AN ACOUSTIC PAINTBRUSH METHOD FOR SIMULATED SPATIAL ROOM IMPULSE RESPONSES

Otto Puomio *

Aalto Acoustics Lab
Dept. of Signal Processing and Acoustics
Aalto University, Espoo, Finland
otto.puomio@aalto.fi

Tapani Pihlajakuja

Nokia Technologies
Espoo, Finland
tapani.pihlajakuja@nokia.com

Tapio Lokki

Aalto Acoustics Lab
Dept. of Signal Processing and Acoustics
Aalto University, Espoo, Finland
tapio.lokki@aalto.fi

ABSTRACT

Virtual reality applications require all kinds of methods to create plausible virtual acoustics environments to enhance the user experience. Here, we present an acoustic paintbrush method that modifies the timbre of a simple room acoustics simulation with the timbre of a measured room response while aiming to preserve the spatial aspects of the simulated room. In other words, the method only applies the measured spectral coloration and alters the simulated and temporal distribution of early reflections as little as possible. Three variations of the acoustic paintbrush method are validated with a listening test. The results indicate that the method works reasonably well. The paintbrushed room acoustic simulations were perceived to become closer to the measured room acoustics than the source simulation. However, the limits of the perceived effect varied depending on the input signal and the simulated and recorded responses. This warrants for further perceptual testing.

1. INTRODUCTION

Immersive audio aims to immerse the listener in an alternative soundscape [1]. The soundscape can range from a recording of an existing place to a simulation of a completely artificial environment, both trying to create a plausible sound environment for the listener. In case we wanted to combine the two, namely emulate an existing space, the safest way would be to run an acoustics simulation. This, however, requires a 3D model of the environment and estimates of its surface materials. The most advanced methods [2] can indeed provide very plausible results, but they are computationally very expensive. Furthermore, it is hard to estimate acoustic material properties correctly and no perfect solution has been presented. Simply put, the approach described above appears infeasible for applications requiring real-time results. In the end, the end result just needs to sound plausible for the given space, even though not physically accurate.

In this paper, we present an acoustic paintbrush method, which offers one solution to create plausible soundscapes without heavy computational load. The acoustic paintbrush method modifies spectral characteristics of a simulated room with the characteristics of another room, while aiming to preserve the spatial properties of the simulation. An ideal end result would therefore be a perceptually plausible reproduction of the modifying space with the spatial cues of the modified one. The method is applied before the auralization, i.e., convolution with the sound signal. Ultimately, the acoustic paintbrush method aims at speeding up the computation by reducing the need for extensive acoustic simulation.

This paper is divided into four sections. First, we present the method and its modifications while connecting them to similar prior work. Then we evaluate the presented methods with a listening test and describe the test arrangements, followed by the analysis of the obtained results. Finally, we discuss the effects found in the data and possible applications of the proposed acoustic paintbrush method.

1.1. Background and similar work in related areas

Ideas similar to the acoustic paintbrush method have been applied before on computer simulated RIRs. Li et al. [3] applied computer vision, bidirectional ray tracing and a RIR recording to create spatial audio for 360° video in post-production. Their approach was to apply room resonances by creating a short modulation filter from the recorded RIR. The filter was designed on the samples up to 5–10 ms after the direct sound and applied on simulated RIRs generated at 50 cm intervals along an approximated camera path.

While the filter described above applies a very similar technique to our method, there are also two significant differences. The first difference is the number of samples used in designing the coloration filter. The first 5 ms is too short to capture the whole timbral change that early reflections cause to the signal. Instead, the presented paintbrush method utilizes the whole early response in the design process. The other difference is that Li et al. apply a zero-phase filter, whereas our method utilizes a minimum-phase filter. While the zero-phase filter does not theoretically affect the phase, it creates a noncausal signal tail before any impulses in the impulse response (IR). Even though the effect is unnoticeable with short filters, it must be accounted for with longer ones.

Surprisingly, the paintbrush method also bears similarities to room response equalization for loudspeakers [4]. In room response equalization, one tries to improve the quality of the reproduced sound by reducing the detrimental effects of the reproduction room. Similar to the method presented in this paper, room response equalization methods utilize a room impulse response (RIR) to design an inverse filter for the room. In particular, room equalizers prefer filter stability and extended sweet spot over a perfectly flat frequency response. The main difference between room response equalization and the presented method is that while the former aims at reducing the effect of the room on audio signal, the latter targets a completely different room response. In addition, the paintbrush method aims at keeping the spatial aspects of the original room. Despite the similarities in the methods, the two methods target completely different

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

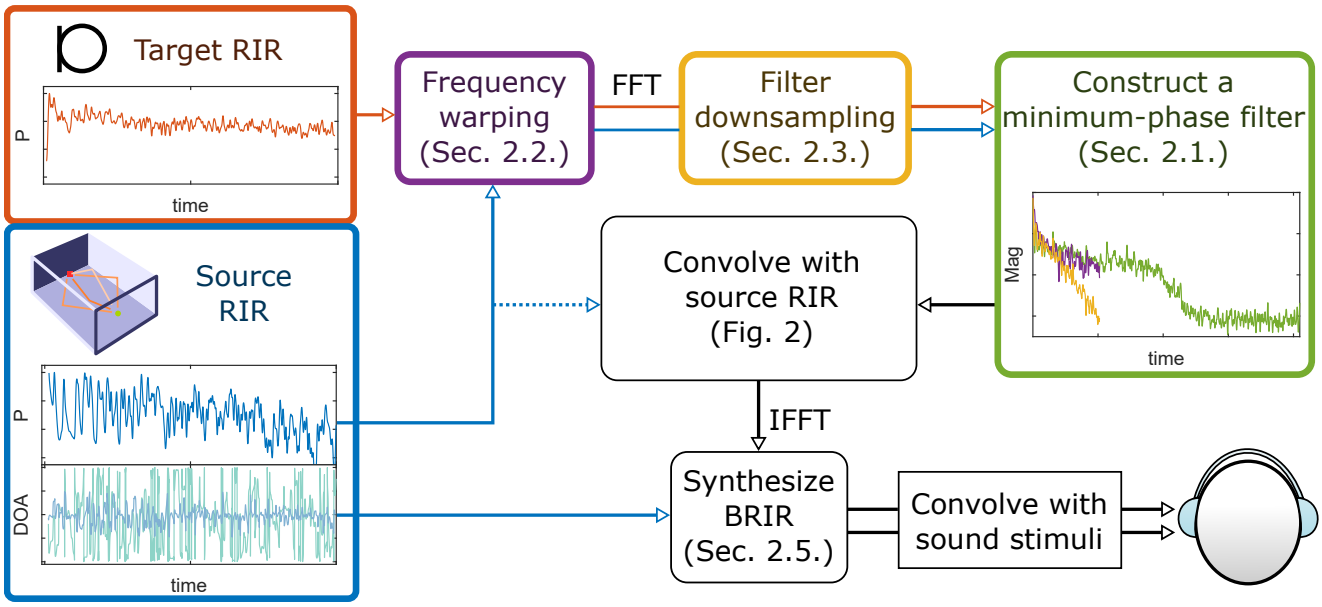

Figure 1: Acoustic paintbrush method flowchart. The source and target RIRs (in blue and red, respectively) are used to construct a minimum-phase filter (green) that colorizes the source RIR for binaural rendering (black). In addition, two extra methods (purple and yellow) may extend the method either together or separately.

scenarios.

2. METHODS

The acoustic paintbrush method is motivated by the virtual acoustic environment creation for virtual reality applications. In many applications, the sound environment might fit its purpose even without a very accurate acoustic simulation. Yet to really succeed in this, one needs computationally light methods that achieve plausible sounding results. The acoustic paintbrush method takes this approach by aiming at offering an alternative to the accurate acoustic modeling.

The paintbrush method is presented in its entirety in Figure 1. The method takes in two RIRs. For clarity, the modifying RIR (red) is henceforth called target and similarly, the RIR being modified (blue) is called the source RIR. The latter is a spatial room impulse response (SRIR), i.e., it contains the metadata indicating the direction-of-arrivals (DOA) of the direct sound and the early reflections. Both the source and the target are referred with their respective colors in all Figures throughout this article; this also applies to the methods and their respective colors introduced next.

The method consists of three procedures: basic procedure (green), frequency warping (purple) and filter length reduction (yellow). Basic procedure can be thought as the core process, while the other two procedures extend it independently. Basic procedure returns a minimum-phase coloration filter that is convolved with the source RIR to get the colored RIRs (henceforth result). The result RIR utilizes the DOAs of the source SRIR to create a binaural room impulse response (BRIR). Finally, the created BRIR is convolved with sound stimuli to get a binaural output for headphones.

There are many ways to apply the paintbrush method. For instance, the source SRIR can be a virtual space simulated with a relatively simple room rendering method, e.g., with the image source method [5, 6]. The target RIR in turn can be an accurate simulation of the same space, a measured RIR of a similar space, or a RIR of a completely different space altogether. The underlying idea in modifying the source is to only avoid simulating a complex acoustic model in real-time.

In the next sections, the three paintbrush method procedures are explained in detail. Section 2.1 presents the basic procedure, section 2.2 extends that with frequency warping, and section 2.3 boosts computational performance by reducing the filter order. Their goodness of fit is examined objectively in section 3.3, and the result RIR spatialization is described in section 2.4.

Note that in this paper and in the listening test (Section 3), only the first 100 ms of the response was colored while the late part was excluded from any procesing. This choice was made because the early part is expected to dominate the perceived coloration of a continuous signal. Additionally, the late part of the signal would mask some of the audible differences of the early responses, therefore the tail was left out for purpose.

2.1. Basic procedure

The first and the most basic procedure implements a simple coloration filter in the frequency domain. Designing the filter is divided into three steps. In the first step, the early part of the response is extracted from both the source RIR $h_s[n]$ and the target RIR $h_t[n]$, n being a discrete time sample. The response cut lengths are measured from the direct sound and the extracted parts are linearly faded out during the last 20 ms of the filter. Henceforth, these extracted early parts of the source and the target are denoted as $h_{s,ER}[n]$ and $h_{t,ER}[n]$, respectively.

In the second step, the early parts define a coloration filter $C[z]$ as follows:

$$C[z] = \frac{|H_{t,ER}[z]|}{|H_{s,ER}[z]|} \tag{1}$$

where $H_{t,ER}[z]$ and $H_{s,ER}[z]$ are discrete Fourier transformed (DFT) versions of $h_{t,ER}[n]$ and $h_{s,ER}[n]$, respectively; and z is

the frequency-domain counterpart of n. In practice, $C[z]$ first whitens the source response with an inverse filter $H_{s,ER}[z]$ and then applies the target coloration with $H_{t,ER}[z]$.

In the final step, the filter is minimum-phased before convolving it into $h_{s,ER}[n]$. This step is done in order to make the filter causal (as discussed at the end of section 1.1). Additionally, the operation aims at reducing phase manipulation effect when compared to a conventional RIR. Minimum-phasing is applied through cepstral domain manipulation. After filtering, the direct sounds of the result filter $h_{r,ER}[n]$ and $h_{s,ER}[n]$ are realigned to compensate for any filter delay. This way, the result RIR and the source DOA vector stay aligned w.r.t. each other for spatialization presented later in this paper.

After the basic procedure, $h_{r,ER}[n]$ is combined with the late part of the response to form the full RIR. The late part can either be simulated (e.g. with a feedback delay network) or extracted from the target recording (in witch case the signals are crossfaded).

2.2. Warped approach

The second procedure applies frequency warping to the first procedure described above. Frequency warping is well studied, for instance, by Härmä et al. [7]. Frequency warping aims to simulate nonlinear resolution of human hearing by modifying the frequency resolution of the filter. When compared to a conventional filter, the warped filter samples the low frequencies more densely, leaving the high-frequency sampling sparser. This is usually implemented with a cascade of all-pass filters that stretch different frequencies by different amounts. This stretching allows us to sample the frequency domain nonuniformly, letting us to focus the filter effort in the frequencies where we hear the differences better. The warped filter can therefore perform better than a conventional finite impulse response (FIR) filter with the same number of taps.

Frequency-warped implementation extends the first procedure described before. Instead of transforming $h_{s,ER}[n]$ and $h_{t,ER}[n]$ to the frequency domain directly, both of them are first transformed to the frequency-warped domain with the warpTB package [7]. These warped signals are then transformed to the Fourier domain and processed as in the first procedure. Finally, the minimum-phased filter is applied to the source RIR in the warped frequency domain, and the result RIR is brought back to (unwarped) time domain before applying it to the signal.

The warping can also be applied in the frequency domain [8]. In this technique, one fits a spline to the frequency domain data. The spline is used as an interpolant to obtain the frequency-warped filter. This approach is more efficient than the time-domain one yet it does not suit for cases where there is no downsampling involved. This is because the conventional DFT samples the frequency domain uniformly. Compared to the frequency warped filter of the same size, the uniformly sampled filter has more sample points on the high frequencies and less on the low ones. Fitting the spline to the uniform data does not improve this resolution, meaning that the low frequencies of the warped filter cannot be more accurate than they are in the original filter. For this reason, the conventional all-pass method was selected to avoid any potential artifacts imposed by the selected warping method.

2.3. Reduced filter length

The third and final procedure aims at reducing filter length created by the first and second procedures. In the earlier procedures, the final filter length is determined by the chosen length of the early part of the response. Therefore, the filter length can span even up to 150 ms, which corresponds to 7200-tap FIR at 48 kHz sample rate. A filter of this length is excessively laborous for e.g. mobile applications and is also deteriorating the performance gain of the paintbrush method. The third procedure addresses this very problem by reducing the filter length through spectral downsampling.

The third procedure can be implemented on both the first and second procedures by resampling the early responses in the frequency domain. In both cases, the early responses are resampled before proceeding to Eq. (1). This means that in case of the second procedure, resampling happens in the warped frequency domain. In the presented implementation, the actual resampling is done through spline interpolation. As the filter lengths heavily depend on analyzed signal length and sample rate, the exact resample ratios are not specified here. Instead, the reader is instructed to consult section 3.2 for exact resample rates for this paper.

As reasoned above, the main reason for spectral downsampling is to reduce computational demand of the paintbrush procedure. By shortening the applied FIR filter length, the number of multiplications drops by same proportion in the frequency domain. On the downside, the filter also shortens in the time domain, reducing its power to apply coloration and reverberation to the signal. This downside would however be easy to circumvent by applying infinite impulse response (IIR) filters, but the implementation of these filters are not discussed further in this paper.

2.4. Spatializing the impulse response

Up to this stage, the paintbrush algorithm has only processed mono impulse responses. These impulse responses do not contain any information about the DOA of the sound. Therefore, DOAs need to be injected to the modified responses when they are spatialized. In this paper, the directions are taken from the source SRIR and combined with the result RIRs to synthesize BRIRs for spatial audio rendering.

The SRIR synthesis is implemented as follows. The synthesis takes a RIR and a DOA vector as an input. For each sample in the impulse response (IR), there is an azimuth and elevation angle describing the DOA for that particular sample. In case of the generated result RIRs, the required DOAs are fetched from the source DOA vector. Such SRIR can be applied to any spatial sound reproduction system.

For multichannel reproduction, the pressure signal samples are distributed to reproduction loudspeakers according to the DOA vector as in the original Spatial Decompostion Method (SDM) [9]. Then the convolution with the actual sound signal is done for each loudspeaker independently. Naturally, these loudspeakers can be "virtualized" by applying head-related transfer functions (HRTF) corresponding to the loudspeaker directions. The virtual loudspeakers effectively create a BRIR, though the sound has only a sparse set of incoming directions. Finally, the BRIR could also be synthesized directly without a discrete number of virtualized loudspeakers. This is possible by matching each sample DOA to the closest HRTF in a dense dataset. The selected HRTF is then scaled with the corresponding IR amplitude. These scaled HRTFs are accumulated to get the synthesized BRIR.

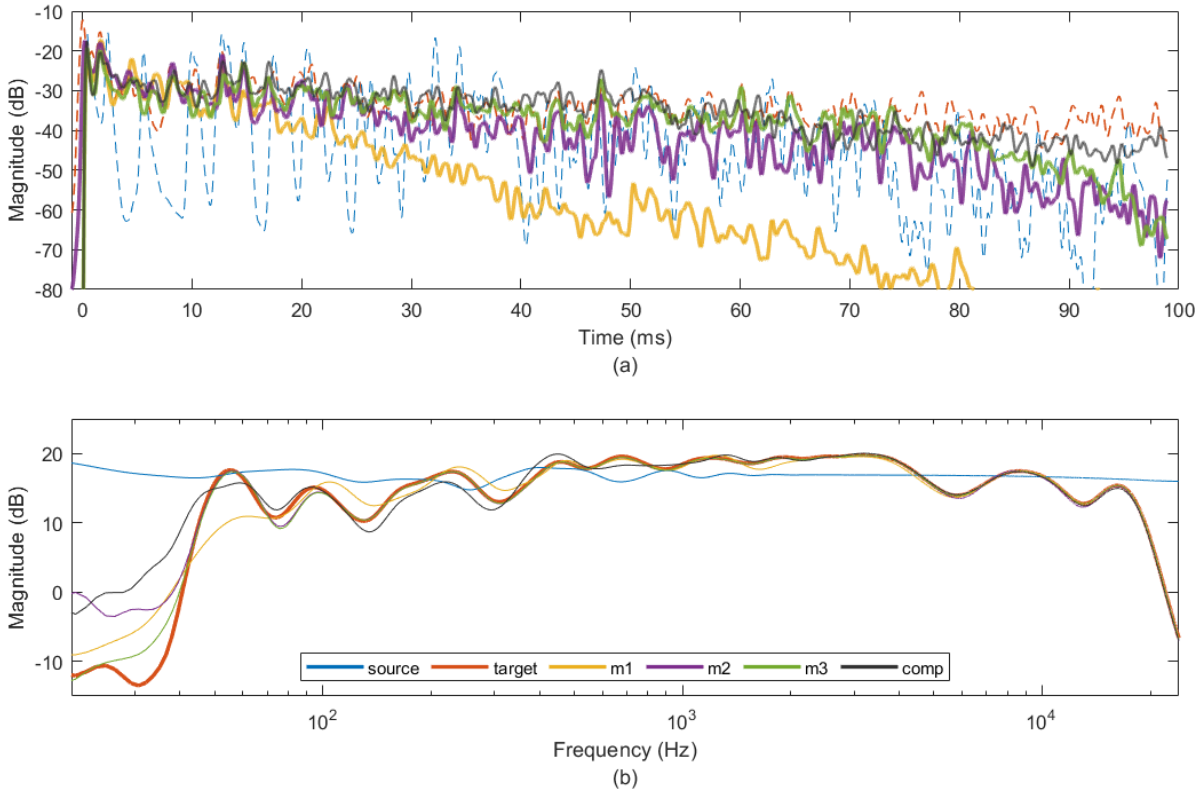

Figure 2: Comparison of different procedures of the paintbrush method that were used in the listening test. (a) smoothened RIR magnitudes (1 ms gaussian window) of different test methods: (green) the basic method (8192 taps); (yellow) the basic method with spectral downsampling (1024 taps); (purple) frequency-warped method with spectral downsampling (1024 taps); and (black) comparison method. Source (blue) and target (red) RIRs are shown for comparison. (b) Frequency spectra of the aforementioned methods with one-third octave bands smoothing.

3. EVALUATION

The performance of the acoustic paintbrush method was evaluated with subjective and objective measures. These measures seek to evaluate how well the method transforms the simulated RIR towards the measured target. The section is organized such that first the simulated and measured rooms and the compared methods are described in detail. This is followed by the objective analysis of the method. The rest of the section then presents the subjective listening test organization and related results.

3.1. Simulated and measured rooms

The acoustic paintbrush method requires the source and target spatial impulse responses, as depicted in Fig. 1.

The target rooms were a small rectangular lecture room (8.6 x 6.4 x 2.3 meters) and a less symmetric coffee room (approx. dimensions of 9.7 x 8.2 x 2.5 meters). Both rooms were measured for spatial impulse responses by using a Genelec 8331 as a sound source and an omnidirectional free-field microphone as a receiver. In the lecture room measurement, the distance between the loudspeaker and the microphone was 2.2 m while the inter-transducer distance was 2.8 m in the coffee room measurement.

The source responses for both rooms were simulated with the image source method utilizing the *roomsim* software [10]. To reduce the number of variables affecting the listening test results, the simulator was configured to use similar room volumes as were in the measured target rooms. For the same reason, the sound source and receiver positions in the model were setup to be as similar as possible to the target measurements. Otherwise, the simulator was configured to execute a simple broadband acoustic simulation with only specular reflections.

To obtain the simulated spatial impulse response, a spatial open microphone array was simulated with the image source method up to 10th order reflections to 7 omnidirectional receivers. The microphone array had one receiver in the center of the array and 3 microphone pairs around it in x, y and z directions. This rather unconventional way to use the image source method was chosen as we wanted to analyze the simulated microphone array impulse responses with the SDM to obtain the required data vectors, i.e., the IR and corresponding DOA vectors. Another reason for this was to save time and avoid implementing a system from scratch for generating the required data.

Both the source and target responses were truncated to contain only early part (up to 100 ms) of the whole IR. This way, the small changes in the early response are not masked by the late reverberation.

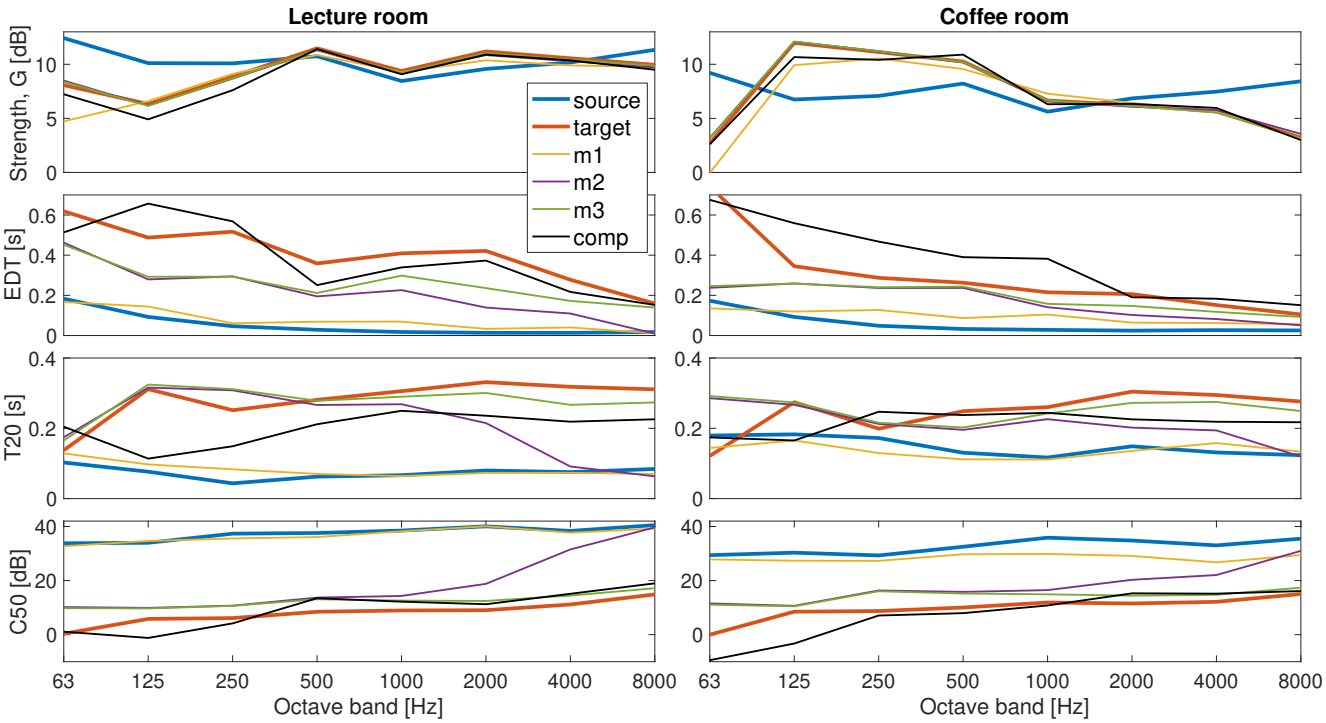

Figure 3: Comparison of objective parameters of the procedures used in the listening test. Strength (G), early decay time (EDT), reverberation time (T20) and clarity (C50) were calculated at octave bands for each test method.

3.2. Tested methods

The listening test samples were prepared with three different versions of the presented paintbrush method and a naive implementation of coloration as a comparison method. The included methods were:

m1 100 ms minimum-phase filter that is downsampled to 1024 taps in the frequency domain (resample rate 8:1).

m2 same as **m1**, but source and target RIRs are frequency-warped before FFT.

m3 100 ms minimum-phase filter, no downsampling (filter length 8192 taps)

comp first 100 ms of target, minimum-phased and convolved with the 100 ms of the source response.

The method **comp** is seen as the "simple" implementation of the paintbrush method. There, the source frequency response is not touched at all before applying the target response as a minimum-phase filter on top of it. Practically, this corresponds to Eq. (1) without the source frequency response in the denominator. This kind of approach relies on the assumption that the simulation is already spectrally white to work well.

In addition to the paintbrushed responses, the listening test included the hidden anchors. The anchors were the first 100 ms of the source and target RIRs of the room under evaluation.

3.3. Objective comparison of methods

Figure 2 illustrates the goodness-of-fit of all the RIRs used in the listening test. The paintbrush method appears to suppress the reflections of the source RIR (blue) while applying target RIR reflections (red) in the result RIR (other colors). The result RIR magnitudes follow the target RIR most of the time. Simultaneously, one can find traces of source RIR magnitude peaks as well. There are two visible exceptions for this behavior. First, **m1** decays faster than the other methods which is caused by the shorter filter length compared to other methods. Second, the other test methods lose energy between 80–100 ms because of the linear fadeout applied on the early part of the source. Overall, all the procedures seem to bring the source RIR closer to the target RIR.

While the differences in the time domain are relatively small, one can find more differences between the procedures in the frequency domain. As expected, **m3** (green) fits the closest to the target, while **m1** (yellow) differs the most out of the four. Also as expected, **m2** (purple) notably improves the fit in the low frequencies when compared to **m1**. Finally, **comp** appears to deviate from the target up to 2 kHz, after which it does not differ from the other methods. Therefore **m3** appears to fit the magnitude spectrum of the target better than **comp**.

One can also compare the methods through objective room acoustical parameters. Four different objective parameters were calculated for the test methods in octave bands, shown in Figure 3. The objective parameters of the target RIR are shown in red, while the test methods introduced in section 3.2 are displayed in other colors. While the strength parameter seems to be similar over all the methods, there is a lot of variance in the other parameters. Test method **m1** appears to differ from the target the most, while **m3** and **comp** gets the closest to it. The performance of the last test method **m2** is between **m1** and **m3**. On 63–500 Hz octave bands, **m2** seems to perform almost identically to **m3**. From 1000 Hz

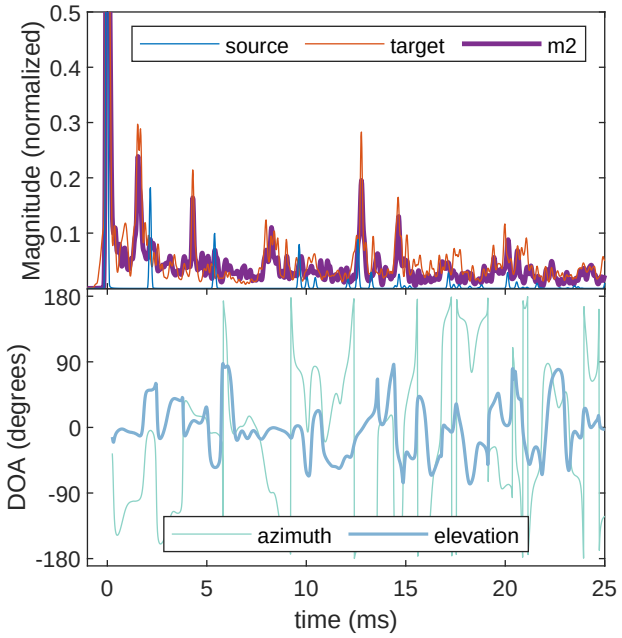

Figure 4: Result RIR and source DOA vector alignment comparison. The direct sounds of the target and **m2** are aligned to the source direct sound, while the source DOA vector is left untouched. The reflections conforming with the source reflections get their original direction while the shifted ones mostly get a more random one.

octave band onwards, the filter starts to depart from **m3**, finally reaching **m1** on 8000 Hz octave band. These results are expected because of the difference in filter length. Basically, **m1** and **m2** apply a FIR filter that is 8 times shorter than in **m3** and **comp**. Also as expected, frequency warping enhances filter performance on the low frequencies, effectively explaining the **m2** performance on sub-500 Hz bands.

Finally, the alignment of the result RIRs and the source DOA vector is compared in Figure 4. It is apparent that the result RIR reflections do get the same direction as the source counterparts if the reflections are time-aligned. For instance, one of these cases occur at the fourth result RIR reflection. In cases the reflections have 'shifted' to target locations, the result reflection seems to get a more random direction. This case is seen in the first reflection. There, the source reflection comes from the ceiling, arriving approximately from $(az = 0, el = 50)$ to the listener. In turn, the shifted result reflection arrives from $(az = -100, el = -10)$, having a completely different direction than the corresponding source reflection.

3.4. Stimulus signals

Three different stimuli were used to determine how signal content affects the method performance. These recordings were 5 – 8 s in length and their contents were the following:

- Dry **cello** music
- Dry **drum** music
- Dry **speech** signal

All together, the listening test consisted of two different rooms, three stimuli and four methods.

3.5. Test organization

Due to the restricting circumstances of the COVID-19 pandemic, the listening test had to be organized as an online listening test. The test was distributed to the subjects by email and run locally on their home computers. After the test, the subjects submitted their results also via email. The spatial sound was reproduced binaurally as explained in Section 2.4.

The chosen listening test methodology was a multiple stimulus test using a continuous grading scale and two hidden reference end points, namely the source and the target. In the test, the subjects were instructed to rank how close each test sample is to the given two references. As explained above, the first reference was the measured target SRIR captured in a real room and the second one was the simulated source SRIR created with the roomsim software. Each subject was asked to grade different procedures of the paintbrush method whether they are closer to either reference. In addition, both references were hidden amongst the test samples, requiring the subjects to identify and grade them accordingly. The presented method was hypothesized to statistically differ from the source reference. In addition, the method was expected to be closer to the target reference, but not necessarily statistically equal to it.

Upon submitting the answers, the subjects were also asked three questions:

1. What cues did you use to discriminate the samples?
2. Which headphones did you use (model)?
3. Any questions/comments/feedback in general?

Answers to these questions were used to identify the most significant discriminators, possible causes for outliers and other criteria that may have affected the results.

In summary, the listening test was a multiple stimulus test with two hidden anchors. Four methods were tested with two different rooms and three sound signals. In addition, each test case was repeated three times. Thus, each test participant performed 18 multiple stimulus gradings, i.e., 3 stimuli x 2 rooms x 3 replications.

3.6. Results

The listening test was completed by 16 subjects, who were researchers in acoustics and could be considered as experienced listeners. First, the subjects were validated by their results to be able to consistently identify the hidden references. On a scale of $[-1 \ldots 0 \ldots 1]$ presented as {'REF1','Neither','REF2'}, the subjects were required to rate both references 0.95 or more at their end of the scale in at least 85 % of the cases. Practically, the subjects were rejected if they failed to rate three or more cases. In the end, this lead to discarding three participants, resulting in N = 13 for further analysis.

After the validation, the preprocessing focused on refining the data. The sample replications were averaged into a single mean value per participant. Finally, hidden references were omitted from the statistical analysis as their scores were saturated to the end points of the grading scale, as was expected.

The grading scores were analyzed with repeated-measures analysis of variance (RM-ANOVA). The within-subject factors were: *room*, *stimulus*, and *method*. First, the data was checked for violations of sphericity assumption with Mauchly's test of sphericity. The test revealed significant effect in interaction *room * method*, $\chi^2(5) = 17.279, p < 0.05, \epsilon < 0.75$. Thus, the data for this interaction is corrected with Greenhouse-Geisser correction for further

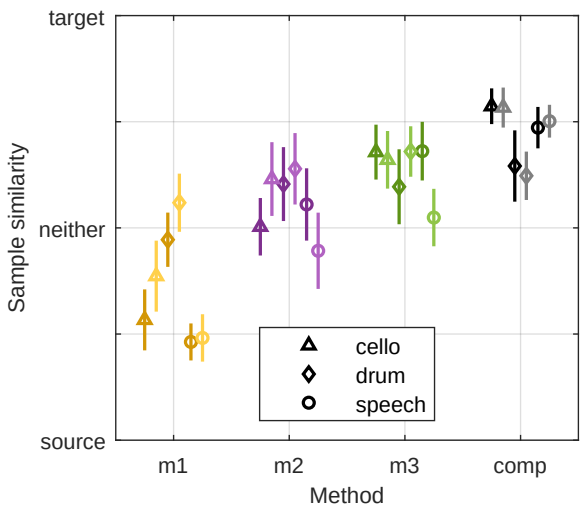

Figure 5: Marginal means and 95% confidence intervals (N = 13) for interaction between *room*, *stimulus*, and *method* in the listening test. Each method is plotted with the same hue as in Figure 2, while the brightness of the color refers to different rooms. Finally, the marker indicates the stimulus used for that particular test sample.

analysis. After the correction, the test for within-subject effects revealed 5 significant effects. These values are shown in Table 1.

As the third-order interaction, *room * stimulus * method*, was significant, further analysis is based on it. The marginal means and their 95 % confidence intervals are shown in Figure 5. This figure does not have any compensations applied for multiple comparisons. Therefore, it should be used only to study trends and not to decide statistical significances.

Figure 5 shows multiple effects that explain the significant test results. First, method **m1** seems to be least consistent over room types and stimulus types. Second, the effect of the room does not seem to have any clear trend when interacting with the other two parameters — but it affects results. Third, the drum sample seems to have more compressed results than the other two sound samples as most of its scores are generally closer to the mid point. Finally, method **m1** seems to be clearly closer to the source reference than the other methods which tend more towards the target.

Table 1: Significant effects ($p < 0.05$) in the repeated-measures ANOVA analysis of test 1 results.

Source	F	Sig.
stimulus	$F(2, 24) = 16.943$	0.000
method	$F(3, 36) = 93.886$	0.000
room * stimulus	$F(2, 24) = 6.213$	0.007
stimulus * method	$F(6, 72) = 24.280$	0.000
room * stimulus * method	$F(6, 72) = 4.299$	0.001

3.7. Written feedback from the subjects

Written feedback was first preprocessed by hand and then analyzed with a bag-of-words algorithm. The analysis reported the top-three discriminator features to be *reverberance*, *timbre* and *perceived source distance*. In addition, three subjects reported on location shift of the direct sound or changes in the perceived stereo image.

The feedback also indicated that there were differences in rating difficulty between different sound samples. Three subjects reported the speech sample to be the easiest to rate, while the drum sample was reported the hardest by two subjects. Cello sample caused mixed feelings among the listeners; one subject reported it as the most difficult sample to rate, while two others indirectly mentioned it to be either harder or easier to rate than another sample.

4. DISCUSSION

The main goal of this paper was to implement an acoustic paintbrush, a method that applies timbre of the target RIR on top of the source RIR. This was approached with a comparison method and three different FIR filter implementations. Based on the listening test results, all the implementations seem to achieve at least some timbral coloration towards the target RIR as expected.

In more detailed inspection, the methods align in the source-target scale as expected. Out of the three methods, **m3** gets the closest to the target, while **m1** performs the worst. Also as expected, **m2** improves the coloring performance of **m1** significantly. These findings are in line with the previous research on frequency-warped filters [7]. What was unexpected is that **comp** performed as well as — if not even slightly better than — **m3**. There are two possible reasons for this. First, **comp** had the same filter length as the best-performing method, therefore its filtering performance is also about the same. Second, by consulting Figure 2, the source RIR appears to have a rather flat frequency spectrum to begin with. One can therefore only speculate how source room inversion affects the coloration result beyond this particular case. Noting that how much closer **m3** was to the target magnitude spectrum than **comp**, this reveal indicates that there is more to the perceived feel of space than the magnitude spectrum alone.

Sound stimulus was observed to affect the subject ratings. The written feedback seems to explain this effect as the rating difficulty was reported to vary between the three signals. This appears in the results as compression or spreading around the mid-axis. The drum was rated the most difficult signal to rate, which shows as all the methods being compressed to almost insignificant differences. The two other signals in turn show a clear uniform trend, which were also easier to rate.

Differences between rooms provided inconsistent effects. Depending on the method and stimulus, there seems to be differences but there are no clear trends that would apply to all combinations. There are multiple causes that could be theorized to affect the result such as: room shape and reverberation time, and difference of the target room from the simulated room. However, these effects simply require more testing with other rooms and signal types to draw conclusive results.

The results in general show quite varying interactions. This can be partly explained by timbre being a complex phenomenon and multiple factors affecting it [11]. The resulting signal has the spectral timbre contributions from the target response but the spatial contribution to the timbre comes from the simulation. This timbral mix, however, seems to be somewhat predictable based on the

listening test results and informal listening, that is, the simulation seems to dominate the spatial perception whereas the paintbrush effect turns the overall timbre into a mix of the two.

One important factor about the listening test should be remembered though — due to restrictions caused by the COVID-19, this test could not represent practical use cases but simply quantified how the acoustic paintbrush performs compared to references. The practical use cases are more varying and there is usually no reference present to which the output could be compared. In these situations, perceptual plausibility is the most important factor. The method can be easily applied to a simple dynamic simulation (i.e., simulation is constantly updated based on listener and source movements) of a virtual space to allow use of a recorded good quality RIR to provide perceptually high-quality timbre for reverberation. It is also possible to extend this to augmented reality (AR) use where it is usually possible to obtain the geometry of the real space using the AR headset. This allows generating a simulation of the space and paintbrush can be applied with a reference RIR from the same space or from a suitable good quality recording. Another use case is the artistic control of rendering. Instead of iterating simulation parameters of the room, the sound designer could select the target RIR they would like the room to sound like. Then during runtime, the room simulation would create the essential spatial cues for immersion while the target response would take care of the timbre.

There are clear topics for future research. First, the current limitations in test facilities did not allow listening tests using head-tracked binaural reproduction or loudspeaker reproduction. In addition, spatial aspects of the method were not formally tested. Based on informal listening, it is expected that the method should preserve the spatial characteristics of the source room, while the objective analysis of the result RIR and source DOA alignment suggested otherwise. Thus, more tests should be performed to evaluate the method. Second, the actual use cases of the method are in real-time rendering and to evaluate the performance properly, a complete rendering system is required. Third, the intended use case for the method is to use a reference RIR of a perceptually pleasant room to color the simple simulation of the virtual space. Formally testing this is not simple, but comparisons to other simulation methods could be performed with preference tests.

5. CONCLUSIONS

This paper presented an acoustics paintbrush method. The method modifies timbre of a simulated spatial impulse response with a measured impulse response while aiming to preserve the spatial and temporal properties of the simulation. The desired timbre is applied on simulated virtual room rendering with a simple filtering step. This approach is beneficial in virtual and augmented reality use cases where perceptual plausibility is preferred over physically exact acoustic simulation.

The presented method was evaluated with a listening test where a reference room was captured and different versions of the paintbrush method were applied on a simple simulation of a similar room. Although the current pandemic situation limited the study to headphone listening at home, the listening test results together with the objective metrics showed that the paintbrush method successfully transforms the rendering of the room from the simulated case towards the reference case. However, the simple method used as a comparison performed similarly to the best-performing method, although its magnitude spectrum differed from the target the most. Also, the results did not allow drawing any conclusions about preserving the spatial properties. Finally, the results showed that there were multiple interactions between the test parameters as the timbre perception is a complex effect.

6. COMPANION PAGE

The companion page for this paper is located at

http://research.spa.aalto.fi/publications/papers/dafx20-pb/.

The page contains sound samples rendered with the paintbrush method.

7. REFERENCES

[1] Agnieszka Roginska and Paul Geluso, *Immersive sound: The art and science of binaural and multi-channel audio*, Taylor & Francis, 2017.

[2] Brian Hamilton and Stefan Bilbao, "FDTD Methods for 3-D Room Acoustics Simulation With High-Order Accuracy in Space and Time," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2112–2124, 2017.

[3] Dingzeyu Li, Timothy R. Langlois, and Changxi Zheng, "Scene-aware audio for 360° videos," *ACM Transactions on Graphics*, vol. 37, no. 4, 2018.

[4] Stefania Cecchi, Alberto Carini, and Sascha Spors, "Room response equalization-A review," *Applied Sciences (Switzerland)*, vol. 8, no. 1, pp. 1–47, 2018.

[5] Jont B. Allen and David A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[6] Jeffrey Borish, "Extension of the Image Model to Arbitrary Polyhedra," *Journal of the Acoustical Society of America*, vol. 75, no. 6, pp. 1827–1836, 1984.

[7] Aki Härmä, Matti Karjalainen, Lauri Savioja, Vesa Välimäki, Unto K. Laine, and Jyri Huopaniemi, "Frequency-warped signal processing for audio applications," *AES: Journal of the Audio Engineering Society*, vol. 48, no. 11, pp. 1011–1031, 2000.

[8] Julius O Smith, *Techniques for Digital Filter Design and System Identification with Application to the Violin*, Ph.D. thesis, Stanford University, 1983.

[9] Sakari Tervo, Jukka Pätynen, Antti Kuusinen, and Tapio Lokki, "Spatial decomposition method for room impulse responses," *Journal of the Audio Engineering Society*, vol. 61, no. 1/2, pp. 17–28, 2013.

[10] Steven M Schimmel, Martin F. Muller, and Norbert Dillier, "A FAST AND ACCURATE " SHOEBOX " ROOM ACOUSTICS SIMULATOR Laboratory for Experimental Audiology , University of Zurich , CH-8091 Zurich , Switzerland," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, , no. May, pp. 241–244, 2009.

[11] Per Rubak and Lars G Johansen, "Coloration in Natural and Artificial Room Impulse Responses," *23rd International AES Conference*, pp. 1–19, 2003.