OPEN ACCESS Freely available online
J. Riionheimo and T. Lokki, "Movie Sound, Part 1: Perceptual Differences of Six Listening Environments"
J. Audio Eng. Soc., vol. 69, no. 1/2, pp. 54–67, (2021 January/February).
DOI: https://doi.org/10.17743/jaes.2020.0066

Movie Sound, Part 1: Perceptual Differences of Six Listening Environments

JANNE RIIONHEIMO,[1] *AES Associate Member,* **AND TAPIO LOKKI,**[2] *AES Fellow*
(janne.riionheimo@aalto.fi) (tapio.lokki@aalto.fi)

[1]*Aalto Acoustics Lab, Department of Computer Science, Aalto University, Espoo Finland*
[2]*Aalto Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*

The soundtracks of movies are composed and mixed in various listening environments and the final mix is reproduced in cinemas. The variation of electroacoustical properties between the rooms could be significant, and mixes do not translate easily from one location to another. This study aims to elicit the audible differences between six different movie listening environments, which are auralized to an anechoic listening room with 45 loudspeakers. A listening test was performed to determine the attributes that describe the alterations in the sound field between the rooms. Experienced listeners formulated a vocabulary and created an attribute set containing 19 descriptive attributes. The most important attribute was the sense of space when dialogue was evaluated. Moreover timbre and especially brightness were important when music was evaluated. Furthermore, the change of width and clarity of the sound field was considered important.

0 INTRODUCTION

Cinemas are sound environments that are expected to have high-quality sound systems and controlled room acoustics. As immersive audio content becomes more common in movies, consumer expectations for better sound quality will increase. From the filmmakers' point of view, a movie's sound image in a cinema should represent what they have aimed at when mixing the movie in a dubbing stage. The authors have heard anecdotal statements among film sound engineers in Finland about the problems in room-to-room consistency between mixing environments and cinemas. Similar problems have also been presented in literature [1–3]. These problems make mixing more difficult as the sound does not easily translate from one location to another. The motivation for our research is to better understand the perceptual aspects of the problems in the translation.

0.1 Background

To overcome the problem of translation, the electroacoustic response of the sound system in a cinema or dubbing stage (mixing room) is usually calibrated according to the SMPTE standard 202:2010 [4] or the ISO equivalent 2969:2015 [5], and the sound pressure levels according to the SMPTE recommended practice 200:2012 [6] or the ISO equivalent 22234:2005 [7]. The goal of standardization is to "have constant perceived loudness and frequency response from installation to installation, and from position-to-position within an installation" [4]. The calibration is performed by passing pink noise through the sound system one loudspeaker at a time and measuring the sound pressure level and steady-state magnitude response in 1/3-octave bands with an omnidirectional microphone at the cinema's reference position or practically inside an area 2/3 of the distance from the screen to the rear wall of the rooms. According to the measurement, the sound level is adjusted to a reference level, and the sound system is equalized to fulfill a specific target curve, the X-curve. The curve has a downward slope of 3 dB/oct at high frequencies from 2 to 10 kHz and 6 dB/oct above 10 kHz. Low frequencies from 50 Hz downward are also attenuated by 3 dB/oct. Additionally, the high-frequency slope is gentler for small cinemas and steeper for very large cinemas. The size is defined as the number of seats, ranging from 30 to 2,000. Allen [8] more closely reviews the origins of the curve, and the evolution of cinema calibration is comprehensively summarized by Gedemer [2].

As mentioned by Newell [1], a growing number of professionals consider the standard obsolete, and a better room-to-room compatibility could be achieved with more state-of-the-art methods. Gedemer [9] has reported similar findings, where the views about the X-curve and translation were surveyed from 35 re-recording mixers from 12 different countries. Although the vast majority of mixing

engineers felt that the translation from the dubbing stage they use to a commercial cinema was above average or excellent, most mixers claim to compensate the X-curve somehow, mainly boosting the high frequencies in music and dialogue. Gedemer wonders if both adaptation and learning help the mixers cope with the translation. The movie sound professionals in this study reported using the stabilization practice [10, pp. 43–45] as a remedy for the problematic translation. The mixing personnel watch and critically listen to an almost finished film cut in a large cinema, after which they fine-tune the mix based on their observations. This practice has been used primarily when working in a new mixing environment for the first time, with a new audio system or audio format. After learning to compensate for the differences, stabilization is not used regularly.

The detailed psycho-acoustical documentation behind the X-curve seems to be missing [11]. SMPTE standard 202:2010 argues that "all published experimenters have found that in a large room, a flat response near-field loudspeaker is subjectively best matched by a distant loudspeaker having an apparent high-frequency roll-off when assessed with steady-state measurements." Still, the only listed reference is an article by Allen [8], where he presents a listening test completed by Dolby in 1971 at the Elstree dubbing stage in the U.K. In the test, near-field monitors with close to flat frequency response were positioned close to the console (1.8 meters). Then the far-field stage loudspeakers 12 meters away were equalized to match timbrally with the flat near-field monitors by listening to both dialogue and music. In that particular dubbing stage, the best match was achieved with high and low-frequency roll-off, like in the X-curve, when the steady-state response was measured. In search of more reliable scientific evidence for the X-curve, Gedemer goes through several related articles in [2] and concludes that "the X-curve remains somewhat shrouded in confusion as to its origins and continued perpetuation."

Current measurement techniques used for cinema calibration can reveal only a steady-state timbral response containing the direct sound and the reflections and reverberation. No information about the temporal, spatial, and directional characteristics of the sound field is elicited. As suggested by Toole [11] and Newell [1], the ear and brain can "hear through" the acoustics of a room as the direct sound plays an essential role in the sound field perception. The steady-state response cannot reveal the response of the direct sound in a reverberant environment. Thus a similar steady-state result in different cinemas is achieved with different direct sound responses. Toole goes through a lot of anecdotal and scientific references [11] and sums up that if the target is a flat direct sound response, the current X-curve high-frequency roll-off results in a dulling of a sound, and the shape of the curve at low frequencies results in the lack of bass in current cinemas and dubbing stages. In [11] Toole reviews the results from the SMPTE report [3], a survey from Newell et al. [12], and the data from Holman [13], where the steady-state responses from 15 cinemas in the USA and Europe calibrated to the X-curve were measured. He finds a trend toward boosted bass, which could result from the calibrator's subjective tweaking inside the X-curve tolerances to compensate the bass weak direct sound. However the high frequencies were attenuated even more than the X-curve recommendation, the cause of which remains unclear.

Considering the previous issues, Newell proposed that the installed cinema speakers should be measured in the near field. Similarly, Toole suggested starting by predicting the final response from the comprehensive anechoic data on the loudspeakers. The flat direct sound response is the target for both writers.

This study aims to elicit sonic differences between mixing rooms and cinemas using a listening test with movie sound professionals. As the number of major studio facilities has fallen in each main media center, the audio production is increasingly done in small units or home studios [14]. The film sound engineers in this study also reported on this kind of trend in Finland. When the creative sound design and pre-mixing are increasingly taking place in smaller units, it is essential to consider the perceptual differences between production rooms, dubbing stages, and cinemas. This information would also help develop simulating tools that allow mixing personnel to listen to how the mix sounds in final listening environments. The approach of this study is practical. Technical choices arise from the need to study authentic movie rooms with authentic movie sound production systems and authentic movie program items.

1 EXPERIMENT SETUP

In order to reliably evaluate the differences in the sound field between different environments, two listening tests were composed so that the same program material could be listened to in six different rooms. Three mixing rooms and three cinemas were chosen for the test. The impulse responses of 5.1 or 7.1 loudspeaker channels were measured in reference positions in the rooms with a microphone array consisting of 6 omnidirectional microphones in a symmetric setup shown in Fig. 1. The impulse responses were analyzed with the Spatial Decomposition Method (SDM) [15], and spatial room impulse responses were synthesized. The program selections from original movie soundtracks with 5.1 or 7.1 audio were auralized using the SDM data and reproduced in an anechoic chamber containing 45 loudspeakers and used as a stimulus set in the listening tests. So-called immersive sound formats like Dolby Atmos or DTS:X have been omitted from the study, as the number of cinemas and mixing rooms with appropriate sound systems were small during the measurements.

This section describes the six rooms, measurement techniques, post-processing, and reproduction method for the listening tests.

1.1 Measured Listening Environments

Generally, the mixing process takes place on a dubbing stage, which is a mixing facility assembled in a cinema or cinema-like environment [16]. However the movie sound

Fig. 1. The measurement microphone array used for the SDM analysis. The array consists of six omnidirectional microphones (G.R.A.S., Type 50 VI-1 with three pairs of 40AI microphones) in a symmetrical setup where the distance between the pair was 50 mm. The top microphone was used as a pressure signal in the analysis.

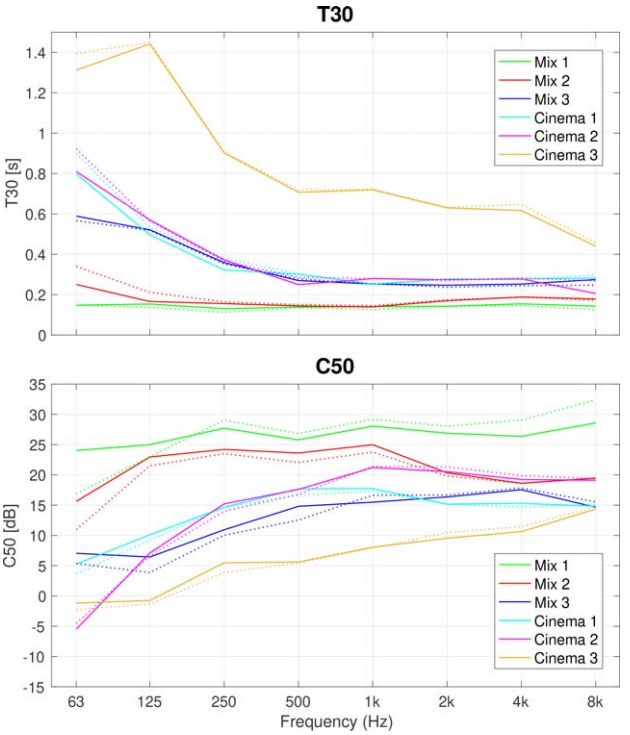


Fig. 2. Reverberation time T_{30} and clarity C_{50} of the measured rooms. Values are average of left, center, and right speakers and measured at the listening position. Reverberation time and clarity from the room auralizations are shown in dotted lines.

production is increasingly done in stages in different-sized units. At least in the Nordic countries, it is common for an engineer to spend just a few days in a large dubbing stage to tweak the details, instead of staying for weeks in a cinema-like environment. It is also possible to do the whole mixing in a small unit. The six rooms that were chosen for this study represent the wide range of mixing rooms as well as cinemas.

Three mixing rooms vary in size from a small dry mixing room to a dubbing stage built in a small old cinema. Three other rooms were cinemas, including the largest one in Finland. The dimensional properties and loudspeaker setups are presented in Table 1. The reverberation time T_{30} and clarity C_{50} of the rooms are shown in Fig. 2.

The smallest mixing room, *Mix 1* with the size of 19 m^2, has near-field monitors at a distance of 1.6 meters and the room acoustics are very dry. *Mix 2* is a film school's mixing room with 4.5-meter listening distance and the largest mixing room, *Mix 3*, is practically the only large-scale dubbing stage in Finland. It is constructed to an old small-scale movie theater with an area of 120 m^2 and 7 meters of listening distance.

Cinema 3 has 635 seats and is the largest cinema in Finland. *Cinema 2* has 257 seats and is 1/5 of the volume of *Cinema 3*. *Cinema 1* is a film school's rehearsal cinema with 50 seats.

In all rooms except *Mix 1*, the 7.1 audio reproduction system was used for the measurements in this study. The small production room *Mix 1* had only a 5.1 system that was measured. *Mix 2*, *Mix 3*, and *Cinema 3* were able to reproduce immersive audio content, but the formats were excluded from the study and only a 7.1 system was used for the measurement.

All of the rooms except *Mix 1* were previously calibrated according to the standard SMPTE ST 202:2010 [4], using the X-curve as a target curve for the electroacoustic response. The target for *Mix 1* was a flat response, which is a normal practice in small rooms according to the interviews with the participants. All rooms were calibrated by the owner and not modified for the measurements.

The source of the picture in *Mix 1* is a 55" TV while in other venues the picture is projected to a cinema screen. In *Cinema 3* the screen speakers are located above the screen and the screen is non-perforated while *Cinemas 1* and *2*, as well as *Mix 3*, have a perforated screen and the screen in *Mix 2* is woven. Screen speakers left, right, and center in *Mix 2*, *Mix 3*, *Cinema 1*, and *Cinema 2* are located behind the cinema screen.

The reference measurement positions according to standard SMPTE ST 202:2010 [4] were selected for the listening position in each room. In the mixing facilities, the mixing position in the centerline of the room was used as a listening position. In the cinemas, the listening position was located in the center line in the 2/3 length from the screen to the projector wall.

1.2 Measurement Technique

The level of the playback system in each venue was calibrated with pink noise with the RMS-level of −20 dBFS producing an 85-dB C-weighed equivalent continuous sound pressure level at the listening position as described in [6] and [7]. Impulse responses were measured with a 7-second logarithmic sine-sweep [17] from 1 Hz to 24 kHz with the peak level of −20 dBFS. The measurement signals were recorded with an external laptop computer with MOTU UltraLite-mk3 soundcard at 24-bit/192-kHz resolution unsynchronized with the source signal. As

Table 1. The dimensions, capacity, and loudspeaker setup of six rooms.

Room	A [m^2]	V [m^3]	H [m]	Seats	D [m]	Format	Surround [ss + sr]
Mix 1	19	40	2.4	···	1.6	5.1	1 + 0
Mix 2	37	110	3.0	···	4.5	7.1	2 + 1
Mix 3	120	400	3.4	···	7.0	7.1	4 + 2
Cinema 1	110	750	6.0	50	7.0	7.1	3 + 3
Cinema 2	250	1,400	7.5	257	11.0	7.1	4 + 3
Cinema 3	825	7,000	12.5	635	20.0	7.1	6 + 4

The area (A), volume (V), height (H), number of seats, listening distance from the center speaker (D), sound format, and surround loudspeaker setup of the six rooms. The height is measured at the screen, which is the highest position in the cinemas with raked seating. The listening distance is measured from the center speaker to the listening position that is the 2/3 length from the screen to the projector wall in the cinemas and the mixing position in the mixing rooms. Surround loudspeaker setup is presented as the number of side surround speakers (ss) and rear surround speakers (sr) at both sides of the room. For instance, *Cinema 3* has six loudspeakers on both sidewalls (a total of 12) and eight loudspeakers on the rear wall (four on the left and four on the right side).

the measurement file and recording were unsynchronized, different digital clocks result in varying lengths between the files. Although the difference is minuscule, between 25. . .561 samples along the overall length of the measurement file (20,928,000 samples), it leads to a 3-ms difference between the length of the files at the longest, so the recordings were later re-sampled according to alignment clicks to match in lengths.

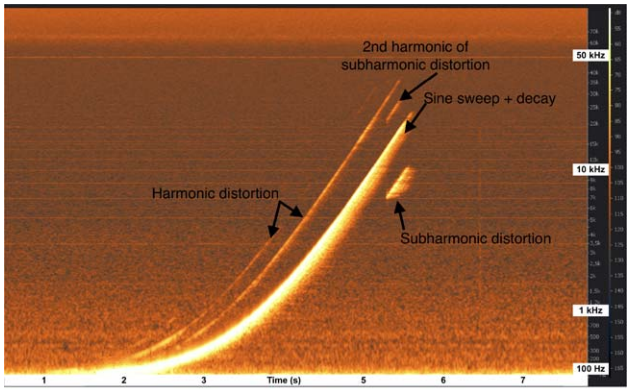
For cinemas, the measurement signals were bundled to a Digital Cinema Package (DCP) that was reproduced by the cinema's audio system for all 7.1 channels successively, i.e., how a movie soundtrack would be reproduced. The measurement DCP-file contained clicks in the beginning and end for aligning the recording with the measurement signals. In mixing facilities, the sweeps were reproduced from the DAW through the audio system as the signals were a movie soundtrack. If the surround speakers were clustered to an array, the measured signal was reproduced by the array, not the individual speaker.

As a result of the measurement process, a spatial impulse response to listener position from each channel of the 7.1 sound system was obtained. These responses were later used in auralization, as explained in SEC. 1.2.2.

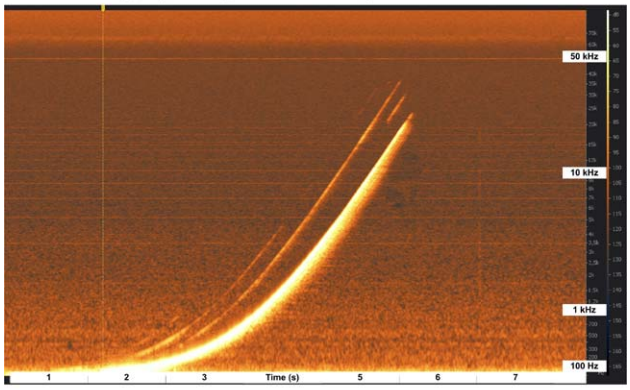
1.2.1 Equalizing and Distortion

The applied SDM method uses pressure values only from one microphone in the array, shown in Fig. 1. Although all the microphones are omnidirectional, the direction of sound incidence affects the frequency response above 10 kHz. In addition, the microphone spacer has a small influence on the response as part of the sound reflects from the spacer even though it is round. As the most critical part of the sound is the direct sound, and the loudspeakers in a different location are mostly in the lateral plane, the impulse response from the top microphone was used as a pressure signal for the SDM analysis and the microphone was equalized to have a linear frequency response at 90 degrees of the sound incident.

As Farina's sweep sine method [17] was used for measuring the impulse response, the harmonic distortion is separated from the impulse response and not incorporated in the convolved sound samples used in the auralizing process.



(a) Spectrogram of a sweep measurement.



(b) Restorated spectrogram of a sweep measurement.

Fig. 3. Visualization of harmonic and sub-harmonic distortions in the spectrogram of a seven-second sweep measurement in *Cinema 2*. In (b) the sub-harmonic distortion is removed with The Izotrope RX 4 software.

The calculated total harmonic distortion of all the measured speakers for the measurement sine sweep signal was below 3% for f $<$ 250 Hz and below 1% for f $\geqslant$ 250 Hz, which is used as an upper limit for a sound production system's harmonic distortion in ITU recommendation ITU-R BS 1116-1 [18]. However, the subharmonic distortion [19] could be found from screen speakers in *Cinema 2*, as can be seen in a spectrogram in Fig. 3(a), resulting in post-ringing in the impulse response [20]. The Izotrope RX 4 software was used to remove the distortion from the recorded sweep

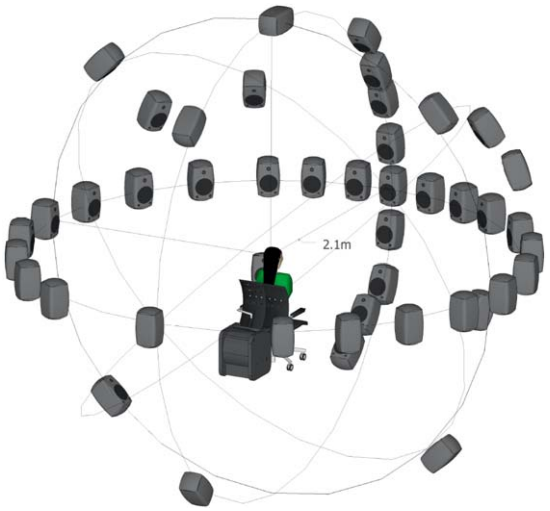


Fig. 4. Wide-frame model of the listening room loudspeaker setup. The 45 loudspeakers are distributed around the listener at 0°, ±10°, ±20°, ±30°, ±45°, ±60°, ±75°, ±90°, ±105°, ±120°, ±135°, ±150°, and 180° on the horizontal plane; at ±10°, ±22°, ±30°, ±45°, 90°, and ±135° on the vertical plane; at ±45° and ±135° at the perpendicular plane; and at (±30°, ±22°) as well as at (±55°, ±22°) positions. A subwoofer is located behind the chair.

[Fig. 3(b)]. The audibility of the subharmonic distortion, as well as the non-linear distortion, is outside of the scope of this study and assumed to be negligible.

1.2.2 Auralization and Listening Room Setup

The SDM was used to analyze the measured impulse responses. The method analyzes the direction of the sound field at every discrete-time sample by means of the six impulse responses. The pressure signal from the top microphone, as explained earlier, was then processed with the directional information to form an SDM-encoded spatial room impulse response for the listening room's loudspeaker setup. The synthesis of the SDM-encoded spatial room impulse responses was implemented with the nearest neighbor (NN) playback, where each SDM sample is played back from the closest loudspeaker concerning the direction parameter associated with the SDM sample. In other words, the SDM analysis takes one impulse response as input and distributes its samples to the reproduction loudspeakers to be used as convolution reverbs.

Instead of the wideband analysis used in the original SDM [15], the analysis was done for octave bands to improve the spatial and timbral accuracy. Moreover the white noise equalization method was used to enhance the tonal characteristics of the sound field [21].

The loudspeaker setup in the anechoic listening room, consisting of 45 loudspeakers (Genelec 8030) and a subwoofer (Genelec 1093A), is depicted in Fig. 4. Loudspeakers are distributed more densely in the front since the hearing is more sensitive in this area, and the main direction of the sound is from the movie screen in cinemas. Since the loudspeakers have a low-frequency limit at 65 Hz, a subwoofer was used to reproduce the frequencies below that. The subwoofer was positioned just behind the listening position, and the calibration of the loudspeaker setup was done to match the frequency responses of auralized rooms to the six real rooms.

The informal listening with two film sound engineers revealed that the horizontal angle of the direct sound varied between the measured rooms and complicated the comparison of other aspects of the sound. Therefore the front speakers (L, C, and R) of all measured rooms were aligned in the same horizontal plane at the ear level. A rotation matrix was formed for each location by locating the loudspeakers from the SDM data and compensating the elevation angles accordingly.

1.2.3 Plausability

The difference in the acoustical data between the original rooms and room auralizations were minor, as can be seen in Fig. 2, where the reverberation time T_{30} and clarity C_{50} from the room auralizations are shown in dotted lines. The average difference in T_{30} is 0.01 seconds between 125 and 8,000 Hz and 0.07 seconds at 63 Hz. The average differences in C_{50} are 0.9 decibels between 125 and 8,000 Hz and 2.9 decibels at 63 Hz. A possible reason for the longer reverberation in the low frequencies is the anechoic chamber's decay on the bass.

Although the anechoic room without image is far from the real theater and the feel of the cinema is missing, the assessors found the auralized acoustical environment very realistic. The identical match between the real rooms and auralizations was not the aim of this study, but the task for the assessors was to elicit the differences between various rooms. In this study, the rooms in question were the auralized ones that meet the criteria for mixing rooms and cinemas. In addition, the acoustical data from the real rooms and auralizations are similar, indicating the auralizations represent the real rooms.

1.3 Program Material

A program material set was selected to contain different spectral, spatial, and dynamical information to reveal different aspects of studied rooms. The criteria for the set was based on the discussions with three film sound engineers. Five different excerpts from movie soundtracks were used as an initial program material set for the listening tests:

Music slow
: Epic orchestral passage from **The Lord of the Rings: Return of the King**; 12 seconds with 5.1 audio

Music fast
: Rhythmic orchestral passage with sound effects from **Tron: Legacy**; 13 seconds with 7.1 audio

Dialogue dry
: Dialogue from **Star Wars: The Force Awakens**; 8 seconds with 7.1 audio

Dialogue wet
: Dialogue with long artificial reverberation from **Jättiläinen**; 11 seconds with 5.1 audio

Ambience

Atmospheric excerpt containing thunder, rain, speech, and interior atmosphere from **Viikossa aikuiseksi**; 23 seconds with 5.1 audio

The selected excerpts differ spatially (center-panned and immersive), dynamically (static and changing dynamics), and spectrally (full and limited spectrum) and have different types and numbers of sound sources (dialogue, music, sound effects, and ambience). All items have information in the low-frequency-effects channel (LFE); however the proportion of the channel is minor in the dialogue excerpts.

The program materials 1–3 were extracted from the Blu-ray, where the lossless DTS-HD Master Audio data was converted to 24-bit, 48-kHz PCM audio data. The final mix PCM audio files were used for the program materials 4–5. The 5.1 audio was converted to 7.1 audio by copying the side surround channels to rear channels and reducing surround channel levels by 3 dB, as in practice most of the movie sound processors do for 5.1 audio that is reproduced in 7.1 sound systems. This upward conversion is also in line with the International Telecommunication Union's recommendation in [22].

As the program materials 1–3 were extracted from the Blu-ray, they could contain some extra audio processing and volume compression comparing the program materials 4–5 that were the original movie soundtracks. However, the program materials were not critically compared to each other, and the absolute timbre or dynamics of the original program materials were not under evaluation. Therefore the applied program materials can be considered equivalent.

The rooms were measured with calibrated signal levels as in the SMPTE recommended practice 200:2012 [6], the aim of which is to equalize the loudness between different listening environments. The aim of this study was to evaluate the differences between the rooms that are calibrated according to standards, so no further loudness equalization was done for the rooms. Only the levels between the program material items were adjusted for comfortable listening ranging from 72 to 78 dB as A-weighted equivalent sound pressure levels as the default level in the listening tests. The levels were verified by informal listening by the authors and the film sound engineer who did not participate in the test. The possible loudness differences between the rooms remained constant.

Each program material was convolved separately with the corresponding 45-channel SDM responses for all 7.1 channels (5.1 in *Mix 1*), forming 30 program material + room combinations used as the listening test stimuli.

2 LISTENING TESTS

The listening tests contained three parts: a pre-test interview, free-elicitation session, and pairwise comparison. The assessors accomplished all the parts within the same session that lasted two to three hours. The assessors were allowed to proceed at their own pace and take a break if needed. The interviews as well as the conversations between the parts were recorded for detailed analysis later on.

2.1 Assessors

Seventeen experienced sound engineers took part in the listening tests. All of them have worked in the Finnish film sound industry in the last five years and together they have been involved in 78% of the mainstream fiction movies (14 out of 18) and 50% of the mainstream documentaries (7 out of 14) that premiered in cinemas in 2017 [23, 24].

All the assessors were interviewed before the actual listening tests to find out the details of their professional careers and what kind of experience with the translation of the mixes they have. The interviews were recorded for further analysis. Many assessors stated that the number of job titles in Finnish movie production depends on the movie and is commonly smaller than listed, for instance in [25]. The tasks could vary from movie to movie and many have jumped from one task to another during their career. In Table 2 the job titles of the assessors at present and during their careers have been listed according to self-reported titles. It is noteworthy that there can be multiple tasks per assessor. The table also shows the length of the professional career of the assessors, averaging 22.3 years.

2.2 Procedure

Two actual listening tests were conducted: free elicitation and pairwise comparison. The user interfaces were created with Max7 software and operated with an iPad in a stand inside the anechoic chamber. A paper questionnaire in a writing tablet was also given to the assessors. The user interface was mirrored to the operator's computer outside the chamber for monitoring possible problems with the test.

2.2.1 Methodology

Several different approaches have been used for eliciting the individual auditory attributes. A comprehensive summary of different methods can be found in [26]. The methods can be divided into two main approaches: direct and indirect elicitation [27]. The indirect elicitation methods separate sensation and verbalization and take advantage of non-verbal methods such as drawing. It would be preferable to communicate features such as locations and represent auditory space [28] and for situations where participants' lexicon is limited. Because this study also includes the timbral, dynamic, and temporal aspects and the assessors were experienced professionals who have been verbalizing auditory events daily, a direct elicitation method was chosen.

The direct elicitation can be further divided into consensus vocabulary techniques and individual vocabulary techniques. In consensus vocabulary techniques, a group of subjects is used to develop attributes and agree on their meaning. The process is time-consuming as multiple group discussions are needed and the goal progressively reaches toward a target. However the result could be a valid and reliable consensus vocabulary as the Audio wheel for reproduced sound presented in ITU-R BS.2399-0 [29] or the assessment parameters presented by the European Broadcast Union [30].

The individual vocabulary techniques allow each participant to develop and employ their own attributes, so no

Table 2. Job titles of the 17 assessors at present and during their career and the length of the professional career.

Job title	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	All
Sound designer/supervising sound editor	e	p	p	p	p			p		p	p					p		8p+e
Audio editor (including adr and dialogue)	e	p	p	e	e			p	e	p	p		e			p		6p+5e
Mixing engineer (music)						p	p							p	p			4p
Re-recording mixer									p	p		p	p					4p
Teacher (lecturer, professor)	p				p	p								p				4p
Production sound mixer	e	p	p	p	e			e	e	e	e					e		3p+7e
Composer			p							e							p	2p+e
Length of the professional career (years)	15	20	23	15	29	20	30	22	30	35	13	39	25	17	15	19	12	22.3

Note. p: present job title; e: earlier job titles.

time-consuming and laborious group discussions and training are needed. The attribute lists are combined and edited to their final form either statistically or through group discussions. The result is a "group" attribute list. Individual vocabulary techniques include Free-choice profiling [31, 32], the Flash profile [33], the Repertory grid technique [34], Audio Descriptive Analysis & Mapping [35, 36], and the Individual vocabulary profile [37].

For this study, a direct elicitation technique with individual vocabulary was chosen, as it is a straightforward and easy to understand concept. First each assessor described the differences in the sound field with their own attributes. After the listening test the attribute lists are combined and edited to their final form, a "group" attribute list. When the vocabulary is not restricted, the attribute pool is likely to be large, so a reduction method for the phrasing has to be used. Also, duplicates, synonyms, and vague terminology have to be handled. In this study the method from [38] is adapted, where the descriptive phrasing was classified in semantic categories emerging from free verbalizations. The standard ITU-R BS.2399-0 [29] has been used as a basis for categorization, but new categories are also formed.

2.2.2 Free Elicitation

During the free elicitation test, the user interface shown in Fig. 5(a) was given to the assessors, where the program materials were arranged row-wise and different rooms column-wise. The order of the rooms has been quasi-randomized and the actual locations marked below the room numbers in Fig. 5(a) were not visible to the assessors. The assessors were able to listen to the 30 different listening test samples at the desired order and pace. The change between samples could be made either by maintaining the position of the playback or by starting from the beginning of the sample. The task was to get familiar with the test samples and describe them freely with familiar vocabulary for the assessor. A form was given to the assessors for the evaluation. Although the assessors were not instructed to compare the listening test samples to each other directly, they were informed about the upcoming pairwise comparison in the second phase of the test and encouraged to also listen to the possible differences between samples. The subjects were told that each of the listening test samples represents either a cinema or mixing facility and the only difference they hear is due to the sound system and acoustics of the venue. The mix remains the same.

The default listening level for the listening test samples was between $L_{Aeq} = 72$–78 dB, but an opportunity was given to the assessors to change the volume at their will. However only one assessor raised the level a few decibels. The duration of the free elicitation session was between 25–40 minutes, depending on the assessor. After the session the form was gone through and clarified if necessary. In addition the participants were given an opportunity to speak freely about the evaluation. The conversations were recorded.

2.2.3 Pairwise Comparison

In the second phase of the test, a pairwise comparison between rooms was performed. The *Dialogue wet* was left out of the pairwise comparison; thus there were four program materials. All the cinemas were compared against all the mixing rooms, forming 36 pairs (3 x 3 x 4). No repetition was used for the pairs. The order of the listening test samples and pairs was fully randomized among the participants. The user interface is shown in Fig. 5(b). The assessors were free to listen to the sample pairs at their own pace. Again, the assessors were able to adjust the volume, but only one assessor raised the level a few decibels.

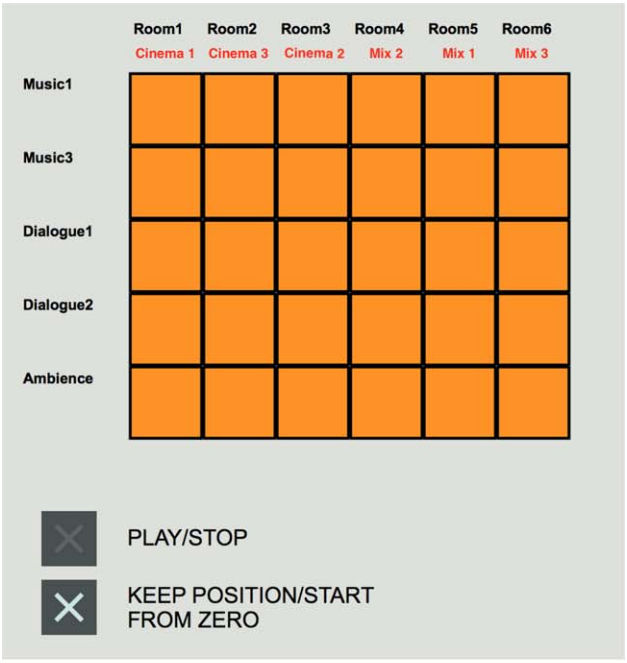
The assessors were told to evaluate the difference between the listening test samples in a pair with 1–4 attributes they are familiar with describing the sound features. The most significant difference was written down first. After evaluating a pair, the ready and next buttons had to be pressed to move to the next pair. The duration of the test was between 25–50 minutes.

3 RESULTS

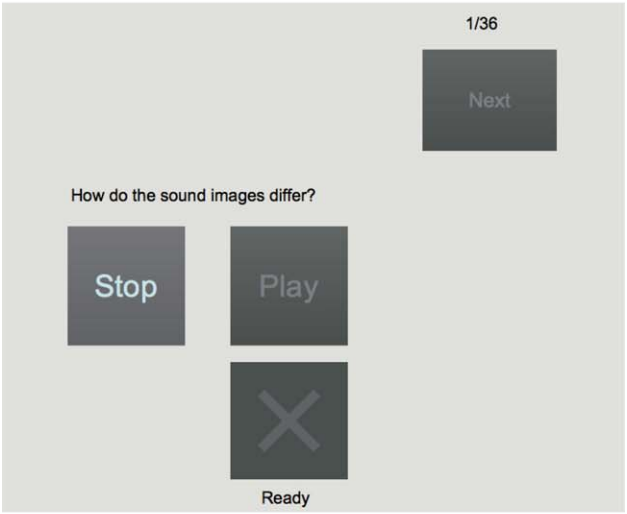
3.1 Free Elicitation

In addition for getting familiarized with the listening environment and stimuli, the task for the test was to describe different rooms freely without any reference. 17 sound engineers gave 825 descriptions for 30 stimuli.

The word count for six rooms and five program materials is presented in Table 3. As can be seen, the assessors used more descriptions for the extremes *Mix 1* with dry acoustics and *Cinema 3* with the longest reverberation. Also, *Mix 3*



(a) Free-elicitation user interface.



(b) User interface for the pairwise test.

Fig. 5. User interfaces for the listening tests. In (a) the room names marked in red below the room numbers were not visible to the assessors. The new listening test sample could be adjusted to start either from the beginning or from the same position as the previous sample with the "keep position/start from zero" button. In (b) the "ready" button had to be pressed before moving on from the "next" button. Assessors wrote the evaluations to the questionnaires.

generated slightly more verbal descriptions than the other rooms.

The descriptions were mostly one or two words long, containing a wide range of content varying from the exact objective description like "pronounced 500 Hz" to highly subjective imagery like "woodish People's House." Mostly the language was technical and exact. The framework for the classification of the descriptions in this study is the attribute list from ITU-R BS.2399-0 [29]. The interviewer

Table 3. Word count in free elicitation for six different rooms and five program material items.

Room	Word count	Stimulus	Word count
Mix 1	158	*Music slow*	208
Mix 2	124	*Music fast*	164
Mix 3	135	*Dialogue dry*	169
Cinema 1	126	*Dialogue wet*	137
Cinema 2	125	*Ambience*	147
Cinema 3	157		
TOTAL	825	TOTAL	825

reviewed all the descriptions with an assessor after the test and asked the assessors to elucidate all vague descriptions.

The 825 descriptions were classified in categories either presented in ITU-R BS.2399-0 or emerging from the descriptions. All descriptions were reduced to their base form and synonyms were grouped. The process was iterative, so some new categories emerged only with one program material item, after which they were added to the category list. If the unclassifiable description was used only once it was excluded from the analysis. In addition nine vague descriptions were abandoned.

Following the method by Samoylenko [39], the attributes were classified into two groups: descriptive and attitudinal, where the attitudinal group is subdivided into "emotional-evaluative" and "naturalness-related" attributes and the descriptive group is divided into unimodal and polymodal attributes. The task of free description did not restrict the use of the attitudinal attributes, so 81 attitudinal attributes were found that make 10% of the whole descriptor pool. These were good (28), natural (25), neutral (14), unpleasant (7), unnatural (6), and artificial (1). Although the attribute natural is included in ITU-R BS.2399-0 under the transparency category, where it is defined as "the sound is similar to the listener's expectation to the original sound without any timbral or spatial coloration or distortion," the descriptions natural and neutral are addressed as attitudinal without any descriptive information in this study. The remaining 744 descriptors were descriptive and unimodal, which is related only to the sense of hearing.

The descriptions for different program material items in all rooms, as well as the descriptions for different rooms with all program material items, are presented in the table in Fig. 6. The attitudinal attributes are included in the table and highlighted in a yellow color.

The free elicitation was accomplished without a reference or an anchor and the assessors were free to choose their listening scheme and the order in which they wanted to listen to the samples. The assessor's listening habits, expectations, and preferences likely influence the individual responses and different context and different bias effects could affect final results [40]. However the assessors were movie sound professionals who listen to movie sound daily and are thus an obvious group to evaluate movie sound analytically. The aim was to form a rich vocabulary, so the personal listening schemes and emphases are beneficial in creating a group vocabulary. When the attribute pool is large

Room Volume	Mix 1 40 m3		Mix 2 110 m3		Mix 3 400 m3		Cinema 1 750 m3		Cinema 2 1400 m3		Cinema 3 7000 m3		All	
Music slow	Bright	19	Good	5	Dark	14	Good	6	Dark	2	Dark	13	Dark 32	15 %
	Narrow	3	Clear	3	Nasal	4	Wide	2	Unclear	2	Nasal	4	Bright 23	11 %
	Intimate	3	Dark	2	Wide	3	Clear	2	Restricted f.r.	2	Boxy	4	Good 12	6 %
	Wide	2	Wide	2	Loud midrange	3	Dry	2	Clear	2	Unclear	3	Nasal 11	5 %
	Restricted f.r.	2	Open	2	Grand	2	Detailed	2	Natural	2	Bright	2	Wide 11	5 %
	Soft bass	2	Grand	2	Boxy	2	Flat f.r.	2			Reverberant	2	Unclear 8	4 %
	Nasal	2	Good bass	2			Natural	2			Loud low midrange	2	Clear 8	4 %
	Good surround	2	Loud midrange	2			Neutral	2					Boxy 7	3 %
	Total words	**42**	**Total words**	**32**	**Total words**	**34**	**Total words**	**35**	**Total words**	**27**	**Total words**	**38**	**Total words 208**	
Music fast	Bright	6	Unclear	2	Unclear	4	Clear	3	Wide	3	Unclear	8	Unclear 18	11 %
	Narrow	3	Grand	2	Dark	4	Unclear	2	Good bass	3	Reverberant	5	Loud surround 7	4 %
	Phasing	2	Soft midrange	2	Loud midrange	2	Soft surround	2	Unclear	2	Distant	2	Dark 6	4 %
	Wide	2	Soft surround	2	Undetailed	2	Boomy	2	Loud surround	2	Loud surround	2	Distant 6	4 %
	Soft bass	2	Narrow	2	Loud surround	2			Punchy	2	Loud low midrange	2	Boomy 6	4 %
			Good	2	Big	2			Precise attack	2	Good surround	2	Reverberant 6	4 %
			Flat f.r.	2							Big	2	Bright 6	4 %
			Phasing	2									Narrow 6	4 %
	Total words	**24**	**Total words**	**28**	**Total words**	**27**	**Total words**	**19**	**Total words**	**29**	**Total words**	**37**	**Total words 164**	
Dialogue dry	Dry	10	Clear	3	Reverberant	5	Natural	4	Clear	4	Reverberant	21	Reverberant 26	15 %
	Narrow	5	Small	3	Dark	4	Room-like	4	Natural	3	Unclear	3	Dry 17	10 %
	Bright	5	Natural	2	Unclear	4	Dry	3	Good	3	Distant	2	Intimate 10	6 %
	Intimate	4	Good	2	Loud surround	2	Distant	2	Intimate	3			Clear 10	6 %
	Small	3	Intimate	2	Room-like	2	Neutral	2	Dry	2			Natural 9	5 %
	Restricted f.r.	3	Dry	2	Big	2	Nasal	2	Neutral	2			Unclear 8	5 %
	Clear	2	Narrow	2			Good	2	Loud upper midrang	2			Narrow 8	5 %
	Unpleasant	2	Detailed	2			Boxy	2						
	Total words	**38**	**Total words**	**23**	**Total words**	**26**	**Total words**	**28**	**Total words**	**24**	**Total words**	**30**	**Total words 169**	
Dialogue wet	Narrow	5	Clear	3	Unclear	4	Nasal	3	Clear	4	Reverberant	9	Reverberant 10	7 %
	Bright	5	Natural	2	Reverberant	3	Natural	3	Intimate	3	Unclear	5	Narrow 10	7 %
	Dry	4	Good	2	Nasal	2	Phasing	3	Narrow	2	Boomy	3	Bright 8	6 %
	Unnatural	2	Metallic	2	Metallic	2	Boomy	2	Natural	2			Clear 8	6 %
	Restricted f.r.	2			Wide	2	Restricted f.r.	2	Loud midrange	2			Natural 8	6 %
	Small	2			Big	2							Unclear 6	4 %
													Boomy 6	4 %
													Nasal 5	4 %
	Total words	**27**	**Total words**	**20**	**Total words**	**23**	**Total words**	**20**	**Total words**	**21**	**Total words**	**26**	**Total words 137**	
Ambience	Bright	4	Clear	2	Nasal	4	Reverberant	2	Narrow	5	Reverberant	4	Nasal 11	7 %
	Narrow	3	Unclear	2	Unclear	3	Bright	2	Intimate	2	Unclear	4	Unclear 11	7 %
	Nasal	3	Well balanced	2	Dark	3	Dry	2	Good	2	Nasal	2	Narrow 9	6 %
	Restricted f.r.	3	Grand	2	Distant	3	Loud midrange	2			Wide	2	Reverberant 8	5 %
	Unpleasant	3			Phasing	2	Loud surround	2			Slap	2	Bright 7	5 %
	Unclear	2			Wide	2							Phasing 6	4 %
	Precise localization	2			Room-like	2							Wide 6	4 %
					Reverberant	2							Restricted f.r. 6	4 %
	Total words	**27**	**Total words**	**21**	**Total words**	**25**	**Total words**	**24**	**Total words**	**24**	**Total words**	**26**	**Total words 147**	
All	Bright	39	25 % Good	12	10 % Dark	25	19 % Natural	10	8 % Clear	12	10 % Reverberant	41	26 % Reverberant 58	7 %
	Narrow	19	12 % Clear	11	9 % Unclear	16	12 % Nasal	8	6 % Natural	9	7 % Unclear	23	15 % Unclear 55	7 %
	Dry	16	10 % Unclear	6	5 % Reverberant	12	9 % Good	8	6 % Narrow	8	6 % Dark	15	10 % Bright 51	6 %
	Restricted f.r.	10	6 % Grand	6	5 % Nasal	10	7 % Neutral	7	6 % Intimate	8	6 % Nasal	7	4 % Dark 49	6 %
	Intimate	9	6 % Phasing	6	5 % Wide	8	6 % Clear	7	6 % Good	7	6 % Distant	7	4 % Narrow 38	5 %
	Nasal	8	5 % Natural	5	4 % Distant	7	5 % Dry	7	6 % Wide	6	5 % Slap	6	4 % Nasal 34	4 %
	Small	7	4 % Narrow	5	4 % Loud midrange	6	4 % Restricted f.r.	5	4 %		Wide	5	3 % Clear 34	4 %
	Unpleasant	7	4 % Detailed	5	4 % Big	6	4 % Room-like	5	4 %		Boomy	5	3 % Dry 30	4 %
	Total words	**158**	**Total words**	**124**	**Total words**	**135**	**Total words**	**126**	**Total words**	**125**	**Total words**	**157**	**Total words 825**	
	Total descriptive	**148**	94 % **Total descriptive**	**106**	85 % **Total descriptive**	**132**	98 % **Total descriptive**	**101**	80 % **Total descriptive**	**106**	85 % **Total descriptive**	**154**	98 % **Total descriptive 744**	90 %

Fig. 6. Free elicitation results. The number of most-used words for all rooms and program material items. Attitudinal words are highlighted in yellow. The percentage of most used words as well as descriptive words are presented in the last row and last column.

the trends of assessments can be investigated by averaging the results.

3.2 Pairwise Comparison

The aim of the pairwise comparison test was the evaluation of the differences between the mixing rooms and cinemas with four program materials. A separation was made between the primary differences that were written first to the questionnaire and the secondary differences that were written to the following rows. The differences were described with one to two-word phrases and a total of 1,145 descriptions were given, of which 544 were the primary descriptions. The word count for the pairwise test is presented in Table 4. On average the 17 sound engineers gave 1.87 descriptions per sample pair.

Again, all the descriptions were classified in categories either presented in ITU-R BS.2399-0 or emerging from the descriptions. All descriptions were reduced to their base form and synonyms were grouped. If a new category emerged with only one specific program material item, the category was added to the category list and the grouping process was done again. The assessors were asked to elucidate any vague descriptions right after the test. Again, the attributes were classified into two groups: descriptive and attitudinal. The question setting in the pairwise comparison limits out the attitudinal attributes effectively, so only three attitudinal descriptors were given: the pleasantness (1 time) and naturalness (2 times) were abandoned in the analysis.

Table 4. Word count in the pairwise comparison for 17 mixing engineers who evaluated 36 sample pairs, in which the 3 mixing rooms were compared to the 3 cinemas with 4 program material items.

Weight	Word count	Stimulus	Word count
1st	544	*Music slow*	294
2nd	411	*Music fast*	311
3rd	157	*Dialogue dry*	259
4th	33	*Ambience*	281
TOTAL	1,145	TOTAL	1,145

Note. 1st: assessors were informed to write the most significant difference first, after which the 2nd, 3rd, and 4th difference could also be written.

Table 5. Attribute list with weighed importance values. Attributes describe the difference between mixing rooms and cinemas. Attributes are presented individually for four different program material items as well as for all program material items.

Attribute	*Music*	*Music*	*Dialogue*	*Ambience*	All
	Slow	*Fast*	*Dry*		
Sense of space	*9%*	**20%**	**53%**	**28%**	**27%**
Brightness	**30%**	**13%**	4%	*9%*	**14%**
Timbre	**15%**	**11%**	*7%*	**11%**	**11%**
Width	**10%**	**10%**	*8%*	*8%*	*9%*
Clarity	*7%*	*9%*	*8%*	*6%*	*7%*
Distance	2%	4%	*5%*	*6%*	4%
Midrange	4%	4%	3%	4%	4%
Surround level	4%	4%	0%	5%	4%
Localization	2%	3%	3%	5%	3%
Bass	1%	6%	2%	1%	3%
Upper midrange	2%	3%	2%	3%	3%
Envelopment	2%	2%	0%	5%	2%
Boxiness	2%	1%	0%	1%	1%
Nasality	3%	0%	1%	1%	1%
Articulation	0%	3%	0%	0%	1%
Presence	0%	0%	2%	1%	1%
Grandiosity	1%	1%	0%	1%	1%
Phasiness	0%	1%	0%	1%	1%
Level	0%	1%	0%	1%	1%

Note. The most important attributes with weighed importance value 10% or more bolded and attributes with weighed importance value between 5. . .9% are in italics.

The procedure mentioned above for classifying and grouping the attributes resulted in 36 attributes. After the attribute list was finished, the total number of occurrences was calculated for each attribute. The primary differences were weighed by factor 1 while the secondary descriptions were weighed by factor 0.5, and the total importance values were calculated for all the attributes. The attributes with fewer than 0.5% importance values were excluded from the analysis. This resulted in 19 attributes total. The attributes as well as the importance values are presented in Table 5.

4 DISCUSSION

4.1 Free Elicitation

In the first test the assessors made no direct comparison between the locations but evaluated all five program items and six rooms freely. A matrix of most used words is presented in the table in Fig. 6.

4.1.1 The Most Common Descriptors

In the free elicitation test the assessors evaluated the rooms most commonly as reverberant, unclear, bright, dark, narrow, nasal, clear, and dry, as can be seen in the last cell of the table in Fig. 6. By combining the opposites, the most common attributes were the timbre (bright and dark), clarity (unclear and clear), and sense of space (reverberant and dry). Looking more closely to individual rooms at the bottom row of the table, it can be seen that the word reverberant is mainly used for *Cinema 3* (71% of the total count 58 for the word) and unclear is mainly used for *Cinema 3* (42%) and *Mix 3* (29%) while bright is used for *Mix 1* (76%) and dark is used for *Mix 3* (51%) and *Cinema 3* (31%). The word narrow is used for *Mix 1* (50%) as well as the descriptor dry (53%). The descriptors nasal and clear are used evenly for all the rooms being almost like opposites: a room is evaluated either nasal or clear.

4.1.2 Rooms

When looking at how the assessors evaluated individual rooms, the extremes *Mix 1* and *Cinema 3* are considered first. The smallest room *Mix 1* with nearfield monitoring is described as bright, narrow, and dry (together 47% of 158 words) while the largest theater *Cinema 3* is described as reverberant, unclear, and dark (together 50% of 157 words). The largest mixing room *Mix 3* is described as dark, unclear, reverberant, and nasal (together 47% of 135 words). The assessors used only a few attitudinal descriptors for these 3 rooms (4% of all 450 words), mainly for *Mix 1*, which was occasionally described as unpleasant and unnatural. Instead, the attitudinal words represent 17% of the total 375 words for the other three rooms *Mix 2*, *Cinema 1*, and *Cinema 2*. Rooms *Mix 1*, *Mix 3*, and *Cinema 3* seem to be more distinctive and evaluated consistently only descriptively as opposed to the other three rooms, which are described both affectively and descriptively. The affective descriptors were, in order of prevalence, clear, good, and natural or less often narrow, dry, or unclear. If we divide rooms into two groups, "distinctive" and "good," it is worth noticing that the assessors used only a few negative attitudinal words like unpleasant for "distinctive" rooms, mainly for the room *Mix 1*, but much more positive attitudinal words for "good" rooms. It seems that more preferred rooms were evaluated more affectively.

In total the assessors used fewer words for "good" rooms (45% of the total 825 descriptors) than for "distinctive" rooms and they also used the different words more evenly. It is notable that although the reverberation time in *Mix 2* is similar in *Mix 1*, only *Mix 1* is evaluated as dry. However, the listening distance in *Mix 2* is 4.5 meters while it is 1.6 meters in *Mix 1*, yet hardly any difference in clarity C_{50} can be seen. *Mix 1* was the only room with a flat frequency response target while other five rooms are calibrated according to the standard SMPTE ST 202:2010 [4], using the X-curve with a downward slope at high frequencies as a target curve for the electroacoustic response. This is the obvious reason for evaluating *Mix 1* as bright and can also lead to describing the room as unpleasant and unnatural alongside other rooms. However the word unnatural comes from the program item *Dialogue wet*, where an artificial reverberation is added to a dry dialogue, which can also be the reason for the description. Although the frequency response contains more high frequencies than in the other rooms, it is also evaluated as restricted.

It is also noteworthy that *Cinema 1* is evaluated as dry and *Cinema 2* is evaluated as intimate while *Mix 3* is evaluated as reverberant and distant even though the reverberation times and listening distances are similar to or even longer than in *Cinema 2*. It is clear that *Cinema 3* is described as reverberant because the reverberation time is more than

twice as long as in the other rooms, but the sense of space seems to also be related to the program material.

4.1.3 The Effect of Program Material

A trend can be seen when looking at the effect of the program material in the last column of the table in Fig. 6. The most common attribute for the slow and epic orchestral passage *Music slow* is the timbre (26%) consisting of the words dark and bright while the most common attribute for the dialogue *Dialogue dry* is the sense of space (25%) consisting of the words reverberant and dry. The other descriptors are more evenly distributed between five program items, unclear being the most used word for the rhythmic orchestral music *Music fast* (11%). It is worth noting that with the *Music slow* the sense of space and with the *Dialogue dry* the timbre are not considered relevant. Looking only at the *Music slow* we can see that the attribute timbre comes mostly from the distinctive rooms; the word dark is used mostly for *Cinema 3* and *Mix 3* while the word bright is used mainly for *Mix 1*. Similarly, in *Dialogue dry*, the sense of space comes from the same distinctive group; the word reverberant is used mainly for *Cinema 3* and *Mix 3* while the word dry is used for *Mix 1*.

4.2 Pairwise Comparison

In the second test, a structured pairwise comparison between three cinemas and three mixing rooms with four program items were carried out. An attribute list containing 19 descriptive attributes in order of importance was composed and presented in Table 5.

4.2.1 The Most Common Attributes

In the pairwise comparison, the five most common attributes were the sense of space, brightness, timbre, width, and clarity. A similar trend to that in the free elicitation can be seen in the program dependency of the attributes. The sense of space represents 53% of all attributes when evaluating *Dialogue dry* while representing only 9% of all attributes when evaluating *Music slow*. On the contrary, attributes brightness and timbre represent 45% of all attributes when evaluating *Music slow* while representing only 11% of all attributes when evaluating *Dialogue dry*. The use of five most-used attributes is more evenly distributed when evaluating *Music fast* while the sense of space represents 28% of all attributes when evaluating *Ambience* containing dialogue and atmospheric sounds.

To find out if the attributes are different without the "distinctive" rooms, the analysis was also done first without extremes *Mix 1* and *Cinema 3*, leaving only four pairs of rooms per program material item and finally without the whole "distinctive" group, leaving only two pairs of rooms to evaluate. The five most common attributes do not change in either of these conditions. Only the percentages and order change, being 14% for the sense of space, 9% for the width, 8% for the brightness, 8% for the timbre, and 7% for the clarity when only the "good" rooms are compared together. When only the extremes are removed from the analysis the five most common attributes and their order remain the same. Also, concerning each program item, the five most common attributes remain the same, except with *Music slow* the clarity is changed with the surround level when "distinctive" rooms are removed.

Assessors evaluated the level difference between the rooms to be insignificant. From this it can be concluded that the SMPTE recommended practice 200:2012 [6] manages to equalize the loudness between rooms of different sizes with different acoustics.

The results from the pairwise comparison match with the free elicitation test. In this study, the free elicitation without a reference seemed to be as reliable as the more structured pairwise comparison, and it also provided descriptive information from the rooms themselves, not just the differences.

4.3 Relation to Other Work

Various studies have proposed descriptive attributes for evaluating the reproduced sound. Gabrielsson [41] investigated the perceived sound quality of loudspeakers, headphones, and hearing aids and suggested eight perceptual dimensions for the sound quality analysis. Eleven attributes were elicited in a study by Berg and Rumsey [42] where the spatial performance of a sound reproducing system was assessed. Choisel and Wickelmaier [43] investigated two different approaches for eliciting auditory attributes and derived eight relevant auditory attributes for evaluating multichannel reproduced sound. Zacharov and Koivuniemi [35, 36] presented eight spatial attributes and four timbral attributes for studying the perceptual nature of spatial sound reproduction systems. Sixteen attributes were developed in a study by Lorho [44], where the quality of spatial enhancement systems for stereo headphone reproduction was investigated. The attributes were divided into three groups relating to localization, space, and timbre. Francombe et al. presented 26 perceptual attributes that contribute to listener preference with experienced and inexperienced listeners when a wide range of spatial audio reproduction methods was investigated [45]. None of the studies have evaluated the perceptual differences between mixing rooms and end listening environments. However all six most important attributes elicited in this study are present in the above studies with the same or an equivalent name. The sense of space and distance is present in all six studies above. Width appears in five and clarity in four studies out of six. Both timbre and brightness appear in three different studies, so one of the two attributes appears in all studies. The significance of an attribute depends on the context of the study.

5 CONCLUSION

This study aimed to elicit sonic differences between movie mixing rooms and cinemas. Two listening tests were performed where experienced listeners evaluated three mixing rooms and three cinemas. First the impulse responses were measured in each room with the sound system used for the movie sound playback. Second the impulse responses were analyzed with the spatial decomposition method and

the spatial impulse responses were synthesized. Finally five program items were auralized to the 45-channel sound system in an anechoic room that enabled the assessors to listen to and compare the rooms sequentially. Two listening tests were performed: a free elicitation and a pairwise test.

The results from both listening tests show that differences in the sense of space, brightness, timbre, width, and clarity as well as in the distance are the most important when comparing cinemas and mixing rooms. The words and attributes used in the listening tests were classified in categories either presented in ITU-R BS.2399-0 or emerging from the descriptions. All 19 attributes in Table 5, except surround level and phasiness, are listed in the ITU-R BS.2399-0 Audio wheel, some with different names. The sense of space (reverberance), brightness (dark-bright), clarity, width, and distance are positioned in the Audio wheel's outermost circumference with the timbre (timbral balance), midrange, and localization in the middle circumference.

The assessors described the differences in the surround level independently, in addition to the width or envelopment. The equivalent for the attribute phasiness is not listed in ITU-R BS.2399-0. The ordinary meaning for the word is that a comb filter's sound resulted from a delayed copy of the sound. It could result from latency in the signal path or a single prominent reflection, for instance, from the mixing console. The attribute is commonly used among mixing engineers. The level differences between the rooms were insignificant, which suggests that the SMPTE recommended practice 200:2012 [6] manages to compensate for the loudness differences between rooms successfully.

The perceptual differences between the rooms are highly dependent on the program material. While the slow and epic full frequency range music is evaluated, the differences in the sound's timbral aspects, especially in the treble, were described as most important. Instead, when a dry dialogue is evaluated, the sense of space was far more critical than the other aspects of the sound. Many of the assessors noticed this difference during the free elicitation test and some even reported that they did not believe that all the samples in one row in the listening matrix in the user interface were from the same room.

Although the largest cinema *Cinema 3* was evaluated mostly with negative words, some of the assessors mentioned that the reverberance made it wider than some other rooms and that the reverberance also uplifts the music although the lack of brightness makes it less preferable. The reverberance of the (large) room is perceived as part of the mix's ambience with the slow music. However, with a dry dialogue, the situation is different, and speech is perceived as too reverberant, unclear, and distant. A listening environment's effect is thus different in different parts of a movie soundtrack, which contains dialogue, music, effects, and ambient sound. The descriptions for the program item *Ambience* were an even combination of the most important words. However the assessors found evaluating the *Ambience* difficult precisely because it contained multiple sound elements. For research purposes a sound sample with explicit content seems to be the best solution.

6 ACKNOWLEDGMENT

This research was partly funded by the Academy of Finland, grant number 296393. We would like to thank all the listening test assessors and Dr. Sakari Tervo for comments and discussions.

7 REFERENCES

[1] P. Newell, K. Holland, J. Newell, and B. Neskov, "New Proposals for the Calibration of Sound in Cinema Rooms," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8383.

[2] L. A. Gedemer, *A New Method for Measuring and Calibrating Cinema Audio Systems for Optimal Sound Quality*, Ph.D. thesis, University of Salford (2017).

[3] SMPTE, "B-Chain Frequency and Temporal Response Analysis of Theatres and Dubbing Stages," *TC-25CSS* (2014).

[4] SMPTE, "Motion-Pictures — Dubbing Theaters, Review Rooms and Indoor Theaters — B-Chain Electroacoustic Response," *Standard 202:2010* (2010).

[5] ISO, "Cinematography – B-Chain Electro-acoustic Reponse of Motion-Picture Control Rooms and Indoor Theatres – Specifications and Measurements," *International Standard 2969:2015* (2015).

[6] SMPTE, "Relative and Absolute Sound Pressure Levels for Motion-Picture Multichannel Sound Systems — Applicable for Analog Photographic Film Audio, Digital Photographic Film Audio and D-Cinema," *Recommended Practice 200:2012* (2012).

[7] ISO, "Cinematography – Relative and Absolute Sound Pressure Levels for Motion-Picture Multi-channel Sound Systems – Measurement Methods and Levels Applicable to Analog Photographic Film Audio, Digital Photographic Film Audio and D-Cinema Audio," *International Standard 22234:2005* (2005).

[8] I. Allen, "The X-Curve: Its Origins and History: Electro-acoustic Characteristics in the Cinema and the Mix-Room, the Large Room and the Small," *SMPTE Motion Imaging J.*, vol. 115, no. 7–8, pp. 264–275 (2006 Jul./Aug.).

[9] L. A. Gedemer, "Evaluation of the SMPTE X-Curve Based on a Survey of Re-Recording Mixers," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), convention paper 8996.

[10] R. Izhaki, *Mixing Audio: Concepts, Practices and Tools*, 2nd ed. (Taylor & Francis, Oxfordshire, England, 2012).

[11] F. Toole, "The Measurement and Calibration of Sound Reproducing Systems," *J. Audio Eng. Soc.*, vol. 63, no. 7/8, pp. 512–541 (2015 Jul./Aug.), https://doi.org/10.17743/jaes.2015.0064.

[12] P. Newell, K. Holland, S. Torres-Guijarro, S. Castro, and E. Valdigem, "Cinema Sound: A New Look at Old Concepts," in *Performing Arts Venue: Balancing the Design,*

Proceedings of the Institute of Acoustics, vol. 32, Pt. 5, pp. 106–126 (Cardiff, UK) (2010).

[13] T. Holman, "Cinema Electro-acoustic Quality Redux," *SMPTE Motion Imaging J.*, vol. 116, no. 5–6, pp. 220–233 (2007 May/Jun.), http://doi.org/10.5594/J11446.

[14] B. Owsinski, *The Mixing Engineer's Handbook*, 3rd ed. (Nelson Education, Toronto, Canada, 2013).

[15] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial Decomposition Method for Room Impulse Responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28 (2013 Jan./Feb.).

[16] Full Sail University, http://getinmedia.com/industry/film-tv.

[17] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion With a Swept-Sine Technique," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), convention paper 5093.

[18] ITU, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," *Recommendation BS.1116-1* (1997).

[19] F. Bolaños, "Measurement and Analysis of Subharmonics and Other Distortions in Compression Drivers," presented at the *118th Convention of the Audio Engineering Society* (2005 May), convention paper 6517.

[20] S. Tervo, J. Saarelma, J. Pätynen, I. Huhtakallio, and P. Laukkanen, "Spatial Analysis of the Acoustics of Rock Clubs and Nightclubs," in *Proceedings of the Ninth International Conference on Auditorium Acoustics* (2015).

[21] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, "Spatial Analysis and Synthesis of Car Audio System and Car Cabin Acoustics With a Compact Microphone Array," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 914–925 (2015 Nov.), http://doi.org/10.17743/jaes.2015.0080.

[22] ITU, "*Multichannel Stereophonic Sound System With and Without Accompanying Picture*," *Recommendation BS.775-2* (2012).

[23] *Suomen Elokuvasäätiö, "Elokuvavuosi 2017 Facts & Figures"* (2012).

[24] IMDb.com, Inc., www.imdb.com.

[25] Full Sail University, http://getinmedia.com/industry/film-tv.

[26] N. Zacharov (Ed.), *Sensory Evaluation of Sound* (CRC Press, Boca Raton, FL, 2019).

[27] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application* (John Wiley & Sons, Hoboken, NJ, 2006).

[28] R. Mason, N. Ford, F. Rumsey, and B. de Bruyn, "Verbal and Non-verbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction," presented at the *109th Convention of theAudio Engineering Society* (2000 Sept.), convention paper 5225.

[29] ITU, "Methods for Selecting and Describing Attributes and Terms in the Preparation of Subjective Tests," *Report BS.2399-0* (2017).

[30] EBU, "Assessment Methods for the Subjective Evaluation of the Quality of Sound Programme Material–Music," Tech Rep. 3286-E (1997 Aug.).

[31] A. A. Williams and S. P. Langron, "The Use of Free-Choice Profiling for the Evaluation of Commercial Ports," *J. Sci. Food Agr.*, vol. 35, no. 5, pp. 558–568 (1984 May), http://doi.org/10.1002/jsfa.2740350513.

[32] A. A. Williams and G. M. Arnold, "A Comparison of the Aromas of Six Coffees Characterised by Conventional Profiling, Free-Choice Profiling and Similarity Scaling Methods," *J. Sci. Food Agr.*, vol. 36, no. 3, pp. 204–214 (1985 Mar.), http://doi.org/10.1002/jsfa.2740360311.

[33] J. Delarue and J. -M. Sieffermann, "Sensory Mapping Using Flash Profile. Comparison With a Conventional Descriptive Method for the Evaluation of the Flavour of Fruit Dairy Products," *Food Qual. Pref.*, vol. 15, no. 4, pp. 383–392 (2004 Jun.), http://doi.org/10.1016/S0950-3293(03)00085-5.

[34] D. M. H. Thomson and J. A. McEwan, "An Application of the Repertory Grid Method to Investigate Consumer Perceptions of Foods," *Appetite*, vol. 10, no. 3, pp. 181–193 (1988 Jun.), http://doi.org/10.1016/0195-6663(88)90011-6.

[35] N. Zacharov and K. Koivuniemi, "Audio Descriptive Analysis & Mapping of Spatial Sound Displays," in *Proceedings of the International Conference on Auditory Display* (2001).

[36] N. Zacharov and K. Koivuniemi, "Unraveling the Perception of Spatial Sound Reproduction: Techniques and Experimental Design," presented at the *AES 19th International Conference: Surround Sound - Techniques, Technology, and Perception, Technology, and Perception* (2001 Jun.), conference paper 1929.

[37] G. Lorho, "Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction," presented at the *119th Convention of the Audio Engineering Society* (2005 Oct.), convention paper 6629.

[38] C. Guastavino and B. F. G. Katz, "Perceptual Evaluation of Multi-dimensional Spatial Audio Reproduction," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1105–1115 (2004), https://doi.org/10.1121/1.1763973.

[39] E. Samoylenko, S. McAdams, and V. Nosulenko, "Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres," *Int. J. Psych.*, vol. 31, no. 6, pp. 255–278 (1996 Dec.), https://doi.org/10.1080/002075996401025.

[40] E. C. Poulton and S. Poulton, *Bias in Quantifying Judgments* (Lawrence Erlbaum Associates, Inc., Oxfordshire, England, 1989).

[41] A. Gabrielsson, "Dimension Analyses of Perceived Sound Quality of Sound-Reproducing Systems," *Scand. J. Psych.*, vol. 20, no. 1, pp. 159–169 (1979 Sept.), http://doi.org/10.1111/j.1467-9450.1979.tb00697.x.

[42] J. Berg and F. Rumsey, "In Search of the Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis of Scaled Verbal Descriptors," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), convention paper 5139.

[43] S. Choisel and F. Wickelmaier, "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound," *J. Audio Eng. Soc.*, vol. 54, no. 9, pp. 815–826 (2006 Sept.).

[44] G. Lorho, "Evaluation of Spatial Enhancement Systems for Stereo Headphone Reproduction by Preference and Attribute Rating," presented at the *118th Convention of the Audio Engineering Society* (2005 May), convention paper 6514.

[45] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 212–225 (2017 Mar.), https://doi.org/10.17743/jaes.2016.0071.

THE AUTHORS

Janne Riionheimo

Tapio Lokki

Janne Riionheimo was born in Toronto, Canada, in 1974. He has studied acoustics and audio signal processing at the Helsinki University of Technology and music technology in Sibelius Academy. He received an M.Sc. degree in 2004 and has since worked as an acoustic consultant and sound engineer. He is currently working as a part-time doctoral candidate at the Department of Media Technology at the Aalto University School of Science under the supervision of Professor Tapio Lokki. In his doctoral research he focuses on movie sound and the role of room acoustics in listening spaces.

•

Born in Helsinki, Finland in 1971, Tapio Lokki has studied acoustics, audio signal processing, and computer science at the Helsinki University of Technology (TKK) and received an M.Sc. degree in 1997 and a D.Sc. (Tech.) degree in 2002. At present Prof. Lokki is an Associate Professor (tenured) with the Department of Signal Processing and Acoustics at Aalto University. Prof. Lokki leads his virtual acoustics team to create novel objective and subjective ways to evaluate room acoustics. In addition the team currently contributes to sound rendering algorithms for virtual reality applications, speech in noise and reverberation-related hearing research, and novel material science, in particular wood fiber-based absorption materials. Prof. Lokki is a fellow of the AES and an honorary member of the Acoustical Society of Finland.