



Perceptual Evaluation of Dynamic Sound Zones, Part 1: Audio Quality of Experience in a Simulated Home Environment

LINE HERMANNSEN,^{1,2,*} AND SØREN BECH,^{1,3} *AES Fellow*
 (line.hermannsen@gmail.com) (sbe@es.aau.dk)

¹*B&O Research, Bang & Olufsen A/S, Struer, Denmark*

²*Department of Computer Science, Aalborg University, Aalborg, Denmark*

³*Department of Electronic Systems, Aalborg University, Aalborg, Denmark*

A sound system was developed to test new innovative designs for creating personal sound zones in a simulated home environment. A listening test paradigm with rating scales was customized and verified as suitable for collecting reliable data describing the basic audio quality and audio quality of experience of sound zones. Based on a previous study, a selection of attributes was evaluated, and all were found to be significantly predicted by the chosen changes in audio content, gain, and bass. A large subwoofer setup was added to the domestic system in some scenarios to reproduce low-frequency components and confirmed to positively affect audio quality of experience. However, the bass enhancement also increased perceived annoyance, which highlights the trade-off between bass quality and sound zone separation. Ways to optimize the experimental protocol are suggested for future studies of perception and affection for audio reproduction in sound zones, including the continued work in the ISOBEL project.

0 INTRODUCTION

0.1 Background

Human ears never sleep and are continuously sensing various auditory stimuli to keep aware of surroundings [1]. Although wanted sounds can positively affect mood and reduce stress and anxiety [2], unwanted sounds can cause stress, sleep disorders, and negative effects on the cardiovascular system [3]. At home and in hospitals, the enhancement of beneficial sounds and auditory masking of noise may be particularly important because these are places that are needed for rest and recovery. Yet, even during these sensitive states, people rely on the ability to sense auditory signals in their surroundings through auditory scene analysis to stay aware and connected with the environment and social contexts.

Personal sound zones may offer a solution to enhance wanted sounds and mask unwanted sounds and at the same time create personal listening experiences for multiple listeners in shared spaces while staying connected.

A personal sound zone is created by reproducing sound with an array of loudspeakers directing signals of interest to one or more listeners. The idea of personal sound is not new [4], but the development of commercial sound zone systems is. Therefore, scientific investigations of how sound zones can be created and optimized to create high-quality listening experiences in real environments is needed.

For consumer evaluations of listening experiences in sound zones, the concept quality of experience (QoE) is useful. QoE focuses on the affective domain by evaluating *the degree of delight or annoyance of a person whose experiencing involves an application, service, or system*. [5]. Yet, to understand how to adapt a reproduction system to allow for dynamic optimization of listening experiences, it is necessary to have a broad overview of how physical factors affect perception (attribute scaling) and affection (audio QoE), which can be found in the filter model ([6] and references herein (Fig. 1).

Changes in the physical factors (e.g., loudness and dynamic range) of a sound will affect perception, which in turn may change the affection. However, even with constant physical factors, perception and affection may change over time due to changes in environment (e.g., physical,

*To whom correspondence should be addressed, e-mail: line.hermannsen@gmail.com.

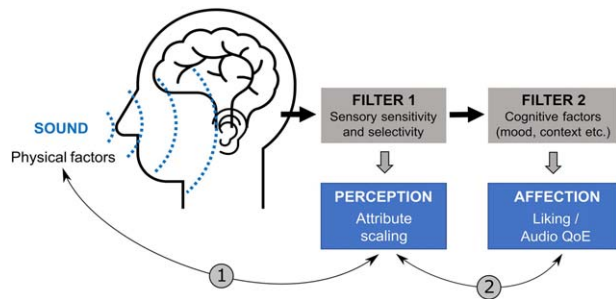


Fig. 1. The filter model (modified after [6]) shows how physical factors of sound relate to human perception and affection. A sound is given by a set of physical factors. When the sound reaches the human auditory system, it will first be filtered by filter 1, resulting in a certain perception of the stimulus. Next, a second filter (filter 2), influenced by human cognitive factors, will result in a perceived affection or audio QoE of a stimulus. In the ISOBEL project, the effects of physical factors on perception (path 1) and how they relate to affection (path 2) are investigated.

temporal, and social contexts), human factors (including expectations, mood, and attention), and/or system influence factors, such as the content that is selected for experience. In turn, all these factors will continuously affect the requirements of the reproduction system to achieve a desired QoE [7, 8]. Optimally, a reproduction system should, therefore, continuously register changes in selected content and context along with physiological state of the user(s) to optimize QoE in real time for a given situation. Adding more complexity to the task, a sound zone system that enables multiple audio experiences simultaneously requires that the adaptations of the system are balanced toward optimal QoE for multiple users and not in favor of one user while decreasing QoE for everybody else.

This present study is part of a series of studies within the Danish ISOBEL project (2019–2023; www.isobel.dk). The ISOBEL project aims to develop dynamic sound zone systems that can adapt in real time to create optimized listening experiences for two users simultaneously in different use cases (with changes in audio contents, sound zone positions, and sound zone widths). For the first time, to the authors' knowledge, commercially viable sound zone systems are being tested in both laboratory and real-life domestic and health care settings to investigate basic audio quality (BAQ), audio QoE, user interactions [9], and user experience [10]. The following gives an overview of the four papers focusing on BAQ and audio QoE of sound zones. The laboratory studies reported in part 0 [11], part 1 (the present paper), and part 2 [3] were conducted first to test and optimize the experimental design before deploying and testing the sound zone systems in real environments, as reported in part 3 [12].

Part 0 [11] contains the preliminary work necessary for the following experiments, including definitions of sound zone scenarios and elicitation of perceptual attributes through expert evaluations and listening tests.

The present paper, part 1, uses elicited attributes for domestic use cases (entertainment and information gathering; defined in part 0) in a customized listening test

asking assessors to rate the attribute intensity in different sound zone scenarios. This study further investigates how bass enhancement with a subwoofer setup affects audio QoE in sound zones. The experimental protocol is evaluated (assessor reliability, attribute correlations, and variance analysis) to ensure reliable and detailed data on how physical factors affect audio perception in domestic environments.

Part 2 [3] focuses on sound zones in a hospital environment and includes all tested use cases (rest, entertainment, and information gathering) to cover a broad range of patients. As in part 1, a customized listening test is used to obtain attribute ratings in different sound zone scenarios. The attribute list is condensed based on results from part 1. A hospital soundscape is added to the setup to improve the ecological validity of the results. The effects of bass are evaluated through removing low-frequency components in the reproduced audio. The experimental protocol is evaluated, and variance analysis is conducted to assess how physical factors affect attribute ratings. Ratings of liking are evaluated with linear regressions to determine the most important attributes for audio QoE.

Part 3 [12] contains the first perceptual evaluations of a commercial sound zone system deployed in real homes. The experimental design was optimized to run a listening test remotely on a laptop in the assessors' homes. The attribute list was focused on domestic scenarios and further reduced based on knowledge from part 2. The listening test was conducted twice for each assessor to investigate how living with the system during a four-week period would affect audio QoE. Due to difficulties in finding test families and a desire to optimize the data collection, a simulated home was also established and verified as viable for further data collection. Models for audio QoE of a domestic sound zone system were evaluated, and the applicability of a real-time distraction model [13] was tested.

The present paper describes the experimental setup and methods including specifications of the sound zone system in SEC. 1. The results from the listening tests are described in SEC. 2 and the findings are discussed in SEC. 3. Finally SEC. 4 summarizes the study conclusions.

0.2 Main Research Questions

In this present study, a sound zone system was used to create two personal sound zones at two positions on a sofa, to assess how sound zones are perceived in a laboratory setup that simulated a domestic environment. The primary research objectives were as follows: 1) to test and optimize a modified multiple stimuli with hidden reference and anchor (MUSHRA) method including assessor scale use for the chosen experimental factors, 2) to evaluate the consistency of ratings¹ within the specific conditions applied, 3) to identify the relevant attributes for evaluation of sound zones within the specific conditions applied, and 4) to establish

¹Consistency of ratings is in the present context defined as an assessor's ability to repeat ratings of identical stimuli.



Fig. 2. The experimental setup with the domestic system placed in front of a three-person sofa. Two subwoofers are placed behind the sofa, and one subwoofer is placed beneath the domestic system (see red arrow). The other five subwoofers are placed around the sofa (see Fig. 3). A tower of amplifiers and sound cards driving the domestic system is seen to the left, and next to it is the control station with a stationary computer and laptop for running the domestic system and the experimental tests. A microphone array on a tripod (seen to the left of the sofa) was used to calibrate the system before the tests.

the effects on BAQ and audio QoE of the physical factors audio content, bass, and gain for combinations of target-to-interferer ratios (TIR).²

1 METHODS

1.1 Laboratory Settings

The listening tests were conducted from winter 2020 to spring 2021 in a sound-proof laboratory at Aalborg University, Denmark.³ The laboratory setup simulated a domestic living room with a sofa and a table placed in front of the sound zone system (Fig. 2).

1.2 Sound Zone System

The prototype version of the sound zone system comprised a custom-built sound system and eight custom-built 10-inch subwoofers with a lower -3 dB cutoff frequency at 50 Hz. The system had a bottom array with 14 midrange drivers and a top array with 96 full-range 1.5-inch drivers. However, a modified version of the system (denoted *domestic system* in the following) had to be developed because of sound card limitations, so the modified version only included the 45 centered full-range 1.5-inch drivers. The optimal crossover frequency was set to 200 Hz on the basis of evaluation by an expert listener. Audio above 200 Hz

²The target-to-interferer ratio is defined as the difference between the target and interferer levels (e.g., LAeq) of the target and interferer signals measured in the same zone. See [14, 13].

³The laboratory has dimensions 7.0 m (width) $\times$ 8.12 m (length) $\times$ 3.0 m (height) and a reverberation time (T60) within the ITU-R BS.1116 targets except that it is 0.2 s in the frequency range 0.5–6.3 kHz, which is below the lower limit (0.25 s) of ITU-R BS.1116.

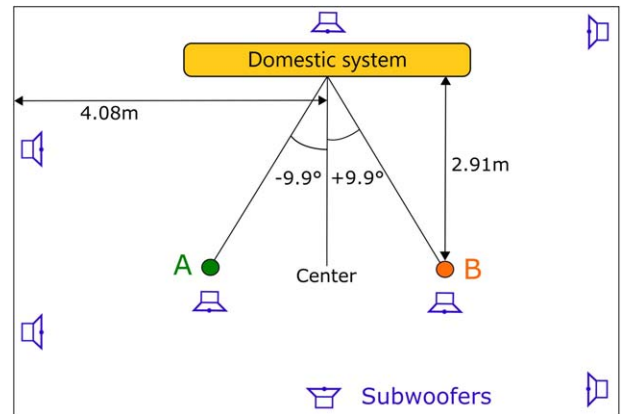


Fig. 3. A schematic overview of the experimental laboratory setup for investigating perception of two sound zones directed toward positions A and B. The assessor was seated in position A. Audio was presented with and without low-frequency components from a subwoofer setup. The wide-angle reproduction included the domestic system and a subwoofer placed right beneath the domestic system. Modified from Martin Bo Møller, Bang & Olufsen A/S.

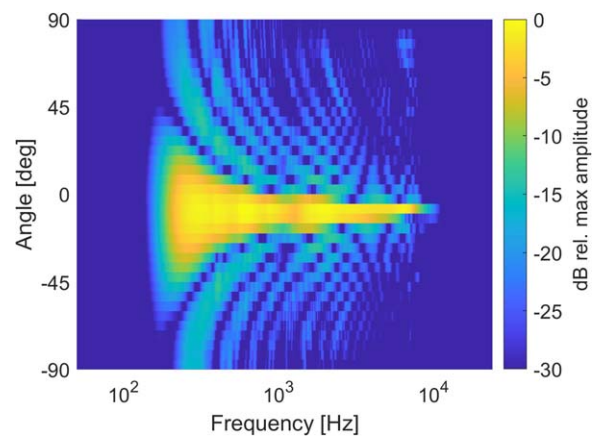


Fig. 4. The simulated free field directivity responses of the domestic system when playing audio in SzA. Note that the reproduction is without the subwoofer system and that the center of the domestic system energy is at an angle of approximately 10° , as shown in Fig. 3. By courtesy of Martin Bo Møller, Bang & Olufsen A/S.

was reproduced in two sound zones directed at an $\pm 10^\circ$ angle (Figs. 3 and 4, center of beam) from the center of the domestic system toward two positions in the room, denoted positions A and B (Fig. 3).

Subwoofers were included to extend the domestic system to frequencies below 200 Hz because bass has a well-established importance for audio quality perception. Yet, the enhancement of bass in sound zones increases the demands on drivers and processing power and may trade off sound zone separation due to the longer wavelengths of low-frequency components. This study, therefore, aimed to investigate this trade-off during laboratory studies where a large subwoofer setup was possible. The subwoofers were placed around the assessor with two woofers right behind the sofa (Figs. 2 and 3).

The filter design for the subwoofer section below 200 Hz was created following the method outlined in [15]. The filter design above 200 Hz followed the delay-and-sum method

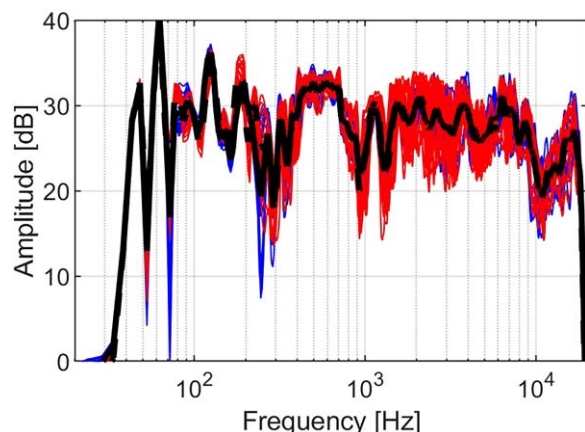


Fig. 5. Predictions of wide-angle frequency responses based on in situ measured room impulse responses to points in zone A (blue) and zone B (red). Note that the results have been smoothed by a one-twelfth-octave moving average filter. The black curve is the average response in zone A. By courtesy of Martin Bo Møller, Bang & Olufsen A/S.

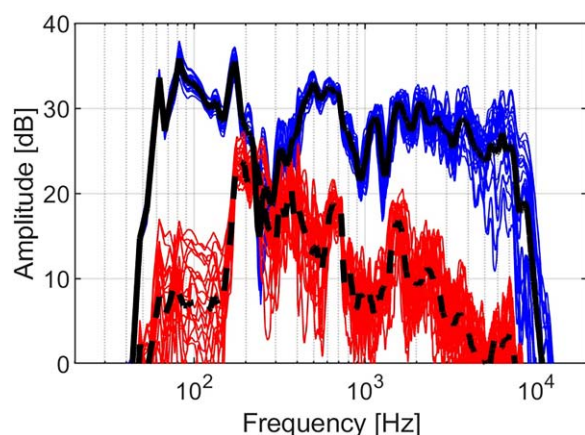


Fig. 6. Predictions of the performance of the domestic system in sound zone mode based on in situ measured room impulse responses to points in zone A (blue) and zone B (red). Note that the results have been smoothed by a one-twelfth-octave moving average filter. The black curves are the average responses. By courtesy of Martin Bo Møller, Bang & Olufsen A/S.

described in [16]. The subwoofer filter calculations were based on room impulse responses measured in a 20 cm by 30 cm square, sampled with ten microphones, and the soundbar filters were computed as delay-and-sum filters based on the assumption of point sources radiating sound into the free field, where the point source locations corresponded to the loudspeaker drivers in the physical prototype. A wide-angle ($\pm 90^\circ$) reproduction mode was also used to simulate a traditional sound bar together with reproduction of low-frequency components from the single subwoofer placed beneath the domestic system (Fig. 2).

The frequency responses of the wide-angle system in zone A (blue) and in zone B (red) at the listening position without an interfering signal are shown in Fig. 5. The frequency responses of the domestic system in sound zone mode are shown in Fig. 6 for zone A (blue) and zone B (red). The contrast as a function of frequency between zone

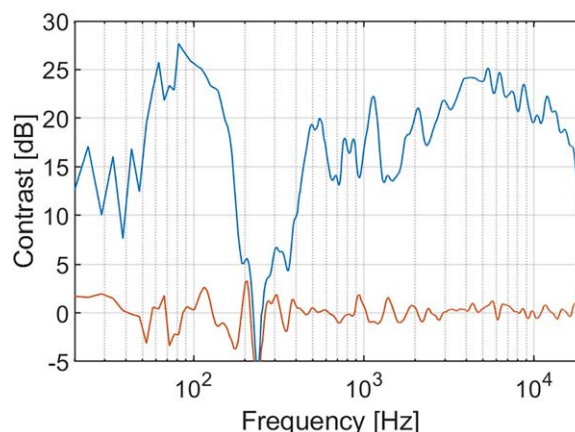


Fig. 7. Contrast as a function of frequency for the domestic system in sound zone mode (blue) and wide-angle mode (red). The contrast is determined as the level difference between the average level in zone A and B. By courtesy of Martin Bo Møller, Bang & Olufsen A/S.

A and B for the domestic system (blue) in sound zone mode and the wide-angle system (red) is shown in Fig. 7. As seen from Fig. 7, the contrast between the zones is between 15 and 25 dB in the majority of the frequency range. Given that the target mean square pressure in both zones was reasonably flat, this translates into a TIR of equivalent magnitude (assuming that the reproduction levels in both zones are identical). Note that the figures show the pressure in both zones when creating a sound zone in zone A with respect to zone B, that is, the difference between the curves indicate the contrast not the TIR.

A single channel (mono) signal was reproduced in each zone or in the wide-angle mode; however, this choice was only based on practical concerns at this stage of system development and is not a general limitation of the system. External processing was conducted in MATLAB on a stationary laptop linked to three MOTU 24 i/o sound cards and multiple seven-channel 150-W amplifiers (ICEpower ICE Theater 7). The sampling rate of the reproduction was 44.1 kHz.

1.3 Evaluating Sound Zone Perception and Affect

Sound zone perception and affection were evaluated with listening tests following a modified (MUSHRA) [17] procedure. The listening test format is described in SEC. 1.3.4, and the experimental protocol is described in SEC. 1.3.5.

Audio content was varied to reflect different use cases (see SEC. 1.3.1). For each target⁴ audio, five scenarios were represented covering two reproduction modes (sound zones and/or wide-angle sound) and different levels of physical factors (Fig. 8). The stimuli were evaluated for attributes linked to BAQ along with attributes in the affective domain, including “liking” as a proxy for audio QoE (see SEC. 1.3.2).

⁴The terms *target* and *interferer* audio are used throughout the text. The target is the audio signal in the sound zone with a listener (zone A). The interferer is the audio signal in the sound zone without a listener (zone B).

Target audio		Interference audio	
		Scenarios	Entertainment
Entertainment	Classical music 1	1-3	No interferer
		4	Pop/rock music 1
		5	Speech
	Classical music 2	1-3	No interferer
		4	Classical music 1
		5	Radio sports
	Rock/pop music 1	1-3	No interferer
		4	Classical music 1
		5	Speech
	Rock/pop music 2	1-3	No interferer
		4	Pop/rock music 1
		5	Radio sports
Information Gathering	Speech	1-3	No interferer
		4	Radio News
		5	Pop/rock music 1
	Radio sports	1-3	No interferer
		4	Podcast
		5	Classical music 1

	Stimulus	Target audio (Wide angle)	Target audio (SzA)	Target audio (SzB)	Interference audio (SzB)
Scenario	1				
	1	Unmodified			
	2	+ bass			
Scenario	2				
	3	Unmodified	Unmodified		
	4	Unmodified	+ 6 dB		
	5	+ bass	Unmodified		
Scenario	3				
	6	Unmodified	Unmodified	Unmodified	
	7	Unmodified	+ 6 dB	Unmodified	
	8	+ bass	Unmodified	Unmodified	
	9	Unmodified	Unmodified	+ 6 dB	
Scenario	4	5			
	10	19	Unmodified		Unmodified
	11	20	+ 6 dB		Unmodified
	12	21	+ bass		Unmodified
	13	22	Unmodified		+ 6 dB
	14	23	Unmodified		+ bass
	15	24	+ 6 dB		+ 6 dB
	16	25	+ 6 dB		+ bass
	17	26	+ bass		+ 6 dB
	18	27	+ bass		+ bass

Fig. 8. (a) An overview of the audio stimuli used in the listening test. The audio content was either music (entertainment) or speech (information gathering), resulting in four primary combinations of target and interfering audio: music-music, music-speech, speech-music, and speech-speech. For each of the six target audio tracks (second green column), the listening test contained five scenarios (1 through 5) as shown in the table to the right (b). The five scenarios (purple) covered modifications of the physical factors gain, and bass and thereby totaled 27 stimuli that the assessor was asked to evaluate (note: the 27 stimuli were distributed on three pages with nine stimuli on each page in the user interface, see Fig. 9). The listening test always consisted of a target audio reproduced as wide-angle sound and/or in SzA. In scenario 3, the target audio was also reproduced in SzB. Scenarios 1, 2, and 3 only contained target audio, whereas scenarios 4 and 5 (stimulus 10 through 27) covered 18 stimuli with audio-on-audio reproduction of the target audio in SzA and the interference audio in SzB. For a more detailed discussion of the experimental design see SEC. 1.3.2.

Table 1. Audio information for the eight audio tracks used in the listening test.

Audio name	Artist and track name, date Time in track (hours:minutes:seconds)
Classical music 1	Brahms, <i>Hungarian Dance No. 18</i> , 00:00:00–00:01:00
Classical music 2	Mahler, <i>Symphony No. 5, 4th Movement</i> , 00:00:20–00:01:15
Pop/rock music 1	The Killers, <i>On Top</i> , 00:01:53–00:02:53
Pop/rock music 2	The Bravery, <i>No Brakes</i> 00:01:59–00:02:54
Speech	The Archimedes Project [20], <i>Anechoic female speech</i> 00:00:00–00:00:59
Radio sports	The Danish Broadcasting Corporation (DR), <i>Liga</i> , June 21, 2020 02:25:38–02:26:38
Radio news	DR P1, <i>Radioavisen</i> , June 22, 2020, 10:00 a.m. 00:00:00–00:01:00
Podcast	DR archives, <i>Om hundrede aar, del 1, 1:5</i> , 2020 00:00:02–00:01:02

1.3.1 Audio Stimuli

This experiment focused on a domestic context with listening tasks entertainment and information gathering [18], and the audio material was therefore selected to fit within these two use cases. Entertainment audio stimuli consisted of four music clips (originally used in [19]) belonging to two genres: classical music and pop/rock. Information gathering stimuli included four audio clips with Danish speech extracted from the public Danish radio [the Danish Broadcasting Corporation (DR)] and the Archimedes project [20]. An overview of audio pieces is given in Table 1. All stimuli were single-channel, 55–60 s long, and looped during the listening tests. Stimuli were initially loudness-matched with a custom-written MATLAB script following the International Telecommunications Union standard BS-1770 [21]. The sound pressure level at the listening position of 55 dB equivalent continuous sound pressure level (Leq) (61-dB Leq with 6-dB gain) was based on a previous experiment, where six assessors were asked to set their optimal listening level in three listening tasks [11]. Before each listening test, the sound pressure level at the listening position was checked with a Brüel & Kjær (B&K) analyzer.

Fig. 8 shows an overview of the audio stimuli used in the listening tests. The specific choice of stimuli and the associated experimental design is discussed in the next section. Scenarios 1 and 3 only included target audio but used both the wide-angle and the sound zone reproduction modes. Scenarios 4 and 5 assessed audio-on-audio scenarios in sound zones by including both target and interfering audio. The target and interference audio were either played back in the loudness-matched but otherwise unmodified version, or they were modified by either increasing the reproduction level by 6 dB or adding subwoofers to the reproduction system.

1.3.2 Experimental Design

The audio stimuli used in the listening tests are shown in Fig. 8. The general use case for scenarios 1, 2, and 3 is a family watching a television with monophonic sound reproduction with one or more of the members needing enhanced speech (e.g., due to impaired hearing) that should not impact the other members listening experience. Thus, the sound system should be able to create a general broadband mono signal with a wide directivity covering the entire room. On top of that, several focused beams (narrow directivity) should be directed toward each of the relevant listeners. The focused beams have limited frequency content (e.g., only speech) and can be processed individually. Scenarios 1, 2, and 3 thus included one audio signal and were reproduced as 1) wideband target audio (stimuli 1 and 2), 2) wideband plus sound zone A (SzA) (stimuli 3, 4, and 5), or 3) wideband plus SzA and sound zone B (SzB) (stimuli 6 through 9). The experimental design for each scenario followed as a simple screening design approach focusing on the main factors audio content, gain, and bass.

The general use case for scenarios 4 and 5 is to be able to create individual sound zones with individual audio content for each member of the family. Stimuli 10 through 27 thus

included both target and interferer audio with different content. A full factorial design including all the factor levels would include 156 stimuli that would be both too time consuming and include unnecessary information at this stage of investigation. Thus, to reduce the experimental effort two and three-way interactions were considered unimportant, and a screening design aimed at exploring only the main factors (audio content, gain, and bass) was employed.

For the entertainment task and music as target (stimuli 10 through 18) the 12 possible music-music combinations (scenario 4)⁵ of four audio clips were reduced to four, representing (target versus interferer) classical versus pop/rock, classical versus classical, pop/rock versus classical, and pop/rock versus pop/rock. The selection of the music pieces was made considering the audibility of changes in the physical factors gain and bass. The music-speech (scenario 5 and stimuli 19 through 27) followed the same logic. The 16 combinations (target versus interferer) of gain and bass changes were reduced to nine for both scenarios 4 and 5 in accordance with a screening design approach. In this way, each target audio was presented as 27 stimuli (Fig. 8). For the information gathering task with speech as the target (speech-speech in scenario 4 and speech-music in scenario 5) the reduction in target-interferer combination followed the same strategy as described for stimuli 10 through 27.

1.3.3 Perceptual Attributes

The initial identification and selection of attributes for sound zone scenarios is described in Part 0 [11]. It included a literature-based list of 52 attributes that was reduced through expert listener evaluations and an experiment assessing the following: 1) how understandable⁶ the attributes were and 2) the relevance of attributes for various sound zone scenarios.

The attribute liking was added to the initial list as a proxy for the audio QoE and assessed for all stimuli on a similar continuous scale as the other attributes. The attribute annoyance, also belonging to the affective domain, was evaluated for entertainment scenarios, whereas distraction was evaluated for information gathering scenarios, based on the investigated relevance for sound zone scenarios in the previous experiment [11]. These two attributes had also been found relevant for audio-on-audio scenarios in Francombe et al. [19].

The attribute width was previously [11] assessed as not being understandable, and reformulation of the definition was suggested. However, because this attribute definition had been reached through multiple elicitations of naïve and expert listeners, the definition was kept, and instead, assessors were asked to call for help if they had any questions or problems understanding any of the attributes. Furthermore, the expert listener assessing relevant attributes for

⁵Combinations including the same audio clip in the target and interferer are excluded, and all clips appear as both target and interferer.

⁶Understandable in this context means how well the assessors understand the given explanation/definition of the attribute.

Table 2. Perceptual attributes evaluated for the two listening tasks entertainment and information gathering (columns).

Entertainment	Information gathering
Loudness [22]	Loudness [22]
Midrange strength [22]	
Nasal [22]	
Canny [22]	
Bass strength [22]	
Boomy [22]	
Width [22]	Distance [22]
Clean [22]	Natural [22]
	Understandable*
Annoyance [19]	Distraction [19]
Liking [†]	Liking [†]

Note: The references listed after each attribute include the description of the attribute.

*It is easy to understand, what is being said in the audio clip.

[†]To which degree do you like the stimuli?

this present study evaluated width as relevant to cover spatial characteristics that would differ between wide-angle audio mode and audio played back in sound zones.

Based on the results described in Part 0 [11] and the above considerations, this present study initially involved 11 attributes for the entertainment task and five attributes for the information gathering task. However, due to the addition of the eight-woofer setup for bass enhancement, the attribute list was reevaluated through evaluations by an expert listener (Tonmeister, employed at Bang & Olufsen A/S) listening to the different audio scenarios in the listening test. For the entertainment task, this led to the reintroduction of a number of attributes: midrange strength, nasal, canny, bass strength, and boomy (all from the DELTA sound wheel [22]). For information gathering, no additional attributes were added. The final attribute list is shown in Table 2.

1.3.4 Assessors

Twelve university students (two women, ten men, age 20–25, fluent in Danish) with measured normal hearing⁷ participated in the listening test. All assessors were presented with a familiarization protocol introducing stimuli and attributes before the listening test. Assessors were paid for their participation. Data were handled to comply with general data protection regulations (GDPR), and assessors were numbered 1 to 12 in the data collection.

1.3.5 Listening Test Format

The listening test was generated in a custom-written *Max* patch as a modified MUSHRA procedure, where the intensity of a given attribute could be compared and rated for a number of stimuli. The intensity scales were shown as vertical bars ranging from 0 to 100 without values, except for two anchor points placed at 15 and at 85 labeled *A little*

⁷Hearing thresholds of participants were tested with an Orbiter 922 (Madsen Electronics) presenting six pure tones in the frequency range 250–4000 Hz at sound pressure levels of 0–40 dB in increments of 5–10 dB. Normal hearing was considered when the hearing level was <20 dB; ISO 7029, 2017.

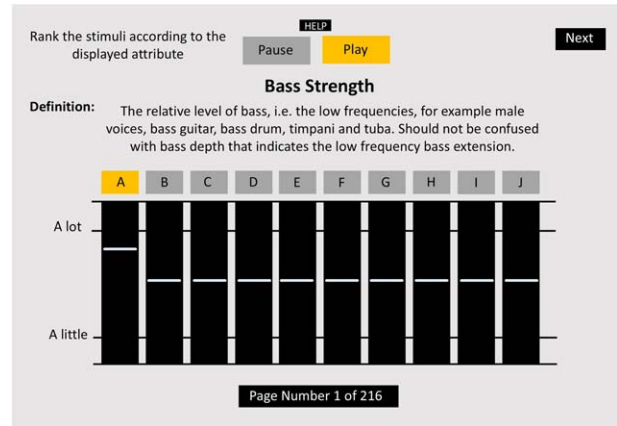


Fig. 9. The user interface of the listening test. Assessors were asked to use the white horizontal slider to indicate their perceived intensity of the given attribute in stimuli A through J. The ten stimuli on each page in the user interface contained nine out of the 27 available stimuli (see Fig. 8), randomly selected, and a hidden repeat of one of the nine stimuli. The intensity bars had a range of 0–100, and the slider was at a default value of 50. The interface shown is translated from Danish.

and *A lot* (translated from Danish), respectively (Fig. 9). A horizontal slider, placed at a default value of 50, was available on each vertical bar to indicate perceived intensity of the given attribute for each stimulus. The 27 audio stimuli for each target audio stimulus were presented on three consecutive pages in the user interface. One random stimuli on each page in the user interface was repeated (so-called *hidden repeat*) to evaluate assessor reliability. Each page thereby contained nine stimuli and one hidden repeat (stimuli A through J; Fig. 9), which follows the MUSHRA standard [17] that recommends nine stimuli under test and a maximum of 12 stimuli per page. No labeled reference and hidden reference were included because these could not be defined. The listening test was mirrored to a tablet interface using MIRA.

1.3.6 Listening Test Protocol

One assessor participated at a time, seated with a tablet in front of them in SzA (Fig. 3) positioned 2.9 m in front of the sound zone system on the left side of the sofa. The neighbor sound zone with the interfering audio was placed 1.1 m to the right of SzA (center to center). The assessor was alone in the laboratory during the test while the experimenter monitored the test from outside the laboratory on a screen linked to a video camera visible in the laboratory. Assessors had given consent to be monitored by camera, which was implemented only to ensure assessor safety and did not log any data.

The listening test was divided into three parts, including a familiarization protocol and two listening tests: one for each of the two listening tasks entertainment and information gathering. The familiarization protocol presented two attributes, rubbing and buzzing ([22]), for 10-s snippets of the twelve combinations of target and interference stimuli (Fig. 8).

The presentation order of the two listening tasks was alternated between the 12 assessors so that each task was presented first six times and presented last six times. The assessor was given a short written introduction to each task to set the scene, following the procedure from Francombe et al. [18] (translated into Danish): *Imagine that you are at home relaxing to music* (entertainment) / *Imagine that you are at home and listening to understand and convey the information in the audio clip* (information gathering).

Within each task, perceptual attributes (see SEC. 1.3.3) were presented in a randomized order, yet always presenting liking last after a short break to minimize the chance of the liking ratings being biased toward the performance of the system on the last attribute. The assessor was instructed in rating an intensity of zero of all stimuli on a page if they found that a presented attribute was not relevant for distinguishing between the stimuli. Furthermore, they were instructed to ask the experimenter for further explanation if they did not understand the description of an attribute. In total, a full listening test included presentation of 240 pages for entertainment stimuli and 72 pages for information gathering stimuli. The assessors were self-paced, and the total duration of the test was in the range 225 to 360 min, with an average of 283 min. To avoid fatigue, assessors were asked to take a short break of 5 min every 30 min; they were offered water and sweets to maintain motivation and concentration.

1.3.7 Missing Data

Due to an error during data collection, some pages in the user interface contained repetitions of audio and attributes that had already been rated, which led to overwriting of previous data and missing data for other combinations of audio and attributes. The data from assessor 6 were removed from the analysis due to multiple connection problems between the tablet and *Max* software, which resulted in the assessor not completing the listening test. For the remaining assessors, the error in the test software led to missing data for 6% (738 out of 11,880 ratings) of the entertainment ratings and 5% (180 out of 3,564 ratings) of the information gathering ratings.

1.3.8 Processing and Analytical Methods

Data were initially preprocessed and structured in MATLAB and then imported into the R environment for statistical analysis using a general 5% criterion for statistical significance. The first part of the analysis focused on evaluation of assessor performance. Data were quality checked with histograms and summary statistics for each assessor to assess scale use and rating consistency. The distributions of scores along the scale were tested with Hartigans' dip test [23], whereas the reliability of the scales were tested by calculating Cronbach's alpha values [24] with annoyance and distraction specified as reversed scores. Assessor consistency was evaluated by comparing ratings of original stimuli and hidden repeats and calculating absolute score differences. To evaluate potential assessor fatigue, assessors were divided into two groups: one group that started with

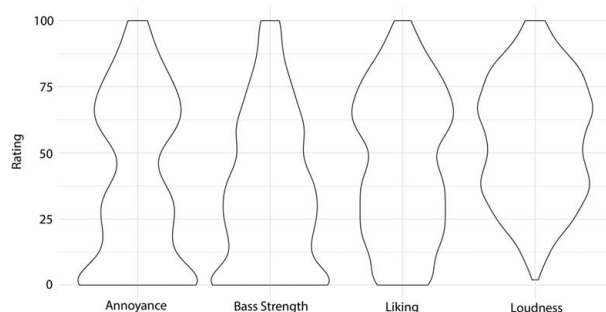


Fig. 10. The distributions of ratings for four attributes shown as violin plots, where the curve width represents the approximate frequency of data points at each rating.

the entertainment task and another group that started with information gathering, hypothesizing that assessors would be more consistent when conducting their first task than when conducting their second task due to fatigue over time. The rating consistency between these two groups was tested for all attributes by running a linear mixed effects model using the linear mixed effects regression (*lmer*) *lmerTest* library [25] with the assessor as a random factor. Next, the model output was used to generate a type III analysis of variance (ANOVA) table with Satterthwaite's method.

The second part of the analysis focused on evaluation of the experimental protocol. Attribute ratings for different modifications of stimuli were summarized with the *summarySE* function from the *Rmisc* library in R [26] to obtain means with 95% confidence intervals (CIs). Means were then compared (Students *t* tests) to assess whether modifications of physical factors were perceivable and led to significant changes in perception, specifically if gain changes were perceivable in ratings of loudness and if subwoofer modifications led to perceptual changes in attributes bass strength and boomy. To investigate the importance of bass for audio QoE, the influence of subwoofer modifications was also evaluated for ratings of liking.

To assess the relevance of the perceptual attributes for sound zone scenarios with multiple audio streams, the last part of the analysis focused on audio-on-audio scenarios (scenarios 4 and 5), where target audio was played in SzA simultaneously with a different audio track playing in SzB. Variance analysis was conducted using *lmer* models (*lmer* function from R library *lmerTest*⁸ for each attribute within each listening task to assess the ability of a given attribute to distinguish between the selected stimuli. Fixed factors were scenario (two levels: scenario 4, scenario 5), bass (four levels: no bass, bass in SzA, bass in SzB, bass in SzA + SzB), gain (two levels: 0 dB, 6 dB), and TargetAudio (see Fig. 8). Assessor was set as a random factor, thereby considering evaluations on a population level rather than an individual level [27]. The initial attribute models were then backward eliminated using the *step* function from the *lmerTest* library to remove nonsignificant fixed factors. Reduced model results were obtained with the function

⁸<https://CRAN.R-project.org/package=lmerTest>

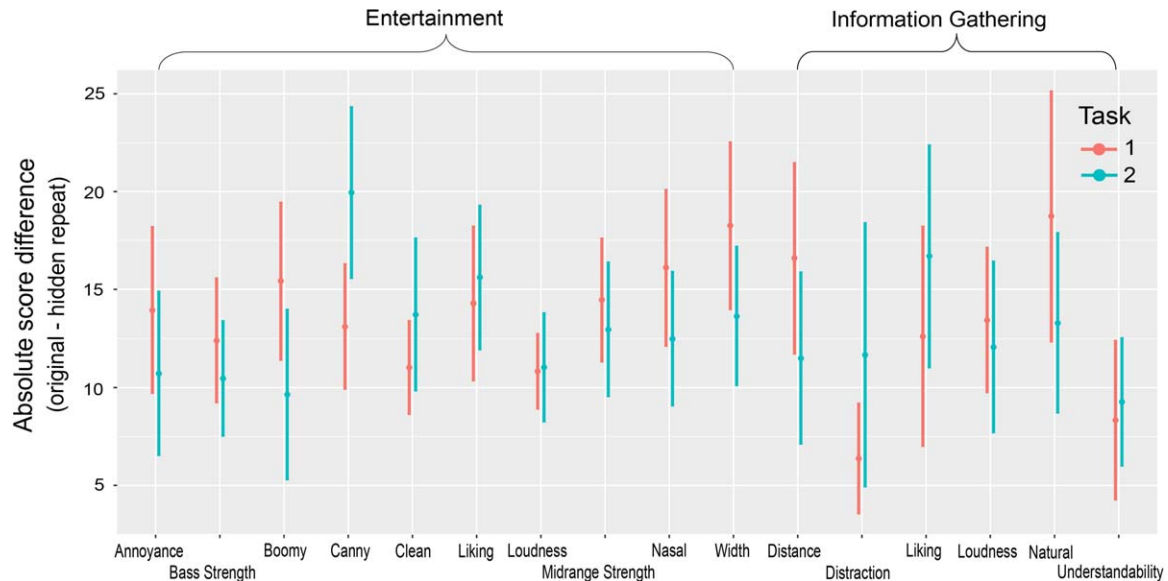


Fig. 11. The rating consistency calculated as the absolute difference in scores for the original stimulus and hidden repeat on each page in the listening test. The consistency is shown as means $\pm$ 95% CIs for both listening tasks across all assessors. For each assessor the presentation order of tasks was randomized, with either entertainment or information gathering scenarios being presented first (task = 1) followed by the other task (task = 2). The difference in consistency for attributes in the first compared with the second task was tested as paired contrasts with the *emmeans* R package and found only a significant difference in rating consistency over time for the attribute canny.

anova to obtain a type III ANOVA table with Satterthwaite's method. To obtain a measure of significant differences between groups, the R library *emmeans*⁹ was used to calculate estimated marginal means and contrasts ("pairs"). The multicollinearity¹⁰ of attributes was tested by calculating Spearman's correlation coefficients (suitable for continuous variables) using R library *PerformanceAnalytics* [28] and function *Chart.Correlation*.

2 RESULTS

2.1 Assessor Scale Use

All but one assessor used the entire range of the scale (0–100). Assessor 4 had a slightly restricted scale use (12–91). Ten out of the 16 attributes were found to have a bimodal distribution of ratings, and seven of them had a clear dip around a central score of 50.¹¹ The density of scores is shown as violin plots in Fig. 10 for four of the attributes. The overall internal scale consistency was found

to be $\alpha = 0.74$ [95% CI (0.72, 0.76)]. Cronbach's alpha values above 0.7 generally reflect a good reliability [29]. If excluding the attribute boomy, the alpha value would increase minimally to 0.75, whereas excluding any of the other attributes would instead decrease the internal scale consistency. Based on this, all attributes could be kept in the analysis while maintaining a good scale reliability.

Across all assessors, zero ratings were given to attributes annoyance and distraction in scenarios without an interfering audio (scenarios 1, 2, and 3). This is an unsurprising result given the absence of interfering audio in these situations. However, these scenarios were in some cases also rated with scores above zero. Ratings of annoyance for scenarios without interfering audio had a median of 0 (interquartile range: 0–10) and mean of 6.5, with 76% of ratings being below the lower anchor point at 15, whereas ratings of distraction had a median of 0 (interquartile range: 0–5.7) and a mean of 8.3, with 80% of ratings being below the lower anchor point at 15.

Additionally, zero ratings were seen for attributes boomy and bass strength in scenarios without subwoofers or only with subwoofers in SzB, again reflecting that assessors understood the rating task. Zero ratings were also seen for attributes canny and nasal when the target audio was classical music and when stimuli had addition of subwoofers. One assessor assigned zero ratings to the majority of stimuli (76%) for the attribute width and had a restricted scale use with a maximum score of 37 for this attribute (contrasting to a full-scale use for all other attributes by this assessor). This was interpreted as the attribute not being understood, and the ratings of width for this assessor were therefore excluded in the further analysis.

⁹<https://CRAN.R-project.org/package=emmeans>

¹⁰Strong linear relationships between attributes are unwanted when fitting regression models because the input variables should be independent.

¹¹The attributes with bimodal distributions are as follows (those that also had a dip around a central score of 50 are indicated with a +) within entertainment: annoyance+, bass strength, boomy, canny+, clean+, liking+, loudness+, nasal, width+; within information gathering: distraction, loudness, understandable+.

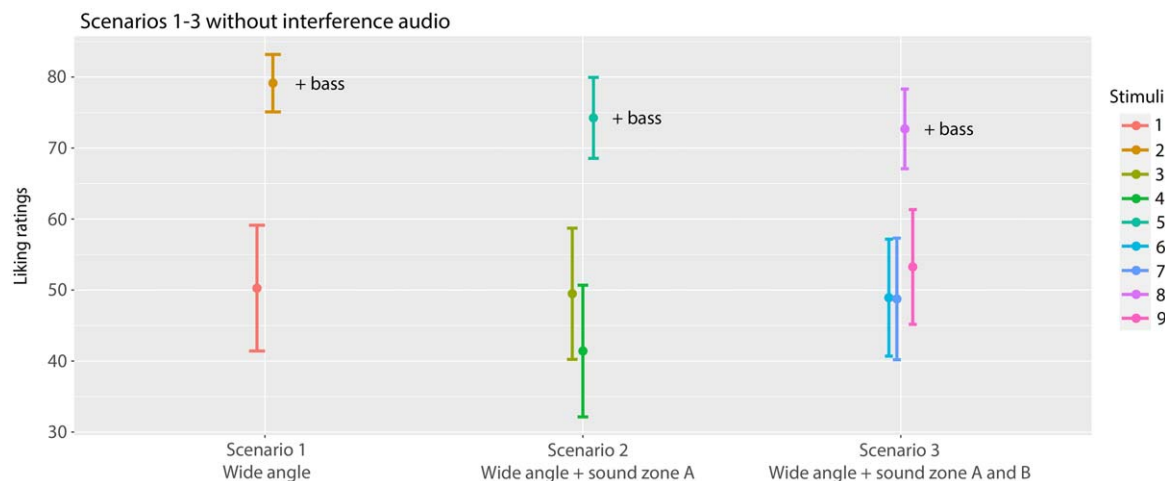


Fig. 12. Perceived liking in scenarios 1, 2 and 3, where only the target audio was present. The three scenarios contained nine stimuli (stimuli 1 through 9) with modifications of reproduction mode (wide-angle, SzA, and/or SzB), subwoofer (bass/no bass), and gain (0/6 dB). For all scenarios, the perceived liking increased when bass was added. Values are shown as means with 95% CIs.

2.2 Assessor Consistency

A prerequisite for obtaining reliable perceptual data is that assessors are consistent in their evaluations. This was tested with a hidden repeat on each page in the user interface that allowed for comparison of ratings of two similar stimuli evaluated within 1–3 min, which was the time assessors spent per listening test page. Rating differences across all assessors and all attributes are shown in Fig. 11, with assessors in general being most consistent when rating distraction and understandable and least consistent when rating canny and natural.

Given the relatively long duration of the full listening test (225–360 min, average 283 min), the potential for assessor fatigue causing a decreased consistency over time was also evaluated. In general, assessors maintained a similar consistency over time; the results showed no significant differences in the rating consistency of two identical stimuli (original and hidden repeat) for all attributes except canny ($p < 0.05$; Fig. 11). When canny was presented in the last part of the listening test, the attribute had less consistent ratings than when it was presented in the first part.

2.3 Physical Factors

In this present study, selected physical factors (audio content, gain, and bass) were modified to assess how these changes affect audio perception and audio QoE of sound zones. The effects were evaluated with summary plots (Figs. 15 and 16) and variance analysis (Table 3). The variance analysis was focused on audio-on-audio scenarios (scenarios 4 and 5). A 5% criterion for statistical significance is indicated as an asterisk (*) ($p < 0.05$). In cases with higher significance levels, this is indicated by a double asterisk (**) ($p < 0.01$) and a triple asterisk (***) ($p < 0.001$).

2.3.1 Audio Content

The audio content, encapsulated by variables target audio and scenario in the variance analysis (Table 3), has

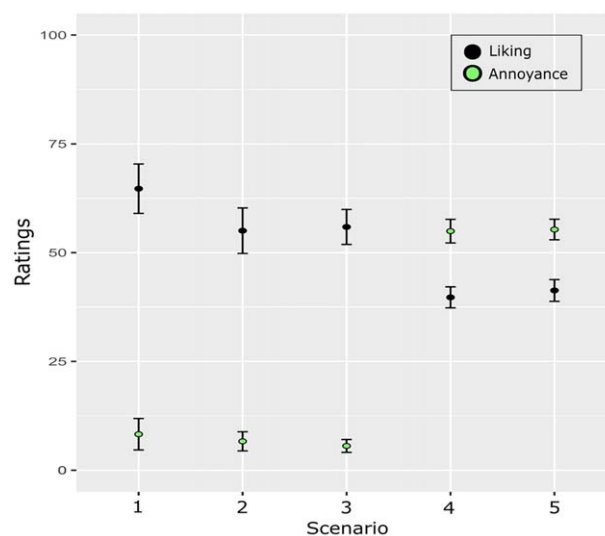


Fig. 13. Perceived liking and annoyance evaluated for scenarios within the entertainment use case shown as means with 95% CIs of the ratings. Scenarios 1, 2, and 3, with only the target audio present, have relatively high ratings of liking, whereas ratings of annoyance are low. Scenarios 4 and 5, with audio-on-audio reproduction, show a decreased liking and significantly higher annoyance ratings than scenarios without interfering audio.

a significant effect on the ratings on all attributes except liking of speech stimuli. Scenarios 1, 2, and 3, with wide-angle audio and without interfering audio have significantly higher ratings of liking and lower ratings of annoyance compared with audio-on-audio scenarios in sound zones (Fig. 13). Liking ratings were highest in scenario 1, significantly lower (Tukey-adjusted $p < 0.05$) in scenarios 2 and 3 and lowest in scenarios 4 and 5 (Fig. 13). In contrast, annoyance ratings were highest in scenarios 4 and 5 and significantly lower (Tukey-adjusted $p < 0.0001$) in scenarios 1, 2 and 3. The effect of scenario interacts with both the gain and subwoofer settings for scenarios 4 and 5 (Fig. 14). The se-

Table 3. Variance analysis models for attributes and liking presented for the two listening tasks (columns).

Entertainment	Information Gathering
Loudness ~ Scenario** + Bass*** + Gain*** + TargetAudio*** + Scenario:TargetAudio* SW-p = 0.090, $R^2c = 0.43$	Loudness ~ Scenario*** + Gain*** + TargetAudio*** + Bass:Gain* SW-p = 0.44, $R^2c = 0.48$
Clean ~ Bass*** + Gain*** + TargetAudio*** SW-p = 0.33, $R^2c = 0.41$	Natural ~ Bass*** + Bass:TargetAudio* SW-p = 0.62, $R^2c = 0.43$
Width ~ Bass*** + Gain*** + TargetAudio*** SW-p = 0.003, $R^2c = 0.19$	Distance ~ TargetAudio*** SW-p = 0.16, $R^2c = 0.31$
Midrange strength ~ Scenario* + Bass** + Gain*** + TargetAudio*** + Scenario:TargetAudio*** + Bass:Gain* SW-p = 0.002, $R^2c = 0.33$	Understandable ~ Scenario*** + Scenario:TargetAudio* SW-p = 4.9e-07, $R^2c = 0.37$
Nasal ~ Bass* + TargetAudio*** SW-p = 0.0004, $R^2c = 0.35$	...
Canny ~ Bass*** + TargetAudio*** + Scenario:TargetAudio*** + Bass:TargetAudio*** SW-p = 0.003, $R^2c = 0.45$	...
Bass strength ~ Scenario** + Bass*** + TargetAudio*** + Scenario:Bass** + Bass:TargetAudio*** SW-p = 4.3e-09, $R^2c = 0.55$	...
Boomy ~ Scenario** + Bass*** + TargetAudio*** + Scenario:Bass*** + Bass:TargetAudio*** SW-p = 2.44e-11, $R^2c = 0.55$	...
Annoyance ~ Bass*** + Gain*** + TargetAudio*** + Scenario:Bass*** + Scenario:TargetAudio*** SW-p = 0.21, $R^2c = 0.38$	Distraction ~ Scenario** + Bass* + Gain*** + TargetAudio*** + Scenario:TargetAudio* SW-p = 0.90, $R^2c = 0.55$
Liking ~ Bass*** + Gain*** + TargetAudio*** + Scenario:Bass* + Scenario:Gain* + Bass:Gain* + Bass:TargetAudio *** SW-p = 0.001, $R^2c = 0.47$	Liking ~ Bass* + Gain** SW-p = 0.01, $R^2c = 0.38$

Note: Data include only scenarios with audio-on-audio (i.e., with interfering audio present). The p values given are from a type III ANOVA with Satterthwaite's method. A 5% criterion for statistical significance is indicated as an asterisk (*) ($p < 0.05$). In cases with higher significance levels, this is indicated by a double asterisk (**) ($p < 0.01$) and a triple asterisk (***) ($p < 0.001$). Interactions are shown as X:Y. Normality of residuals of the reduced models was tested with a Shapiro-Wilk test; p values (SW-p) are shown in the table together with explained variance calculated as conditional R^2 values (function *r.squaredGLMM* from R library *MuMIn*). For nonnormal model residuals, estimates of skewness (acceptable ± 3) and kurtosis (acceptable ± 10) were found acceptable and therefore not included here. The intercept and error term were included in each model, but not shown here.

lected audio content reproduced as the target audio within entertainment scenarios also interacts with the bass settings for ratings of bass strength, liking, and annoyance when interfering audio is present (Fig. 16).

2.3.2 Gain

The gain of the audio material is a significant factor of perceived loudness (Table 3) in both listening tasks. An added gain of 6 dB to the target audio increased the per-

ceived loudness (Fig. 15) in both use cases, which verifies that the added gain was perceivable for the selected audio material. A 6-dB increase in gain of the interfering audio also affected perceived loudness but by decreasing the loudness perception of the target audio (Fig. 15). Besides loudness, variance analysis for the audio-on-audio scenario (Table 3) shows that gain adjustment is a significant factor in the ANOVA models for attributes clean, width, midrange strength, annoyance, and liking for the entertainment task

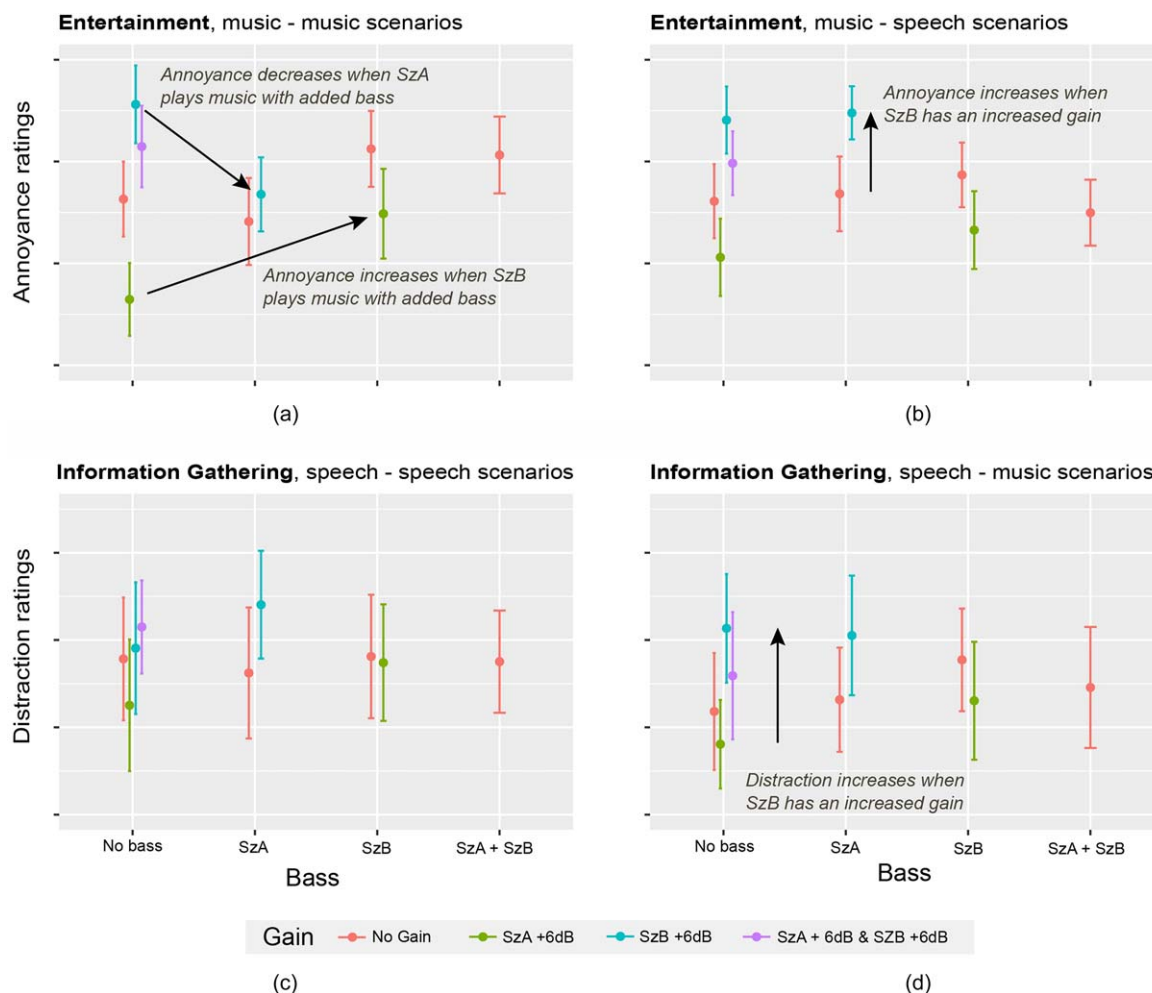


Fig. 14. The effects of audio content (subplots), bass (x-axis), and gains (colors) on interference in scenarios (4 and 5) with two active sound zones. Ratings of interference are represented by the attribute annoyance for entertainment use cases (scenario 4) with music as the target audio [(a) and (b)], whereas distraction represents interference in information gathering use cases (scenario 5) where speech is the target audio [(c) and (d)].

and for distraction in the information gathering task. For perceived liking in the entertainment task, the gain interacts with the scenario. The gain is not a significant factor for how understandable speech stimuli are (Table 3, right column).

2.3.3 Bass Enhancement

To enhance bass perception and evaluate the effects of bass on overall liking of stimuli, a subwoofer setup was added in some scenarios. The variance analysis for each attribute for audio-on-audio scenarios (Table 3) shows that the modifications of subwoofers had a significant effect on the perception of all attributes for entertainment scenarios. For information gathering scenarios, all attributes except distance and understandable include “subwoofer” as a significant factor.

The addition of subwoofers to SzA or both sound zones increased the perceived bass strength compared with having no active subwoofers ($p < 0.001$) or only subwoofers in SzB ($p < 0.001$), although subwoofers in SzB also resulted in significantly higher perceived bass strength ($p < 0.01$)

than having no subwoofers (Fig. 16 top). However, when the target audio was classical music 2 (Mahler), the subwoofer conditions had a less pronounced effect. The perceived bass strength was found to be significantly lower for target audio Mahler compared with the other tracks, including Brahms ($p < 0.05$) when subwoofers were added in SzA. When the bass enhancement was added in both sound zones, the perceived bass strength of Mahler was significantly lower than the perceived bass strength of pop/rock 1 (The Killers) and pop/rock 2 (The Bravery) ($p < 0.05$) but not than Brahms. The differences between target audio content were found to be insignificant for the other subwoofer settings.

When evaluating liking across subwoofer settings (Fig. 16, middle), it is seen that liking is highest when subwoofers were added to SzA or SzA + B. In contrast, the lowest liking ratings are seen when there were either no subwoofers or only subwoofers in SzB in combination with the target audio being a pop/rock music track. The modifications of subwoofers have less of an effect if the target audio was instead a classical music piece. Ratings of annoyance were found to be significantly higher ($p < 0.05$) when subwoofers were

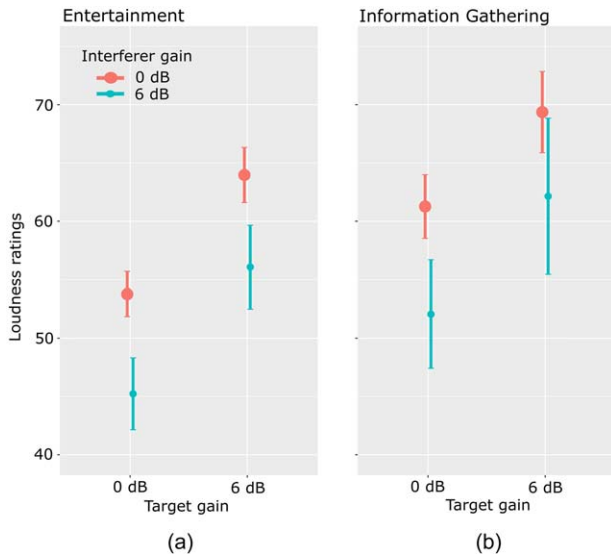


Fig. 15. Perceived loudness in the entertainment (a) and information gathering (b) tasks shown as means with 95% CIs of the scores for variations of target and interferer gains. Note that the y-axis represents the rating scale 0–100 and not decibel values.

added in SzB or both sound zones compared with having only subwoofers in SzA (Fig. 16, bottom), although the perceived annoyance varied with the target audio content, as seen by the significant interaction ($p < 0.001$) between bass and target audio (Table 3).

2.4 Attribute Correlations

2.4.1 Entertainment Use Cases

Across all audio scenarios within the entertainment use case (scenarios 1 through 5; Fig. 8), liking is significantly correlated with all attributes. Positive correlations were found for all other attributes, with the highest positive correlations for clean, width, and bass strength (Spearman's correlation coefficients $\rho = 0.52, 0.41, \text{ and } 0.39$, respectively; $p < 0.001$). Negative correlations were found with canny, nasal, and annoyance ($\rho = -0.40, -0.27, -0.22$; $p < 0.001$). Annoyance was also found to be negatively correlated with width, clean, and midrange strength ($\rho = -0.15, -0.33, -0.37$; $p < 0.001$).

For audio-on-audio scenarios, ratings of liking were found to be significantly correlated with all attributes except annoyance ($\rho = -0.013$; insignificant p value). Liking has negative correlations with nasal and canny ($\rho = -0.14$ and -0.28 ; $p < 0.001$) and positive correlations with all other attributes, with the highest positive correlation being with clean ($\rho = 0.42$; $p < 0.001$). Bass strength and boomy were positively correlated ($\rho = 0.56$; $p < 0.001$), with bass strength having a higher correlation with liking ($\rho = 0.32$; $p < 0.001$) compared with boomy ($\rho = 0.25$; $p < 0.001$). Annoyance was negatively correlated with clean ($\rho = -0.15$; $p < 0.001$) and slightly positively correlated with width ($\rho = 0.098$; $p < 0.05$) and boomy ($\rho = 0.082$; $p < 0.05$) but otherwise not significantly correlated with other attributes.

2.4.2 Information Gathering Use Cases

When correlating attributes for all audio scenarios (scenarios 1 through 5; Fig. 8) within the information gathering use case, ratings of were are positively correlated with ratings of natural, loudness, and understandable ($\rho = 0.25, 0.26, \text{ and } 0.33$, respectively; $p < 0.001$) and negatively correlated with distraction ($\rho = -0.34$; $p < 0.001$). Distraction was also significantly negatively correlated with loudness and understandable ($\rho = -0.43$ and -0.28 , respectively; $p < 0.001$). For audio-on-audio scenarios, liking is significantly correlated with understandable ($\rho = 0.30$; $p < 0.001$), natural ($\rho = 0.20$; $p < 0.001$), loudness ($\rho = 0.13$; $p < 0.01$), and distance ($\rho = 0.12$; $p < 0.05$), whereas distraction was not significantly correlated with liking ($\rho = -0.054$). Instead, distraction was found negatively correlated with understandable ($\rho = -0.14$; $p < 0.01$) and loudness ($\rho = -0.15$; $p < 0.01$) and positively correlated with distance ($\rho = 0.14$; $p < 0.01$). Loudness was positively correlated with understandable ($\rho = 0.28$; $p < 0.001$).

3 DISCUSSION

3.1 Optimizing the Experimental Method

3.1.1 Assessor Scale Use and Consistency

Assessors maintained their motivation and rating consistency (Fig. 11) throughout the listening tests, despite a relatively long test duration of 225–360 min. This was likely achieved through a combination of factors including the following: 1) assessors being able to determine the pace of the listening test themselves, 2) the option to pause the test and have breaks approximately every 30 min, 3) availability of sweets and water, 4) time-based payment for their participation, and 5) sound exposure levels set to an optimal loudness that would allow for listening in longer periods without fatigue.

All assessors used the scale outside the anchor points at 15 and 85 score points, suggesting that the anchor points were successful in allowing the assessors to navigate the scale without restricting scale use. However, scale use may have been slightly restricted by the default position of the slider because rating distributions were found to be bimodal for most attributes (see SEC. 2.1). This may be caused by assessors either choosing to drag the slider up or down from the default at a central score of 50 and seldom staying or coming back to the default position. Also, the slider for each stimulus had to be activated by movement before being able to continue to the next stimulus.

Yet, bimodality is natural and can be seen in many processes and populations, and as all attributes were prone to the same effects of the user interface, the data still provide the possibility to investigate relative perceptual differences of the presented stimuli and correlations between attributes. To investigate further whether the bimodal distributions are an effect of the given audio scenarios and not influenced by the experimental setup, by the experimental setup, the slider could be placed at either end of the scale or randomly in future listening tests.

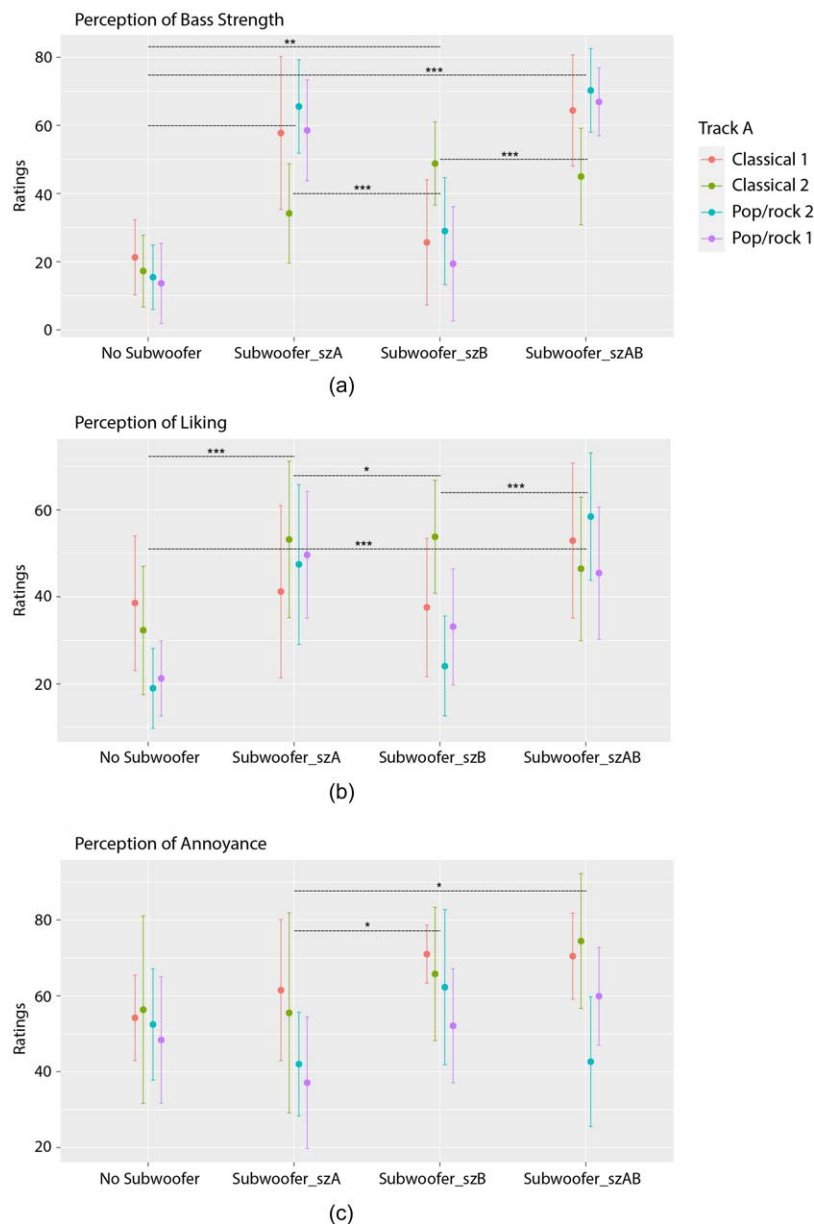


Fig. 16. Perceived bass strength (a), liking (b), and annoyance (c) shown as means with 95% CIs for variations of bass and target audio content in SzA. The target audio for each bass condition from left to right is: Brahms, Mahler, The Bravery, and The Killers (see also color legend on right). Results are for music-music scenarios (scenario 4) within the entertainment listening task and only include perception of stimuli without added gain (i.e., 0 dB gain in both sound zones). Horizontal dashed lines indicate Tukey-adjusted p value significance between subwoofer modifications calculated as paired contrasts between estimated marginal means averaged across target audio. A 5% criterion for statistical significance is indicated as an asterisk (*) ($p < 0.05$). In cases with higher significance levels, this is indicated by a double asterisk (**) ($p < 0.01$) and a triple asterisk (***) ($p < 0.001$).

3.1.2 Attributes Relevant for Sound Zones

Based on attribute correlations, bass strength and boomy are found to be moderately correlated ($\rho = 0.56$; $p < 0.001$). Bass strength has the highest correlation with ratings of liking ($\rho = 0.32$; $p < 0.001$), suggesting that this attribute is the better representative for audio characteristics focused on bass. Furthermore, the test of scale reliability (Cronbach's alpha) also indicate a slight improvement if removing boomy from the attribute list, also highlighting that bass strength is preferable. Within midrange attributes (Table 2), canny is found to be more correlated with lik-

ing ($\rho = -0.28$) than midrange strength and nasal, which are all within the timbral attributes in the DELTA sound wheel [22], suggesting that canny could be representative of timbral characteristics in sound zones in a condensed test paradigm.

Attributes annoyance and distraction are found to be negatively correlated with liking when evaluating ratings across all scenarios (SECS. 2.4.1 and 2.4.2). However, when focusing on audio-on-audio scenarios in sound zones, these two attributes are not significantly correlated with liking. The lack of a correlation may be due to scenarios without in-

terfering audio (scenarios 1, 2, and 3) occupying the lower part of the scale, which may have resulted in a clustering of ratings for stimuli with interfering audio. This effect could likely be avoided by including only audio-on-audio scenarios in future listening tests.

The attribute width was previously shown not to be understandable [11] yet evaluated as relevant for sound zone scenarios and included here with the original description to test the attribute in this customized listening test setup. Here, width is significantly predicted by the physical factors (Table 3), which indicates that the attribute is relevant for evaluations of sound zone scenarios. Furthermore, the attribute has one of the highest positive correlations with liking (SEC. 2.4.1). Yet, the inconsistent ratings by one assessor still point to this attribute being difficult to comprehend for some assessors.

3.2 Effects of Physical Factors on Basic Audio Quality and Audio Quality of Experience

All evaluated attributes are found to be significantly influenced and thus sensitive to the physical changes in the sound field. These attributes are therefore relevant when describing the perceptual changes in the field. Overall, scenarios 1, 2, and 3, with wide-angle audio and no interferer, have more positive ratings (e.g., higher liking and lower annoyance; Fig. 13) than when an interfering audio was present in sound zones. It is not possible to evaluate how sound zones may improve the listening experience compared with wide-angle audio because only the sound zone scenarios included an interfering audio. This will be investigated further in part 2 [3]. Within the scenarios, the attribute ratings also depended on the audio content, gain, and bass enhancement. The variance analyses (Table 3) conducted for all attributes show explained variance of 19–55% by the physical factors, indicating that additional factors that are not included here influence the attribute ratings.

3.2.1 Gain

As expected, an increase in gain, and therefore reproduction level, of the target audio caused a higher perceived loudness (Fig. 15). However, when the interfering audio had an added gain of 6 dB, the perceived loudness was also affected and resulted in a decreased perception of loudness for both music and speech scenarios. As a result, when both audio streams were amplified by 6 dB, the perceived loudness was lower than when only the target audio had an added gain. This could reflect a masking effect of the interferer resulting in the target audio being perceived as played at a lower level. For speech scenarios, masking can lead to decreased speech intelligibility and could thereby be reflected in lower ratings of understandable. Yet, the variance analysis shows that gain is not a significant factor of understandable, which suggests that the speech intelligibility is not compromised. This is likely due to the loudness of stimuli being within an optimal range, resulting in a sufficient TIR to decode the speech signals regardless of gain settings. The changes in gain significantly influenced ratings of attributes clean, width, and midrange strength

and annoyance/distraction and liking within both use cases (Table 3), highlighting that TIR and reproduction level are important for assessing both BAQ and audio QoE.

3.2.2 Bass Enhancement

Adding bass through subwoofers to the reproduction of target audio in SzA increased perception of bass strength and boomy (Table 3). For both attributes, the effect of the subwoofer setup interacted with the target audio content and scenario. For music-music scenarios (Fig. 16), it is seen that, when adding subwoofers to SzA or both sound zones, perceived bass strength is lower for the classical target audio Mahler. This highlights that the effects of bass enhancements depend on the availability of low-frequency components in the audio material. The fact that perceived bass strength and liking is comparable when adding subwoofers to only SzA and to both sound zones whereas annoyance increases with bass enhancement in both sound zones ($p < 0.05$) points to the trade-off between including bass for optimizing audio quality and minimizing leakage of interfering audio into the listener's sound zone. The attributes annoyance and distraction are both significantly affected by the changes in subwoofer settings, but subwoofer has a larger effect on annoyance evaluated for music scenarios ($p < 0.0001$) than on distraction evaluated for speech scenarios ($p < 0.01$) (Table 3), likely due to the more broadband characteristics of the music stimuli compared with the speech stimuli.

4 CONCLUSION

The primary research objectives were as follows: 1) to test and optimize a modified MUSHRA method, including assessor scale use for the chosen experimental factors, 2) to evaluate the consistency of ratings within the specific conditions applied, 3) to identify the relevant attributes for evaluation of sound zones within the specific conditions applied, and 4) to establish the effects on BAQ and audio QoE of the physical factors audio content, bass, and gain for combinations of TIRs.

The first research goal was to test a modified MUSHRA method. The results have shown that a listening test paradigm, such as a modified MUSHRA where assessors are asked to rate attributes on a scale 0–100 is suitable for capturing perceptual and affective differences for sound zones systems and scenarios similar to the ones investigated. Indications were found that the starting position of the slider on the rating scales may influence attribute rating distributions, and therefore, it is suggested that the slider be placed randomly or at either end of the scale. Because a starting position in the middle of scale seems to produce bimodal distributions, future experiments within the ISO-BEL project will use a starting position at 0. The effects of a starting position at 50 will need further investigation because the starting position of the slider was not included as a factor in the experiment.

The second research question was to evaluate the consistency of the assessor ratings. It was found that assessors

in general maintained a similar consistency over time; the results showed no significant differences in the rating consistency of two identical stimuli (original and hidden repeat) for all attributes except canny. When canny was presented in the last part of the listening test, the attribute had less consistent ratings than when it was presented in the first part.

The third research question was to identify the relevant attributes for the investigated scenarios. The results have shown that all evaluated attributes were found relevant for the chosen sound zone scenarios, which included modifications of audio content, gain, and bass enhancement for a range of TIR values. However, it was found that attributes boomy, nasal, and midrange strength can be excluded because these characteristics are already well represented with bass strength and canny. The attributes annoyance and distraction are found significantly important for describing the chosen changes of physical factors in sound zone scenarios. The selection of attributes for the future experiments within the ISOBEL project will be based on these results.

The fourth research question was to quantify the effects of the physical factors tested on BAQ and audio QoE. The results show that, to obtain high-resolution data with the aim of optimizing audio QoE in sound zones, the selection of stimuli should be focused on audio-on-audio scenarios in future studies. Furthermore, sound zone scenarios should be evaluated together with wide-angle audio-on-audio scenarios to better understand the perceptual benefits of focusing sound in zones. The results also show that the TIR of the target and interfering audio influence the perceived loudness, but the selected reproduction level maintains the intelligibility of speech stimuli regardless of gain settings. Furthermore, adding bass to the target audio positively affects liking, whereas adding bass to the interfering audio increases annoyance, pointing to the trade-off between bass quality and sound zone separation. The following experiments within the ISOBEL project [3][12] will thus include further examination of all of the above discussed physical factors.

5 ACKNOWLEDGMENT

This project was funded by the Innovation Fund Denmark (Project Number 9069-00038B) and the ISOBEL project partners. Thanks to Dr. Martin Bo Møller (Bang & Olufsen) and Claus Vestergaard Skipper (Aalborg University) for technical support with the audio setup, Dr. Geoff Martin (Bang & Olufsen) for listening and eliciting attributes, Anne S. S. Larsen (Bang & Olufsen) and Peter M. Nielsen (Bang & Olufsen) for carrying out the experiments, the assessors for their dedicated participation, and Dr. Jon Francombe (Bang & Olufsen) and two anonymous reviewers for their valuable feedback on the manuscript.

6 REFERENCES

- [1] World Health Organization (WHO), "Environmental Noise," in *Compendium of WHO and Other UN Guidance on Health and Environment, 2022 Update Update*, (World Health Organization, Geneva, Switzerland, 2022).
- [2] Global Council on Brain Health, "Music on Our Minds: The Rich Potential of Music to Promote Brain Health and Mental Well-being," <https://www.GlobalCouncilOnBrainHealth.org>. (2020).
- [3] L. Hermannsen and S. Bech, "Perceptual Evaluation of Dynamic Sound Zones, Part 2 - Audio Quality of Experience in a Simulated Hospital Environment," unpublished document to be submitted to *Journal of the Audio Engineering Society* (2025).
- [4] W. F. Druyvesteyn and R. M. Aarts, "Personal Sound," *J. Acoust. Soc. Am.*, vol. 96, no. 5-Supplement, pp. 3281 (1994 Nov.). <https://doi.org/10.1121/1.410932>.
- [5] A. Raake and S. Egger, "Quality and Quality of Experience," in S. Møller and A. Raake (Eds) *Quality of Experience: Advanced Concepts, Applications and Methods*, T-Labs Series in Telecommunication Services, pp. 11–33 (Springer, Cham, Switzerland, 2014).
- [6] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application* (John Wiley & Sons, Hoboken, NJ. 2006). <https://doi.org/10.1002/9780470869253>.
- [7] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, et al., "Factors Influencing Quality of Experience," in S. Møller and A. Raake (Eds), T-Labs Series in Telecommunication Services, pp. 55–72 (Springer, Cham, Switzerland, 2014). https://doi.org/10.1007/978-3-319-02681-7_4.
- [8] T. Walton and M. Evans, "The Role of Human Influence Factors on Overall Listening Experience," *Qual. User Exp.*, vol. 3, p. 1 (2018 Jan.). <https://doi.org/10.1007/s41233-017-0015-4>.
- [9] R. Jacobsen, K. Skov, S. Johansen, M. Skov, and J. Kjeldskov, "Living with Sound Zones: A Long-term Field Study of Dynamic Sound Zones in a Domestic Context," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* p. 405 (Hamburg, Germany) (2023 Apr.). <https://doi.org/10.1145/3544548.3581535>.
- [10] K. F. Skov, P. A. Nielsen, and J. Kjeldskov, "Tuning Shared Hospital Spaces: Sound Zones in Healthcare," in *Proceedings of the 18th International Audio Mostly Conference*, pp. 63–70 (Edinburgh, UK) (2023 Sep.). <https://doi.org/10.1145/3616195.3616198>.
- [11] L. Hermannsen and S. Bech, "Perception of Audio Quality and Audio-on-Audio Interference in Sound Zones," in *Proceedings of the 14th International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–4 (Lippstadt, Germany) (2022 Sep.). <https://doi.org/10.1109/QoMEX55416.2022.9900892>.
- [12] L. Hermannsen and S. Bech, "Perceptual Evaluation of Dynamic Sound Zones, Part 3 - Audio Quality of Experience in Real and Simulated Home Environments," unpublished document to be submitted to the *Journal of the Audio Engineering Society* (2024).
- [13] J. Rämö, S. Bech, and S. H. Jensen, "Validating a Real-time Perceptual Model Predicting Distraction Caused by Audio-on-Audio Interference," *J. Acoust.*

Soc. Am., vol. 144, no. 1, pp. 153–163 (2018 Jul.). <https://doi.org/10.1121/1.5045321>.

[14] J. Rämö, S. Marsh, S. Bech, R. Mason, and S. H. Jensen, “Validation of a Perceptual Distraction Model in a Complex Personal Sound Zone System,” presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), paper 9665.

[15] M. B. Møller and M. Olsen, “Sound Zones: On Envelope Shaping of FIR Filters,” in *Proceedings of the 24th International Conference on Sound and Vibration*, pp. 613–620 (London, UK) (2017 Sep.). <http://dx.doi.org/10.2514/6.2017-1428>.

[16] F. Olivieri, F. M. Fazi, P. A. Nelson, M. Shin, S. Fontana, and L. Yue, “Theoretical and Experimental Comparative Analysis of Beamforming Methods for Loudspeaker Arrays Under Given Performance Constraints,” *J. Sound Vib.*, vol. 373, pp. 302–324 (2016 Jul.). <https://dx.doi.org/10.1016/j.jsv.2016.03.005>.

[17] ITU-R, “Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems,” *Recommendation ITU-R BS1534-3* (2015 Oct.).

[18] J. Francombe, R. Mason, M. Dewhurst, and S. Bech, “Determining the Threshold of Acceptability for an Interfering Audio Programme,” presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), paper 8639.

[19] J. Francombe, R. Mason, M. Dewhurst, and S. Bech, “Elicitation of Attributes for the Evaluation of Audio-on-Audio Interference,” *J. Acoust. Soc. Am.*, vol. 136, no. 5, pp. 2630–2641 (2014 Nov.). <https://doi.org/10.1121/1.4898053>.

[20] V. Hansen and G. Munch, “Making Recordings for Simulation Tests in the Archimedes Project,” *J. Audio Eng. Soc.*, vol. 39, no. 10, pp. 768–774 (1989 Oct.).

[21] ITU-R, “BS1770-5 (11/2023), Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level,” *Recommendation ITU-R BS1770-5* (2023 Nov.).

[22] T. Pedersen and N. Zacharov, “The Development of a Sound Wheel for Reproduced Sound,” presented at the *138th Convention of the Audio Engineering Society* (2015 May.), paper 9310. <https://api.semanticscholar.org/CorpusID:106836416>.

[23] M. Maechler, “diptest: Hartigan’s Dip Test Statistic for Unimodality - Corrected,” <https://CRAN.R-project.org/package=diptest> (accessed April 10, 2024).

[24] W. Revelle, “psych: Procedures for Psychological, Psychometric, and Personality Research,” <https://CRAN.R-project.org/package=psych> (accessed March 21, 2025).

[25] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, “lmerTest Package: Tests in Linear Mixed Effects Models,” *J. Stat. Softw.*, vol. 82, no. 13, pp. 1–26 (2017 May). <https://doi.org/10.18637/jss.v082.i13>.

[26] R. M. Hope, “Rmisc: Ryan Miscellaneous,” <https://CRAN.R-project.org/package=Rmisc> (accessed May 2nd, 2022).

[27] A. Kuznetsova, R. Christensen, C. Bavay, and P. B. Brockhoff, “Automated Mixed ANOVA Modeling of Sensory and Consumer Data,” *Food Qual. Pref.*, vol. 40, part A, pp. 31–38 (2015 Mar.). <https://doi.org/10.1016/j.foodqual.2014.08.004>.

[28] B. G. Peterson and P. Carl, “Performance Analytics: Econometric Tools for Performance and Risk Analysis,” <https://CRAN.R-project.org/package=PerformanceAnalytics> (accessed Dec. 9, 2020).

[29] Kline, *The Handbook of Psychological Testing* (Routledge, London, UK, 2000), 2nd ed.

THE AUTHORS



Line Hermannsen

Line Hermannsen has an M.Sc. and a Ph.D. from the Department of Bioscience, Aarhus University, Denmark (Marine Bioacoustics Laboratory and Marine Mammal Research) in underwater noise pollution and effects on marine mammals, including acoustic impacts on mammalian hearing. From 2020 to 2023, she was a postdoctoral fellow at Bang & Olufsen Research and affiliated with Department of Computer Science at Aalborg University, Denmark, working in the ISOBEL project to investigate perception and optimization of sound zones that can create better acoustic environments. Since 2023, Dr. Hermannsen has been working as a postdoctoral researcher at the Technical University of Denmark (Section of Coastal Ecology, DTU Aqua, National Institute of Aquatic Resources) and as a consultant at Sweco Denmark, specializing in marine biology, coastal ecology, marine pollution, bioacoustics, and acoustic impacts on marine life.

•



Søren Bech

Søren Bech retired in 2024 as Director of Research at Bang & Olufsen and is now Director of Research at Systematic. He is a Professor of Audio Perception at Aalborg University. He received an M.Sc. and a Ph.D. from the Department of Acoustic Technology, Technical University of Denmark. Dr. Bech has authored over 52 peer-reviewed journal articles and more than 100 conference contributions. He has been principal investigator (PI) of 19 funded [European Union (EU) and national] international collaborative research projects. He is a Fellow of the Acoustical Society of America and Audio Engineering Society (AES) and is a past Governor and Vice President of the AES and associate technical editor of the AES journal. His research interest includes psychoacoustics and, in particular, human perception of reproduced sound in smaller rooms. Other interests include experimental procedures and statistical analysis of data from sensory analysis of audio and video quality.