JASA ARTICLE

Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages

Antti Kuusinen[a] and Tapio Lokki[b]
Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT:
This article presents a listening experiment in which the listeners' task was to recognize the acoustics of a seat in a specific concert hall. Stimuli included two short passages extracted from a Beethoven symphony and samples of a solo violin auralized to four real concert halls. In each trial, listeners were presented with a reference and four alternatives with one correct match. In the "same" condition, the reference and the alternatives contained the same source sound. In the "different" condition, the source sounds were different musical passages but always of the same sound type, that is, symphonic music or solo violin. Results show that on average listeners could recognize the halls when the task was performed with the same source sound but had difficulty when listening to different sounds. The patterns of erroneous responses exhibited confusion between particular hall pairs and corresponded well to the values and just-noticeable-differences of the traditional objective room acoustic parameters. While the type of music is previously well known to influence the perception of concert hall acoustics, the present results indicate that even minor changes in the source sound content may have a strong impact on the ability to recognize the acoustics of individual halls. © *2020 Acoustical Society of America*. https://doi.org/10.1121/10.0001915

(Received 24 March 2020; revised 11 August 2020; accepted 16 August 2020; published online 10 September 2020)
[Editor: Brian F. G. Katz] Pages: 1380–1390

I. INTRODUCTION

One of the less studied aspects in the perception of room acoustics concerns our ability to recognize a particular acoustic space when we listen to different sounds, i.e., different signals that excite the space. The successful recognition of a particular acoustic space when listening to different sounds would require the listener to be able to "hear through" and extract the acoustical characteristics of the room despite the differences in the excitation signals. This aspect has previously received only a little attention in research and to the authors' knowledge, the ability to recognize concert halls (or rooms in general) with different signals has not been directly tested before. Thus, the aim of the concert hall matching experiment presented here is to gather some first-hand evidence to better understand our ability to recognize different concert halls based on their perceptual characteristics.

Previous studies concerning the perception of concert hall acoustics have been performed either by evaluating a set of perceptual attributes *in situ* without the possibility of direct and instantaneous comparisons between halls (e.g., Barron, 1988; Hawkes and Douglas, 1971) or by using reproduced stimuli in the laboratory where direct comparisons are possible. In terms of experimental design, being able to control the properties and the presentation of the stimuli offers many advantages over *in situ* listening, but there is concern for the generalization of the results from the laboratory to natural conditions.

One issue is that laboratory experiments are often performed by directly comparing acoustic characteristics while listening to the same sound in different conditions (e.g., Lokki *et al.*, 2012; Schroeder *et al.*, 1974). Direct comparisons using exactly the same source sounds enables discrimination of even minor perceptual changes in acoustic conditions. Perceptual studies with such direct comparisons have indicated, for instance, that individual halls (and seating positions) can be associated with a unique perceptual profile that distinguishes individual acoustic conditions from each other (Lokki *et al.*, 2012). However, one may ask whether these results are realistic in the sense that they reflect the differences people are actually able to perceive on their own by attending different concerts in real life. Are listeners able to recognize particular concert hall acoustics when the music and the orchestras vary from concert to concert and from venue to venue?

Literature provides only indirect evidence for this matter. It has been observed that the absolute amount of perceived reverberation differs between different signal types, e.g., vocal stimuli are typically perceived more reverberant than stimuli without vocals (Frissen *et al.*, 2009; Teret *et al.*, 2017). Such inaccuracy in perceiving the absolute levels of room acoustic features with different source signals also possibly makes recognizing individual rooms more difficult.

To shed some light on these questions, this article presents a concert hall acoustics matching experiment, where listeners were required to find a match to a reference stimulus among a set of four auralized concert halls while listening to the same or different source sounds. Anechoic sounds included two passages of a Beethoven symphony

[a]Electronic mail: antti.kuusinen@aalto.fi, ORCID: 0000-0002-8106-0421.
[b]ORCID: 0000-0001-7700-1448.

 © 2020 Acoustical Society of America

and short samples of a solo violin. In each trial, listeners were presented with a reference and four alternatives with one correct match. In the "same" condition, the reference and the alternatives contained the same source sound. In the "different" condition, the source sounds the different musical passages of the Beethoven symphony or different samples of the solo violin, but always of the same sound type, that is, symphonic music or solo violin. Beethoven and violin stimuli were not compared to each other. Auralizations were produced from the measurements of four real concert halls with the loudspeaker orchestra (Pätynen, 2011) and reproduced in a three-dimensional (3-D) multichannel loudspeaker setup using the spatial decomposition method (SDM) (Tervo *et al.*, 2013).

II. METHODS AND MATERIALS

A. Anechoic sounds

Anechoic sounds consisted of two source types which were (1) symphonic music with multiple instruments playing together and (2) the sounds of a single violin. In order to avoid any confusion considering the design of the experiment (the details are presented in Sec. II E), it is worthwhile to underline that the symphonic music passages and the violin sounds were never compared to each other in the listening experiment. The "different" condition in this study refers only to listening to different musical passages or samples with each type of signal. The following now describes the anechoic sound samples in more detail.

For the symphonic music, two musical excerpts of six seconds each were taken from the anechoic recordings of *L. van Beethoven* (1770–1827), Symphony No. 7, movement I (Pätynen *et al.*, 2008). Both excerpts correspond approximately to two bars in the score and six seconds in duration. The first part referred to as "Beethoven 1 (Beet1)" correspond to bars 26 + 2/4–28 + 2/4 and the second part "Beethoven 2 (Beet2)" to bars 29–30. There is only half a bar (less than one second) between these parts, but they still differ in their content because cellos and basses are present only in the latter part. Otherwise, the composition of the orchestra and the musical dynamics are similar in both parts and all the natural relative changes in sound level or otherwise between these parts were retained in the experiment. Figure 1 illustrates the difference in the frequency content between these excerpts where the maximum magnitude over both signals have been normalized to 0 dB for easy comparison.

Auralizations of the Beethoven excerpts were produced by first convolving the anechoic instrument tracks individually with the SRIRs corresponding to the source positions in the loudspeaker orchestra and then combining the convolved multichannel signals together for reproduction. This same auralization procedure has also been used in our previous studies of concert hall acoustics (e.g., Lokki *et al.*, 2016).

For the violin, a total of 16 different samples were selected from the sample bank of segmented anechoic orchestra recordings (Kuusinen, 2014). These samples were all less than five seconds long but otherwise varied in duration and composition and included short phrases as well as some single tones. Figure 2 illustrates the differences and variability in the frequency composition between these samples. Using a set of 16 different violin samples instead of only one or two was considered as adding variability in the matching task and making the results more generalizable. Listening to exactly the same sounds over and over again was also considered very repetitive for the listener, who may then lose interest and concentration on the task.

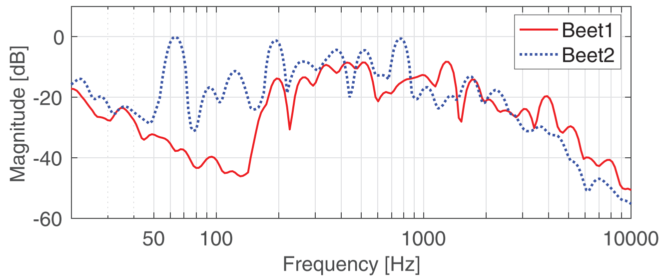

FIG. 1. (Color online) Magnitude spectra of the two Beethoven excerpts ("Beet1" and "Beet2"). The maximum magnitude over both signals has been normalized to 0 dB for easy comparison. The lines have been smoothed over one-third octave bands.

Auralizations of the violin sounds were produced with the spatial room impulse responses (SRIRs) from the source location of the first violin. It is noteworthy that in this source location, the measurements were performed with a pair of loudspeakers: one on a stand directed towards the audience and another one placed on the floor and directed upwards. Compared to a single loudspeaker with forward directed radiation characteristics, this configuration achieved better correspondence with the directionality of a real violin, especially in the upper hemisphere where a real violin radiates sound the strongest. More details of the measurement configuration and the comparison of power responses of a real violin and this loudspeaker pair are described by Pätynen (2011).

B. Concert halls and objective parameters

Concert halls in this experiment included two classic rectangular, so-called shoebox shaped halls: Amsterdam Concertgebouw (AC) and Munich Herkulessaal (MH) as well as Berlin Philharmonie (BP), which is a vineyard, and Cologne Philharmonie (CP), which is a fan shaped with a steeply rising audience section. The blueprints and general data of the concert halls are presented in Fig. 3. Room impulse response (RIR) measurement location was at a 19 meter distance from the position of the first violin. This set of two shoeboxes and two other shapes was selected to evaluate whether listeners are able to identify individual halls or if they only distinguish between halls (or hall types) in a more general way.

The single number averages for the objective parameters (ISO, 2009) are presented in Fig. 4. The vertical lines correspond to ±1 just-noticeable-difference (JND) reported in the standard (note that the JND for L_J is not known).

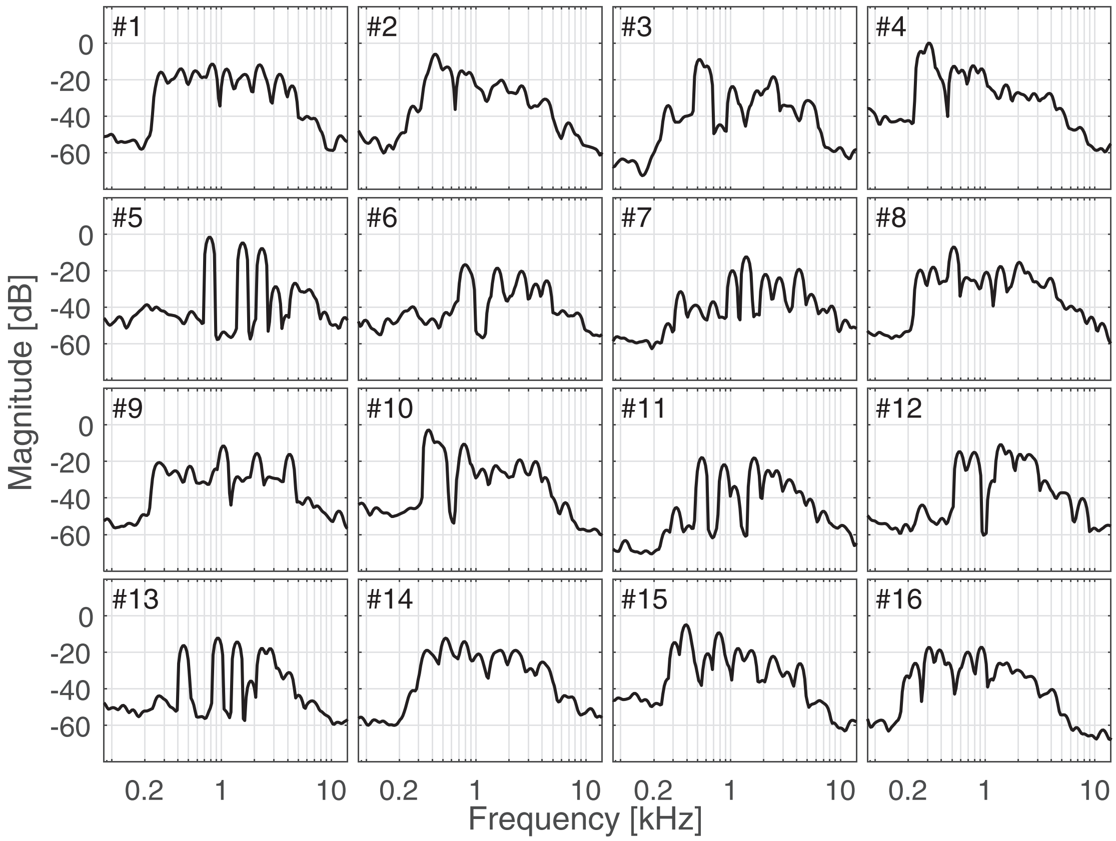

FIG. 2. Magnitude spectra of the 16 violin samples. The maximum magnitude over all 16 samples has been normalized to 0 dB for easy comparison. The lines have been smoothed over one-third octave bands.

Because the two passages of Beethoven symphony included slightly different instrumentation and only a single source was used in the auralizations of the solo violin, the objective parameter values are presented separately for each source configuration used in the auralizations. For each of the two Beethoven configurations, only those sources/RIRs (i.e., instrument positions on the stage) that had activity in the respective musical passage were used in the calculations. The parameter values represent the average calculated from these RIRs. Note also that the RIRs were measured with the loudspeaker orchestra (i.e., directive sources) instead of an omnidirectional source specified in ISO 3382–1:2009. Nevertheless, the sources were identical in the measurements of different halls and, therefore, the relative changes

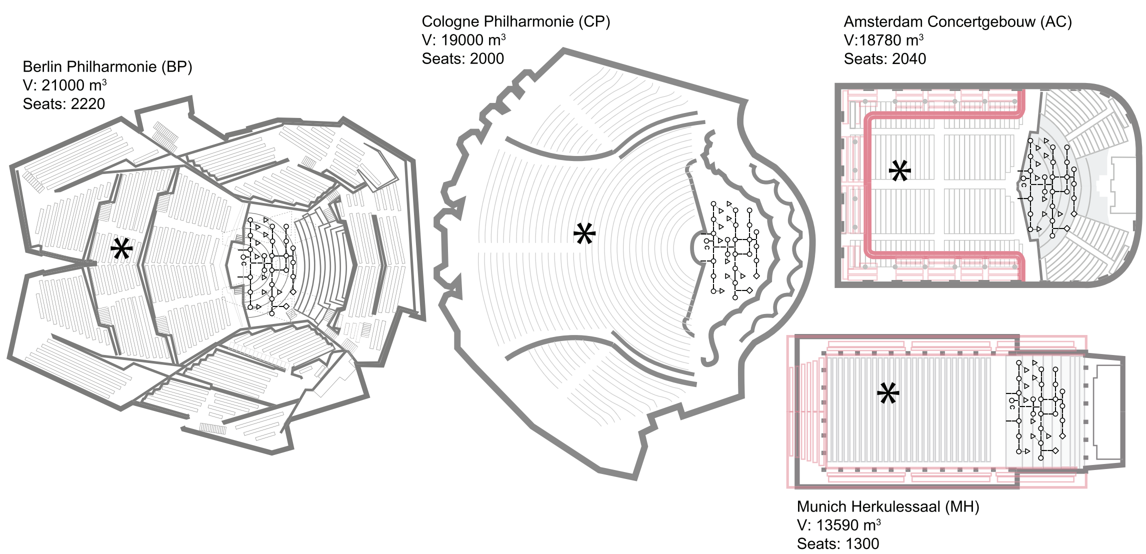

FIG. 3. (Color online) Blueprints of the concert halls. Star indicates the measurement positions at 19 meter distance from the position of the first violin.

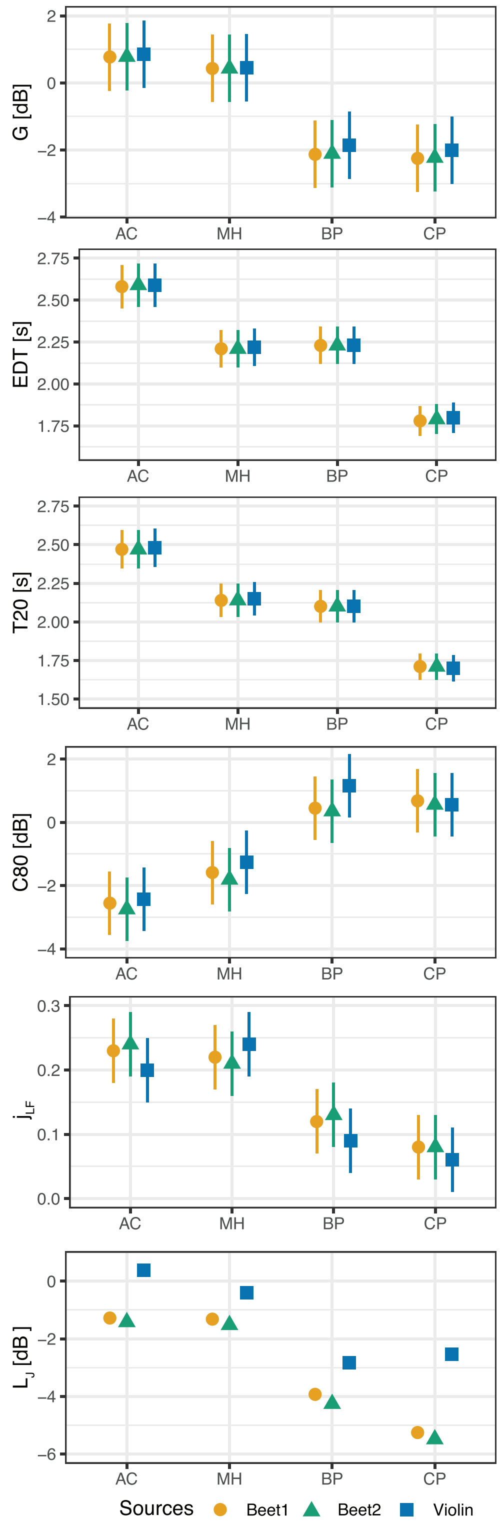

FIG. 4. (Color online) Objective parameter values calculated by using only the sources that correspond to the instrumentation in the sound samples. The vertical lines represent the JNDs.

between the halls are meaningful, although the absolute values might differ from other measurements.

The parameter values in Fig. 4 show that the differences in the source configurations have very little influence regarding the acoustic parameters calculated from the RIRs. For G, EDT, and $T20$, the values are almost exactly the same and the values are also within the JNDs for $C80$ and j_{LF}. Only for L_J can one note the difference between the values calculated from the single RIRs corresponding to the violin source position and the averages derived from the multiple RIRs of the Beethoven sources that are very close to each other. However, it is worthwhile to remember that Beethoven and Violin stimuli were not compared to each other in the experiment.

Considering the objective parameters, the strength factor G shows a distinct division between the two shoebox halls (AC and MH) and the other two halls (BP and CP). This same division is notable also with $C80$, j_{LF}, and L_J, but perhaps not as clearly as with G. If recognizing the halls was based on these aspects, it is probable that one may perceive the difference between the two pairs of halls but may not distinguish the individual halls in the pair. However, the reverberation time parameters EDT and $T20$, in contrast, show a division where MH and BP are very similar, whereas AC and CP are at the opposite ends of the scale. Thus, if recognition was based solely on the perceived length of reverberation, AC and CP would be easy to recognize, whereas one may have difficulty in distinguishing between the acoustics of MH and BP. Overall, if the listener is able to use the differences between the halls in all aspects, or for instance, use the combination of differences in strength (G) and reverberance (EDT or $T20$), one could, in theory, be able to recognize the acoustics of each individual hall in a consistent way.

C. Auralization technology and the listening setup

Halls were measured with the loudspeaker orchestra (Pätynen, 2011), which consists of 33 loudspeakers on the stage connected to the measurement system through 24 channels. SRIRs were measured at the receiver position with an open microphone array, which enables the analysis of spatial information in an impulse response with the SDM (Tervo *et al.*, 2013). SDM exploits the time-difference-of-arrival of sound waves between all the microphone pairs to estimate the direction of incidence for each sample in a RIR. Based on the spatial metadata, the RIR in one omnidirectional microphone is distributed to the reproduction loudspeakers around the listener.

The listening setup consisted of 44 active loudspeakers (Genelec Ones 8331A) in a 3-D setup in an anechoic chamber. Finally, the anechoic instrument sounds (described in Sec. II A) were convolved with these spatially distributed impulse responses resulting in the final set of 44-channel sound files (48 kHz sampling rate, 24 bits) for playback.

The experiment was implemented in MATLAB and run on an iMac Pro desktop computer. The computer was connected to RME ADI-6432 audio interface via a RME MADIface XT external soundcard module. The RME ADI-6432 sends the 44-channel audio signals to the loudspeakers in the anechoic listening space. The loudspeaker array was calibrated according to the manufacturer's recommendation using their proprietary software (GLM 3). This procedure consists of measuring sweeps from all individual loudspeakers at the listening position. The system optimizes the loudspeaker levels and signal delays so that the sound will arrive at the listening position at the same time and with the same level from all loudspeakers. Furthermore, the frequency responses of the loudspeakers are analyzed and equalized to ensure that the reproduction is neutral in the frequency spectrum and undesired coloration of sound due to room acoustics and the locations of the individual loudspeakers in the room is minimized. A-weighted background noise level in this space is –2.1 dB when the loudspeakers are turned off, and 11.6 dB when the loudspeakers are turned on.

D. Playback sound levels

The playback of the stimuli was set to a comfortable listening level, in which temporal and spatial aspects were observed to be clearly audible for both authors. Note, however, that the relative differences between the concert halls were still maintained in the experiment. A-weighted sound pressure levels (L_{ASmax}) tabulated in Table I were measured at the listening position. L_{ASmax} values vary approximately between 63 and 71 dB depending on the hall and the signal. One may note that there are some discrepancies between the L_{ASmax} values and the values of strength factor G, especially regarding the halls AC and MH. These differences are most likely due to the properties of the source signal combined with the A-weighting applied in the sound pressure level measurements.

E. Experimental design

The listening experiment was implemented as a four alternative forced choice test (4-AFC). When stimulus alternatives can be presented in parallel to each other, 4-AFC paradigm has been found to be more efficient than, for instance, two alternative forced choice (2-AFC), which would require much more trials with the same number of stimuli (Jäkel and Wichmann, 2006). Employing the 4-AFC paradigm enabled us to include repetitions of trials that had not been otherwise possible without markedly increasing the duration of the test.

The experiment was designed around the following two main experimental conditions: In the "same" condition, the reference stimulus and the alternative stimuli were auralizations produced with exactly the same excitation signals. In the "different" condition, the alternative stimuli were auralizations produced with an excitation signal that was different from the signal in the reference position. Considering the two different anechoic signal types (Beethoven and violin) presented above, the experiment was arranged as follows:

The two excerpts of Beethoven (Beet1 and Beet2) were paired into four reference–alternative pairs: Beet1–Beet1, Beet2–Beet2, Beet1–Beet2, and Beet2–Beet1, that is, two cases of the "same" and two cases of the "different" condition. Each of the four concert halls were presented in the reference position once in each of these four conditions, resulting in 16 separate trials for the Beethoven signal.

Beethoven excerpts were not paired with the violin or vice versa, so two signal types represent separate experimental blocks in this design.

Sixteen violin samples were randomly divided into eight cases for the "same" and eight cases for the "different" condition. With eight cases of "same" and "different" conditions, each of the concert halls now appeared two times in the reference position.

Now, it is possible that among this set of violin samples there are some "easy" and some more "difficult" samples. If all listeners were to listen to exactly the same set of cases, there could be a systematic bias in the results between the concert halls. In order to circumvent the possibility for such a bias, the combinations of individual violin samples with halls in the reference position were randomized between the test subjects. Therefore, listeners did not receive exactly the same set of cases, but their individual sets differed in terms of violin sample-hall combinations. To be clear, all subjects did listen to all violin samples, but how the samples were paired with the halls were randomized between the subjects. In statistical terms, the variance between the test subjects was confounded with the variance within the set of the violin samples.

In order to reduce the total duration of the experiment, we did not include the repetition of the test and, therefore, could not obtain a direct measure of the consistency or the reliability of the test subjects. Finally, the presentation order of all 32 trials (16 Beethoven + 16 violin) was randomized separately for each test subject.

F. Test procedure

Graphical user interface (GUI) is shown in Fig. 5. The listener's task was to listen to the reference stimulus (on the left side in the figure) and to find the correct match among

TABLE I. A-weighted maximum sound pressure levels [L_{ASmax} (dB)] of all samples measured at the listening position. Values have been rounded to the nearest integer.

	Beethoven		Violin																	
Hall	Beet1	Beet2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Max	Min
AC	70	70	67	67	65	68	70	66	67	67	67	67	64	68	66	65	65	64	70	64
BP	68	69	65	66	64	67	68	64	68	66	65	64	64	66	67	65	63	63	68	63
CP	68	66	66	66	64	66	71	65	66	65	64	66	63	65	65	66	66	64	71	63
MH	69	70	68	68	64	68	71	66	68	66	67	67	64	66	68	65	66	65	71	64

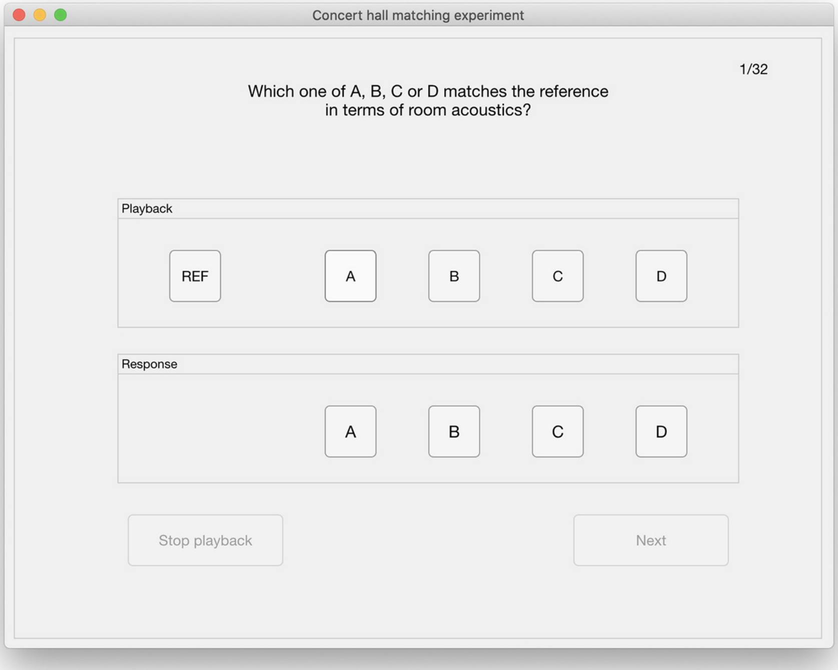

FIG. 5. (Color online) The GUI used in the matching task. The samples could be listened to by pressing the buttons on top and responses were given by the buttons below.

four alternatives ("A," "B," "C," and "D") (on the right) in terms of room acoustics (note that the term "reference" is used to point to the stimulus in the reference position, and the term "alternative" is used as reference to four alternatives). In this experiment, we were also interested in the manner these stimuli were potentially confused with each other and in this respect, the experiment can be viewed as a type of multi-class classification task (Giannakopoulos and Pikrakis, 2014). Such classification tasks are common in machine learning, where various performance metrics have been developed for the analysis of the results.

In practice, participants were first given written and verbal instructions about the task and how to use the GUI. They were told that one of the alternatives always matched the reference in terms of room acoustics and their task was to find out which one. Participants were also told that the experiment was about concert hall acoustics and that the samples were auralizations of different concert halls. To familiarize the listeners with the GUI and the task, they first completed a training set of four trials, which included one "same" and one "different" trial for both signal types. In this training set, listeners were given feedback on whether their responses were correct or incorrect. The experimenter also made sure that the participants understood the task correctly.

The training set was followed by the main experiment of 32 trials. Participants were free to listen to all samples as many times as they desired and to find the match using their own strategies. The listener could not select smaller segments and the stimuli always started from the beginning when listened to. Subjects did not receive any feedback on the correctness of their responses during the main experiment. The duration of the test depended on the performance of each individual. The fastest listener took only 21 min to complete the test, while the slowest one took 58 min. On average, completing the test took 34 min.

After completing the experiment, participants were interviewed for the perceptual attributes that they used in discriminating the samples from each other.

G. Participants

Twenty-five listeners (6 women, 19 men) under 40 years old participated in the study. Participation was voluntary and listeners were informed in writing that they could stop the test at any time they wanted to with no reason. We did not carry out audiometric testing in this study, but participants were explicitly asked if they suffered from any known hearing impairments; however, none of them reported doing so. The subject group was heterogeneous in terms of their level of experience and expertise in room acoustics, spatial sound, and listening experiments in general. Some participants were totally inexperienced in critical listening, others had background in music and/or room acoustics and a few already had some previous experience in listening experiments. Eight of the subjects were recruited from the staff of Aalto Acoustics Laboratory, and these can be considered highly experienced in critical listening.

III. RESULTS

Overall results are illustrated in Fig. 6. The discrepancy between the "same" and "different" conditions is apparent and can be quantified with the z-test for two proportions: the overall performance in the "same" condition is at 73.5% and

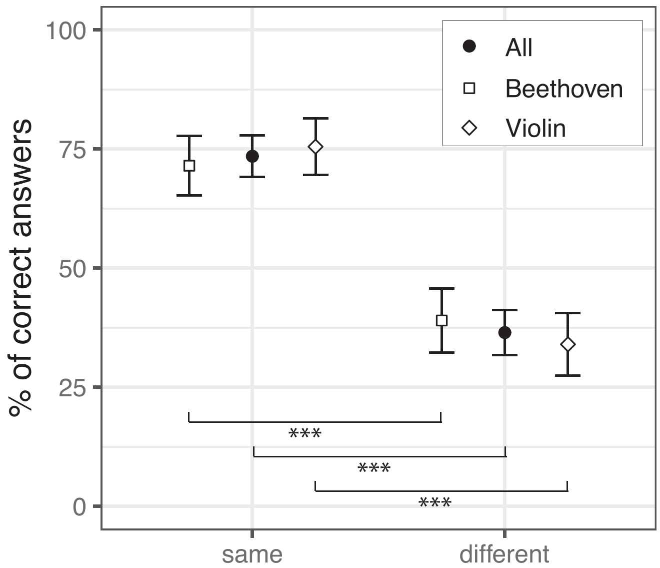

FIG. 6. Overall results: The mean proportion of correct answers with confidence intervals derived from the binomial distribution with normal approximation. Data are pooled over all individuals and each observation is treated as an independent binary (correct/incorrect) observation. Brackets and asterisks (***$p < 0.001$) denote significant differences in the proportions of correct responses between the conditions.

it is significantly better [$X^2(1) = 109.1, p < 0.001$] than the performance for the "different" condition, which is at only 36.5%. Exact binomial tests and the corresponding 95% confidence intervals in Fig. 6 show that while the recognition rates in the "different" condition are significantly above the chance level of 25%, they are only barely so, especially when compared to the "same" condition. There are also only little differences between Beethoven and violin signals, indicating that the performance did not depend much on the signal type.

The individual results in Table II indicate that there are listeners who did not perform above chance level, not even within the "same" condition (AS06, AS07, AS13, AS15, AS20; denoted by asterisks in the table). Using the exact binomial test as an indicator of statistical significance, eight correct responses (out of 16) are required to attain 95% confidence level ($p < 0.05$) that individual performance is above chance in the "same" and "different" condition. Using this threshold, 19 listeners performed above the chance level in the "same" condition, but only five individuals (AS01, AS05, AS08, AS18, and AS22) performed above this level in both the "same" and "different" conditions. It is worthwhile to mention that we conducted an alternative analysis where the individuals who did not perform above the chance level in the "same" condition (AS06, AS07, AS13, AS15, AS20) were excluded from the data. However, removing these individuals did not impact the overall results and, therefore, it was decided to keep them included. Consider also that while the eight highly experienced listeners from Aalto Acoustics Lab (denoted by superscript A in the table) performed quite consistently in the "same" condition; only two of them scored eight or more in the "different" condition and none of the participants could make all correct responses.

As mentioned in Sec. II F, the recognition performance in this task can be analyzed with tools developed for multi-class classification data sets (Giannakopoulos and Pikrakis, 2014). One common approach is to derive confusion matrices (or "error matrices") illustrated in Fig. 7, where the columns correspond to the hall in the reference position and the rows correspond to the given responses. The number and the color of each cell represent the normalized proportion of responses for the corresponding reference-response pair. The values range from 0 to 1 and the color from dark to light, respectively. The proportions of correct responses are represented on the diagonals and a perfect performance would result in ones on the diagonal and zeros elsewhere. Otherwise, the values in one column add up to one. The raw number of correct responses is presented on the "Sum" row in Table II. The more the numbers and the color are dispersed over the matrix, the more confusion there has been between the halls and thus, poorer performance. This way, confusion matrices can be used to examine the overall accuracy as well as the pattern of erroneous judgments, for instance, if there is some particular pair or pairs of halls that have been confused with each other more than with the rest.

Looking at Figs. 7(a) and 7(c), the matrices again illustrate the good performance in the "same" condition. However, the results are not perfect and confusions between the halls seem to arise, particularly between AC and MH and between BP and CP. Whenever the listener has made an incorrect response in the "same" condition, this misclassification seems to happen particularly within these two pairs. In the "different" conditions illustrated in Figs. 7(b) and 7(d), there are, in general, more confusions across all four halls, but grouping of the two pairs of halls is still observable to some extent.

In order to examine this pairing of responses more closely, the data was reclassified to represent this grouping: The responses were reclassified as "correct" whenever the response corresponded to either one within the pair. For instance, when AC or MH was the reference, a "correct" response would now be either one of the two. If the observed misjudgments did actually result from the confusion within these pairs, this reclassification should markedly increase the proportions of correct answers in the data set. The result of this reclassification is presented in Fig. 7 by the numbers in the corners of the confusion matrices with horizontal and vertical lines across the matrices indicating the grouping. Considering these values, for instance, in the Beethoven "different" condition [Fig. 7(b)], reclassification clearly improved the results (increase from 0.5 and 0.3 to 0.8), which indicates that mistakes were made especially between these particular hall pairs. Overall, the proportion of correct answers in the "different" condition increases to 71%, which is significantly ($p < 0.001$) above the change level (in this case 50%). It is worthwhile to mention that this same reclassification was performed also by using pairings AC–BP vs MH–CP as well as AC–CP vs BP–MH.

TABLE II. All results. Numbers are formatted as "same-diff," i.e., the number of correct matches when the matching was performed with the same signals and the number of correct matches when the matching was performed with different signals. Hall columns (AC, MH, BP, CP) indicate the reference hall to match. The number in parentheses at the top of each column indicates the maximum possible score. Asterisks (*) on the subject identifier indicate that the performance in the "same" condition was not significantly above the chance level. Superscript *A* denotes people from Aalto Acoustics Lab who are highly experienced in listening experiments.

	Beethoven (/2)				Violin (/2)				Signal (/8)		Sum (/16)		Total (/32)	
ID	AC	MH	BP	CP	AC	MH	BP	CP	Beethoven	Violin	Same	Diff.	*n*	%
AS01	1-1	2-1	1-0	2-2	2-2	2-1	1-1	2-1	6-4	7-5	13	9	22	69
AS02	0-1	2-0	1-0	1-0	1-0	2-2	2-0	2-0	4-1	7-2	11	3	14	44
AS03	2-2	2-1	2-1	2-0	2-0	1-0	1-1	1-1	8-4	5-2	13	6	19	59
AS04	1-0	2-1	2-0	1-0	0-1	2-1	1-1	1-2	6-1	4-5	10	6	16	50
AS05	2-2	1-1	1-1	1-2	2-0	2-2	1-2	2-2	5-6	7-6	12	12	24	75
AS06*	1-0	0-0	1-0	0-0	2-2	2-1	0-1	0-0	2-0	4-4	6	4	10	31
AS07*	0-0	1-2	0-2	1-0	2-0	1-0	0-1	2-0	2-4	5-1	7	5	12	38
AS08	2-2	1-1	2-0	1-1	2-1	2-1	2-1	2-1	6-4	8-4	14	8	22	69
AS09	1-1	2-1	2-0	1-0	2-0	2-0	2-0	1-1	6-2	7-1	13	3	16	50
AS10	2-1	2-0	2-0	2-1	2-0	2-1	2-1	2-0	8-2	8-2	16	4	20	62
AS11	1-1	2-2	1-0	2-0	1-0	2-2	0-1	2-1	6-3	5-4	11	7	18	56
AS12	1-1	1-0	2-0	0-1	1-0	2-0	0-0	2-1	4-2	5-1	9	3	12	38
AS13*	1-1	1-1	0-1	0-0	1-0	0-0	1-0	2-1	2-3	4-1	6	4	10	31
AS14	2-1	1-1	2-1	2-1	1-0	1-1	2-0	2-1	7-4	6-2	13	6	19	59
AS15*	1-0	0-0	0-0	1-2	0-0	1-2	2-0	2-0	2-2	5-2	7	4	11	34
AS16	2-0	2-1	1-0	2-1	2-0	2-0	2-0	2-1	7-2	8-1	15	3	18	56
AS20*	0-0	0-2	0-2	1-2	0-1	0-0	1-0	1-2	1-6	2-3	3	9	12	38
AS 17^A	2-1	2-0	2-0	2-1	1-1	1-0	2-0	2-0	8-2	6-1	14	3	17	53
AS 18^A	2-2	2-1	2-1	2-1	2-1	2-2	2-0	2-0	8-5	8-3	16	8	24	75
AS 19^A	2-2	2-0	2-1	2-0	2-0	2-0	2-1	2-1	8-3	8-2	16	5	21	66
AS 21^A	2-0	2-0	2-0	2-1	2-1	2-1	0-1	2-1	8-1	6-4	14	5	19	59
AS 22^A	2-2	2-2	2-2	2-1	2-2	1-0	2-0	2-2	8-7	7-4	15	11	26	81
AS 23^A	1-1	2-1	1-0	1-1	2-0	2-0	2-1	2-0	5-3	8-1	13	4	17	53
AS 24^A	2-1	2-0	2-1	2-1	2-0	1-2	1-0	0-2	8-3	4-4	12	7	19	59
AS 25^A	2-2	2-0	2-1	2-1	2-0	2-1	2-1	1-1	8-4	7-3	15	7	22	69
Sum	35-25	38-19	35-14	35-20	38-12	39-20	33-14	41-22	143-78	151-68	294	146	440	Ave.
/N	/50				/50				/200	/200	/400	/400	/800	55 %

However, in those cases, the proportion of correct responses in the "different" condition did not increase above the chance level indicating further that the observed pairings are not arbitrary.

There are also various performance measures available for multi-class classification data (Giannakopoulos and Pikrakis, 2014), such as "precision" and overall "accuracy" presented in Table III. Precision equals the number of correct responses (i.e., correct "positive" responses) out of the total number of responses given to that particular hall, i.e., the overall number of times that particular hall has been selected. Accuracy is in contrast calculated as the number of correct responses out of all responses, i.e., the proportion of correct answers. While accuracy reflects the performance

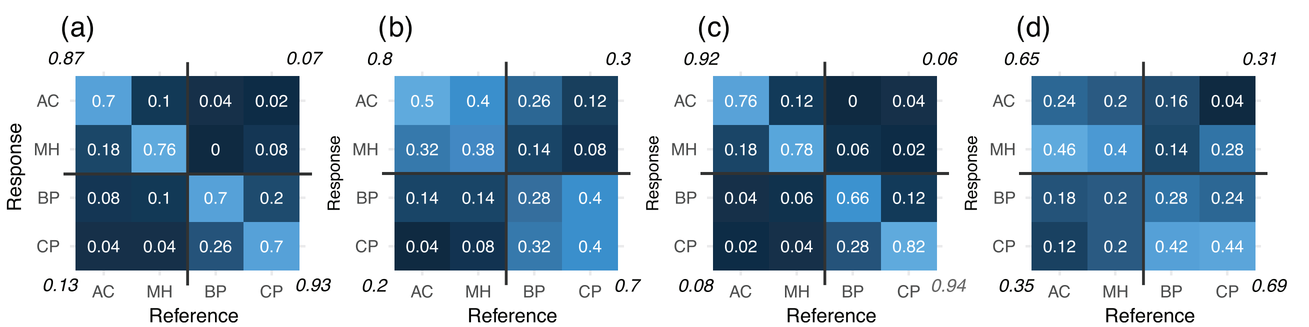

FIG. 7. (Color online) Confusion matrices (or "error" matrices) where each cell depicts the proportions (between 0 and 1) of responses given in the matching task. For example, in (a) whenever a reference auralization corresponded to AC, 70% of the responses were correct, while in 18% the response was incorrectly MH. The values in the corners present the results of the reclassification of the correct responses with horizontal and vertical lines across the matrices indicating the grouping (i.e., AC–MH vs BP–CP, see text for details). (a) Beethoven "same" condition; (b) Beethoven "different" condition; (c) violin "same" condition; (d) violin "different" condition. Color range is scaled from 0 (dark blue) to 1 (white).

TABLE III. Performance metrics precision and accuracy calculated from the listening test results. Hall columns (AC, MH, BP, CP) indicate the reference hall to match. Precision equals the number of correct responses (i.e, correct "positive" responses) out of the total number of times that particular hall was selected. Accuracy is the proportion of correct answers and can be directly calculated from Table II.

	Precision										
	Beethoven				Violin				Accuracy (%)		
	AC	MH	BP	CP	AC	MH	BP	CP	Beethoven	Violin	Total
All	0.56	0.59	0.48	0.59	0.64	0.51	0.53	0.54	54	55	55
Same	0.81	0.75	0.65	0.67	0.83	0.75	0.75	0.71	72	76	74
Diff.	0.39	0.41	0.29	0.48	0.37	0.31	0.31	0.37	39	34	37

over the whole confusion matrix, precision could reveal if some particular hall was very distinct from the rest. Here, the precision values are similar across the halls in different conditions, indicating that no hall did clearly stand out from the rest. In the "same" conditions, AC has a little better precision than the rest and CP seem to stand out a little in the Beethoven "different" condition, but otherwise, there are no clear differences between the halls in this respect.

A. Interviews

After completing the experiment, listeners were asked to shortly describe the perceptual features they had listened to in the test. People reported that they mainly listened to reverberation (14 occurrences), width (10), and distance (7) but also other spatial aspects such as the locations of instruments, the size of the room, spatial appearance, and envelopment, as well as "reflections" in general. Some people specifically mentioned bass or low frequencies (six occurrences) and many used timbre related attributes such as coloration, balance, ringing, brightness, and muddiness. Interestingly, volume or loudness was mentioned only by three people, and this implies that loudness differences were not prominent in this stimulus set. Finally, it can be noted that practically all participants commented that the matching task was very difficult when listening to different signals.

IV. DISCUSSION

Matching performance was considerably better when listening to the same excitation signals than when the signals were different. When matching was done with the same signals, listeners were mostly able to recognize the sound of each hall on an individual level, but this level of discrimination was lost in the "different" condition. It is also important to note that the sounds listened to in the "different" condition were not in fact completely different from each other. Beethoven signals were part of the same musical sequence and the violin samples represented just a single instrument and were, therefore, in many ways similar. In real concert halls and in our daily life in general, we are exposed to much greater differences in the sound and perhaps the recognition of room acoustics is also that much harder.

We believe that the current results reflect the well (young) population in general because the participants in this study included listeners with different levels of expertise in room acoustics and listening experiments. The results did indicate that expertise and previous experience in listening tests plays a role in how well people performed in this task, at least in the "same" condition. The experienced listeners typically made only a few mistakes in the "same" condition, while others made more or less errors depending on the individual. It is interesting, however, that in the "different" condition, the more experienced listeners did not perform consistently better than the others. This result indicates that it was difficult to recognize the halls in this condition irrespective of the previous experience in listening experiments or room acoustics.

Audiometric screening was not carried out in this study due to practical reasons. Therefore, it is possible that some participants unknowingly suffer from elevated hearing thresholds, although they did not report on any hearing impairments themselves. It is an interesting open question whether and how much a possible hearing impairment would affect the performance in this particular task.

One essential aspect of this experiment is that the matching task with the same or different excitation signals requires listeners to attend very different aspects in the stimuli. When excitation signals, i.e., the source sound contents are the same, the listener can concentrate on even very minor differences in any room acoustic feature(s) (or any auditory cue in general), that may give away that the stimuli do not represent the same hall. It is well known that observable perceptual differences, for instance in width and timbre, can be caused by minor changes in the reflection patterns and when allowed to switch back and forth between acoustic conditions with the same signal, these changes can be perceived. This sensitivity to changes when listening to the same signals is one of the reasons why many previous experiments have been conducted with this approach. For instance, in the descriptive profiling studies (e.g., Lokki *et al.*, 2012), the direct comparisons of halls while listening to the same piece of music enabled the discrimination of perceptual characteristics of each individual hall and resulted in detailed perceptual profiles for the studied halls.

When the excitation signals are different, the differences in the spectral and temporal characteristics do not allow the direct one-to-one perceptual comparison of the stimuli. The task is different in nature because it now requires the listener to "hear through" the differences in the signals and to extract those characteristics in the stimuli that are caused by the acoustics of the rooms. It is an open question, whether our auditory system is sensitive to different acoustic conditions to the extent that we are able to recognize certain conditions even in situations when rooms are excited with signals that share only little similarities in their content. With the same signals, the comparison can be made with reference to all characteristics of the stimuli as a whole, while with different signals, the reference needs to be made only to the relevant subset of the characteristics that are needed to complete the task. Thus, there is much less information for the listener to base their judgment on in the different condition than in the same condition.

Previously, the ability to recognize different rooms by their acoustics has received very little attention, and in the authors' knowledge, there is no direct precedence for the current experiment. In the development of the room acoustic quality inventory, Weinzierl *et al.* (2018) concluded that "…listeners are obviously able to identify the room and its acoustical properties as a consistent cognitive object." The current study has now put this abstract statement on an empirical basis, but according to the evidence here, the recognition of rooms by their acoustic properties seems to be rather elusive and approximate, especially when room acoustics are perceived in variable sound contexts. Weinzierl *et al*. (2018) did report that the test–retest reliability in their study was generally poor, which may be related to the present observation. However, they attributed this reliability issue to "time-varying situational factors" such as attention-, emotion-, and personal-trait related aspects which are often impossible or very difficult to control in the experiment. The results here, in turn, highlight the fact that an important part of these situational factors are the changes in the source sound content which can strongly influence the perception of room acoustics.

Nevertheless, the current results do also partly support the assertion that listeners can identify particular room acoustics as "a consistent cognitive object," but in a broader sense. The results showed that incorrect responses were made especially between particular hall pairs: on one hand, two shoebox shaped halls, AC and MH, were confused with each other, and on the other hand BP and CP were often mixed up. This result indicates that, on average, listeners may not be able to recognize the specific qualities of each individual hall or to distinguish between rooms that are similar in some or many perceptual aspects, but can recognize broad acoustic features and differences between acoustic conditions.

Because objective room acoustic parameters are derived from RIR measurements, they do not depend on the signal. However, the source configuration may have an influence on these measures and more importantly on the subjective perception of acoustic characteristics; however, as indicated at least by the parameter values in Fig. 4, this is hardly the case here, especially because the Beethoven passages and the solo violin were not compared to each other at any stage. Considering the values illustrated in Fig. 4, they appear to be very much in-line with the matching test results. For instance, AC and CP are the furthest away from each other especially in terms *EDT* and $T20$ and, accordingly, they are the ones that are the least confused with each other. Moreover, the pairing AC and MH as well as BP and CP are reflected in the values of G, $C80$, j_{LF}, and L_J. The values of reverberation time parameters *EDT* and $T20$ are very similar for BP and MH, but these halls were not so heavily mixed up with each other in the results. Note that participants reported reverberation as one of the main aspects that they concentrated on, but perhaps reverberation, in general, was the most obvious answer. Regarding the pattern of *EDT* and $T20$ values, it seems that reverberation time was not used as the main determinant among these parameters. Overall the objective parameters reflect the results quite well.

Regarding the interviews, only two people mentioned loudness as a cue, which is interesting because loudness or strength is one of the most fundamental aspects of hall acoustics and it has invariably come up as one of the main discriminative factors also in our previous sensory profiling experiments (e.g., Lokki *et al.*, 2016). Considering the current experiment, we think that recognition in the "same" condition could have been based on loudness differences, but then again, there were perhaps even more prominent cues, such as differences in timbre or reverberation. Loudness differences between the alternatives are probably not a good perceptual cue in the "different" condition because the listener is only given one reference stimulus and can only make comparisons with this single stimulus. This may be the reason why the differences in loudness did not appear to be an important aspect in this study, although the relative loudness differences between halls were real.

Given that these results represent only a limited set of concert halls and only a few types of source sounds, there is a number of interesting questions to be addressed in future studies. For instance, how would listeners perform with a set of halls that were all completely different (or similar) in their acoustic parameters? What about matching across different signal types that may exhibit very different excitation characteristics, for instance, violin vs trombone, flute vs cello, or speech sounds with different voices? Finally, it would also be interesting to extend this study to a wider range of different spaces.

V. CONCLUSIONS

This concert hall acoustics matching experiment showed that most of the listeners were able to find the matching stimuli consistently when the task consisted of listening to the same excitation signals. The task was considerably more difficult when they listened to auralizations made with different signals as a large majority of the listeners did not perform above the chance level in this condition. Moreover, the signals employed in this "different" condition were either subsequent short excerpts of the same musical sequence in a Beethoven symphony or samples of the same solo violin. While the type of music is previously well known to influence the perception of concert hall acoustics, this study showed that even minor changes in the source sound content may have a strong influence on the perception of room acoustics.

When listeners were not able to distinguish the halls on an individual level, the patterns of erroneous responses were not arbitrary, but confusions occurred most often within those pairs of halls which were also similar in terms of the objective parameter values of G, $C80$, j_{LF}, and L_J. This observation indicated that despite listening to different signals, people were able to recognize broad acoustic features and acoustic differences. These results are important for the perceptual evaluations of room acoustics, for instance, in *in situ* listening scenarios where large differences in the source sound content from situation to situation may exist.

ACKNOWLEDGMENTS

We thank Laura McLeod for her work on an earlier version of this experiment and we thank all our test subjects for participating in this experiment. This research was partly funded by the Academy of Finland, Grant No. 296393. An interested reader is encouraged to participate in a shorter online version[1] of this experiment that has been implemented with binaural stimuli.

[1]For more information see https://anttikuusinen.shinyapps.io/RoomAcousticsMatchingExperiment/.

Barron, M. (**1988**). "Subjective study of British symphony concert halls," Acta Acust. united Ac. **66**(1), 1–14.

Frissen, I., Katz, B. F., and Guastavino, C. (**2009**). "Effect of sound source stimuli on the perception of reverberation in large volumes," in *Auditory Display* (Springer, New York), pp. 358–376.

Giannakopoulos, T., and Pikrakis, A. (**2014**). "Chapter 5—Audio classification," in *Introduction to Audio Analysis*, edited by T. Giannakopoulos and A. Pikrakis (Academic Press, Oxford, UK), pp. 107–151.

Hawkes, R., and Douglas, H. (**1971**). "Subjective acoustic experience in concert auditoria," Acta Acust. united Ac. **24**(5), 235–250.

ISO 3382-1 (**2009**). "Acoustics - measurement of room acoustic parameters - Part 1: Performance spaces" (International Standards Organization, Geneva, 2009).

Jäkel, F., and Wichmann, F. A. (**2006**). "Spatial four-alternative forced-choice method is the preferred psychophysical method for naive observers," J. Vision **6**(11), 1307–1322.

Kuusinen, A. (**2014**). "An anechoic audio corpus for room acoustics and related studies," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, April 3–5, Berlin, Germany, pp. 113–118.

Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S. (**2012**). "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," J. Acoust. Soc. Am. **132**(5), 3148–3161.

Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S. (**2016**). "Concert hall acoustics: Repertoire, listening position and individual taste of the listeners influence the qualitative attributes and preferences," J. Acoust. Soc. Am. **140**(1), 551–562.

Pätynen, J. (**2011**). "A virtual loudspeaker orchestra for studies on concert hall acoustics," Ph.D. thesis, Aalto University School of Science, Espoo, Finland.

Pätynen, J., Pulkki, V., and Lokki, T. (**2008**). "Anechoic recording system for symphony orchestra," Acta Acust. united Ac. **94**(6), 856–865.

Schroeder, M., Gottlob, G., and Siebrasse, K. (**1974**). "Comparative study of European concert halls: Correlation of subjective preference with geometric and acoustics parameters," J. Acoust. Soc. Am. **56**(4), 1195–1201.

Teret, E., Pastore, M. T., and Braasch, J. (**2017**). "The influence of signal type on perceived reverberance," J. Acoust. Soc. Am. **141**(3), 1675–1682.

Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (**2013**). "Spatial decomposition method for room impulse responses," J. Audio Eng. Soc. **61**(1/2), 16–27.

Weinzierl, S., Lepa, S., and Ackermann, D. (**2018**). "A measuring instrument for the auditory perception of rooms: The room acoustical quality inventory (RAQI)," J. Acoust. Soc. Am. **144**(3), 1245–1257.