Riionheimo, Janne; Lokki, Tapio

Movie Sound, Part 2: Preference and Attribute Ratings of Six Listening Environments

Published in:
Journal of the Audio Engineering Society

DOI:
10.17743/JAES.2020.0065

Published: 01/02/2021

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Riionheimo, J., & Lokki, T. (2021). Movie Sound, Part 2: Preference and Attribute Ratings of Six Listening Environments. *Journal of the Audio Engineering Society*, *69*(1/2), 68-79.
https://doi.org/10.17743/JAES.2020.0065

OPEN ACCESS Freely available online
J. Riionheimo and T. Lokki, "Movie Sound, Part 2: Preference and Attribute Ratings of Six Listening Environments"
J. Audio Eng. Soc., vol. 69, no. 1/2, pp. 68–79, (2021 January/February).
DOI: https://doi.org/10.17743/jaes.2020.0065

Movie Sound, Part 2: Preference and Attribute Ratings 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*

In this study, the assessors evaluated the alterations in the sound field of six movie listening environments. The sound fields of the listening environments were auralized to an anechoic listening room with 45 loudspeakers so that assessors could compare the rooms with each other directly. 31 experienced listeners evaluated five descriptive attributes on a continuous scale for each room with two program material items, dialogue and music. The preference ratings for the rooms were also collected. The perceptual evaluations were compared to the objective electroacoustic data of the rooms. The sense of space, clarity, and distance match the measured clarity C_{50} at the middle frequencies, while the brightness matches the level of the high frequencies in the electroacoustic response above 4 kHz. No psychoacoustical support was found for the current standards, according to which the high frequencies should be attenuated more in large cinemas with longer reverberation than in small cinemas. It turned out that the movie sound professionals do not prefer either too dead or too live listening environments.

0 INTRODUCTION

Sound engineers mix the movie soundtrack in a mixing room, and the audience listens to it in a cinema. The sonic difference between these listening environments can be significant, and the mix does not always translate easily from one location to another. The recommendations for acoustics of these rooms are given in various books and design guidelines [1–5], and the requirements for the electroacoustic response and sound pressure levels are defined in the standards [6–9]. So far, many papers have presented opinions on how, why, and to what extent sound technicians should follow the recommendations for equalizing the electroacoustic response according to the so-called X-curve [10–13] described in SMPTE standard 202:2010 [7].

Various studies have proposed descriptive attributes for evaluating reproduced sound [14–17, 17–20]. However, none of the studies have evaluated the perceptual differences between mixing rooms and end listening environments. Previously the present authors studied the perceptual differences of six movie sound listening environments [21]. In that study, 17 experienced movie sound professionals elicited 19 attributes to describe the perceptual differences between three mixing rooms and three cinemas. Music, dry dialogue, wet dialogue, and ambient sound were used as the program material. The assessors evaluate the same six listening environments with the five most important attributes from the previous study in the present study. The evaluations are then compared with measured data from the room auralizations. The research questions are:

- How do the evaluations match to the measured parameters from the room auralizations?
- How different are the rooms?
- What kind of rooms do the assessors prefer?
- How to improve the translation between rooms?

The results are discussed with respect to current recommendations and standards. Finally, suggestions are made based on the results.

1 EXPERIMENT SETUP

In this study, the assessors evaluated five descriptive attributes and the preference rating with two program material items, music and dialogue. The auralized listening environments consisted of three cinemas and three mixing rooms. A microphone array consisting of six omnidirectional microphones in a symmetric setup was used to capture the impulse responses of all 5.1 or 7.1 channels in reference positions in the different listening environments. Impulse responses were then analyzed by the Spatial Decomposition

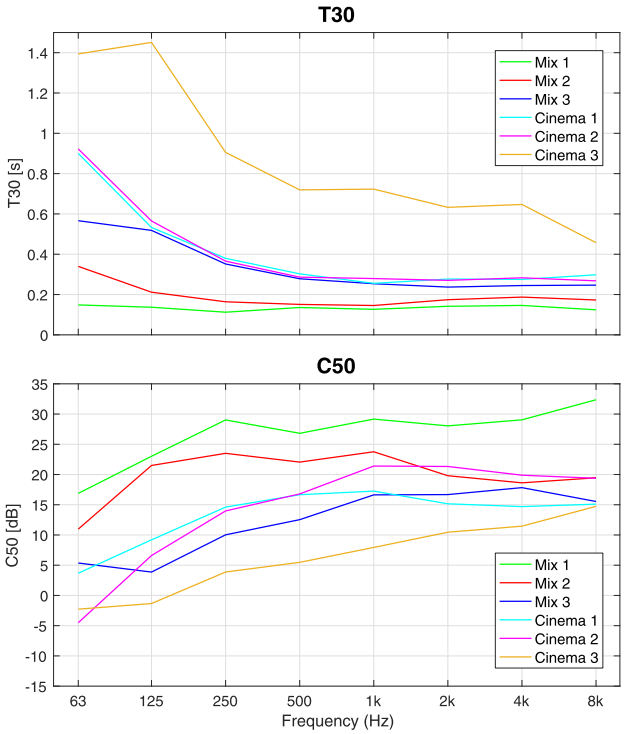

Fig. 1. Reverberation time T_{30} and clarity C_{50} of the auralized rooms. Values are average of left, center, and right speakers and measured at the listening position.

Method (SDM) [22]. These responses were synthesized to spatial impulse responses, which were used to auralize the program selections from original movie soundtracks for an anechoic listening room that had 45 loudspeakers. A more detailed review of the method and data processing can be found in a previous study by the authors [21].

1.1 Listening Environments and Auralizations

Six listening environments were the same as used for evaluating the attributes in the previous study by the authors [21]. Three of the environments were mixing rooms and the other three cinemas. The rooms vary from small mixing room with nearfield monitoring to large-scale cinema with 635 seats. The dimensional properties of the rooms are presented in Table 1.

Because the assessors evaluated the rooms in the anechoic chamber, where the sound field from real rooms was rendered, the acoustics of room auralizations instead of real rooms is presented. The acoustical parameters, electroacoustic response, and spatio-temporal visualizations from room auralizations were measured similarly to the real rooms. The measurement signals were rendered for all room auralizations and reproduced in the anechoic chamber while the microphone probe was capturing the sound in the listening position. The recorded spatial impulse responses were used to calculate acoustical parameters, electroacoustic response, and spatio-temporal visualizations. The reverberation time T_{30} and clarity C_{50} of the auralized rooms are shown in Fig. 1.

Table 1. The dimensions, capacity, and listening distance of six rooms.

Room	A[m^2]	V [m^3]	H [m]	Seats	List. dist. [m]
Mix 1	19	40	2.4	···	1.6
Mix 2	37	110	3.0	···	4.5
Mix 3	120	400	3.4	···	7.0
Cinema 1	110	750	6.0	50	7.0
Cinema 2	250	1,400	7.5	257	11.0
Cinema 3	825	7,000	12.5	635	20.0

The area (A), volume (V), height (H), number of seats, and listening distance 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.

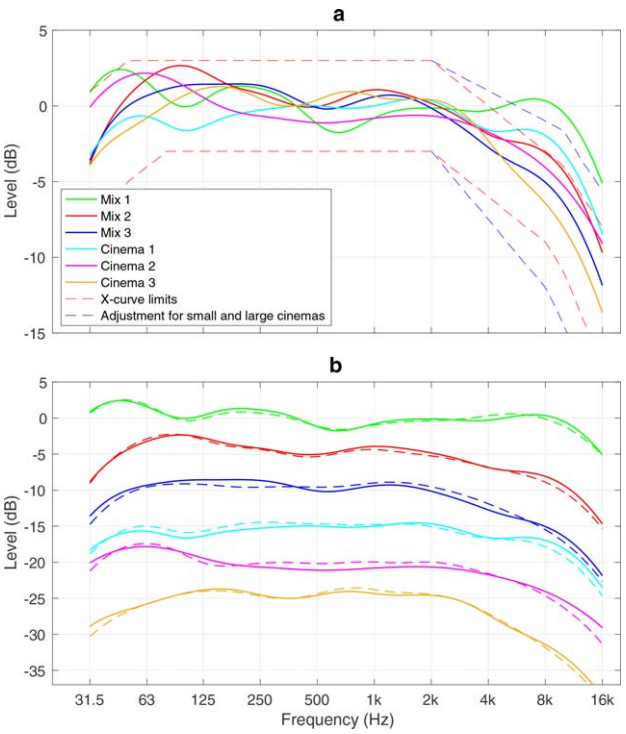

Fig. 2. (**a**) The electroacoustical responses of the auralized rooms and the X-curve limits from the standard SMPTE ST 202:2010 [7]. The electroacoustical responses are average of left, center, and right speakers and measured at the listening position. The limits for medium-sized theatres (500 seats) are presented in red and the upper limit for small (30 seats) and the lower limit for large theatres (2,000 seats) are presented in blue color. The blue upper limit for small theatres is used for all the mixing rooms and *Cinema 1*. (**b**) The electroacoustical responses of the auralized rooms (solid line) and real rooms (dashed line). The responses are offset downward by 5 dB from the preceding trace to distinguish individual responses and the differences more easily.

All of the rooms, except *Mix 1*, were previously calibrated according to standard SMPTE ST 202:2010 [7], using the X-curve as a target curve for the electroacoustic response. The target for *Mix 1* was a flat response. The rooms were calibrated by the owner and not modified for the measurements. The electroacoustical responses of the room auralizations are presented in Fig. 2(a) with the X-curve data

consistent with the standard. The responses are calculated in the listening position inside the anechoic chamber, where the real rooms were evaluated. The reference measurement position according to standard SMPTE ST 202:2010 [7] was used as a listening position in each real room. The mixing rooms were measured in the mixing position, while in the cinemas the measurement position was located in the center line in the 2/3 length from the screen to the projector wall. The height of the measurement position was 1.2 meters from the floor at the listening position. In Fig. 2(b), the electroacoustical responses from the real rooms and the auralizations are presented. The differences are minor.

1.1.1 Spatio-temporal Visualization

The recorded spatial impulse responses are visualized using the spatio-temporal visualization method presented in [23]. The spatial impulse responses present the sound field as pressure and direction of arriving sound at each sample. By integrating the sound energy with regards to the estimated direction of arrival beginning from a specific time moment until the end of the impulse response, the energy from a specific direction in a specified time interval can be plotted. Spatio-temporal visualizations of all the loudspeakers are presented in Fig. 3. The individual speaker volumes were previously calibrated. Five different time windows are used. The sound is analyzed from 0, 5, 15, 50, and 100 ms after the arrival of the direct sound, until the end of the impulse response.

The strong, clear, and precise direct sounds can be seen as sharp blue spikes in the visualizations, as in Fig. 3(a), when the time interval is 0...1,000 ms, as it is the only interval containing the direct sound. The enveloping late energy is presented as an even and more circular pattern.

As can be seen in Fig. 3(d), there was a problem in the auralization of the rear speakers in *Cinema 1*. The sound from the rear loudspeakers is arriving only from the center of the rear wall, unlike in real *Cinema 1*, where the sound from the rear speakers arrived evenly throughout the area of the rear wall. Interestingly, none of the assessors nor the authors noticed or reported anything anomalous in the sound field. The preferences of different reproduction methods were studied in [18] and it was shown that the presence of spatial content itself increases the listener's preference but adding more loudspeaker channels does not. According to the results of this study, the exact location and number of rear speakers are not relevant, at least for the program material used.

1.2 Program Material

Two program material items from the previous study by the authors [21] were used as listening test samples:

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

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

The selected excerpts were found to be most explicit in the previous study by the authors [21], containing a simple composition of sonic elements. Music is spectrally full and spatially wide and the mix itself contains reverberation, while dialogue is more monophonic, dry, and clear.

The program material items were extracted from the Blu-ray and converted from DTS-HD Master Audio data to 24-bit, 48-kHz PCM audio data. The original 5.1 audio from the Blu-ray was converted to 7.1 audio by copying the side surround channels to rear channels and reducing surround channel levels by 3 dB. The default volume in the listening test was adjusted to 72 dB for the dialogue and 78 dB for the music as an A-weighted equivalent sound pressure level.

2 LISTENING TESTS

2.1 Assessors

31 experienced listeners took part in the listening tests. 17 of the assessors were professional sound engineers with 18.8 years of career on average, 12 of whom were also involved in the attribute evaluation in the previous study by the authors [21]. In addition, eight sound engineer students and six other sound experts attended the listening test. All of them were accustomed to listening critically, either while mixing or participating in listening tests.

2.2 Procedure

One listening test was conducted in which five attributes and the preference were evaluated. The attributes were the brightness, sense of space, width, clarity, and distance. The definition of the attributes was discussed before the test. The assessors were mainly experienced professionals who are used to verbalize and characterize sound on a daily basis, while students are motivated to familiarize themselves with the audio vocabulary as part of their education. The attributes are part of the audio wheel in standard ITU-R BS.2399-0 [24], and the definitions from the standard were used to familiarize assessors with the terminology.

Twelve of the assessors took part in the previous study by the authors [21] and were familiar with the audio reproduction environment and listening test samples, so they started the listening test immediately. The rest of the assessors familiarized themselves with the environment and program material by listening to the samples freely. The interface for free listening as well as for the listening test was created with Max7 software and operated with an iPad in a stand next to the listening spot in the anechoic chamber.

At first, the assessors were instructed to select the attribute modules in optional order on the main page of the user interface in Fig. 4(a). After completing the five attributes the preference module could be opened. When a module button was pressed a graphical interface of the actual rating task was opened. The rating module for the sense of space is shown in Fig. 4(b). The task of the modules was to give a rating to a listening test item according to the attribute in question on a continuous scale. The assessors were instructed to use the whole scale, so for instance, when evaluating the sense of space [Fig. 4(b)], the most live room

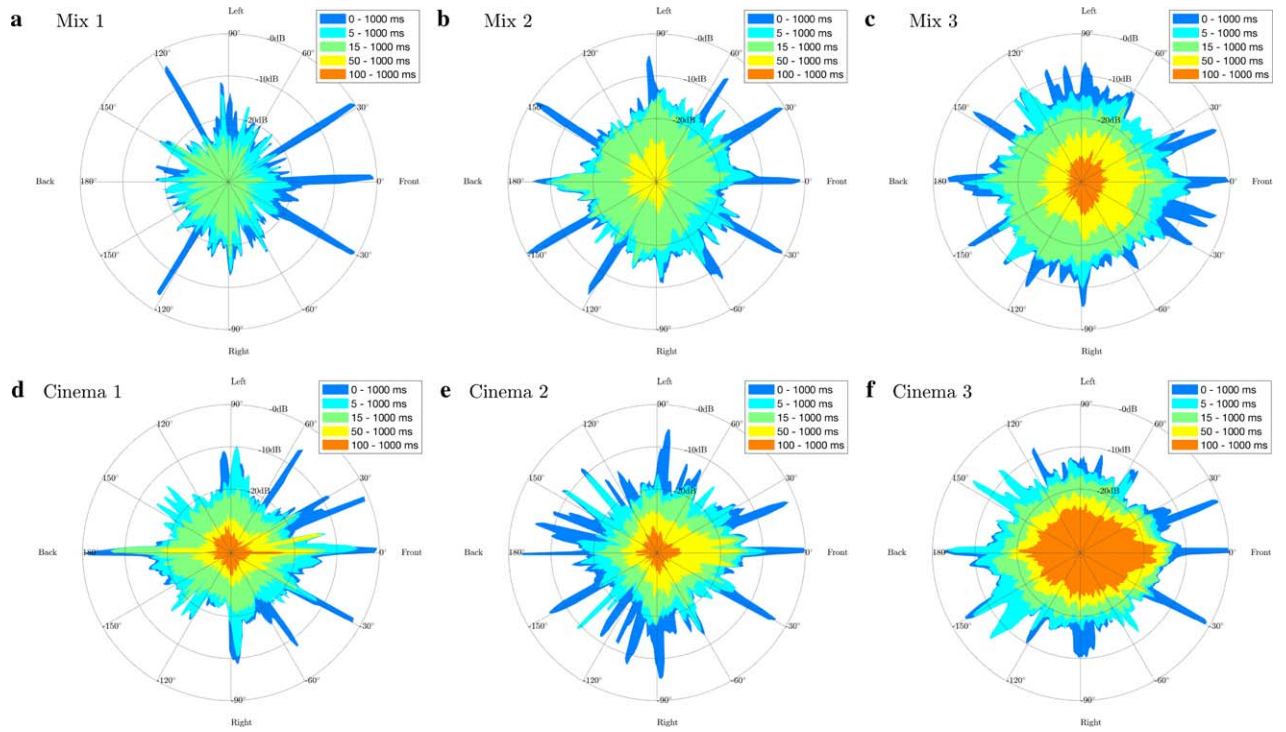

Fig. 3. The spatio-temporal visualizations of the auralized rooms with all loudspeaker channels. The spatial impulse responses are windowed with five time windows: 0. . .1,000 ms shown in blue, 5. . .1,000 ms in cyan, 15. . .1,000 ms in green, 50. . .1,000 ms in yellow, and 100. . .1,000 ms in orange. Due to an auralization problem, the sound from the rear loudspeakers is arriving only from the center of the rear wall in *Cinema 1*, unlike in the real room, where the sound from the rear speakers arrived evenly throughout the area of the rear wall.

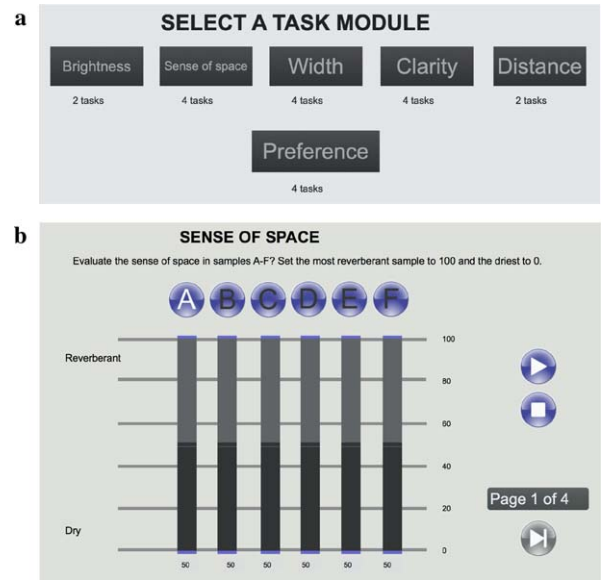

Fig. 4. User interfaces for the listening test. (**a**) The main page where the assessors selected the attribute module to evaluate in optional order. The preference was evaluated last. (**b**) The rating module for the sense of space.

should be rated to 100 and driest room to 0. The rest of the rooms were to be rated between the extremes with an estimated value for the attribute. The system was looping the listening test samples, and playback could be started and stopped at any time from the buttons. When switching the room, the position of the playback was maintained. The preference was asked at the end of the whole listening test to ensure that the rating of the descriptive attributes was not disturbed with affective preference rating.

The evaluation of six rooms in one interface could be challenging, so the task was tested with two audio professionals before the actual test. Following the suggestions from them, the assessors were instructed to do the rating relatively fast and intuitively in a few minutes, after which they could listen to the samples in rated order to ensure the progressive change and finally finalize the ratings.

The rooms were measured at calibrated level according to [6], leading to insignificant loudness differences between the rooms as concluded in the previous study by the authors [21]. The listening level for different program material items was adjusted between $L_{\mathrm{Aeq}} = 72$–78 dB by default. However an opportunity was given to the assessors to change the listening level at their will. Still, the level balance between the rooms remained the same.

The order of the rooms in the interface was randomized in each turn. Each attribute and program material item was rated twice. The brightness was evaluated only for *Music slow* and the distance for *Dialogue dry*. Together 20 tasks were completed ($2 + 2 \times 2 + 2 \times 2 + 2 \times 2 + 2 + 2 \times 2$). The duration of the rating session varied from 25 minutes to 60 minutes depending on the assessor.

3 RESULTS

The sound of three mixing rooms and three cinemas was evaluated with five attributes with two program material items. All rooms were rated according to all attributes and preference with a continuous scale from 0 to 100.

3.1 Reliability of the Assessors

The reliability of the assessors was tested by checking how accurately they can replicate their ratings. As all attributes were rated twice, the correlation of the ratings can be checked by calculating the RV coefficient with the Pearson type III approximation [25] for two matrices, each containing one rating. The p-value of an RV coefficient indicates if the correlation is significant or not. RV coefficients and p-values were calculated with the FACTOMINER package [26].

The correlation p-values of replications for each of 31 assessors are presented in the table in Fig. 5. At first, in columns 1–6, the p-values are calculated for the five attributes and the preference individually containing the evaluations of both program material items and six rooms. Then, in columns 7–8, the p-values are calculated for the two program material items, each one containing the evaluations of the six rooms and five attributes plus the preference. Finally, the p-values for the whole data were calculated in the last column, where the brightness and distance are combined in the same column in the matrix, as they were rated only with one program material item, resulting in two 6×12 matrices for all assessors [6 attributes $\times$ (6 rooms $\times$ 2 program material items)].

It can be seen that when the entire data is used for the analysis (column *All*), $p < 0.05$ for all assessors, meaning that the attribute ratings were replicated consistently and reliably. The correlation p-values for music and dialogue alone are slightly higher, which is due to the decrease in the amount of data. The correlation is not significant for four assessors when rating the music.

The number of insignificant p-values, i.e., inaccurate replications, vary between attributes from 6% with sense of space to 39% with width. This figure probably reflects the difficulty of the evaluation. If so, the evaluation of music was more difficult than dialogue, the width was the most challenging attribute to evaluate, and the sense of space was the easiest. Discussions with the assessors after the test support this conclusion.

The assessor numbers in the first column for the seventeen professional sound engineers are marked in gray. The assessors with most inaccurate replications were numbers 9 and 28, followed by numbers 1 and 6, which was the only professional sound engineer with inaccurate replications with three attributes. On average, the percentage of inaccurate replications were 12% for the professional sound engineers and 26% for the others, if only the evaluations for five attributes and the preference are taken into account. However, as the correlation of the whole data is significant, all the assessors are included in the analysis. Consequently the final rating for each room was a mean of the two ratings.

3.2 Attribute Ratings

The ratings were evaluated without any reference or anchor points, and the extremes were forced to opposite sides of the scale, so the results are non-normally distributed, which was also confirmed with a Shapiro-Wilk normality check that failed for the data. Therefore, the ratings are compared with the Friedman test, which is a non-parametric alternative for repeated measures analysis of variances. The Friedman test shows significant differences between the rank means for all attributes and the preference at the significance level of 0.05. The results from the pairwise comparison of the rooms are presented in the table in Fig. 6, which shows if the rank between two rooms for five attributes and the preference is significant or not. The significance level is 0.05 and the significance values are adjusted by the Bonferroni correction for multiple tests.

The results for ratings are presented as box plots with dots representing individual ratings in Fig. 7 for the five attributes and in Fig. 8 for the preference. The rank means are marked above the room name on the X-axis.

4 DISCUSSION

4.1 Sense of Space

One presumption was that it is easier to evaluate the sense of space with the dry dialogue than with the music containing reverberation in the mix. The box plots in Figs. 7(a) and 7(b) support the presumption: the dispersion of the ratings is larger with music. Also, the driest and the most live rooms are more clearly at the extreme ends of the scale with dialogue. The reverberation is easier to perceive when listening to the dialogue, giving a greater difference between *Cinema 3* and other rooms. *Cinema 3* seems to sound more distinctive with dialogue. The order of rooms is different for music and dialogue, but the extremes are evaluated similarly.

When comparing the measured reverberation time and the rating, the results are incoherent. Although the measured reverberation times of *Mix 1* and *Mix 2* are near to each other in the middle frequencies, as can be seen in Fig. 1, *Mix 2* is evaluated as much more reverberant than *Mix 1* especially with music, where *Mix 2* is ranked even above *Cinema 1*. Similarly, the measured reverberation times of *Mix 3*, *Cinema 1*, and *Cinema 2* almost coincide in Fig. 1, but *Mix 3* is evaluated to be more reverberant and alike to the *Cinema 3*, especially with music.

The rank of the dialogue matches well with the C_{50} in the middle frequencies (average over 500 Hz and 1 kHz bands) in Fig. 1. Supposedly, the perception of the sense of space cannot be derived directly from the reverberation time alone but is formed by many qualities such as the listening distance and the clarity of the room.

4.2 Clarity

The clarity ratings are consistent between music and dialogue except for the order of *Cinema 1* and *Cinema 2*.

Assessor	Sense of space	Clarity	Width	Brightness	Distance	Preference	Music	Dialogue	All
1	0,038	0,022	0,179	0,006	0,085	0,470	0,077	0,004	< 0,001
2	0,010	0,007	0,002	0,010	0,006	0,010	0,004	0,006	< 0,001
3	0,005	0,027	0,163	0,002	0,001	0,041	0,055	0,007	< 0,001
4	0,006	0,006	0,040	< 0,001	< 0,001	0,005	0,010	0,005	< 0,001
5	0,009	0,008	0,008	0,008	< 0,001	0,009	0,005	0,002	< 0,001
6	0,030	0,026	0,064	0,138	NA	0,052	0,016	0,009	< 0,001
7	0,009	0,047	0,033	< 0,001	0,052	0,012	0,021	0,006	< 0,001
8	0,009	0,028	0,130	< 0,001	0,022	0,008	0,016	0,006	< 0,001
9	0,086	0,157	0,258	0,169	0,063	0,011	0,026	0,035	< 0,001
10	0,012	0,010	0,030	0,032	0,068	0,008	0,009	0,012	< 0,001
11	0,013	0,122	0,071	0,035	0,008	0,009	0,008	0,025	< 0,001
12	0,010	0,008	0,010	< 0,001	0,014	0,008	0,004	0,005	< 0,001
13	0,012	0,009	0,038	0,011	0,002	0,011	0,011	0,006	< 0,001
14	0,007	0,017	0,015	0,001	< 0,001	0,002	0,006	0,006	< 0,001
15	0,007	0,007	0,018	0,016	0,002	0,009	0,004	0,005	< 0,001
16	0,007	0,006	0,006	0,005	< 0,001	0,005	0,006	0,006	< 0,001
17	0,013	0,027	0,315	0,003	0,001	0,393	0,013	0,008	< 0,001
18	0,007	0,008	0,013	< 0,001	< 0,001	0,014	0,008	0,006	< 0,001
19	0,004	0,006	0,035	< 0,001	< 0,001	0,025	0,005	0,006	< 0,001
20	0,019	0,022	0,053	0,008	0,002	0,032	0,006	0,007	< 0,001
21	0,011	0,010	0,007	< 0,001	0,004	0,016	0,011	0,005	< 0,001
22	0,018	0,071	0,019	0,008	0,006	0,008	0,021	0,007	< 0,001
23	0,114	0,007	0,086	< 0,001	0,045	0,032	0,010	0,005	< 0,001
24	0,005	0,005	0,244	< 0,001	0,007	0,004	0,001	0,006	< 0,001
25	0,022	0,009	0,011	< 0,001	0,004	0,019	0,010	0,003	< 0,001
26	0,014	0,007	0,020	0,001	< 0,001	0,009	0,006	0,003	< 0,001
27	0,033	0,026	0,010	0,021	0,059	0,083	0,093	0,008	< 0,001
28	0,036	0,110	0,523	0,706	0,101	0,068	0,165	0,032	< 0,001
29	0,007	0,005	0,013	0,003	< 0,001	0,006	0,006	0,006	< 0,001
30	0,005	0,006	0,290	0,021	0,002	0,019	0,008	0,005	< 0,001
31	0,005	0,019	0,012	0,015	0,160	0,794	0,030	0,009	< 0,001
Insignificant	6 %	13 %	39 %	14 %	32 %	19 %	13 %	0 %	0 %

Fig. 5. The p-values of RV coefficients per 31 assessors between first and second ratings. The p-values for five individual attributes and the preference in columns 1. . .6 are calculated from the ratings of six rooms and two program material items (*Music slow* and *Dialogue dry*). The p-values for music and dialogue in columns 7. . .8 are calculated from the ratings of five attributes and the preference and six rooms. The p-values in the last column 9 are calculated from all ratings: five attributes and the preference, six rooms, and two program material items. Significant RV coefficient, i.e., accurate replications, are presented in green ($p < 0.05$) and insignificant RV coefficients in red. The percentage of insignificant replications are presented in the last row of the table. In the assessor column, the seventeen professional sound engineers are marked in gray. Assessor 6 inadvertently failed to estimate the distance, which is marked NA in column 5.

The ratings match well with the middle frequency values (average over 500 Hz and 1 kHz bands) of the measured C_{50} in Fig. 1, especially with dialogue. Interestingly, the measured and rated clarity of *Mix 3* is weaker than in *Cinema 2*, although the reverberation time is similar, and the listening distance is shorter. The explanation for this may be that the volume of *Cinema 2* is three times larger than *Mix 3*, so its surfaces must be more absorbent to obtain a similar reverberation time. Thus the reflections are weaker and the direct sound is more prominent in *Cinema 2* than *Mix 3*.

4.3 Width

The dispersion of the evaluation of the width for music is wide for each room, as can be seen in Fig. 7(e), which results in average ratings getting centered in the middle of

	Sense of space				Clarity				Width				Brightness		Distance		Preference			
Room	Music	p-value	Dialogue	p-value	Music	p-value	Dialogue	p-value	Music	p-value	Dialogue	p-value	Music	p-value	Dialogue	p-value	Music	p-value	Dialogue	p-value
Mix 1	Mix 2	0.017	Mix 2	0.263	Mix 2	1.000	Mix 2	1.000	Mix 2	0.000	Mix 2	0.030	Mix 2	0.001	Mix 2	1.000	Mix 2	0.042	Mix 2	0.042
	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	1.000	Mix 3	0.315
	Cinema 1	0.081	Cinema 1	0.000	Cinema 1	0.376	Cinema 1	0.000	Cinema 1	0.001	Cinema 1	0.000	Cinema 1	0.030	Cinema 1	0.000	Cinema 1	0.218	Cinema 1	1.000
	Cinema 2	0.01	Cinema 2	0.000	Cinema 2	0.005	Cinema 2	0.315	Cinema 2	0.239	Cinema 2	0.000	Cinema 2	0.000	Cinema 2	0.063	Cinema 2	0.860	Cinema 2	1.000
	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.002	Cinema 3	0.000
Mix 2	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	0.000	Mix 3	1.000	Mix 3	0.000	Mix 3	0.002	Mix 3	0.000	Mix 3	0.010	Mix 3	0.000
	Cinema 1	1.000	Cinema 1	0.001	Cinema 1	1.000	Cinema 1	0.000	Cinema 1	1.000	Cinema 1	0.002	Cinema 1	1.000	Cinema 1	0.012	Cinema 1	1.000	Cinema 1	0.000
	Cinema 2	1.000	Cinema 2	0.860	Cinema 2	0.735	Cinema 2	1.000	Cinema 2	0.073	Cinema 2	1.000	Cinema 2	1.000	Cinema 2	1.000	Cinema 2	1.000	Cinema 2	1.000
	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	1.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000
Mix 3	Cinema 1	0.000	Cinema 1	1.000	Cinema 1	0.000	Cinema 1	1.000	Cinema 1	0.447	Cinema 1	1.000	Cinema 1	0.000	Cinema 1	0.937	Cinema 1	0.065	Cinema 1	1.000
	Cinema 2	0.001	Cinema 2	0.003	Cinema 2	0.053	Cinema 2	0.001	Cinema 2	0.002	Cinema 2	0.000	Cinema 2	0.447	Cinema 2	0.001	Cinema 2	0.315	Cinema 2	0.002
	Cinema 3	1.000	Cinema 3	0.180	Cinema 3	0.180	Cinema 3	0.073	Cinema 3	1.000	Cinema 3	1.000	Cinema 3	0.148	Cinema 3	0.131	Cinema 3	0.008	Cinema 3	0.010
Cinema 1	Cinema 2	1.000	Cinema 2	0.530	Cinema 2	1.000	Cinema 2	0.164	Cinema 2	1.000	Cinema 2	0.148	Cinema 2	0.134	Cinema 2	0.486	Cinema 2	1.000	Cinema 2	0.021
	Cinema 3	0.000	Cinema 3	0.001	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	1.000	Cinema 3	1.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.001
Cinema 2	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.021	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000	Cinema 3	0.000

Fig. 6. Pairwise comparison of the rooms with the Friedman test. The significant differences ($p < 0.05$) are presented in green. The significance values are adjusted by the Bonferroni correction for multiple tests.

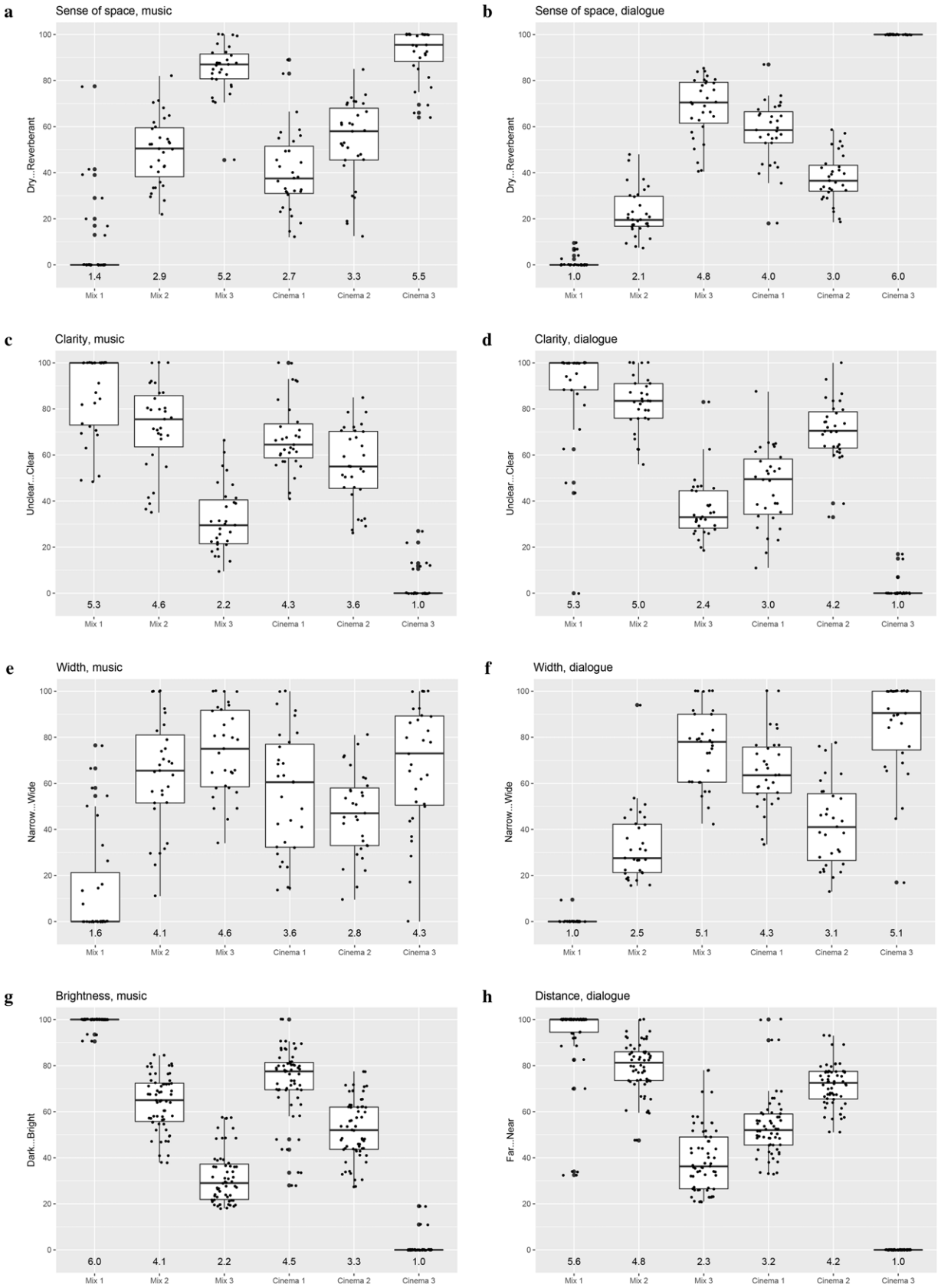

Fig. 7. The evaluations for the five attributes of the six rooms with two program items. The evaluations are presented as box plots showing the median, interquartile range, and range of the distribution of ratings. Individual ratings are represented as dots.

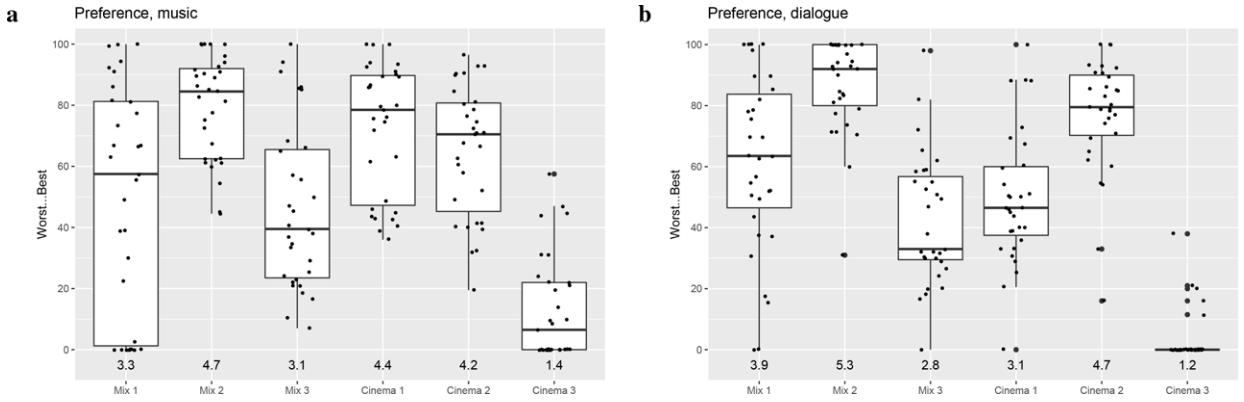

Fig. 8. The evaluations for the preference of the six rooms with two program items. The evaluations are presented as box plots showing the median, interquartile range, and range of the distribution of ratings. Individual ratings are represented as dots.

the scale, especially with music, except the rating of *Mix 1*. As suggested earlier, based on the reliability measures, the width was the most challenging attribute to evaluate. The meaning of the attribute width is likely to be ambiguous, at least with cinema sound, the aim of which is to reproduce immersive sound to the audience. In addition to the width, the attributes for the spatial extent in the audio wheel in standard ITU-R BS.2399-0 [24] are depth, envelopment, and balance. Evaluating only the width could be difficult without confusing it to other attributes.

The width of the dialogue matches well to the evaluations of the sense of space in Fig. 7(b), which indicates that the cause of the width is the reverberation and reflections of the room when the dry dialogue is evaluated. The dialogue is reproduced only from the center channel with some weak ambience from other channels. However, the cause of a wide or narrow sound image is different when evaluating the music that is mixed immersively around the listener. The mix mainly determines the width of the music, and thus it also depends on the alignment and calibration of the sound system in the room.

4.4 Brightness

The ratings for the brightness in Fig. 7(g), which was evaluated only with the music, are unambiguous, and dispersion is relatively narrow. The ratings match well with the level of the high frequencies above 4 kHz of the electroacoustical responses in Fig. 2(a). All rooms, except *Mix 1*, were calibrated according to the standard SMPTE ST 202:2010 [7], which gives 6...10 decibel tolerances for the calibration. In addition, the high-frequency roll-off is steeper for large theatres and gentler for small theatres, allowing the calibrator to adjust the response according to the size of the theatre and finally by ear. In the standard three arguments are given for the different amounts of high frequency in different-sized theatres:

1. The air absorption proportional to signal path length.
2. Increasing amount of reverberation in larger theatres, which is attenuated in high frequencies due to air absorption. As a result, larger and more reverberant spaces sound naturally duller with steady-state signals.
3. A psychoacoustical phenomenon according to which "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."

In addition, it is argued in the standard that all published experiments have confirmed this phenomenon; however only a paper by Allen [12] is cited. According to Allen, one listening test was done in the Elstree dubbing stage in the UK in the early 70s that was used to support the shape of the X-curve. Allen also argues that although the steady-state response shows falling high-frequency characteristics, the direct sound is still flat because the naturally dull reverberation "dulls" the response. Consequently, if the steady-state response was to be tuned flat, the loudspeakers should be tuned over-bright, which can result in short duration sounds like consonants sounding too bright.

A similar view is presented by Toole and Newell, who have suggested in [13] and [10] that the ear and brain can "hear through" the acoustics of a room and that the direct sound plays an essential role in the sound field perception. According to them, the reverberated energy should not be taken into account in the calibration process. The modern-day analyzer software like SMAART and SysTune could window out most of the reflections with frequency-dependent window length [27, 28] and aim to access only the direct sound. In the bass where the wavelength is long, the direct sound and first reflections are intermingled, but in the middle and high frequencies, the direct sound is accessible at least in the larger rooms where no obstacles are present along the wave path.

According to the results in this study, the perception of brightness is not affected by the size of the hall but depends on the amount of high-frequency energy above 4 kHz of the electroacoustical impulse response when the music is evaluated. Brightness was not evaluated with dialogue, because it was not assessed to be relevant enough in the previous study by the authors [21]. No support for the high-frequency roll-off that varies with the size of the theatre was found.

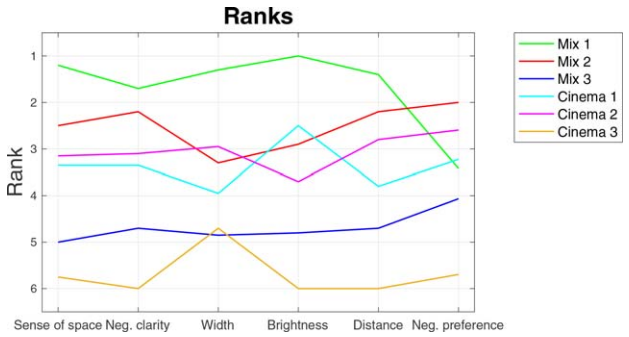

Fig. 9. The ranks of rooms of five attributes and the preference for six rooms. The scale for the clarity and preference is inverted, presenting the unclarity and negative preference.

The evaluations were performed based purely on the auditory information. The influence of visual information on auditory perception is outside the scope of this study. However, the effect of the size of a movie screen on the realistic combination of sound and vision was studied in [29]. It was stated that the amount of low frequencies necessary for realistic sound increases when the screen size increases, which is typical in large theatres. A similar phenomenon was not found at high frequencies, which is in line with the results of this paper.

4.5 Distance

The distance was evaluated only with dialogue. The ranks match the actual listening distance with *Mix 1*, *Mix 2*, and *Cinema 3*, but the order of the rest of the rooms does not. *Mix 3* is evaluated as more distant than *Cinema 1* and *2*; similarly it was evaluated as more reverberant with dialogue than *Cinema 1* and *2*. The order of all rooms matches well with the measured clarity C_{50} in the middle frequencies (average over 500 Hz and 1 kHz bands) in Fig. 1, as does the sense of space with dialogue.

4.6 Preference

The evaluations for the preferences are presented in Fig. 8. As can be seen, the assessors preferred *Mix 2* and evaluated *Cinema 3* as the weakest room, both with music and dialogue. The preference ratings are highly dispersed when rooms are evaluated with music, especially for *Mix 1*. The driest and brightest room divides opinions; some ranked it as the best room and some the worst. It seems that, on average, the assessors prefer a clear, intimate, and somewhat dry sound field that is not too bright and totally dead.

4.7 Rooms

The ranks of five attributes and the preference for six rooms averaged over music and dialogue are presented in Fig. 9, where the scale for the clarity and preference is inverted. By looking at the closeness of the lines, the similarity of the rooms can be evaluated and three groups can be identified. *Mix 3* and *Cinema 3* are more reverberant and distant, darker, unclearer, and wider than other rooms. *Mix 1* is dry, clear, narrow, bright, and intimate, whereas the other three rooms are quite alike. *Mix 3* is aimed for final mixing of the movie soundtracks, so it is the last step before the movie is showed in cinemas. Although *Mix 3* is smaller than all three cinemas in this study, its sound is a good average of small and large cinemas that makes it a good reference for a generic cinema.

These results are consistent with the results in the previous study by the authors [21], where one task was to describe rooms freely. The most used words for *Cinema 3* and *Mix 3* were reverberant, unclear, dark, and nasal. These rooms were also described as distant and wide, whereas *Mix 1* was described as bright, narrow, dry, and intimate as well as unpleasant. The most used word for *Mix 2* was good, followed by clear. *Cinema 1* was described as natural and *Cinema 2* as clear.

4.8 Room to Room Translation

Of the five attributes evaluated in this study, room acoustics are an essential factor in terms of the sense of space, clarity, and distance, while the equalization of the speaker system controls the brightness. The width seems to be related to both the loudspeaker system and room acoustics. The electroacoustic response in rooms is determined in the SMPTE standard 202:2010 [7] and the absolute sound pressure levels and balance between the loudspeakers in the SMTPE recommended practice 200:2012 [6]. One goal of the SMPTE standard ST 202:2010 is to have a constant perceived frequency response from installation to installation. As found in this study, the rooms calibrated according to ST 202:2010 sound different in the dark-bright scale. An alternative method for measuring the frequency response and especially the brightness is needed.

The perceived sense of space, clarity, and distance seem to match the measured clarity C_{50} in the middle frequencies that is controlled by the room acoustics of the rooms. There are no recommendations for the clarity in cinemas, but the reverberation time recommendations can be found in various books and cinema design guidelines, for instance in [1–5]. According to the recommendations, the reverberation time at 500 Hz should approximately double when the volume of the room is ten-folded, which roughly corresponds to the situation where the average absorption coefficient of the room remains the same while the volume of the room increases. So, if the surfaces of the room are acoustically similar, the reverberation in a larger room is longer than in a smaller room. The reverberation time recommendation limits at 500 Hz versus room volume, as well as the six rooms of this study, are shown in Fig. 10.

The three most preferred rooms, *Mix 2*, *Cinema 1*, and *Cinema 2*, fall slightly below the recommendation; *Mix 1* and *Mix 3* are just above the lower limit of the recommendation; and the least preferred room, *Cinema 3*, is in the middle of the recommendation. A shorter reverberation and higher clarity values improve the preference of the room to a certain degree. The assessors prefer clearer and drier rooms as recommended. For the cinemas in general, it looks like increasing the average absorption coefficient while the volume of the room increases would improve

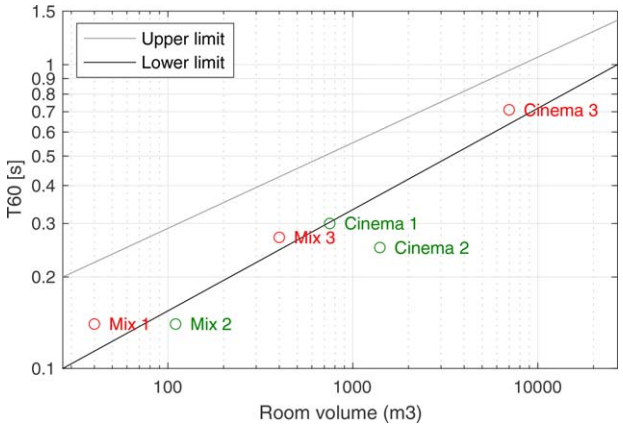

Fig. 10. Acceptable reverberation time versus room volume at 500 Hz according to [1–5] and the six rooms of this study. Three most preferred rooms are marked green.

the preference of the room among sound engineers. At the same time, translation between rooms should improve as the differences in reverberation time decrease.

Theoretically, the reverberation can also be reduced by filtering the audio signal with the inverse of the room impulse response (RIR). However, as all the positions in the room have different RIR, this approach is limited only to a single listening spot. Another approach has been presented in [30], where the room reverberation of a listening room is suppressed with filtering the input signal with "spectrogram inversion." The method works in a larger area as long as the direct-to-reverberant ratio is not significantly changed. The suppressor still reduces room reverb farther away from the source, but when closer to the source, the audio signal itself is dereverberated.

In the guidelines, the design of interior acoustics aims to firstly ensure high dialogue intelligibility and sharp imaging, secondly control (flutter) echoes and reflections between the screen and the loudspeaker baffle wall, and lastly minimize boominess on bass. Ideally the sense of space in a movie is incorporated into the soundtrack itself and reproduced through speakers distributed around the room, without being affected by the room's acoustics. However very short reverberation times may degrade intelligibility [31], and a small amount of reverberation could be useful in achieving even sound coverage [32], so the totally anechoic listening environment is not recommended. In this study, the rooms are evaluated only in the reference positions, and the evaluations may change according to the listening spot. Determination of the minimal amount of room reverberation to ensure adequate sound coverage and acceptable speech intelligibility is left to future studies.

In the previous study by the authors, the movie sound professionals reported using the stabilization practice [33, 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 the observations. A good option for the practice would be to simulate the acoustics of different cinemas in a mixing room. Current immersive sound formats require various loudspeakers distributed around the room, also in the ceiling. The same speaker setup could also be used for the acoustic simulation.

5 CONCLUSION

In this study, the sound and acoustics of six rooms were evaluated perceptually with music and dialogue. 31 assessors, including 17 professional movie sound engineers, gave continuous ratings between 0–100 for five descriptive attributes and the preference. The ratings were compared with the measured data. The results showed that:

- The perceived sense of space matches better with the measured C_{50} (scale inverted) in the middle frequencies than the measured reverberation time T_{30}.
- When dialogue was used as a program material, the sense of space was easier to evaluate, and the perceptual difference between the most reverberant cinema and other rooms were larger.
- Perceived and measured clarity match well.
- The width was difficult attribute to evaluate, indicating the word itself is ambiguous, especially with surround sound. The width of the soundscape is affected by the angle of the left and right screen speakers, the volume of the surround speakers, and the envelopment caused by the room reverberation.
- The ratings for brightness match well with the level of the high frequencies above 4 kHz of the electroacoustical responses.
- When music was evaluated, no psychoacoustical effect was found, suggesting that the high frequencies should be attenuated more in large cinemas with longer reverberation than in smaller cinemas, as suggested in the standard [7].
- The perceptual distance matches better with measured clarity C_{50} in the middle frequencies than the actual listening distance.
- The ratings for the smallest and largest room were similar for both music and dialogue. However the ratings for all other rooms depended on the program material.
- The ratings were more dispersed with music than dialogue, suggesting that at least some attributes of the sound field could be evaluated more easily and accurately with dialogue.
- The assessors preferred somewhat clear and dry sound over reverberant and distant; however the room should not be totally dead nor too bright.

The reverberation times of the rooms were compared to the recommendations, and it turned out that the assessors preferred shorter reverberation times as recommended. Shorter reverberation time would improve the translation between the rooms as the difference between rooms is reduced. The use of acoustic simulation of large cinemas in the mixing rooms was suggested.

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] Dolby Laboratories, Inc., "Technical Guidelines for Dolby Stereo Theatres" (1994).

[2] JBL Professional, "Cinema Sound System Manual" (2003).

[3] Theatre Operations, a division of Lucasfilm Ltd., "THX Sound System Program, Instruction Manual, Architect's and Engineer's Edition," (1987).

[4] Meyer Sound Laboratories, Inc., "Acoustic Guidelines for Cinemas" (2010).

[5] Dolby Production Services, "Studio Approval Requirements For Mixing All Dolby Theatrical Formats, (Issue 18)" Dolby Laboratories, Inc. (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] SMPTE, "Motion-Pictures — Dubbing Theaters, Review Rooms and Indoor Theaters — B-Chain Electroacoustic Response," *Standard 202:2010* (2010).

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

[9] 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).

[10] 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.

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

[12] 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.).

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

[14] 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/10.1111/17743/j.1467-9450.1979.tb00697.x.

[15] 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.

[16] 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.).

[17] 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.

[18] 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.

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

[20] 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.

[21] 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 Jan./Feb.), https://doi.org/10.17743/jaes.2020.0066

[22] 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.).

[23] J. Pätynen, S. Tervo, and T. Lokki, "Analysis of Concert Hall Acoustics via Visualizations of Time-Frequency and Spatiotemporal Responses," *J. Acoust. Soc. Am.*, vol. 133, no. 2, pp. 842–857 (2013), http://doi.org/10.1121/1.4770260.

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

[25] J. Josse, J. Pagès, and F. Husson, "Testing the Significance of the RV Coefficient," *Comp. Stat. Data Analy.*, vol. 53, no. 1, pp. 82–91 (2008 Sept.), http://doi.org/10.1016/j.csda.2008.06.012.

[26] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R Package for Multivariate Analysis," *J. Stat. Softw.*, vol. 25, no. 1, pp. 1–18 (2008).

[27] Rational Acoustics, "Smaart v8 User Guide" (2016).

[28] AFMG Ahnert Feistel Media Group, "AFMG SysTune, Software Manual, Rev. 5," (2014).

[29] P. Newell, K. R. Holland, B. Neskov, S. Castro, M. Desborough, S. Torres Guijarro, A. Pena, E.

Valdigem, D. Suarez Staub, J. Newell, L. Harris, and C. Beusch, "The Effect of Visual Stimuli on the Perception of "Natural" Loudness and Equalisation," in *Proceedings of the 24th Conference on Reproduced Sound: Immersive Audio, Proceedings of the Institute of Acoustics*, vol. 30, Pt. 6, pp. 1526 (Brighton, UK) (2008 Nov.).

[30] C. Faller, "Modifying Audio Signals for Reproduction with Reduced Room Effect," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), convention paper 10226.

[31] J. S. Bradley, H. Sato, and M. Picard, "On the Importance of Early Reflections for Speech in Rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, pp. 3233–3244 (2003), http://doi.org/10.1121/1.1570439.

[32] D. Liu, "Acoustic Design of Wide Dynamic Range, High-Impact Cinema," presented at the *6r Convention of the Audio Engineering Society* (1996 Aug.), convention paper 4318.

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

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 Aalto University School of Science under 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 and received an M.Sc. degree in 1997 and 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.