McKenzie, Thomas; Schlecht, Sebastian J.; Pulkki, Ville

Auralisation of the Transition between Coupled Rooms

Published in:
2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)

DOI:
10.1109/I3DA48870.2021.9610955

Published: 10/09/2021

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
McKenzie, T., Schlecht, S. J., & Pulkki, V. (2021). Auralisation of the Transition between Coupled Rooms. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (pp. 1-9). Article 9610955 IEEE.
https://doi.org/10.1109/I3DA48870.2021.9610955

Auralisation of the Transition between Coupled Rooms

Thomas McKenzie
Dpt. Signal Processing and Acoustics
Aalto University
Espoo, Finland
thomas.mckenzie@aalto.fi

Sebastian J. Schlecht
Dpt. Signal Processing and Acoustics
and *Dpt. Media, Aalto University*
Espoo, Finland
sebastian.schlecht@aalto.fi

Ville Pulkki
Dpt. Signal Processing and Acoustics
Aalto University
Espoo, Finland
ville.pulkki@aalto.fi

***Abstract*—The perceptual experience of the transition between coupled rooms remains a little investigated area of research. This paper presents a pipeline for auralising the transition between coupled rooms, utilising a time-varying partitioned convolution for fast position-dependent switching between spatial room impulse responses (SRIRs) and parametric binaural rendering over highly acoustically transparent headphones, with in-situ calibration to the corresponding real-world acoustics. The system is verified by an in-situ listening test with both real and virtual stimuli, conducted in six degrees-of-freedom virtual reality with three-dimensional visuals from measured room models. Results show that the auralisation is rated as highly natural, equalling the naturalness of the corresponding real world auditory stimuli. This pipeline is therefore appropriate for testing of coupled room transition algorithms and SRIR interpolation techniques, as well as non-in-situ testing.**

***Index Terms*—Spatial room impulse response, six degrees-of-freedom, coupled rooms, room transition**

I. Introduction

Room acoustics provide important spatial cues for the human auditory system. The presence of early reflections and reverberation causes auditory objects to be externalised [1], [2], as well as allowing the inference of the geometry and size of the space to the listener [3]. Several characteristics of reverberation are dependent on the source and receiver positions in a room, such as direct-to-reverberant ratio, early reflections and modal coupling, while reverberation time remains largely constant.

However, when coupled rooms are considered, features such as double-slope decays in the room response, edge diffraction and portalling effects occur [4], [5], all of which vary with inter-room position and coupling aperture size (for example, see Fig. 1). Portalling is used in this paper to refer to scattering and diffraction around the coupling aperture, which gives the perception of the sound source location at the coupling aperture [6]. The transition between coupled rooms is therefore a highly complex interaction, and poses significant challenges for acoustical modelling [7]. While some research has been undertaken in understanding the effects of coupling aperture size on propagation and diffusion effects of coupled rooms [8], little has been published on the perceptual experience of travelling between coupled rooms. When a listener moves in a single shoebox room, acoustical changes are smooth and gradual. In room transitions, however, rapid changes in acoustics occur with small positional changes.

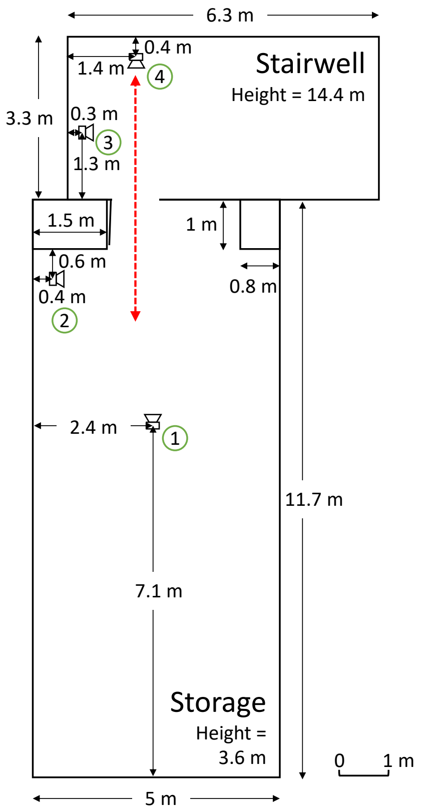


Fig. 1. Room geometry and loudspeaker locations of the coupled room transition. Measurement trajectory denoted by red arrow. Loudspeakers are numbered in green: 1 and 4 retain a continuous line-of-sight between the loudspeakers and microphone for all measurement positions, whereas 2 and 3 feature occlusion at some measurement positions.

Measurements of room acoustics are widely used in virtual reality systems offering six degrees-of-freedom (6DoF) immersive experiences, dereverberation algorithms for more efficient speech recognition [9] and even aiding reconstruction of historic monuments [10]. The measurement of the acoustic response of a room, the room impulse response (RIR), is

The authors would like to thank the Human Optimised XR (HumOR) Project for funding this research.

typically made using a loudspeaker as the sound source playing an exponential sine sweep excitation signal, and a microphone as the receiver [11]. RIRs measured with spherical microphone arrays, which use the principles of Ambisonics to encode microphone signals into spherical harmonic (SH) format [12], are known as spatial room impulse responses (SRIRs). These allow for greater flexibility post measurement, as they can be analysed using directional approaches, and can be reproduced over both loudspeaker arrays and headphones.

Recent literature has investigated how RIRs change with different receiver positions inside a single room, both for virtual reality [13], [14] and dereverberation applications [15]. While it is possible to interpolate between measured RIRs [16], [17], the perceptual requirements for inter-measurement distance vary with auditory stimuli, whereby sounds with limited frequency bandwidth can forgive larger distances between measurements [18], and the greater diffuseness of late reverberation allows for different measurement distances for different parts of the impulse response [13]. However, given the greater complexity of coupled room acoustics, these findings may indeed be inapplicable [7].

In a previous study, a dataset of 101 SRIRs was recorded using a fourth-order spherical microphone array in 5 cm intervals from 2.5 m inside one room to 2.5 m inside the adjacent room. Four coupled room transitions were measured, each repeated with four different source positions, inside each room, both with and without a continuous line-of-sight (CLOS) between the source and receiver [19]. Analysis of the measurements showed clear trends: direct-to-reverberant ratio decreases when the source and receiver are in opposing rooms with no CLOS, and increases for less reverberant rooms. These effects are greater when the difference in reverberation between the two rooms is larger, and change depending on the source position. Directional analysis showed that the reflection patterns are generally consistent in each room, but become intricate in the region around the coupling aperture. Additionally, it shows the presence of strong reflections, sometimes with a greater amplitude than the occluded direct path, especially around the coupling aperture. The portalling effect commonly occurs when the source and receiver are in opposing rooms with an occluded direct sound.

As the transition between coupled rooms is highly complex, it will most likely require a greater accuracy in reproduction. For interpolation between two RIR measurements, this is therefore likely to be a more demanding task than for two measurements inside the same room. This paper presents a pipeline for dynamic auralisation of SRIRs via a time-varying convolution method that allows for fast switching of filters to compensate for changes in listener position, with parametric binaural rendering over acoustically transparent headphones. The system is calibrated in-situ to the response of the corresponding real-world loudspeaker and acoustics, and features virtual reality visuals from three-dimensional room models, and user tracking. The pipeline is evaluated through an in-situ perceptual listening test including both real and virtual stimuli, conducted in 6DoF virtual reality, to assess how natural the auralisation is perceived when compared to the corresponding real world scenario. For this reason, acoustic measurements are preferred to simulations, which can produce plausible auralisations, but not authentic when compared to corresponding impulse response measurements [20].

The paper is laid out as follows: Section II details the auralisation pipeline, the in-situ calibration and the acquirement of room models for virtual reality visual rendering. Section III presents the methodology and results of a perceptual listening test to evaluate the naturalness of the system. Section IV discusses the results of the evaluation, before final remarks along with further work are concluded in Section V.

II. Methods

This section details the dynamic auralisation system for reproducing the transition between coupled rooms using a dataset of SRIRs. A time-varying convolution solution for real-time rendering of the measured rooms for the given source and receiver positions is first presented, using SH filters to allow for dynamic binaural rendering. The reproduction and calibration stages are then covered in detail, to bring the auralisation as close to the corresponding real world sound as possible, and to make appropriate direct comparisons between the binaural auralisation and loudspeaker rendering. Additionally, the acquirement of corresponding room models for three-dimensional virtual reality visuals is also described.

In the real time system, audio is processed on Cycling 74 Max, version 8.1.8, delivered using an Apple Macbook Pro with a Fireface UCX audio interface, which has software controlled input and output levels. Four Genelec 8331A coaxial loudspeakers are used, with the central point of the loudspeaker coaxial drivers at a height of 1.5 m, corresponding to the approximate average height of the human mouth [21]. Visuals are processed on Unity, version 2020.3.11f1, and delivered to the user using a Lenovo Legion laptop and an Oculus Rift S virtual reality headset. All audio in this study is 24-bit resolution with a sample rate of 48 kHz. Sound field orientation and positional data, for dynamic binaural rendering, is obtained from the virtual reality headset, sent via Open Sound Control (OSC) from Unity to Max.

The impulse responses used in this study are from the dataset of measured SRIRs for the transition between coupled rooms [19], available under a Creative Commons license[1]. The room transition investigated in this paper is from a storage space to a stairwell, with background noise levels of 32.8 dBA and 35.2 dBA, respectively, measured using a Sinus Tango sound pressure level meter at a position of 2.5 m from the coupling aperture. The RT60s of the storage space and the stairwell are 0.29 s and 0.73 s, respectively, at a distance of 2.5 m from the coupling aperture. These are calculated from the omnidirectional channel of the measured SRIRs as the mean of calculations at 500 Hz and 1 kHz freq bands, extrapolated from an RT30 measurement from -5 dB to -35 dB, using the IOSR MATLAB toolbox[2]. An illustration

[1] http://doi.org/10.5281/zenodo.4095493

[2] https://github.com/IoSR-Surrey/MatlabToolbox

of room geometries and loudspeaker locations is presented in Fig. 1. Loudspeakers 1 and 4 retain a CLOS with the listener for all positions, whereas loudspeakers 2 and 3 have an occluded line-of-sight at some positions.

A. Time-varying Convolution with Switching of Filters

Input signal convolution with the SRIRs is achieved through a virtual studio technology (VST) plugin, written in MATLAB, version R2020a. The plugin uses fast convolution in the frequency domain [22] with the overlap-add method. A challenge is in updating the filters in real time without audible artefacts, whilst minimising latency, computational load, and RAM usage. To achieve this, partitioned convolution was employed. Fig. 2 presents a block diagram of the time-varying convolution algorithm.

The SRIRs are 1.5 seconds long, at a sampling rate of 48 kHz, so the impulse response length $N = 72000$ samples. The partition block size $K = 1024$ samples, which equates to 21.3 ms. The filters are divided into B blocks of K samples, so in this case $B = 71$, with the final part of the impulse response zero padded to complete the final block. Each block is then zero padded by K and the frequency domain equivalent is computed using a discrete Fourier transform (DFT). To reduce computational load, only the first half of the DFT result is saved.

For each block of input signal in the process loop, the convolution method is as follows. The signal matrix is circular rotated by one column. This puts the previous block of input signal into the second column. The current block of input signal is zero padded by K and the DFT is taken, before which the first half of the DFT result is saved into the first column of the signal matrix.

Each block of the signal matrix is then multiplied by each block of the filter matrix corresponding to the chosen filter selection, and the results of each block are then summed. The block is then duplicated, flipped and the complex conjugate is taken to rebuild the second half of the frequency domain signal, before being converted into the time domain using an inverse DFT. The first half of the convolution result is the output, which is summed with an overlap v made from the second half of the previous block convolution result.

With no changing of filters, this overlap-add method would be sufficient for artefact free convolution. However, when changing the filter in real time using this basic algorithm, interference occurs due to the mismatch between the overlap and the convolution result [23]. To mitigate this, the convolution process is repeated with the filter selection of the previous input signal block. This previous filter selection is the one used as the signal output in most cases. In the event that the filter is changed, the overlap from the new filter is used instead, which avoids the overlap mismatch interference.

While this produces no overlap mismatch, there can still be a jump from one block to the next due to the abrupt changing of filter. This is mitigated by repeating the convolution stage a third time, for the filter of two input signal blocks previous, and implementing a linear crossfade between the time-domain output signals. Again, in the event that the filter is changed, this convolution is not used, and the crossfade would be between the first and second convolution. Although the computational cost is increased by repeating the convolution operation three times, the resulting signals do not suffer from audible glitches when the filter selection is changed.

While the filter selection-dependent overlap introduces a delay of one block length between the convolved filter selection and the chosen filter, this does not equate to additional signal latency, as the signal matrix is not delayed. However, this potential mismatch between filter selection and user position is unlikely to be perceivable, as the block size used is 21.3 ms, and 100 ms is generally considered the threshold of audio latency [24], although it can be as low as 60 ms [25].

Due to computational limitations, the SRIRs were truncated to a length of 0.75 s. However, this is still longer than the measured RT60 time of the two rooms and so should not be perceivable. Additionally, the SRIRs were truncated to a SH order of 3, due to the limited reproduction order of the parametric binaural decoding method used.

B. Binaural Rendering

The convolved SH signals are delivered to the user over headphones, using the higher-order Directional Audio Coding (HO-DirAC) binaural decoding VST plugin from the Sparta plugin suite [26], [27], and non-individualised Neumann KU 100 head-related transfer functions (HRTFs) from the SADIE 2 database [28].

The choice of binaural rendering method was made by listening and auditioning different options. A linear decoding method, such as binaural Ambisonic rendering with an HRTF pre-processing method [29], would reproduce the low frequency sound field with physical accuracy, but at high frequencies the sharpness and localisation of sources may be imprecise and blurred, and suffer from timbral errors. On the other hand, parametric decoding methods such as HO-DirAC and COMPASS [30] offer higher accuracy in localisation and timbre at high frequencies, though they make psychoacoustic assumptions in the processing [31]. Preliminary listening, comparing the binaural decoding options in-situ with the loudspeakers, showed a clear preference for the parametric methods, with greater externalisation, localisation and sharper source width. A recent study showed reverberation can be perceptually sufficient for SH orders of three or higher, provided a very high order of rendering for the direct sound and accurate early reflection rendering [32]. This supports the choice to prioritise the localisation and timbre and choose a parametric binaural decoding method.

In order to directly compare loudspeaker audio with the auralisation, the headphones chosen for binaural delivery had to feature a fast transient and even frequency response, and be as close as possible to free-air equivalent coupling (FEC) [33], meaning they should be acoustically transparent. The Mysphere 3.2 headphones, a modern day sequel to the AKG K1000, feature a flat frequency response and highly open over-

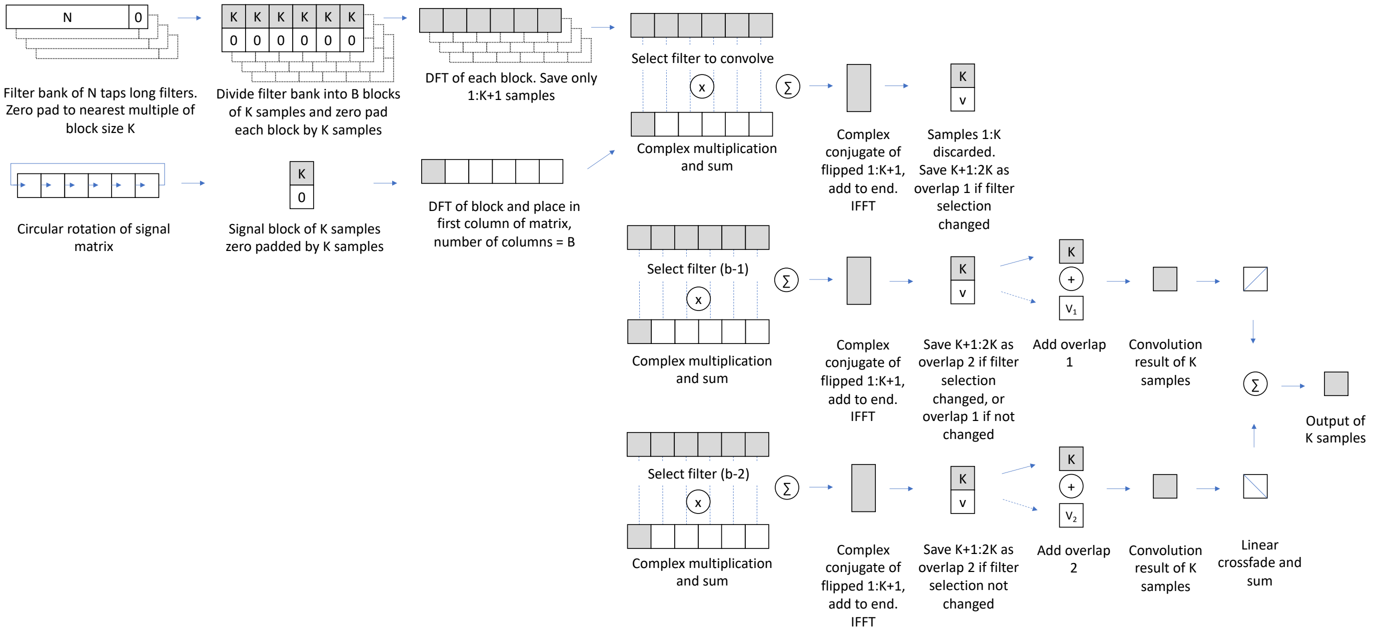

N
0
Filter bank of N taps long filters. Zero pad to nearest multiple of block size K
K K K K K K
0 0 0 0 0 0
Divide filter bank into B blocks of K samples and zero pad each block by K samples
DFT of each block. Save only 1:K+1 samples
Select filter to convolve
x
∑
Complex multiplication and sum
Complex conjugate of flipped 1:K+1, add to end. IFFT
K
v
Samples 1:K discarded. Save K+1:2K as overlap 1 if filter selection changed
Circular rotation of signal matrix
K
0
Signal block of K samples zero padded by K samples
DFT of block and place in first column of matrix, number of columns = B
Select filter (b-1)
x
∑
Complex multiplication and sum
Complex conjugate of flipped 1:K+1, add to end. IFFT
K
v
Save K+1:2K as overlap 2 if filter selection changed, or overlap 1 if not changed
K
+
V₁
Add overlap 1
Convolution result of K samples
∑
Output of K samples
Select filter (b-2)
x
∑
Complex multiplication and sum
Complex conjugate of flipped 1:K+1, add to end. IFFT
K
v
Save K+1:2K as overlap 2 if filter selection not changed
K
+
V₂
Add overlap 2
Convolution result of K samples
Linear crossfade and sum

Fig. 2. Block diagram of the overlap-add time-varying convolution workflow for fast switching between filters, which in this study are fourth-order spatial room impulse responses. This represents the processing for one block of input signal b.

ear design whereby the sound frames are not in contact with the ear. They therefore appeared suitable to this study.

To test the transparency of the headphones alongside two other headphone models, a G.R.A.S. KEMAR dummy head was placed in the anechoic spherical 45 loudspeaker array at Aalto University, and HRTF measurements were taken for all loudspeakers. The measurements were then repeated with two configurations of the Mysphere headphones, as well as the Sennheiser HD 650 and AKG K702. To obtain an overall estimation of transparency, the magnitude difference between the mean of the FFT of all HRTFs taken wearing no headphones versus wearing the four headphone configurations was calculated. Fig. 3 presents frequency plots of the calculated magnitude differences for the four configurations tested (left ear). Whereas the Mysphere headphones disturb the sound field at high frequency magnitudes of up to 4 dB and 12 dB with open and closed frames, respectively, the Sennheiser HD 650 and AKG K702 headphones produce perturbations of up to 22 dB and 16 dB, respectively. The plots therefore demonstrate the improved transparency of the Mysphere headphones and thus their suitability to applications comparing real and virtual sound fields.

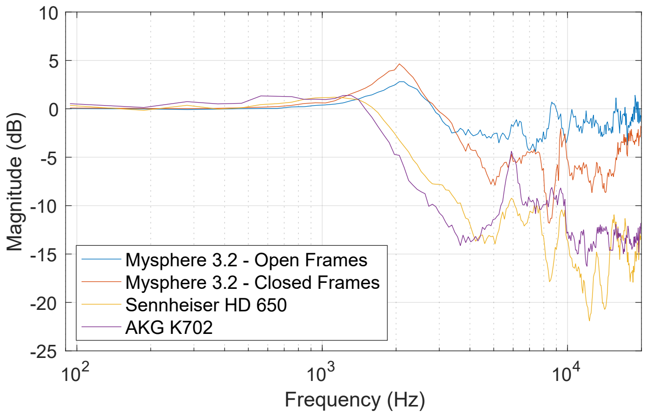
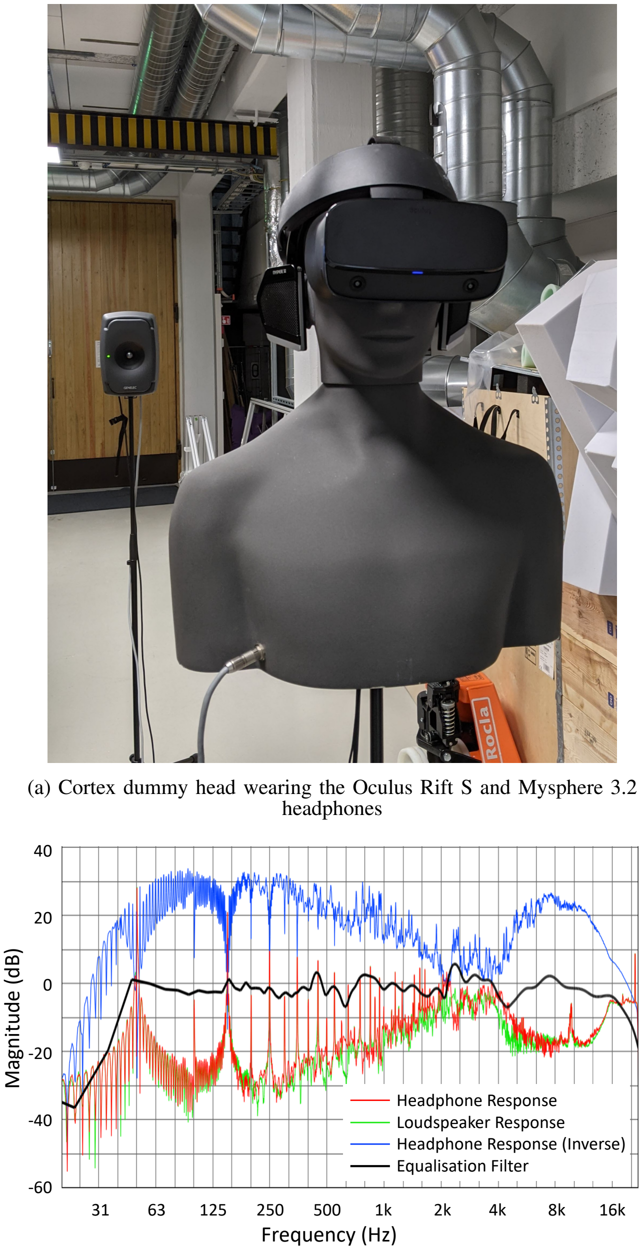


Fig. 3. Frequency plots of the magnitude difference between KEMAR HRTFs measured first wearing no headphones and then with four headphone configurations. Average of 45 locations around the sphere (left ear).

C. In-situ System Calibration

The auralisation system is calibrated in-situ to the response of one of the loudspeakers. This is achieved using a Cortex dummy head, placed 2.5 m inside the storage space, which has the headphones fitted and the virtual reality headset on, and loudspeaker 1 (see again Fig. 1). The system is aligned such that the virtual orientation and position matches the real orientation and position, then two binaural RIRs are measured using the exponential sine sweep method [11]. The impulse responses are measured and processed using elements from the HISSTools toolbox for Cycling 74 Max [34]. The sweeps have a 5 second duration and a frequency range of 20 Hz to 22.05 kHz.

The first sweep is played back through the loudspeaker, and the second is played back through the auralisation system and delivered over the Mysphere headphones. The second sweep is therefore being convolved with the corresponding SRIR measurement, being rotated to the same position as the real loudspeaker is, and convolved with HRTFs. The headphone response is then inverted using regularisation parameters from 0 dB at 20 Hz to −40 dB at 60 Hz, and then from −40 dB at 10 kHz to 0 dB at 22 kHz, and convolved with the loudspeaker response to produce the calibration filter, which is smoothed using 1/8 octave smoothing, made minimum phase to avoid introducing latency, and truncated to a length of 4096 samples.



(a) Cortex dummy head wearing the Oculus Rift S and Mysphere 3.2 headphones

(b) Frequency plot of the system equalisation

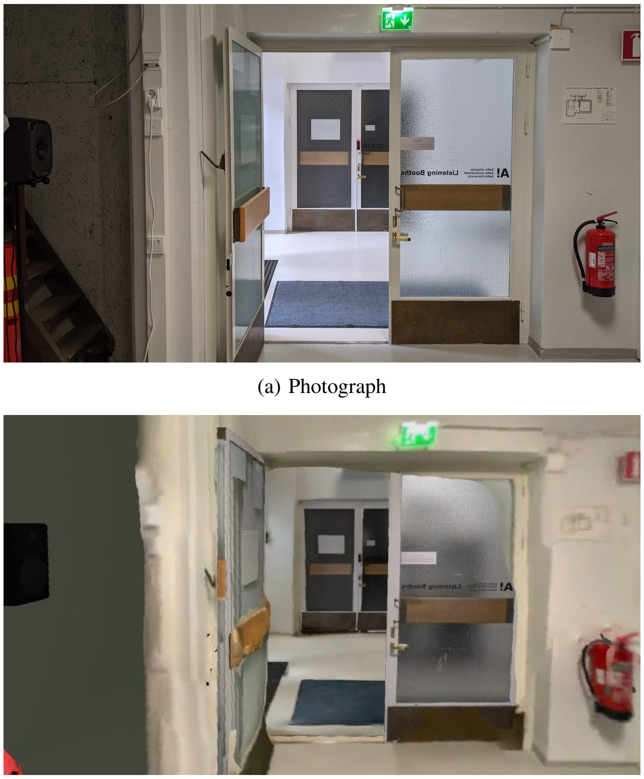
Fig. 4. System calibration for real and virtual comparison. The headphone response, with the auralisation, was equalised to the loudspeaker response.

The calibration setup and measured frequency responses are presented respectively in Fig. 4. The similarity between the measured headphone and loudspeaker responses demonstrates how close the system already is, before calibration. The equalisation filter is relatively flat, with gains lower than ±10 dB for frequencies between 60 Hz and 16 kHz.

Finally, the output of the auralisation is convolved with this filter. This produces the loudspeaker frequency response over headphones through the auralisation system. It should be noted that as the calibration uses the Cortex dummy head, the equalisation will not perform as well with other listeners. Theoretically, if the auralisation system was perfect, one calibration filter should be sufficient for all loudspeakers, positions and orientations. In practice, however, it is likely that this equalisation will perform the best at the measured position, and may be less accurate elsewhere, due to the many sources of system error in the auralisation pipeline, such as the limited spatial resolution of the spherical microphone array used to measure the SRIRs, non-individualised HRTFs and headphone fitting.

D. Room Scan Measurement and Visual Rendering

To display the room transition in virtual reality, three-dimensional models of the two rooms were captured using light detection and ranging (LIDAR) technology from an Apple iPad Pro. The models were refined and reduced in file size in Blender (version 2.92) to improve real-time rendering performance, before being imported into Unity where certain features were enhanced, such as the doors and windows, with higher resolution two-dimensional textures and sharper edges being added. A comparison of the corresponding real and virtual environments is shown in Fig. 5. The loudspeaker model is movable in the environment, such that whichever loudspeaker is currently playing is displayed (as determined in Max, and sent to Unity via OSC). User position and orientation data, for convolution filter selection and sound field rotation, respectively, is sent from Unity to Max via OSC.



(a) Photograph

(b) 3D scan

Fig. 5. Comparison of an in-situ photograph of the room transition to the corresponding visuals in Unity.

III. Evaluation

To evaluate the perceived realism of the auralisation system, a listening test was conducted in-situ, with virtual reality visuals and both loudspeaker and headphone audio. Participants were presented with one condition and loudspeaker combination at a time, and were asked to walk the transition and rate the sound quality in terms of 'naturalness'.

A. Listening Test Design

Four conditions were investigated. These are labelled as:

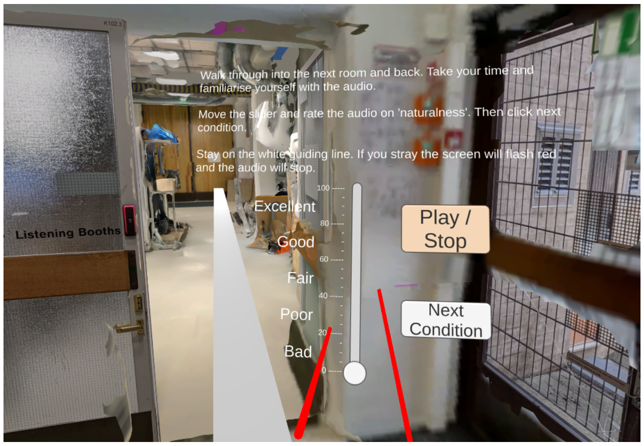
- *Real*
- *Virtual Full*
- *Virtual X Fade*
- *Virtual Direct Sound*

where the *Real* condition is the in-situ loudspeakers. In all three auralisation (*Virtual*) conditions, the first 1 ms of the convolution used the full resolution of 101 measurements over the 5 m distance, as accurate direct sound has been shown to be crucial in auralisations [32]. Note that the direct sound of the SRIRs is time-aligned. The *Virtual Full* used the full resolution of measurements for the whole auralisation system. The *Virtual X Fade* used the full resolution of measurements for the first 1 ms of convolution, as in the full system, but past 1 ms of the SRIRs, the convolution was a linear interpolation between the first and last measurement, whereby listener positions between the two extremes would be a mix of the two. This condition is intended to test a scenario whereby only fixed measurements inside each room are available, with no measurements of the transition. Finally, the *Virtual Direct Sound* features just the first 1 ms of the SRIRs in full resolution, with no reverberation. This is intended as an anchor condition.

The test stimulus was a dry recording of a drumkit, chosen for the inclusion of transients, sharp attacks, and a wide range of frequency content. The four loudspeakers and four test conditions made a total of 16 trials. Trial ordering was randomised and blind.

The listening test instructions and control panel were placed in the virtual environment: the position of these was controlled by the Oculus left hand controller, and interactions made using the trigger on the Oculus right hand controller. A screenshot of the Unity scene control panel is presented in Fig. 6. To ensure participants stay inside the bounds of the measured impulse responses, a guiding line is placed at 1.2 m above the ground in the Unity scene, from 2.5 m inside the storage space to 2.5 m inside the stairwell, corresponding to the position of the measurements. Should the participant stray more than 25 cm from the guiding line in the X or Z axis, the screen flashes red and the audio cuts out.

Listening tests were conducted on 15 participants aged between 24 to 40 (13 male, 2 female) with self reported normal hearing and prior critical listening experience (such as education or employment in audio or music engineering). A photograph of a participant taking part is presented in Fig. 7.



Walk through into the next room and back. Take your time and familiarise yourself with the audio.

Move the slider and rate the audio on 'naturalness'. Then click next condition.

Stay on the white guiding line. If you stray the screen will flash red and the audio will stop.

Fig. 6. Screenshot of the virtual reality visuals, displaying the control panel and instructions for the test, as well as a white line designating the allowed path to follow. Should the participant stray more than 25 cm from the guiding line in the X or Z axis, the screen flashes red and the audio cuts out.

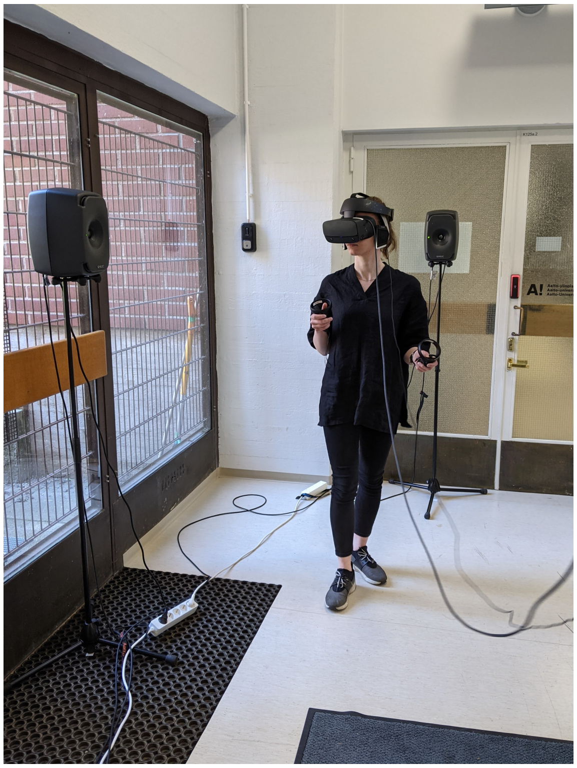


Fig. 7. Test participant performing the task.

B. Results

The results of the listening test are presented as violin plots in Fig. 8. Violin plots were chosen as they display the density trace and box plot in a single illustration, which shows the structure of the data more than a traditional box plot [35]. The width of the violins show the density of data, median values are presented as a white point, the interquartile range is marked using a thick grey line, the range between the lower and upper adjacent values is marked using a thin grey line, and individual results are displayed as coloured points. The data

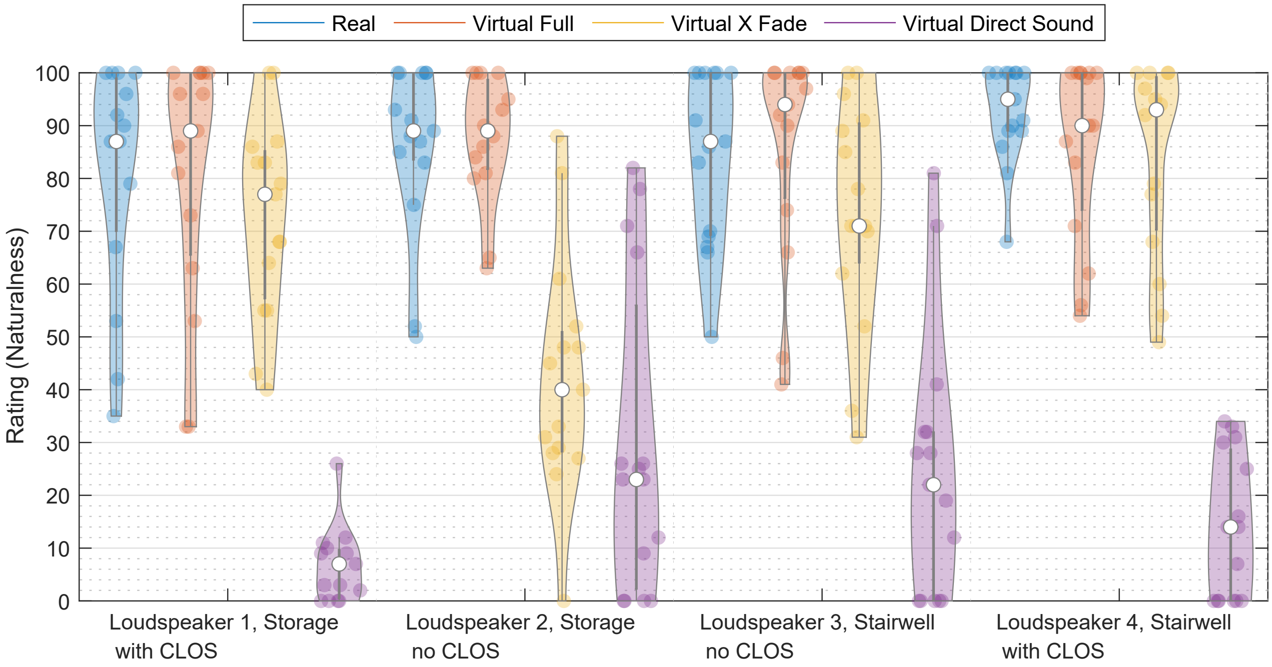


Fig. 8. Violin plots of the listening test results. CLOS refers to a continuous line-of-sight between the loudspeaker and listener for all listener positions (refer to Fig. 1 for loudspeaker positions and room geometries). Median values are presented as a white point, the interquartile range is marked using a thick grey line, the range between the lower and upper adjacent values is marked using a thin grey line, and individual results are displayed as coloured points.

was tested for normality using the Kolmogorov-Smirnov test, which showed all data as non-normal. Therefore, statistical analysis was conducted using non-parametric methods.

The *Real* and *Virtual Full* conditions were in general perceived as highly natural for all four tested loudspeaker positions, with median values between 87 and 95. To assess whether the results between the *Real* and *Virtual Full* conditions were statistically significant, a Friedman's Analysis of Variance (ANOVA) was conducted, which showed no significance at a 95% confidence interval: $\chi^2(7) = 3.23, p = 0.86$. This is a meaningful outcome as it shows that the auralisation system was perceived at similar naturalness levels as the real loudspeakers.

The *Virtual X Fade* condition was not perceived as natural as the *Virtual Full* condition, except for at loudspeaker 4. It was perceived as significantly less natural for loudspeakers 2 and 3, which have no CLOS for all listener positions. The *Virtual Direct Sound* was rated overall the least natural for all loudspeaker positions. However, some participants rated the condition highly for loudspeakers 2 and 3. When informally asked how they made their judgements, some claimed the sound level and locatedness felt relatively natural in these two trials.

IV. Discussion

The evaluation has shown that the auralisation system is capable of producing a natural auditory experience of the transition between coupled rooms. This is achieved through the combination of high resolution SRIR measurements with a time-varying convolution method, delivered over highly transparent headphones using parametric binaural rendering, with sound field position and orientation compensation informed from the virtual reality headset. The system is calibrated in-situ to the response of a real loudspeaker, which reduces further any spectral inaccuracies.

With the exception of the *Virtual Direct Sound* condition, all tested conditions produced high levels of perceived naturalness for most loudspeaker positions. An interesting result comes from the *Virtual X Fade* condition, which achieved generally high perceived levels of naturalness for the loudspeakers 1 and 4, with a continuous line-of-sight (CLOS) to the listener, but lower for loudspeakers 2 and 3, with no CLOS. This is likely due to the loudspeaker position. For loudspeakers 1 and 4, with CLOS to the listener, the direction of the loudspeaker is similar for both the first and last measurement. When fading between the two measurements during user positional changes, the direction of the source does not change. For loudspeakers 2 and 3, however, the direction of the loudspeaker differs greatly from the first to the last measurement. Therefore, when fading between these two measurements, the virtual source can appear to come from the wrong position. Another aspect to note is that, for loudspeakers 1 and 4 that have CLOS to the listener, both rooms are excited more evenly. The lower relative differences in energy between the two rooms may reduce the perceived differences.

In contrast to the *Virtual X Fade* condition, the *Virtual Direct Sound* condition was rated as more natural for loudspeakers 2 and 3, and less natural for 1 and 4. A likely explanation for this is that the lack of reverberation made it highly unnatural for loudspeakers 1 and 4, however the accurate locatedness and amplitude changes, along with the accurate portalling effect due to the lack of CLOS, caused loudspeakers 2 and 3 to be perceived as more natural.

Two participants noted the *Virtual Direct Sound* condition felt localised inside the head at some points, whereas the other conditions were always externalised. This supports previous research, which suggests reverberation, and in particular early reflections, are important for a sense of externalisation [1], [2].

There was no statistically significant difference between the *Real* and *Virtual Full* conditions, which suggests the auralisation system was capable of matching the naturalness of the real loudspeakers. However, it is surprising that the median result of the *Real* condition was lower than the *Virtual Full* for loudspeakers 1 and 3. When asked how their answers were decided, some participants noted some conditions felt more spatious and lateralised than others. It is possible that the binaural rendering may in some cases over-exaggerate lateralisation, which leads to a hyper-real experience in comparison. Further testing is required to investigate this. Another possible reason could be if there were small positional differences between the virtual reality visuals and the real location of the loudspeakers.

V. Conclusions

This paper has presented an auralisation method for the transition between coupled rooms. A time-varying partitioned convolution method allows for real-time switching of higher-order spatial room impulse responses, and dynamic rendering is achieved using acoustically transparent headphones and parametric binaural decoding with head position and orientation data from the virtual reality headset. The system is calibrated to the corresponding real-world loudspeaker response in-situ, and virtual reality visuals are made possible from three-dimensional models from room scans using LIDAR technology.

The system was evaluated in a listening test including both real-world loudspeakers and the binaural auralisation. The auralisation was rated as highly natural, and produced statistically similar results to loudspeakers. Therefore, the system can be recommended for use as a natural sounding reproduction of the transition between coupled rooms, and the system can be used for non-in-situ testing.

A simplified auralisation, whereby the reverberant signals were interpolated from single measurements in each room, was also tested. This produced high perceived naturalness for some loudspeaker positions but poor naturalness for others. Future work will look at improved methods of interpolation between two measurements, to produce a more natural auralisation for sound source locations with no continuous line-of-sight between the source and listener for all listener positions. The results of this study could also be used to inform future room simulation engine development.

Acknowledgment

The authors thank Niklas Alenius for measuring the three-dimensional scans of the rooms.

References

[1] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916, 2001.

[2] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," *J. Acoust. Soc. Am.*, vol. 138, no. 2, pp. 1154–1167, 2015.

[3] D. Khaykin and B. Rafaely, "Acoustic analysis by spherical microphone array processing of room impulse responses," *J. Acoust. Soc. Am.*, vol. 132, no. 1, pp. 261–270, 2012.

[4] N. Xiang, Y. Jing, and A. C. Bockman, "Investigation of acoustically coupled enclosures using a diffusion-equation model," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1187–1198, 2009.

[5] A. Billon, V. Valeau, A. Sakout, and J. Picaut, "On the use of a diffusion model for acoustically coupled rooms," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2043–2054, 2006.

[6] N. Raghuvanshi and J. Snyder, "Parametric directional coding for precomputed sound propagation," *ACM Trans. on Graphics*, vol. 37, no. 4, pp. 1–14, 2018.

[7] P. Luizard, B. F. G. Katz, and C. Guastavino, "Perceptual thresholds for realistic double-slope decay reverberation in large coupled spaces," *J. Acoust. Soc. Am.*, vol. 137, no. 1, pp. 75–84, 2015.

[8] N. Xiang, J. Escolano, J. M. Navarro, and Y. Jing, "Investigation on the effect of aperture sizes and receiver positions in coupled rooms," *J. Acoust. Soc. Am.*, vol. 133, no. 6, pp. 3975–3985, 2013.

[9] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Sig. Proc. Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[10] B. N. J. Postma and B. F. G. Katz, "Acoustics of Notre-Dame cathedral de Paris," in *Int. Cong. on Acoustics*, Buenos Aires, 2016, pp. 1–10.

[11] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th Conv.*, Paris, 2000, pp. 1–23.

[12] M. A. Gerzon, "Periphony: with-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[13] A. Neidhardt, A. I. Tommy, and A. D. Pereppadan, "Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets," in *AES 144th Conv.*, Milan, 2018, pp. 1–11.

[14] E. Stein and M. M. Goodwin, "Ambisonics depth extensions for six degrees of freedom," in *AES Int. Conf. on Headphone Technology*, vol. 2019, San Francisco, 2019, pp. 1–10.

[15] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *IEEE Int. Conf. on Digital Sig. Proc.*, Santorini, 2009, pp. 1–5.

[16] N. Antonello, E. De Sena, M. Moonen, P. A. Naylor, and T. Van Waterschoot, "Room impulse response interpolation using a sparse spatio-temporal representation of the sound field," *IEEE/ACM Trans. on Audio, Speech and Lang. Proc.*, vol. 25, no. 10, pp. 1929–1941, 2017.

[17] K. Müller and F. Zotter, "Auralization based on multi-perspective Ambisonic room impulse responses," *Acta Acustica*, vol. 6, no. 25, pp. 1–18, 2020.

[18] A. Neidhardt and B. Reif, "Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis," in *AES 148th Conv.*, Online, 2020, pp. 1–10.

[19] T. McKenzie, S. J. . Schlecht, and V. Pulkki, "Acoustic analysis and dataset of transitions between coupled rooms," in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, Online, 2021, pp. 481–485.

[20] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, vol. 145, no. 4, pp. 2746–2760, 2019.

[21] M. Roser, C. Appel, and H. Ritchie, "Human height," 2013. [Online]. Available: https://ourworldindata.org/human-height

[22] F. Wefers, "Partitioned convolution algorithms for real-time auralization," Ph.D. dissertation, RWTH Aachen University, 2014.

[23] Ø. Brandtsegg, S. Saue, and V. Lazzarini, "Live convolution with time-varying filters," *Appl. Sci.*, vol. 8, no. 1, pp. 1–29, 2018.

[24] A. Lindau, "The perception of system latency in dynamic binaural synthesis," in *Fortschritte der Akustik: Tagungsband der 35. DAGA*, no. 1, 2009, pp. 1063–1066.

[25] D. S. Brungart, A. J. Kordik, and B. D. Simpson, "Effects of headtracker latency in virtual audio displays," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 32–44, 2006.

[26] A. Politis, J. Vilkamo, and V. Pulkki, “Sector-based parametric sound field reproduction in the spherical harmonic domain,” *IEEE J. on Selected Topics in Sig. Proc.*, vol. 9, no. 5, pp. 852–866, 2015.

[27] A. Politis, L. McCormack, and V. Pulkki, “Enhancement of Ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing,” in *IEEE Workshop on Applications of Sig. Proc. to Audio and Acoustics*, 2017, pp. 379–383.

[28] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, “A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database,” *Appl. Sci.*, vol. 8, no. 11, pp. 1–21, 2018.

[29] T. Mckenzie, D. T. Murphy, and G. Kearney, “An evaluation of preprocessing techniques for virtual loudspeaker binaural Ambisonic rendering,” in *EAA Spatial Audio Sig. Proc. Symp.*, 2019, pp. 149–154.

[30] A. Politis, S. Tervo, and V. Pulkki, “COMPASS: coding and multidirectional parameterization of Ambisonic sound scenes,” in *IEEE Int. Conf. on Acoust., Speech and Sig. Proc.*, 2018, pp. 6802–6806.

[31] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time – frequency domain spatial audio*, 1st ed. John Wiley and Sons, Ltd, 2018.

[32] I. Engel, C. Henry, S. V. A. Garí, P. W. Robinson, and L. Picinali, “Perceptual implications of different Ambisonics-based methods for binaural reverberation,” *J. Acoust. Soc. Am.*, vol. 149, no. 2, pp. 895–910, 2021.

[33] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, “Transfer characteristics of headphones measured on human ears,” *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 203–217, 1995.

[34] A. Harker and P. A. Tremblay, “The HISSTools impulse response toolbox,” in *Int. Comp. Music Conf.: Non-cochlear Sound*, Llubljana, 2012, pp. 148–155.

[35] J. L. Hintze and R. D. Nelson, “Violin plots: a box plot-density trace synergism,” *The American Statistician*, vol. 52, no. 2, pp. 181–184, 1998.