PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Methodologies for Assessment of Speech and Audio for Optimized Quality of Experience

Daniel P Darcy[1], Alex Brandmeyer[2], Rich Graff[3], Nathan Swedlow[4], and Poppy AC Crum[5]

[1] Dolby Laboratories, San Francisco, CA

ABSTRACT

Optimized Quality of Experience (QoE) in audio depends on preservation of creators' artistic intentions, from dialog intelligibility to spatial arrangement of sound. Numerous challenges exist in measuring audio sensation over new or advanced systems such as mobile audio, spatial audio playback systems, and augmented or virtual reality environments. The impact of these systems on dialog, which is critical to the success of creative storytelling, requires strategies to measure and optimize characteristics such as clarity, naturalness, and intelligibility of speech. We describe a range of experimental approaches to assess and drive audio algorithm development. First, we recognize the importance of and expand on classical methodologies such as forced-choice signal detection, methods of adjustment, and magnitude estimation to assess or validate system performance, and to understand end user preferences and the perceptual metrics underlying them. Second, we have developed a framework for objectively capturing spatial and timbral properties of audio used to optimize spatial audio systems along salient perceptual parameters most critical for high QoE. Finally, we utilize physiological measurements such as pupillometry, EEG, thermal imaging, and electrodermal activity to make direct estimations of the relationship between auditory perception and biological and neurophysiological states, such as cognitive load during speech comprehension tasks.

1. INTRODUCTION

Goals for successful audio delivery include preserving the content creators' artistic decisions and intent during both the origination and presentation of media, and the maximizing of a consumer's Quality of Experience (QoE) across a range of diverse endpoints, from cinemas and high-end home theaters to mobile devices and headphones. This is challenging due to the high variability of presentation hardware capacity and, frequently, the requirement that media be routed to systems with different fundamental configurations. For instance, multi-channel immersive audio may be played over a high-fidelity multichannel system, or downmixed for 2-channel transmission over small devices like mobile loudspeakers, or downmixed and binauralized for headphone playback, an increasingly popular consumption use case.

This review summarizes some of the methods we employ to assess audio QoE, system performance, and the preservation of creative intent. Because experimental objectives will often not include a known reference condition against which direct comparisons of quality can be made, precise measuring and full visualization data that pertains to different qualitative or quantitative attributes of sound is valuable and often critical for understanding system and QoE attributes. When measured and analyzed successfully, these data can lead to insights that help tune system performance and audio delivery to maximize listener QoE.

2. METHODOLOGIES FOR THE ASSESSMENT OF AUDIO QoE

2.1 Traditional Methodologies

Traditional psychophysical methodologies for assessing audio QoE include longstanding experimental techniques such as forced-choice detection, limit detection to determine just noticeable

[1] dan.darcy@dolby.com
[2] alex.brandmeyer@dolby.com
[3] rich.graff@dolby.com
[4] nathan.swedlow@dolby.com
[5] poppy.crum@dolby.com

differences (JNDs), mean opinion scores (MoS), methods of adjustment (MoA), and magnitude estimation (ME). These well-characterized methods can be powerful approaches to generating robust measurements of attributes contributing to audio QoE.

Forced-choice detection between systems, where a listener is presented stimuli as a two-alternative or two-interval selection between which a choice is selected, can be the most precise approach to signal detection of differences between systems (Green, 1966). When paired systems are presented in the context of what system is preferred, a determination can be made that one or the other system elicits a higher audio QoE in the listener for the given audio signal and conditions. This subjective evaluation of preference is a sensitive and direct measurement of audio QoE that can be associated with objective queries corresponding to audio attributes of interest, or those that are hypothesized to play a role in enhancing audio QoE, discussed further below.

For measurements of one or more parameters that need to be assessed in order to identify the preferred value providing the highest QoE, experiments using a method of adjustment (Stevens, 1975) framework are useful. Often, these parameters fall along a unitary dimension such as amplitude or distance, and this type of experiment allows a listener to map their preferred conditions for parameters presented on a continuum. Data are collected to determine optimal values, as well as upper and lower values representing acceptable boundaries. Since experiments that directionally adjust along a parameter space are prone to hysteresis effects, stimuli are randomly presented in ascending and descending order, defining the extent and boundaries of hysteresis.

Magnitude estimation (ME) is a useful technique for characterizing relationships among the QoEs of multiple systems that may be difficult to compare using other approaches such as forced-choice tasks, either because the systems are extremely different such that comparisons are confounded by lack of overlapping QoE, or often because the number of systems is too large to usefully collect pairwise comparison data. In ME experiments, a listener assigns values or ratios corresponding to perceived audio QoE for various systems (Stevens, 1956). Using this approach, an understanding of the landscape of QoEs across multiple systems and parameters can be developed, with the added advantage that additional systems can be subsequently measured and interpolated in relationship to any of the system measurements previously collected.

2.2 Extensions to Traditional Methodologies

We extend the utility of traditional methodologies to complement and enrich visualization of the psychophysical data sets we collect, provide additional insights into a given system's performance, and ultimately use these combinations of techniques to optimize audio QoE.

Forced-choice data can be complemented with measurements of confidence or conviction in the choice task being performed. We typically collected ask a listener to rate their degree of confidence in their decision on a 0 (no confidence) to 3 (high confidence) scale. These values can be used along with forced-choice or rank ordered visualization to indicate the overall confidence in system performance selections, or to weight the forced-choice data with confidence measurements (Figure 1, numerical values on individual bar graphs). Further, these data can be used to validate that systems overwhelming preferred also tend to be favored with high confidence. Interesting insights can occur when pairing these data, such as when two systems are preferred about equally. If the confidence data reflects that these choices are being made with low confidence, it may be the case that there is little differentiating the two. On the other hand, if they are equally preferred with high confidence, this may reflect strong listener preferences that are breaking down across different bases for preference judgement, something that may warrant further investigation.

Another method we typically leverage is to collect forced-choice data capturing metrics with objective qualities, such as spatial properties of height, width, spaciousness, externalization, etc., in addition to forced-choice data of subjective preference (Figure 1, left to right columns). The objective qualities are chosen as the basis for hypotheses about what audio characteristics may contribute to QoE. All these data are collected as a panel of questions pertaining to the same trial stimuli being presented for a forced-choice task and are analyzed for correlations as discussed below.

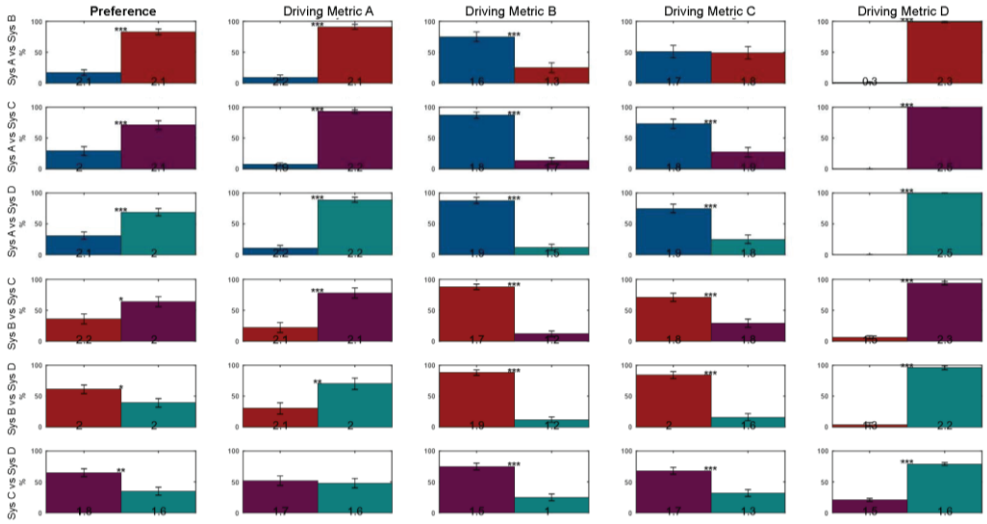

Figure 1 – An example forced-choice experiment with driving metrics and confidence ratings. Data are from 4 systems compared pairwise. Questions about subjective preference, and hypothesized driving metrics that may relate to preference, are collected. Additionally, confidence ratings on a 0-3 scale are collected and used to label the forced-choice outcome at the base of each bar plot.

We often apply rank ordering models to forced-choice data when more than two systems are compared. One common approach for this is the Bradley-Terry (BT) probability model (Bradley, 1952) that predicts the outcome of a paired comparison between two given systems for which pairwise data has been measured (Figure 2, left five columns). When complete forced-choice comparison data are sampled as pairwise comparisons, as we routinely target when testing fewer than around seven systems, the probability model is equivalent to the forced-choice data, but the models can also be fit using incomplete pairwise sampling. Other probability models similar to BT are also common, such as Thurstone Case V, widely used in imaging studies.

In analysis, per-listener correlation between preference and the hypothesized driving metrics leading to preference provides a distribution that gives insight and validation of those relationships (Figure 2, right). For example, if spaciousness is considered to be a primary contributor to QoE, then the expectation may be that each time a system is judged as more spacious, it is the preferred system. This will be revealed as a distribution tending towards a correlation of 1 or 100%. Similarly, if the correlation distribution between driving metrics and preference centers around equal chance at 0.5, it may be inferred that the metric in question is unrelated to preference or that other differences in the systems play a dominant role in judgement of preference.

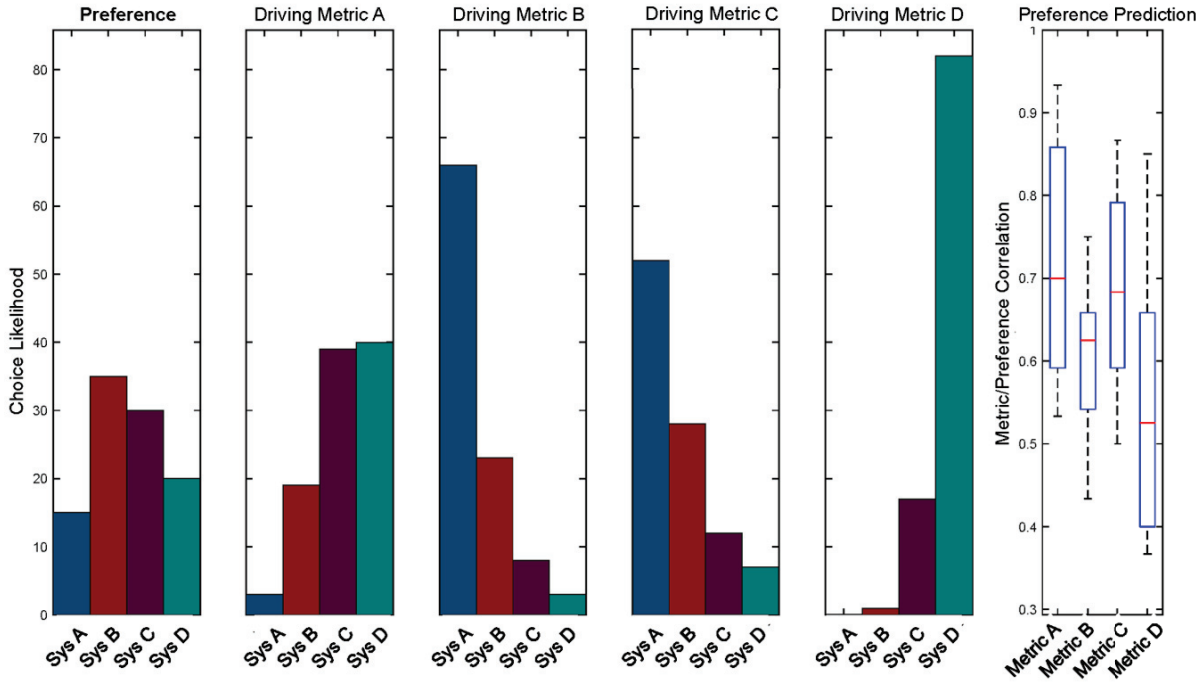

Figure 2 – The colored bar graphs in the left five columns represent Bradley-Terry probability modelling of the forced-choice data in Figure 1 to obtain a rank ordered visualization of preference and driving metrics for the four systems. The 'preference prediction' panel at right describes the correlation between the distributions driving metrics and preference. This graph plots the distribution of chance that a system chosen for a given driving metric would also be the preferred system.

2.3 Assessment of Spatial Attributes in Audio QoE

Spatial assessment of audio and its contributions to QoE are particularly relevant today as the prevalence of multichannel and object-based audio playback and delivery systems increases, in addition to the evolution of techniques to effectively binauralize sound for immersive audio over headphones and dual-channel devices. Although traditional psychophysics have powerful applications, increasingly the spatial assessment of sound requires innovative approaches to measure the variability and nuances of a listener's unique perception of spatial attributes. We use an objective capture framework called ADA ("ADaptive Audio") to characterize spatial parameters of sound and derive performance criteria that can be optimized to deliver maximum QoE (Darcy, 2016).

The ADA system consists of conducting listening experiments using an app running on a tablet in a room with visual landmarks, typically concentric rings placed at 18" intervals that indicate distance from the listener. The listening subject is situated in the center of the rings holding the tablet. Images on the tablet correspond to the room the listener is in, including a representation of the visual landmarks (Figure 3). Spatial audio stimuli are presented, over headphones or multi-channel loudspeaker systems, and the listener is instructed to visually draw his or her impression of sound within the room-representation in the tablet. Tools are provided in the app to draw an unrestricted representation of sound over time and in 3D space using both orthographic and perspective views. Additionally, a score is collected for the timbre of the spatialized sound relative to the non-spatial mono downmix version of the content in order to measure effects of spatialization that degrade the spectral naturalness of sound (Figure 4, left).

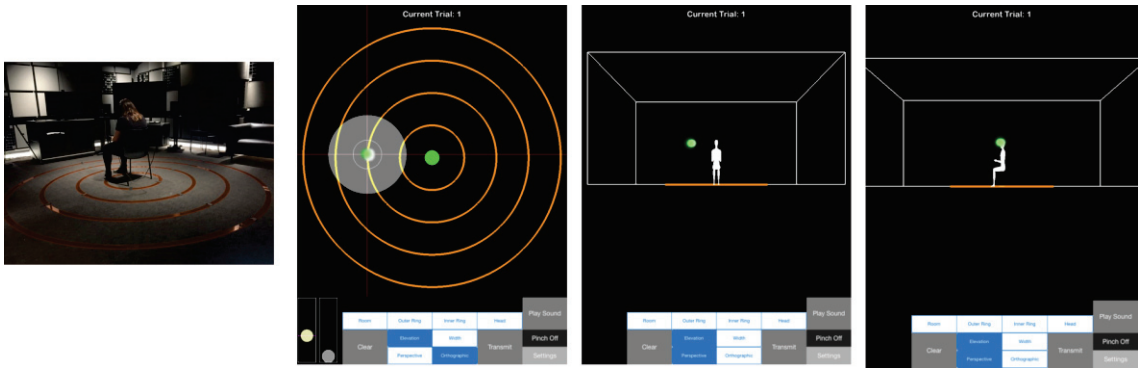

Figure 3 – The ADA listening room configuration and app. Listeners are seated in the center of visual landmarks presented as concentric rings (left). Sound is presented over headphones or loudspeakers and the listener enters their spatial percept of sound on a tablet app that has corresponding orthographic and 3D perspective views of the room.

Using this system, high-dimensional data in the form of the freely entered drawings, along with the temporal dynamics of entry, are captured as the listener's perceptual response to the presented audio. These data can be fit, for instance with ellipsoids, to reduce the dimensionality of the full input of a complex cloud of points in space. These fitted data can be further analyzed to derive properties like location, and other spatial attributes of sound such as width, dispersion and diffuseness. Trajectories can also be captured and analyzed with spline fits.

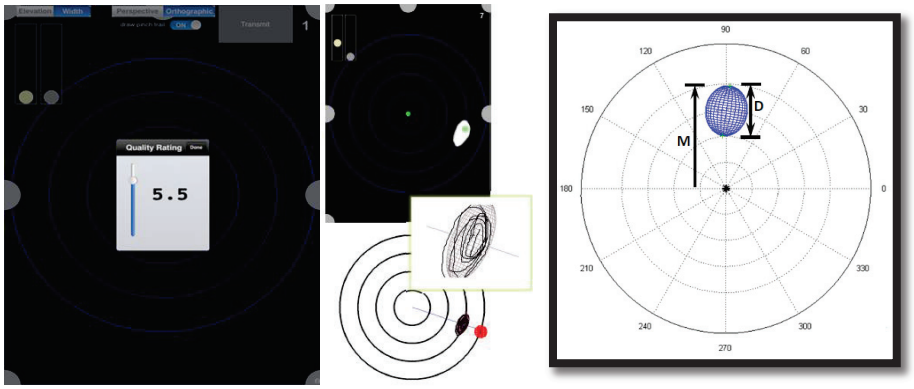

Figure 4 – In the ADA framework, listeners also report a rating for spectral naturalness of spatialized sound compared to a non-spatial mono reference (left). Analysis reduces the dimensionality of the user input by, for example, fitting ellipsoids to the spatial data (center). These objects can be furthered analyzed to derive metrics of interest such as maximum forward externalization (M) or dispersion (D) (right).

These captured metrics are then related using an optimization function designed with specific objectives in mind. For example, for high QoE of a binaural virtualizer designed to reproduce object-based audio played over headphones, a function would be created that increases with localization accuracy, maximal or realistic externalization outside the head, and minimal spatial smearing effects that create unwanted perceived dispersion. This function could penalize other undesirable effects such as front-back reversals or spatial instability that often lead to a low QoE of headphone virtualizers. Visualization and analysis of listener data using this platform demonstrates a high degree of heterogeneity of listener experiences (Figure 5) which can be related via such optimization functions to deliver high audio QoE for a given individual.

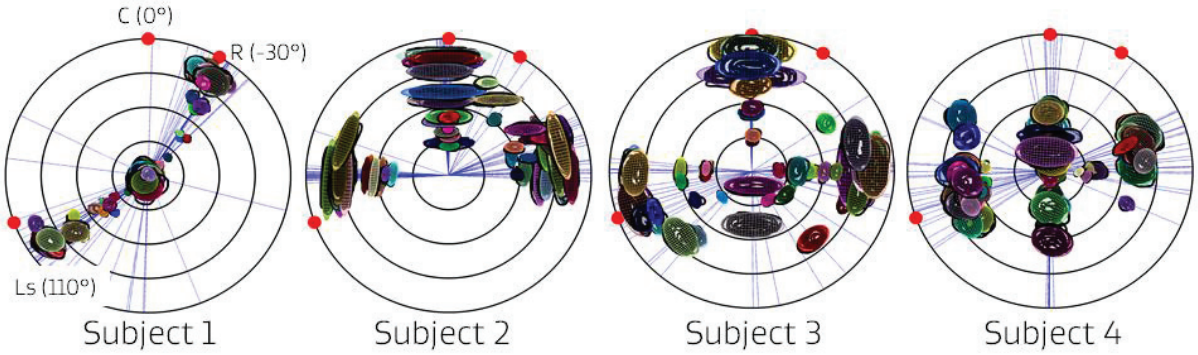

Figure 5 – An experiment illustrating listener data for four individuals listening to a headphone virtualizer that is externalizing sound at three azimuthal positions: 0, -30 and 110 degrees. Data from multiple trials are overlaid. Notable features include high user variability of localization and externalization, good right and left surround localization but a complete lack of center channel externalization for Subject 1, and numerous front-back reversals in Subjects 3 and 4.

2.4 Physiological Assessment

Task-based experiments that involve conscious effort to measure subjective or objective responses to audio QoE share the drawback that they are not reading out the same perceptual and physiological state of the listener that they would experience when listening without regard to task. The discontinuity created by needing to make and report a conscious decision may be critically altering the QoE of the listener.

Physiological measurements directly capture the state of a listener in ways that can provide insight into his or her experience that is complementary to psychophysical task-based experiments. Because audio QoE consists of both bottom-up perceptual mechanisms as well as high-level cognitive and emotional concepts, the implicit readout of physical and neural states promises to contribute greatly to our understanding of QoE and to contextualize what is presently known based on psychophysics.

Physiological measurements can be made from the Central Nervous System, for example using electroencephalogram (EEG) recordings, and the Autonomic Nervous System, where electrodermal (EDR) or electrocardiogram (ECG) readings can indicate states of heightened physical excitement or anticipation. We use both of these approaches to capture physiological and cognitive reactions to audio presentation, leveraging the fact that many of these biological signatures are universal and invariant of a listener's individual background or understanding of the experimental context, which is often not the case for a psychophysical test.

Other methods of physiological measurements we utilize for understanding QoE include thermal imaging and emotion recognition based on optical imaging, both of which can track changes in subject responses and engagement during audio listening. Pupillometry, the study of pupil fluctuations that track cognitive processes affected by attention, engagement, and exertion, provides an extremely sensitive measure of cognitive load (Kahneman, 1966) and is discussed further below.

2.5 Considerations for Speech in Audio QoE

One area where we utilize both traditional psychophysics and physiological measurements is in exploration of speech and content dialog. Optimal speech-to-noise ratios (SNR) for cinematic content can be assessed using a combination of forced-choice experiments to narrow the parameter space of possible SNR manipulations, for example, boosting speech, lowering background (noise), or both. Method of adjustment experiments can be further conducted to characterize optimal dialog boost levels for different dialog isolation methods, and for varied content at different starting SNRs. Using these approaches have allowed us to characterize preferred dialog level settings for different content and environment conditions. Further, experiments can easily be done using these methodologies to understand interactions under differing real-world use cases, for instance using mobile phone

loudspeakers or headphones in environments with high ambient noise such as planes or crowded congregation areas.

Pupillometry is a well-suited measurement for understanding audio QoE in particular, because experiments can be performed in steady-light conditions that do not contribute to light-based changes in pupil diameter. It is also non-contact and non-invasive, and precision measurements can be made with relatively simple, low-cost equipment. In a representative experiment, we captured pupillometry responses during a speech comprehension task using three headphone conditions, single-talker presented as dual mono, multi-talker presented as two speakers mixed together in dual mono, and multi-talker as the two speakers each hard-panned to left or right (Figure 6). The listener is instructed to attend to the single talker, which is also one of the voices used in the multi-talker condition. The effect of attending to the assigned voice during the multiple talker condition without the benefits of spatial separation of the voices produces increased pupil dilation over the single talker condition, indicating greater cognitive load due to higher exertion required to understand the designated speech. When simultaneous multi-talker presentation is separated by hard panning each of the two speakers to the left or the right channel, cognitive load as measured by pupil dilation is only slightly higher to what is observed in the single talker condition.

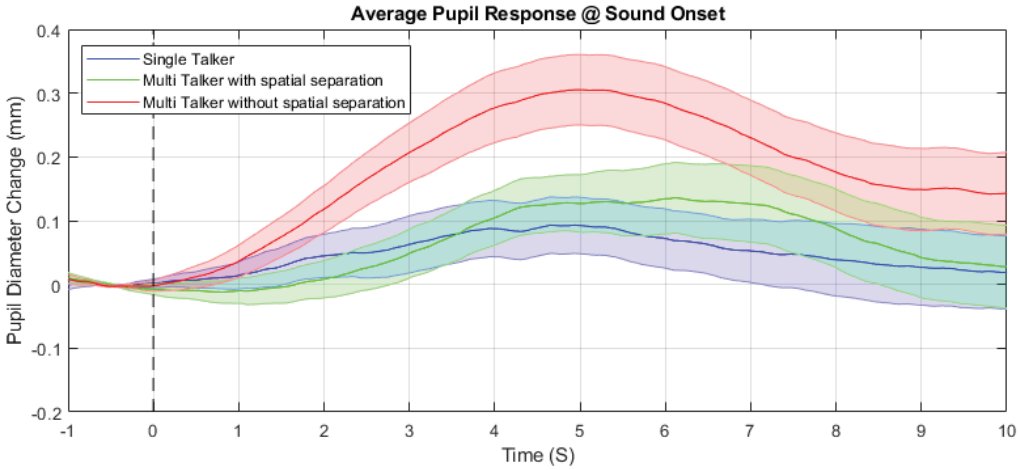

Figure 6 – Data from a headphone experiment measuring pupil responses to three conditions: single-talker (dual mono), multi-talker without spatial separation (dual mono mix of two voices), and multi-talker with spatial separation (two voices each hard panned to left or right). The listener is instructed to attend to the single-talker voice which is one of the voices in the multi-talker condition. From the baseline condition of single-talker, cognitive load measured by pupil diameter is increased in the multi-talker condition without spatial separation. Load measured by pupil diameter returns to baseline when the voices are separated into respective left/right channels.

3. CONCLUSIONS

Consumers have entered an era in which audio is experienced across a massive range of content, environments, equipment and individual use cases. Increasingly, our understanding of audio perceptions based on traditional psychophysics is being complemented by new advances in physiological measurements and techniques for spatial assessment of audio. Insights derived from this combination of approaches can be used both to maximize the consumer's QoE, and to preserve the integrity of artistic and creative decisions intended to be conveyed with high fidelity from content generation through distribution and delivery.

These combined techniques will become increasingly relevant in characterizing audio QoE especially as massive amounts of physiological data become available via wearable and mobile device sensors. This will allow experience to be understood in the context of real-life use cases, and will lead to new and effective ways to optimize QoE.

REFERENCES

1. Green DM, Swets JA. 1966 Signal Detection Theory and Psychophysics. New York: John Wiley
2. Stevens SS. 1975. Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects. New York: John Wiley.
3. Stevens SS. The Direct Estimation of Sensory Magnitudes: Loudness. The American Journal of Psychology 1956; 69(1):1-25.
4. Brandley RA, Terry ME. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons 1952; 39(3/4):324-345
5. Darcy DP, Terry KB, Davidson GA, Graff R, Brandmeyer A, Crum PAC. Methodologies for High-Dimensional Objective Assessment of Spatial Audio Quality 2016; AES 140th Convention, Paris, France
6. Engelke U, Darcy DP, Mulliken GH, Bosse S, Martini MG, Arndt S, Antons JN, Chan KY, Ramzan N, Brunnström K. Psychophysiology-based QoE assessment: a survey. IEEE Journal of Selected Topics in Signal Processing 2017; 11(1):6–21.
7. Kahneman D, Beatty J. Pupil diameter and load on memory. Science 1966; 154(3756):1583-1585