *spatial clarity*.

5 Conclusions

Advances in spatial audio technology enable an improved and more complex rendering of sound source movement. The created sound scenes fall under the category of spatially dynamic sound scenes, perceptually characterized by changing spatial cues and perceptual factors that can only be identified over time (such as trajectory, velocity, and motion in relation to the listener's head). The goal of this article was to determine

whether current audio quality evaluation methodologies are sufficient for evaluating the perceptual quality of SDSS.

The literature review showed that currently available, standardized methodologies for sound quality evaluation are unfit for evaluation of stimuli that are dynamic. In Section 3.2, limits of existing methods were identified. Standardized methods for audio quality evaluation have relied on perceptually constant stimuli. This is problematic for SDSS for which it is necessary to evaluate dynamic properties. One problem, for example, is the difficulty for participants of attaching a single value to a dynamic scene. Such retrospective ratings are also subject to a number of possible biases.

Consequently, methodologies and new metrics should be explored. Using temporal evaluation methods to record continuous observations of quality changes is one of the possible solutions to the challenges. Reviewing existing methodologies for temporal quality evaluation in Section 3.3 showed that there has been relatively little research on temporal quality evaluation of multimedia content and systems, and even less in the audio domain. In contrast to that, research of food evaluation has established procedures for temporal evaluation of perceptual attributes. Experimental work using methodologies from food science applied to spatial sound perception has showed good results but should be investigated further.

5.1 Future work

Multiple continuous methodologies that would possibly be suited to evaluating temporal quality changes in audio quality evaluation were reviewed. Future work could explore how well they work for evaluation of perceptual sound quality in practice.

Additionally, one of the main goals in future work should be the determination of potential descriptive attributes of spatially dynamic sound scenes. The attributes chosen by Gil et al. [17, 18] were based on available literature [49] and experimental selection without an elicitation process, as the available list of attributes for spatial sound quality assessment is currently lacking attributes defining perceptual spatial movement [49, 59]. Therefore, it is possible that more attributes are relevant for SDSS and could be elicited by combining TCATA with a preliminary elicitation process (e.g., by using the CATA method on multiple possible attributes first). A further step of evaluation, as suggested by Gil et al. [17], is a scaling operation (e.g., time-intensity methodology for sensory evaluation [14]) to investigate the perceptual intensity of attributes over time.

Moreover, Gil et al. [17]'s preliminary results indicate the possibility of identifying perceptual attributes describing dynamic changes in SPSS even when only used with a stereo loudspeaker setup. As domestic-environments are one of the main use-case scenarios for SDSS, the approach could further be tested in more ecologically valid environments and with "real" stimuli. Additionally, such a setup would permit a implementation of a matching audio-visual cues (e.g., a television screen). The literature reviewed in Section 4 suggests that the visual component is an additional important aspect. Therefore, future research should consider audio and visual modalities alongside each other.

6 Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 956369.

References

[1] Saberi, K., Dostal, L., Sadralodabai, T., and Perrott, D., "Minimum audible angles for horizontal, vertical, and oblique orientations: lateral and dorsal planes," *Acta Acustica United with Acustica*, 75(1), pp. 57–61, 1991.

[2] Grantham, D. W., Hornsby, B. W., and Erpenbeck, E. A., "Auditory spatial resolution in horizontal, vertical, and diagonal planes," *The Journal of the Acoustical Society of America*, 114(2), pp. 1009–1022, 2003, doi:10.1121/1.1590970.

[3] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopco, N., "Sound externalization: a review of recent research," *Trends in Hearing*, 24, 2020, doi:10.1177/2331216520948390.

[4] Baumgartner, R., Reed, D. K., Tóth, B., Best, V., Majdak, P., Colburn, H. S., and Shinn-Cunningham, B. G., "Asymmetries in behavioral and neural responses to spectral cues demonstrate the generality of auditory looming bias," *Proceedings of the National Academy of Sciences*, 114, pp. 9743–9748, 2017, doi:10.1073/pnas.1703247114.

[5] Hassager, H. G., Gran, F., and Dau, T., "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment," *The Journal of the Acoustical Society of America*, 139 5, pp. 2992–3000, 2016, doi:10.1121/1.4950847.

[6] Zhang, W., Samarasinghe, P. N., Chen, H., and Abhayapala, T. D., "Surround by sound: a review of spatial audio recording and reproduction," *Applied Sciences*, 7(5), pp. 532–550, 2017, doi:10.3390/app7050532.

[7] Shirley, B. and Oldfield, R., "Clean audio for TV broadcast: an object-based approach for hearing-impaired viewers," *Journal of the Audio Engineering Society*, 63, pp. 245–256, 2015, doi:10.17743/jaes.2015.00017.

[8] Walton, T., Evans, M., Kirk, D., and Melchior, F., "Exploring object-based content adaptation for mobile audio," *Personal and Ubiquitous Computing*, 22(4), pp. 707–720, 2018, doi:10.1007/s00779-018-1125-6.

[9] Ward, L. and Shirley, B. G., "Personalization in object-based audio for accessibility: a review of advancements for hearing impaired listeners," *Journal of the Audio Engineering Society*, 67(7/8), pp. 584–597, 2019, doi:10.17743/jaes.2019.0021.

[10] Gasull Ruiz, A., Sladeczek, C., and Sporer, T., "A description of an object-based audio workflow for media productions," in *Audio Engineering Society 57th International Conference: The Future of Audio Entertainment Technology—Cinema, Television and the Internet*, Hollywood, USA, 2015.

[11] Kaplanis, N., Bech, S., Lokki, T., Waterschoot, T., and Jensen, S., "A novel method for perceptual assessment of small room acoustics using rapid sensory analysis," *Journal of the Acoustical Society of America*, 139, pp. 2202–2202, 2016, doi:10.1121/1.4950565.

[12] Kaplanis, N., Bech, S., Sakari, T., Pätynen, J., Lokki, T., van Waterschoot, T., and Jensen, S., "A rapid sensory analysis method for perceptual assessment of automotive audio," *Journal of the Audio Engineering Society*, 65(1/2), pp. 130–142, 2017, doi:10.17743/jaes.2016.0056.

[13] Lawless, H. T. and Heymann, H., *Sensory evaluation of food: principles and practices*, New York: Springer, 2010.

[14] Cliff, M. and Heymann, H., "Development and use of time-intensity methodology for sensory evaluation: a review," *Food Research International*, 26(5), pp. 375–385, 1993, doi:10.1016/0963-9969(93)90081-S.

[15] Pineau, N., Schlich, P., Cordelle, S., Mathonnière, C., Issanchou, S., Imbert, A., Rogeaux, M., Etiévant, P., and Köster, E., "Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time–intensity," *Food Quality and Preference*, 20(6), pp. 450–455, 2009.

[16] Castura, J. C., Antúnez, L., Giménez, A., and Ares, G., "Temporal Check-All-That-Apply (TCATA): a novel dynamic method for characterizing products," *Food Quality and Preference*, 47, pp. 79–90, 2016, doi:10.1016/j.foodqual.2015.06.017.

[17] Gil, J., Bech, S., and Christensen, F., "Evaluation of the perceived sound quality in spatially dynamic sound environments: a literature study," in *Audio Engineering Society 152nd Convention (accepted)*, The Hague, Netherlands, 2022.

[18] Gil, J., Bech, S., and Christensen, F., "Temporal evaluation of sound quality using TCATA," in *Audio Engineering Society 152nd Convention (accepted)*, The Hague, Netherlands, 2022.

[19] Woodcock, J., Davies, W. J., and Cox, T. J., "Influence of visual stimuli on perceptual attributes of spatial audio," *Journal of the Audio Engineering Society*, 67(7/8), pp. 557–567, 2019, doi:10.17743/jaes.2019.0019.

[20] Iwamiya, S.-i., "Interactions between auditory and visual processing when listening to music in an audiovisual context: 1. Matching 2. Audio quality," *Psychomusicology: A Journal of Research in Music Cognition*, 13(1-2), pp. 133–153, 1994, doi:10.1037/h0094098.

[21] Blauert, J., *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997, doi:10.7551/mitpress/6391.001.0001.

[22] Ballas, J. A., Brock, D., Stroup, J., and Fouad, H., "The effect of auditory rendering on perceived movement: loudspeaker density and HRTF," in *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, 2001, doi:1853/50653.

[23] Perrett, S. and Noble, W., "The contribu-

tion of head motion cues to localization of low-pass noise," *Attention, Perception, & Psychophysics*, 59, pp. 1018–1026, 1997, doi:10.3758/BF03205517.

[24] Yost, W. A., Zhong, X., and Najam, A., "Judging sound rotation when listeners and sounds rotate: sound source localization is a multisystem process," *The Journal of the Acoustical Society of America*, 138(5), pp. 3293–3310, 2015, doi:10.1121/1.4935091.

[25] Carlile, S. and Leung, J., "The perception of auditory motion," *Trends in Hearing*, 20, 2016, doi:10.1177/2331216516644254.

[26] Perrott, D. R. and Strybel, T. Z., "Some observations regarding motion without direction," in R. H. Gilkey and T. R. Anderson, editors, *Binaural and spatial hearing in real and virtual environments*, pp. 275–294, New Jersey: Lawrence Erlbaum Associates, 1997.

[27] Letowski, T., "Sound quality assessment: concepts and criteria," in *Audio Engineering Society 87th Convention*, New York, USA, 1989.

[28] Raake, A. and Wierstorf, H., "Binaural evaluation of sound quality and quality of experience," in J. Blauert and J. Braasch, editors, *The technology of binaural understanding*, pp. 393–434, Cham: Springer, 2020, doi:10.1007/978-3-030-00386-9_14.

[29] Bech, S. and Zacharov, N., *Perceptual audio evaluation—theory, method and application*, John Wiley & Sons, 2007.

[30] Zacharov, N., *Sensory evaluation of sound*, CRC Press, 2019.

[31] ITU-R, "Methods for the subjective assessment of small impairments in audio systems," Rec. BS.1116, International Telecommunication Union, 2000.

[32] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," Rec. BS.1534, International Telecommunication Union, 2000.

[33] Rumsey, F., "Subjective assessment of the spatial attributes of reproduced sound," in *Audio Engineering Society 15th International Conference: Audio, Acoustics, & Small Spaces*, Audio Engineering Society, Copenhagen, Denmark, 1998.

[34] Cowan, N., "On short and long auditory stores," *Psychological Bulletin*, 96(2), pp. 341–370, 1984, doi:10.1037/0033-2909.96.2.341.

[35] Hands, D. S. and Avons, S., "Recency and duration neglect in subjective assessment of television picture quality," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 15(6), pp. 639–657, 2001, doi:10.1002/acp.731.

[36] Narbutt, M., Skoglund, J., Allen, A., Chinen, M., Barry, D., and Hines, A., "AMBIQUAL: towards a quality metric for headphone rendered compressed ambisonic spatial audio," *Applied Sciences*, 10(9), pp. 3188–3208, 2020, doi:10.3390/app10093188.

[37] Thery, D. and Katz, B. F., "Auditory perception stability evaluation comparing binaural and loudspeaker Ambisonic presentations of dynamic virtual concert auralizations," *The Journal of the Acoustical Society of America*, 149(1), pp. 246–258, 2021, doi:10.1121/10.0002942.

[38] Weiss, B., Guse, D., Möller, S., Raake, A., Borowiak, A., and Reiter, U., "Temporal development of quality of experience," in S. Möller and A. Raake, editors, *Quality of experience*, pp. 133–147, Cham: Springer, 2014, doi:10.1007/978-3-319-02681-7_10.

[39] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Rec. BT.500, International Telecommunication Union, 2002.

[40] Hamberg, R. and de Ridder, H., "Continuous assessment of perceptual image quality," *Journal of the Optical Society of America A*, 12(12), pp. 2573–2577, 1995, doi:10.1364/JOSAA.12.002573.

[41] ITU-T, "Continuous evaluation of time varying speech quality," Rec. P.880, International Telecommunication Union, 2004.

[42] Köster, F., Guse, D., Wältermann, M., and Möller, S., "Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech," in *Fortschritte der Akustik, DAGA*, Nuremburg, Germany, 2015.

[43] Bouch, A. and Sasse, M. A., "Case for predictable media quality in networked multimedia applications," in *Proc. SPIE vol. 3969: Multimedia Com-*

puting and Networking 2000, pp. 188–195, 1999, doi:10.1117/12.373521.

[44] Pinson, M. H. and Wolf, S., "Comparing subjective video quality testing methodologies," in *Proc. SPIE vol. 5150: Visual Communications and Image Processing 2003*, pp. 573–582, 2003, doi:10.1117/12.509908.

[45] Kokotopoulos, A., "Subjective assessment of a multimedia system for distance learning," in *European Conference on Multimedia Applications, Services, and Techniques*, pp. 395–408, Milan, Italy, 1997, doi:10.1007/BFb0037365.

[46] Jumisko-Pyykkö, S., MV, V. K., and Korhonen, J., "Unacceptability of instantaneous errors in mobile television: from annoying audio to video," in *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Dervices*, pp. 1–8, Espoo, Finland, 2006, doi:10.1145/1152215.1152217.

[47] Borowiak, A., Reiter, U., and Svensson, U. P., "Quality evaluation of long duration audiovisual content," in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 337–341, Las Vegas, USA, 2012, doi:10.1109/CCNC.2012.6181113.

[48] Henrique, N., Deliza, R., and Rosenthal, A., "Consumer Sensory Characterization of Cooked Ham Using the Check-All-That-Apply (CATA) Methodology," *Food Engineering Reviews*, 7, pp. 265–273, 2015, doi:10.1007/s12393-014-9094-7.

[49] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A spatial audio quality inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014, doi:10.3813/AAA.918778.

[50] Jones, E. and Powell, T., "An anatomical study of converging sensory pathways within the cerebral cortex of the monkey," *Brain*, 93(4), pp. 793–820, 1970, doi:10.1093/brain/93.4.793.

[51] King, A. J., Hammond-Kenny, A., and Nodal, F. R., "Multisensory processing in the auditory cortex," in A. Lee, M. Wallace, A. Coffin, A. Popper, and R. Fay, editors, *Multisensory Processes*, pp. 105–133, Springer, 2019, doi:10.1007/978-3-030-10461-0_6.

[52] Lee, A. K. and Wallace, M. T., "Visual influence on auditory perception," in A. Lee, M. Wallace, A. Coffin, A. Popper, and R. Fay, editors, *Multisensory Processes*, pp. 1–8, Springer, 2019, doi:10.1007/978-3-030-10461-0_1.

[53] Alais, D. and Burr, D., "Cue combination within a Bayesian framework," in A. Lee, M. Wallace, A. Coffin, A. Popper, and R. Fay, editors, *Multisensory Processes*, pp. 9–31, Springer, 2019, doi:10.1007/978-3-030-10461-0_2.

[54] Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., and Matusz, P. J., "The multisensory function of the human primary visual cortex," *Neuropsychologia*, 83, pp. 161–169, 2016, doi:10.1016/j.neuropsychologia.2015.08.011.

[55] Alais, D. and Burr, D., "The ventriloquist effect results from near-optimal bimodal integration," *Current Biology*, 14(3), pp. 257–262, 2004, doi:10.1016/j.cub.2004.01.029.

[56] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices," *Nature*, 264(5588), pp. 746–748, 1976, doi:10.1038/264746a0.

[57] Bernstein, J. G. and Grant, K. W., "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," *The Journal of the Acoustical Society of America*, 125(5), pp. 3358–3372, 2009, doi:10.1121/1.3110132.

[58] Szychowska, M., Hafke-Dys, H., Preis, A., Kociński, J., and Kleka, P., "The influence of audio-visual interactions on the annoyance ratings for wind turbines," *Applied Acoustics*, 129, pp. 190–203, 2018, doi:10.1016/j.apacoust.2017.08.003.

[59] Zacharov, N., Pedersen, T., and Pike, C., "A common lexicon for spatial sound quality assessment—latest developments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, IEEE, Lisbon, Portugal, 2016, doi:10.1109/QoMEX.2016.7498967.