Sound rendering with early reflections extracted from a measured spatial room impulse response

1st Otto Puomio
Dept. of Computer Science and
Dept. of Signal Processing and Acoustics
Aalto University,
Helsinki, Finland
ORCID 0000-0001-8749-2674

2nd Tapani Pihlajakuja
Nokia Technologies
Espoo, Finland
tapani.pihlajakuja@nokia.com

3rd Tapio Lokki
Dept. of Signal Processing and Acoustics
Aalto University
Helsinki, Finland
ORCID 0000-0001-7700-1448

***Abstract*—The acoustic properties of early reflections are not trivial to model when one creates plausible sounding virtual acoustic environments. This paper proposes a novel solution to be used with existing room acoustic modeling systems, such as image-source method, where the synthetic filters for reflection surfaces are replaced with captured real reflection responses. The reflections are extracted from a measured spatial impulse response by analyzing the sound pressure level and the direction-of-arrival to detect prominent individual early reflections. These detected reflections are extracted with a suitable window and are applied in room acoustic modeling algorithms as FIR filters. The presented concept is validated with a listening test, suggesting the material filter length must be over 2 ms. However, an all-round filter length could not be determined due to the dependencies on sound signal content and the room. Further research is therefore required to study the found dependencies, to find the optimal extraction parameters and to validate the filters in practical use-cases.**

***Index Terms*—room acoustics, spatial room impulse responses, early reflections, material filters**

I. Introduction

Virtual reality (VR) audio rendering requires a plausible simulation of the spatial room impulse response (SRIR). Typically, a SRIR consists of directional early reflections (ER) and more diffuse late reverberation. From these two, ERs unconsciously inform the listener about many aspects of the room, for instance the room size and shape as well as wall materials. When all these properties support the visual feedback of the VR headset, they contribute to the overall immersion of the audiovisual experience [1].

Room acoustics modeling aims at simulating the SRIR for the auralization to create the immersive soundscapes [2]. There are dozens of different computational methods to obtain the SRIR [3], [4] and one can listen to the auralizations directly with multichannel systems or binaurally with head-tracked headphones. Recently, researchers have started to call such static auralizations as 3 degrees-of-freedom (DOF) rendering. Basically, it means that a listener is in a static position in a virtual sound environment and can only rotate or tilt her head. Naturally, 3 DOF methods also include virtualizing the multichannel setups with head-related transfer functions (HRTFs).

In many VR applications, the user can navigate, move and turn freely in any direction in a virtual world. Therefore, the room acoustics modeling and spatial sound rendering have to cope with all possible movements of the listener. The acoustic simulation has to be dynamic and spatial sound rendering needs to react to all movements. Thus, the rendering system requires to support 6 DOF that also takes the user's translational movements into account. In fact, 6 DOF rendering is not a novel idea and solutions have already been presented over 20 years ago [5]. Many similar systems have been introduced since those early days. Most of them apply the principle that the direct sound and ERs are dynamically updated according to the movements of the listener and the late reverberation is more or less static within one virtual environment.

This paper adds one more method to modify the well-known 6 DOF sound rendering approach presented in [5]. The presented method replaces the synthetic material reflection filters with real measured ones extracted from SRIRs. The material filter approach is straightforward and do not need any information on the materials. The presented method is validated with a listening test using static binaural rendering.

A. Modeling the early reflections with material properties

It is challenging to model an ER accurately as the reflecting surface can be flat or rough and hard or soft [4]. Therefore, each room acoustic modeling algorithm treats material properties differently [4]. One traditional way is to create a material filter for each reflection using the measured octave band absorption data [6]. With such filters, the frequency-dependent sound attenuation can be incorporated to the image-source method. A rough surface scatters the sound, spreading the energy of a reflection both in time and space. Although the image-source method is only valid for specular reflections, it can be extended to also take rough surface reflections into account [7].

Octave band absorption data can be found from material data tables. If a particular material data is needed, it can be measured with in-situ technique [8]. However, in-situ measurements cannot be done in many cases as the materials

This work has been supported by Nokia Technologies and the Academy of Finland [296393]

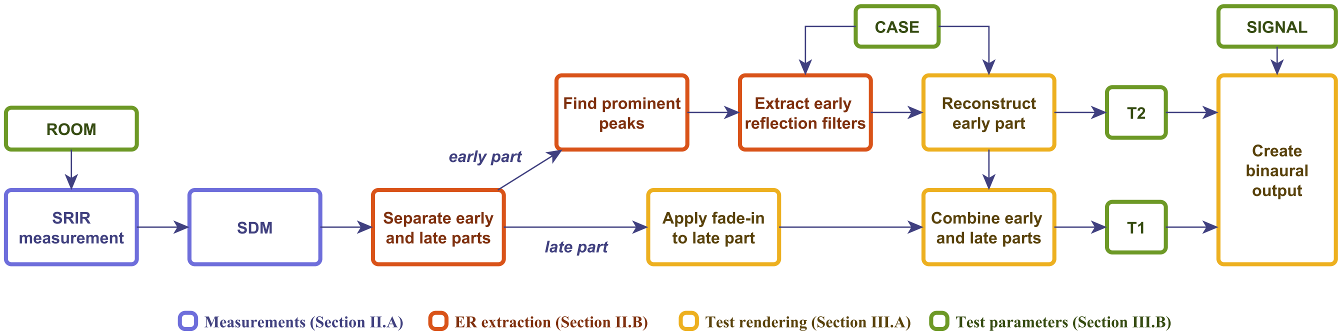

CASE
SIGNAL
ROOM
Find prominent peaks
Extract early reflection filters
Reconstruct early part
T2
early part
Create binaural output
SRIR measurement
SDM
Separate early and late parts
Apply fade-in to late part
Combine early and late parts
T1
late part
Measurements (Section II.A)
ER extraction (Section II.B)
Test rendering (Section III.A)
Test parameters (Section III.B)

Fig. 1. Test setup flowchart.

of the virtual world do not exist or are not available. Naturally, reflections could also be extracted from traditional impulse response measurements with an omnidirectional microphone, but it is really hard to detect a reflection in a measured room impulse response (RIR). To improve the reliability of the detection, this paper also utilizes the spatial information of the measured SRIR.

In brief, this paper proposes a method that aims at extracting material filters from a measured SRIR. The method utilizes a microphone array and the Spatial Decomposition Method (SDM) [9] to detect the ERs and their directions-of-arrival (DOA). The method then extracts the material filters autonomously from the data. The extracted reflections can be used as material filters in any image source based simulation. The assumption is that a measured filter provides a plausible sounding reflection in a simulation of a virtual space.

The paper is organized as follows. First, we introduce the reflection detection method step-by-step in Section II, followed by a description of the proof-of-concept rendering method. Then we evaluate the plausibility of the ER extraction and the proof-of-concept rendering with a listening test, which is described in Section III. The results of the listening test are reported in Section III-D and discussion of the presented method and the results follows in Section IV.

II. Method

The sound rendering is based on the image-source method [10], [11] to model the ERs of a virtual space. The concept presented here replaces the material filters with measured ERs, enabling very simple and straightforward room acoustics modeling. In practice, the image-source method is first used to locate the ERs in terms of timing and direction. Then, measured ER filters are simply inserted to those locations, effectively replacing any simulated reflections and taking reflection properties into account. If the image source visibility is updated according to the user's movements in the virtual world, the presented approach enables both 3 DOF and 6 DOF rendering frameworks. The method enables to collect a database of measured reflections, which can be used in sound rendering in various applications.

A. Measurements

The presented concept requires a spatial impulse response to be able to extract the ERs from the measurement. For this, the method applies the SDM to obtain the required data. To validate the concept, a small lecture room was measured with an open microphone array consisting of 6 omnidirectional microphones in pairs in x, y, and z direction. The distance between the microphones in each pair was 25 mm. In addition, the A-format impulse responses were measured with 192 kHz sampling rate to increase the time resolution of the capture. As the SDM algorithm applies cross-correlation to estimate the time difference of arrivals between microphone pairs, using high sampling rate is beneficial when the distance between microphones is small [12]. The measurement was analyzed with the SDM Toolbox [13] to obtain one omnidirectional impulse response and a sample-wise DOA vector.

B. Early reflection extraction

The ER extraction algorithm is presented in Fig. 1. The algorithm works in two stages. First, ERs are detected from the impulse response by finding prominent peaks. After this, the found peaks are extracted from the pressure signal by using an asymmetric temporal window. The extraction is also limited to the first 100 ms of the impulse response. Discrete ERs can be only found in the beginning of the response as the density of incoming wall reflections increase towards the end of the signal. When the reflection density increases, it also becomes less probable to find non-overlapping reflections. Therefore, analyzing the whole impulse response would only waste computational resources for a marginal gain.

Peak prominence detection follows the procedure presented in [14]. In short, the procedure calculates peak prominence (a.k.a. topographic prominence) and DOA stability for each peak in the SRIR. Peak prominence is determined from RIR sound level smoothed with a 0.125 ms Gaussian window and evaluates as the relative height of the sound levelpeak w.r.t. the neighboring peaks. The more the peak stands out from the signal, the higher its peak prominence. DOA stability in turn measures the reliability of the DOA estimate by counting the number of samples the DOA estimate stays stable. The reliable DOA estimates stay in one place for longer, therefore

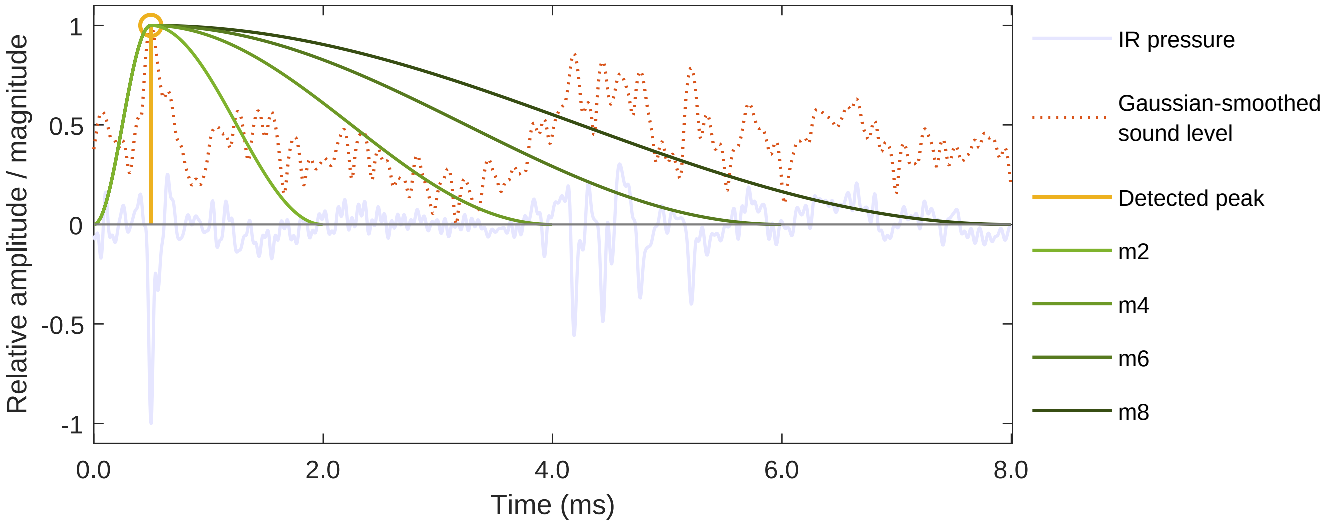


Fig. 2. Early reflection extraction with an asymmetric extraction window. A part of the impulse response (light blue) and the corresponding Gaussian-smoothed sound level (red) are normalized here for better visualization. All extraction windows (shades of green) start 0.5 ms prior to the detected peak (yellow), while total window lengths vary between 2–8 ms.

scoring a higher DOA stability value. After calculating the two features, the detected peak count is reduced by setting a prominence threshold for the analyzed peaks. As a result, one obtains a list of indices describing the ER peak locations that fulfill the given peak prominence and DOA stability criteria.

In the second stage, the prominent ERs are extracted from the pressure signal with an asymmetric temporal window. The reason for an asymmetric window lies in the behavior of arriving sound energy. The majority of this energy arrives via the shortest path from the source to the wall (or walls) and to the listener. In addition, the wall structure may delay the arrival of energy in two ways. First, the surface of the reflecting wall may be rough, spreading the reflection energy in time. Second, walls with a multi-layer structure have typically longer travel times for low frequencies. Those frequencies pass through the first wall layer only to be reflected from the second wall behind it. In addition to the mentioned delay, the energy delay may be also caused by the measurement loudspeaker. The loudspeaker may have prolonged impulse response in the low frequencies, further spreading the sound energy in time. In the end, most of the reflection-related energy mainly arrives after the sound level peak and samples extracted before would not contribute to the material filter design as desired.

The asymmetric window design is presented in Fig. 2. The applied asymmetric window consists of a fixed 0.5 ms hann-window fade-in and $T - 0.5$ ms hann window fade-out, where T is the length of the extraction window. The reflection to be extracted is windowed so that the maximum is located at the maximum of the extraction window.

III. Listening test

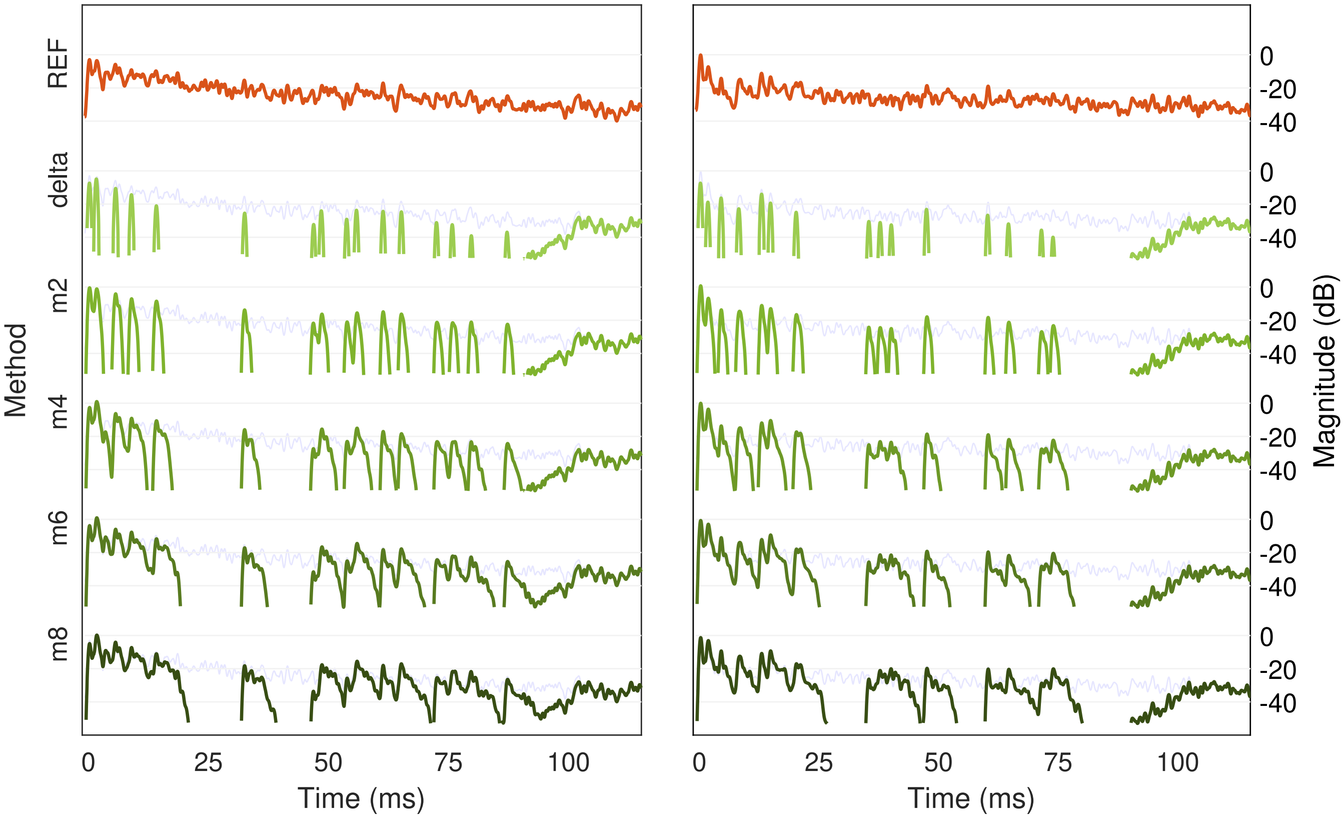
The presented ER extraction method was evaluated with a listening test aiming at discovering two things. First, the test aimed at validating the presented concept and second, it attempted to uncover how the extraction filter length affects the perceived quality of the renderings. As rendering the materials dynamically would have made obtaining consistent results difficult, we decided to implement a simple renderer to generate static but consistent test signals. Furthermore, in the static case the original measurement can be used as a reference against which the renderings are compared.

Due to the restricting circumstances of the COVID-19 pandemic, the listening test was organized remotely. The test participants were experts and enthusiasts in the audio field contacted via the laboratory email list. The listening test was sent to the subjects and they ran it using their own computers and headphones. When the test was completed, the subjects sent the results back to the author by email.

A. *Test rendering*

The implemented test renderer is presented in Fig. 1 in yellow. The renderer reconstructs the early part of the response from the extracted ER filters, optionally attaches reverberation tail to it and finally synthesizes binaural output with a given sound signal. This section describes the actual rendering process, while the following section focuses on the parameters affecting the output.

The early response of the RIR was reconstructed by back-assigning the material filters to an empty vector. The ER filter locations were determined by their extraction locations. For each location, the corresponding filter was selected and normalized. Then the filter was amplitude-modulated so that the maximum pressure value of the extracted filter matched the pressure value of the matched peak. These modulated filters were then accumulated to a single vector that formed a synthesized early response of the RIR. Finally, the equivalent sound level of synthesized response was equalized to match the equivalent sound level of the measured early response. The



(a) Coffee room (ROOM1)

(b) Lecture room (ROOM2)

Fig. 3. Reference RIR (red), delta reconstruction (light green) and reconstructed RIRs for different extraction filter lengths (darker shades of green). The responses are plotted on magnitude scales (in dB) and the late reverberation fades in at 100 ms.

obtained filter formed the no-tail version in the listening test as described later in section III-B.

Obviously, also higher-order ERs are needed to synthesize a plausible sounding SRIR. If the 3D model of the measured room was available, one could reconstruct the higher-order reflections by combining the material filters of the reflecting walls. However, the model was not available for the test renderer; instead, also the prominent higher-order ERs were extracted from the SRIR as material filters. As seen later however, their DoA stability is usually lower than the first-order reflections have.

The listening test also used sound samples rendered with the reverberation tail of the room. For this, the late part of the RIR was attached to the synthesized early parts. First, a fade-in was applied to the late response, starting at t=80 ms and exponentially increasing to full gain at t=100 ms. Then the processed late part was combined with the synthesized early responses, forming the tail versions of the listening test samples.

Finally, the obtained mono impulse responses were spatialized by combining them with the DOAs of the measured impulse response. In this paper, the response is spatialized by generating a synthesized binaural room impulse response (BRIR). The method was chosen by its simplicity, although the RIR could have also been spatialized to a virtual loudspeaker setup [15]. The BRIRs were synthesized with an artificial HRTF dataset [16] by selecting one HRTF for each SRIR DOA [9] and by modulating it with the corresponding pressure value. When accumulated to one vector, these modulated HRTFs synthesized the desired BRIR.

B. Test parameters

These cases were applied under different test conditions and convolved with anechoic recordings. Thus, the independent variables were **ROOM**, **SIGNAL**, **TAIL** and **CASE**. These cases are described below.

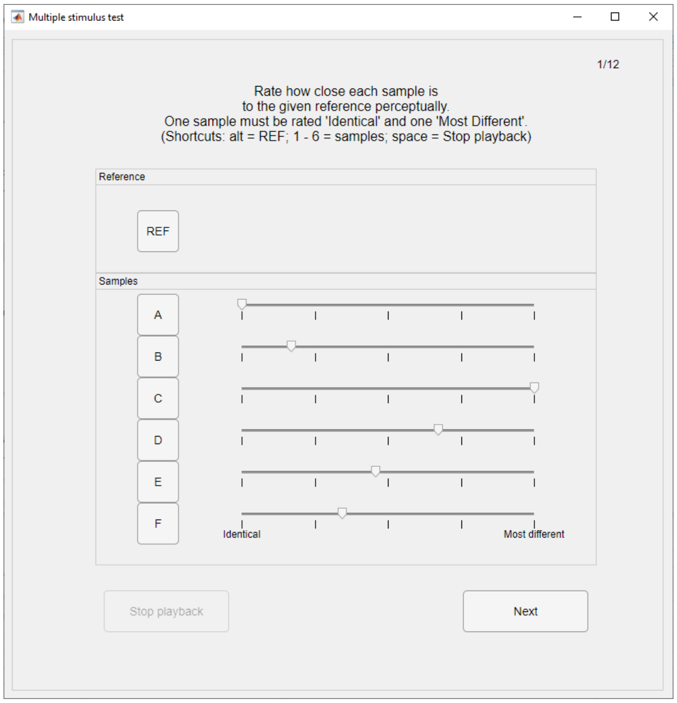
ROOM Two rooms were measured for the listening test as explained in section II-A. The rooms were a concave coffee room with a volume of 199 m^3 (ROOM1) and a small rectangular lecture room of 127 m^3 (ROOM2). The peak prominence threshold was 23 dB for ROOM1 and 25 dB for ROOM2, while DOA stability was left unfiltered in order to catch the higher-order reflections. Found ERs, their peak prominences and DOA stabilities are shown in Tables I and II. The reader should note that the first-order reflections have considerably higher DOA stability values than higher-order reflections. The tabulated ERs were then used to synthesize the SRIRs shown in Fig. 3. The synthesized impulse responses shown in shades of green are noticeably sparser than the reference impulse response shown in red.

SIGNAL The test used three different anechoic recordings: cello, drum and speech. Each recording was 5 – 8 s long. Each signal represented a different sound event; cello represented a continuous signal; drum a transient one; and speech a normal human communication.

TAIL Room reverberation may affect the artifacts the listener can perceive in the sound signal. The test samples were therefore rendered with (T1) or without (T2) the reverberation tail. In brief, T1 applied the late part of the SRIR as described in section III-A. T2 in turn only equalized the sound level of the early part of the response.

CASE Each listening test set consisted of six cases. The cases can be divided into three categories. First, one of the sampels was the reference (**REF**) that was a direct rendering of the measured SRIR either with or without **TAIL**. Second, there was an anchor formed by substituting the found reflection peak positions with delta impulses (**delta**). Third, the rest four cases were formed by extracting and back-assigning ER filters of length 2, 4, 6 and 8 ms (**m2**, **m4**, **m6**, **m8**, respectively). It should be noted that all cases used the exact same DOA vector, meaning that the directional information was not modified between the cases.

C. Listening test implementation



Rate how close each sample is
to the given reference perceptually.
One sample must be rated 'Identical' and one 'Most Different'.
(Shortcuts: alt = REF; 1 - 6 = samples; space = Stop playback)

Fig. 4. Graphical user interface of the listening test.

The listening test was designed as a customized version of the multiple stimulus test with a hidden reference and anchor (MUSHRA). By using the user interface presented in Fig. 4, the subjects were asked to rate the six test samples (CASEs) based on their perceptual distance to the given reference sample. One of the samples had to be rated as ’identical’ and another sample as ’most different’. The subjects ranked each test condition once, resulting in total 12 test sets (2 ROOMs x 3 SIGNALs x 2 TAILs).

The participants were required to complete a training session before the actual listening test. During this session, the subjects evaluated four test sets similar to the actual test. The four sets were selected so that they included both ROOM and

TABLE I
FOUND EARLY REFLECTIONS FOR ROOM1. THE ACTUAL MATERIAL FILTERS HAVE DOA STABILITY VALUE (SECTION II-B) OF MORE THAN 10 SAMPLES (@192 KHZ).

Peak	**TOA** (ms)	**Peak prominence** (dB)	**DOA stability** (samples)
direct sound	0.00	62.1	134
1	1.43	35.1	75
2	5.45	36.3	47
3	8.75	23.3	51
4	13.94	23.2	1
5	32.34	25.2	0
6	46.71	23.9	3
7	48.31	29.9	42
8	53.60	24.4	4
9	55.64	25.3	58
10	61.18	36.0	2
11	64.99	30.8	52
12	72.31	23.9	37
13	75.59	23.8	3
14	79.56	24.7	6
15	87.00	29.5	8

TABLE II
FOUND EARLY REFLECTIONS FOR ROOM2. THE ACTUAL MATERIAL FILTERS HAVE DOA STABILITY VALUE (SECTION II-B) OF MORE THAN 50 SAMPLES (@192 KHZ).

Peak	**TOA** (ms)	**Peak prominence** (dB)	**DOA stability** (samples)
direct sound	0.00	61.9	286
1	1.56	32.0	79
2	4.32	27.4	75
3	8.01	26.8	31
4	12.90	38.8	111
5	14.65	25.3	85
6	20.01	25.2	12
7	35.22	26.1	14
8	37.58	25.7	13
9	39.91	26.0	11
10	47.28	39.7	60
11	60.07	29.2	27
12	64.41	28.2	18
13	71.31	25.6	8
14	73.73	30.2	0

TAIL cases and each SIGNAL at least once. Furthermore, the subjects were instructed to adjust the volume to a comfortable level during the training and keep the volume constant during the actual test.

After the test, the subjects returned their answers via email. Along with their results, they were asked to answer the following questions:

1) What cues did you use to discriminate the samples?
2) Which headphones did you use (model)?
3) Which audio card did you use (model, or 'default' if not specific)?
4) How would you rank the sound samples in terms of difficulty?
5) Any questions/comments/feedback in general?

D. Results

The listening test was completed by 11 subjects. All the participants reported using either spatial cues or sound coloration to discriminate the sound samples. Furthermore, the subjects reported speech the easiest and cello the hardest to grade. One of the reported headphones was found to be considerably lower quality than the others. These headphones had a bass frequency roll-off starting from 100 Hz and 10 – 20 dB gain boost from the baseline on 1.5 – 9 kHz frequencies. For this reason, the subject was excluded from the results. The rest of the headphones were found to have good quality, thus the authors have no reason to assume that the remaining results would have been significantly different if the test would have been conducted in the laboratory. In the end, 10 subjects were accepted for further data analysis.

The listening test results were analyzed with a four-way repeated-measures analysis of variance (RM-ANOVA). The within-subjects factors were the aforementioned **ROOM**, **SIGNAL**, **TAIL**, and **CASE**. The analysis was performed in two steps. The data was first checked for violations on assumptions of sphericity with Mauchly's Test of Sphericity [17]. The test indicated that two interactions listed in Table III violated the assumption significantly ($p < 0.05$, $\varepsilon < 0.75$). Those interactions were therefore adjusted with Greenhouse-Geisser correction during the second step. In that step, the data was tested for significant effects with RM-ANOVA analysis, resulting in eight cases reported in Table IV ($p < 0.05$). There one can find that the fourth-order interaction was not significant but there are two significant third-order interactions, namely room * signal * case and signal * tail * case. Thus, further analysis focuses on these cases.

The two third-order interactions have been visualized in Figs. 5 and 6. In all figures, the CASEs are shown along the x-axis while the grading scale is shown along the y-axis. To clarify the content, each interaction has been divided into two separate plots. Namely, the data is divided by ROOM in Fig. 5 and by TAIL in Fig. 6. The aforementioned parameters also define the colors of the curves (blue and red). To make the comparison between the plots easier, each curve set is drawn lightly behind the other set. The second parameter, namely SIGNAL, determines the shade of the color as well as the marginal mean marker. As none of the figures has no compensations for multiple comparisons, they are only applicable for studying the trends in the data and not to infer statistical significance. To make the trends easier to see, different CASEs are finally connected by a dashed line.

TABLE III
MAUCHLY'S TEST FOR SPHERICITY RESULTS.

Effect	χ^2	df	p	ε
room * case	38.749	14	0.001	0.542
room * tail * case	38.116	14	0.001	0.539

CASE appears to be the strongest predictor for the grading. This can be seen most clearly in ROOM2 (Fig. 5b) and with TAIL2 (Fig. 6b). There, the longer material filters appear to generally grade better. The subjects have also found delta and REF consistently as expected.

However, the grading appears to become SIGNAL-dependent when either ROOM or TAIL is changed. In ROOM1 (Fig. 5a), the results are found to saturate to a different gradings. CELLO appears to perform better the longer the material filter is. In fact, the subjects started to mix up the m8 filter with REF. The second best grades are given to SPEECH starting at m4, and DRUMS performs the worst, getting similar grades for all material filters.

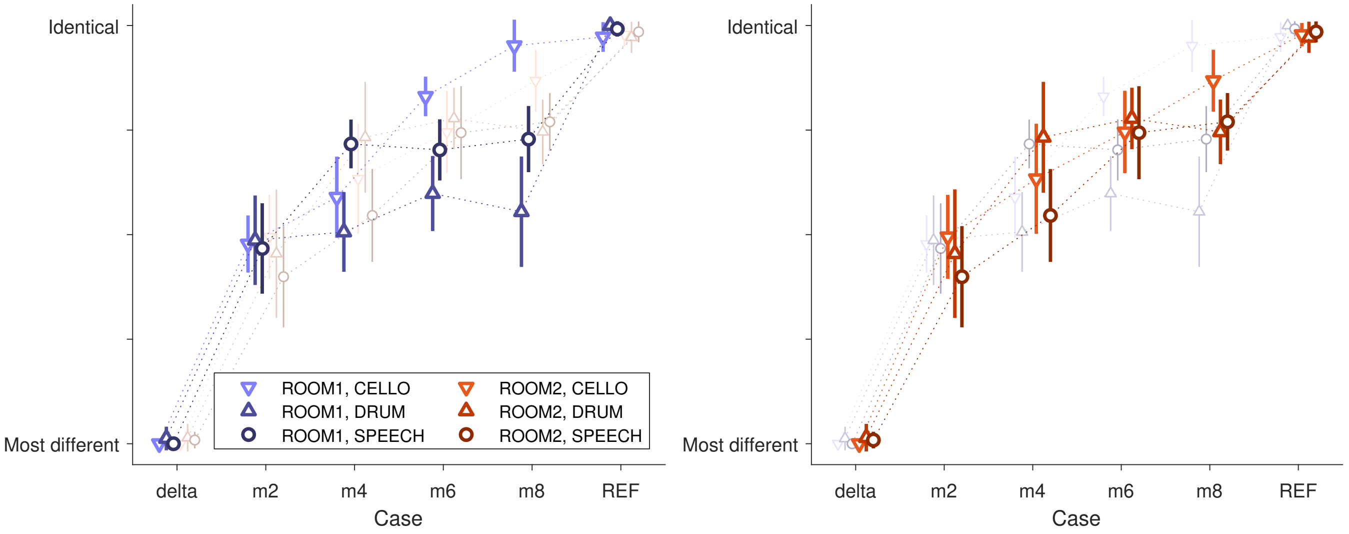
T1 (Fig. 6a) in turn is found to mostly grade slightly better than T2. Also here, CELLO starts to mix up with REF while, surprisingly, DRUM performs worse with m8 than m6. Otherwise, the filters appear to grade similarly from m4 and up.

IV. Discussion

In this paper, the authors presented a method that aimed at extracting material filters from a measured SRIR. As a proof-of-concept, the material filters were extracted with different lengths and then used to synthesize SRIRs. The obtained responses were compared in a listening test that indicated a significant improvement in authenticity over a delta reference. However, other acoustic properties also appeared to affect the results, making it more difficult to draw universal recommendations about an optimal material filter length.

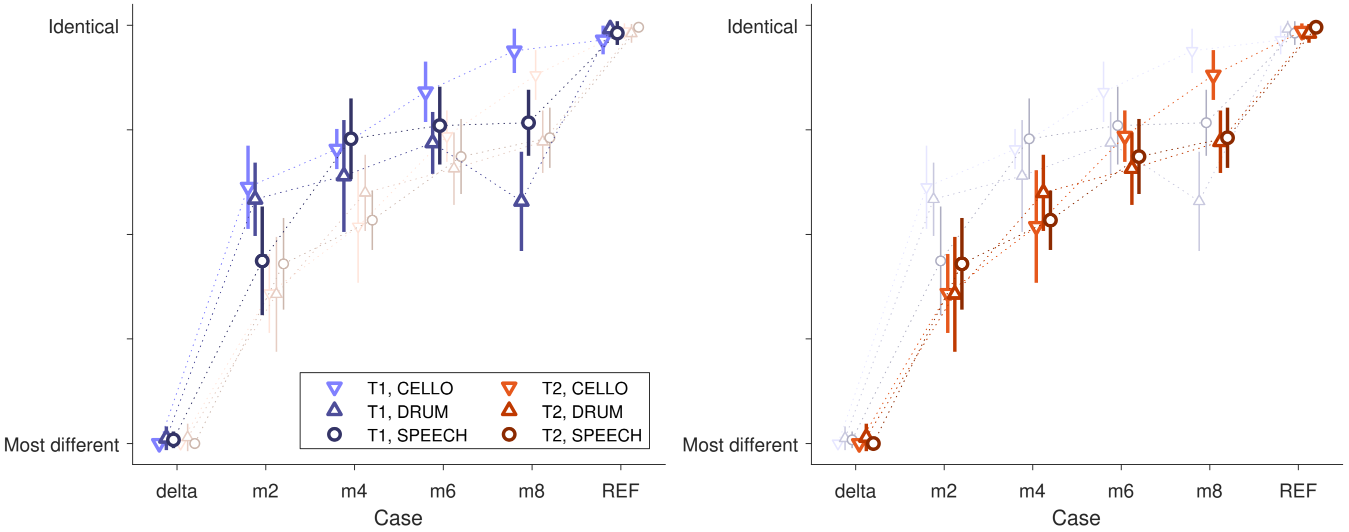
TABLE IV
SIGNIFICANT EFFECTS ($p < 0.05$) IN THE REPEATED-MEASURES ANOVA ANALYSIS OF THE TEST RESULTS.

Effect	df	F	p
signal	(2, 18)	3.989	0.04
tail	(1, 9)	17.024	0.00
case	(5, 45)	676.806	0.00
room * signal	(2, 18)	6.490	0.01
room * tail	(1, 9)	9.440	0.01
signal * case	(10, 90)	5.647	0.00
room * signal * case	(10, 90)	4.866	0.00
signal * tail * case	(10, 90)	3.737	0.00



(a) (b)

Fig. 5. Marginal means and 95 % confidence intervals (N = 10) for room * signal * case interaction.



(a) (b)

Fig. 6. Marginal means and 95 % confidence intervals (N = 10) for signal * tail * case interaction.

The optimal material filter length appeared to vary depending on SIGNAL in different ROOM and TAIL conditions. Nevertheless, the authors suggest that the material filter length should be longer than 2 ms in order to not lose quality. This is due to the fact that ranking still improved after m2 in most cases. One should also note that making the filter longer does not necessarily improve the perceived quality. This is especially apparent in Fig. 6a where T1 DRUM rating drops from m6 to m8. However, the long filters do not always degrade the output either; m8 still appears to benefit T1 CELLO even though the other signals would not. The found differences are probably caused by different signals revealing different properties in the generated SRIRs. The transient sounds of DRUM reveal the differences more clearly than the other two stimuli. In contrast, CELLO is more continuous and harmonic, which probably hides the subtle differences in the SRIR more easily. Finally, SPEECH is somewhere between these two; it has continuous harmonic parts like CELLO does, but also consonants giving more broadband content like DRUM has. To conclude, 6 ms material filters seem to render reasonable quality in the studied rooms, yet the result cannot be generalized without further studies.

There is also a possibility that longer window lengths also capture acoustic events not belonging to that particular material. For instance, longer window lengths might have captured multiple reflections unintentionally, which would definitely affect the timbre. Furthermore, longer filter also uses the measured DOAs at longer time span. When combined with unintentionally captured reflections, also the spatial image would change. Removing the extra reflections could be possible by, for instance, averaging the reflection over multiple measurement positions. This, however, is left for further research.

As seen in Fig. 3, the reconstructed RIRs did not definitely use all the possible reflections in the room. This is due to the prominent reflection policy that leaves out the reflections that have no sufficient peak prominence. The reconstructed SRIRs are therefore imperfect and degrade them further away from REF. In the listening test, this was apparent in ROOM2 where a missing reflection were reported to cause a shift in perceived sound source location. However, all reconstructed CASEs can be assumed to be affected by same amount as they used the same reflection locations. Therefore, one can assume that these kind of errors affect the maximum grade each CASE can get. On the other hand, the error can be assumed to not affect the order of the CASEs graded in the same test set.

The implemented listening test did not follow the use scenario described in the introduction of this paper. The original motivation was to simulate a room with the image-source method by applying extracted early reflections as material filters. Moreover, there could indeed be a library of extracted early reflections from which the suitable ones could be used according to defined criteria. Such rendering case was not tested here because the reference would have been impossible to define. Furthermore, we did not find out a convenient way to judge the quality of different renderings.

In the end, the reflection filters do improve the authenticity of the rendering when compared to delta impulses. In addition, the listening test results raise many questions that need further research. Many research directions have been suggested above; in addition, the material filters should definitely be tested in 6 DOF rendering.

V. Conclusions

This paper introduced a method to replace simulated material properties with measured early reflections in an acoustic simulation. With this, we aimed at creating a plausibly simulated room impulse response, assuming that applying measured material filters provides us with more acoustic detail than simulating material filters from scratch.

The extracted filter performance was compared to a corresponding delta impulse response in the listening test with varying filter extraction lengths. We found that extracting filters from a measured room impulse response do capture the material properties of the room. However, the required filter length appears to be also heavily room and stimulus dependent, which calls for more research in the future. Many other research directions were also pointed out, such as applying filters to a dynamic 6 degrees-of-freedom simulation. In the end, the extracted material filters were found to be a promising step towards authentically replicating a room for a 6 DoF rendering.

VI. Companion page

The companion page includes the sound tracks used in the listening test. The page is located at

http://research.spa.aalto.fi/publications/papers/i3da21-er/.

References

[1] M. Slater, “Immersion and the illusion of presence in virtual reality,” *Br. J. Psychol.*, vol. 109, no. 3, pp. 431–433, 2018.

[2] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, “Auralization – an overview,” *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861–875, 1993.

[3] U. P. Svensson and U. R. Kristiansen, “Computational modeling and simulation of acoustic spaces,” in *AES 22nd Int. Conf. on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002, pp. 11–30.

[4] L. Savioja and U. P. Svensson, “Overview of geometrical room acoustic modeling techniques,” *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 708–730, 2015.

[5] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, “Creating interactive virtual acoustic environments,” *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 675–705, 1999.

[6] J. Huopaniemi, L. Savioja, and M. Karjalainen, “Modeling of reflections and air absorption in acoustical spaces — a digital filter design approach,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA’97)*, Mohonk, New Paltz, New York, 1997.

[7] S. Siltanen, T. Lokki, S. Tervo, and L. Savioja, “Modeling incoherent reflections from rough room surfaces with a beam tracer,” *J. Acoust. Soc. Am.*, vol. 131, no. 6, pp. 4606–4614, 2012.

[8] Y. Takahashi, T. Otsuru, and R. Tomiku, “In situ measurements of surface impedance and absorption coefficients of porous materials using two microphones and ambient noise,” *Appl. Acoust.*, vol. 66, no. 7, pp. 845 – 865, 2005.

[9] 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. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=16664

[10] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[11] J. Borish, “Extension of the image model to arbitrary polyhedra,” *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[12] S. Tervo, J. Pätynen, N. Kaplanis, M. Lydolf, S. Bech, and T. Lokki, “Spatial analysis and synthesis of car audio system and car-cabin acoustics with a compact microphone array,” *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 914–925, 2015.

[13] S. Tervo, “SDM toolbox,” Matlab toolbox. [Online]. Available: https://se.mathworks.com/matlabcentral/fileexchange/56663-sdm-toolbox

[14] O. Puomio, N. Meyer-Kahlen, and T. Lokki, “Locating image sources from multiple spatial room impulse responses,” *Appl. Sci.*, vol. 11, no. 6, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/6/2485

[15] O. Puomio, J. Pätynen, and T. Lokki, “Optimization of virtual loudspeakers for spatial room acoustics reproduction with headphones,” *Appl. Sci.*, vol. 7, no. 12, p. 1282, 2017. [Online]. Available: http://www.mdpi.com/2076-3417/7/12/1282

[16] T. Huttunen and A. Vanne, “End-to-end process for HRTF personalization,” in *Audio Eng. Soc. Convention 142*. Berlin, Germany: Audio Engineering Society, 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=18723

[17] J. W. Mauchly, “Significance test for sphericity of a normal n-variate distribution,” *Ann. Math. Stat.*, vol. 11, no. 2, pp. 204–209, 1940.