First experiments on how to blend playback and recording rooms in headphone based reproduction of multichannel sound

Karlheinz Brandenburg[1,2], Nils Merten[1], Renato Profeta[1], Christian Schneiderwind[2], Ulrike Sloma[1], Thomas Thron[1], Jeramey Tyler[1]

[1] *Brandenburg Labs GmbH, 98693 Ilmenau, Germany, Email: us@brandenburg-labs.com*

[2] *Technische Universität Ilmenau, 98693 Ilmenau, Germany, Email: karlheinz.brandenburg@tu-ilmenau.de*

Abstract

It has been shown that plausible reproduction of sound sources in a room is possible if the rendering algorithm uses the most important cues for localization of audio sources. This includes fast and accurate Six-Degrees-of-Freedom (6DOF) tracking of the listeners' current and constantly changing position. Another important cue is a model of the listening environment which accurately represents the sound source and receiver relation through a Binaural Room Impulse Response (BRIR). It has been found that for the reproduced BRIRs, a simple room model, the direct sound, a few early reflections, and a common late reverberation model are good enough. In addition, a good generalized Head Related Transfer Function (HRTF) is needed.

Such systems enable the reproduction of multichannel loudspeaker systems over headphones, e.g. 9.1.6 with 16 loudspeakers, without the need to have that number of physical speakers available. For this setup, a basic problem arises: According to the room divergence effect, a model of the actual room and not of another recording environment should be used for the most plausible and externalized reproduction. The paper will explain this problem and show some first anecdotal evidence of how to blend the current and the remote room acoustics to achieve a convincing binaural reproduction.

Motivation

Immersive audio via headphones lives in a contradiction: The room divergence effect (see [1] and, for an overview, [2]) makes it necessary to match the binaural rendering to the actual room the listener is in. On the other hand, when mixing audio content, e.g. from a recording in a concert hall or a large church, it should sound like this environment. It is clear from anecdotal evidence and experience, e.g. from using wave field synthesis, that it is easier to recreate the sound of a reverberant room with many reflections in an acoustically drier room. Can this be extended to plausible binaural synthesis? Can the sound of another room be recreated by binaural synthesis without losing plausibility or externalization?

Recap: Immersive binaural synthesis

The experiments and first anecdotal evidence described in this paper use a binaural rendering system as described in [3]. The basic ideas and concepts for this system came from [1, 4–7]. Other binaural rendering systems have been proposed by [8] and in the proposed MPEG-I Audio Standard, see e.g. [9].

To achieve a good acoustical illusion, the requirements usually are externalization and plausibility.

- Externalization is defined by the localization of virtual sound sources clearly outside of the head and at a distance from the listener's head. This can be viewed as a binary choice (in head versus out of head localization) or on a scale, e.g. from 0 (in head) to 1 (possibly very far from the listening position).

- Plausibility in its most simple definition means that the listener cannot distinguish the virtual source inside the room from an internal reference (how the source should sound inside that room). Plausibility can also be subdivided into room related, source related or content related plausibility.

In the system used for this paper, the goal is to achieve a high degree of plausibility. This includes a very good externalization of the virtual sound source.

To be specific:
A Binaural Room Impulse Response (BRIR) for a virtual sound source is rendered using a measurement of a Room Impulse Response (RIR) onto which reflections are superimposed. The RIR is measured from a real source, e.g. a loudspeaker, towards a receiving position with an omnidirectional microphone. To adapt the rendering to different listener positions, fast and accurate 6DOF tracking is used. The BRIR is derived from parametric extrapolation according to the relation of sound source pose and position towards the current listener pose and position in the room. The metadata for the early reflections are gathered from a simplified shoebox model of the listening room. According to times and directions of arrival of the reflections, Head Related Transfer Functions (HRTFs) are convolved with portions of the RIR and blended with a modeled binaural diffuse part of the measured RIR to create the final rendering filter. In the end, the BRIR filter is applied to the input of the virtual source.

It has been shown in a listening test (see [10]) that this type of rendering creates a plausible reproduction of a virtual sound source which, in terms of plausibility and externalization, is second to no other algorithm. Anecdotal evidence from more than a thousand audio professionals at different conferences agrees with this result.

The original proof of concept demonstrated at a number of conferences and exhibitions just had two real and two virtual sound sources (at the same position). The current system allows rendering of up to 16 virtual sources. For that, according to the original paradigm the RIRs from all of these sources to the measurement microphone would need to be measured. [11] describes the measurement of a much smaller number of positions (e.g. 5 or 7) and the extrapolation of the virtual source RIRs to other positions to achieve a higher number of sound sources in the virtual setup. This way with limited amount of work a standardized multi-channel signal can be played back to the headphones.

Plausible reproduction of room acoustics differing from the listening environment

During demonstrations of the binaural rendering system, sound mixers frequently pose the following request:

"I want to listen to the recording environment and not to my studio" or, differently phrased "I want to listen to my mix at home but get the same acoustic impression as in my studio".

As shown by the paradigm of Room Divergence [1], this is not a trivial question. Some global observations hint that this might be possible:

- Via loudspeakers, this works to some degree: The listener always listens to the environment in the recording (original or added later via reverb) and the reflections of the room he is in.
- When listening in a home environment, the listening room acoustics are always mixed with the room acoustics of the studio, esp. the artificial reverb put on the recording.
- Sound engineers know that mastering has to be done with the ultimate listening space in mind.

The main topic for the work reported is: How can these facts be adapted to listening via headphones?

Approaches to solve these problems

A number of approaches have been tried to make the room sound different but still plausible. The possibilities below are an incomplete list of how to modify the RIRs to make the room sound like a different, larger environment. These are first experiments with the clear goal to keep the plausibility, externalization, and timbre of the sound.

- Stretching/shrinking late reverberation, i.e. modifying the average RT60. The idea is to change the late reverberation of the current room's RIR to match the average RT60 of the target room's RIR. [12]
- Stretching/shrinking late reverberation wrt the band wise RT60. While the former version was applied in a broad band sense, an alternative could be to change the late reverberation of the current room's RIR to match the band wise RT60 of the target room's RIR.
- The early reflections have an impact on localization. Still a more radical approach is to replace early reflections. This means swapping out the early reflections from the current room's RIR with the early reflections from the target room's RIR. This could be done for parts of the RIR only.
- To keep more information of the target room, a possibility is to geometrically map the early reflections from the target room onto the current room.
- As a compromise, replacing the current RIR with the target RIR but retaining the current room's geometrical mapping. This can be done for a part of the RIR, too. This is the approach taken at the DAGA demo, see below.

Test Design and Anecdotal Evidence

Testing the success of these room blending approaches is a challenging task due to the direct missing reference of the target room. One possible approach is to rely on a stable inner reference and ask participants in a listening test to rate the similarity between a room and another well-known room. This is of high interest but is also dependent on the listener, and the individual results are not directly comparable. Direct AB comparisons are almost impossible to achieve within the scope of this research and would be prone to inadvertently priming test subjects by highlighting minute, irrelevant issues during AB testing.

A proposed test design involves using the system described in the reference room, selecting a few target rooms, and applying the proposed parameter changes. The task for the listeners would be to listen to the resulting virtual sound sources and ask them for:

1. Quality, e.g. plausibility, externalization, audio visual coherency.
2. Similarity with the inner reference of the target room.

Experimental Setup @ DAGA and Results

At the DAGA conference, some participants had the opportunity to listen to one of the approaches of parameter changes. Unfortunately, this could not be tested during busy times, so just a few listeners really tested the system for that.

For the approach, two rooms were chosen: first, the exhibition room as the listening room and second, a room located in Ilmenau as the target room. For the target room, three acoustically different rooms with different reverberation patterns, were available. The approach was to take the RIR of the listening room and blend over to the target rooms RIR after a certain blending time. The blending time differed. Around the blending time, the energy was compared block-wise, and the gain of the target RIR was adapted to yield a smooth transition. For the calculation of the resulting RIRs for the other listening positions, the geometry of the listening room was chosen. Figure 1 shows the magnitude spectrogram of an RIR created with the described procedure. As a result, the listeners could listen to several RIRs from an RIR with solely the listening room over slowly, in six steps, adding more and more of the RIR of the target room to solely the target room RIR.

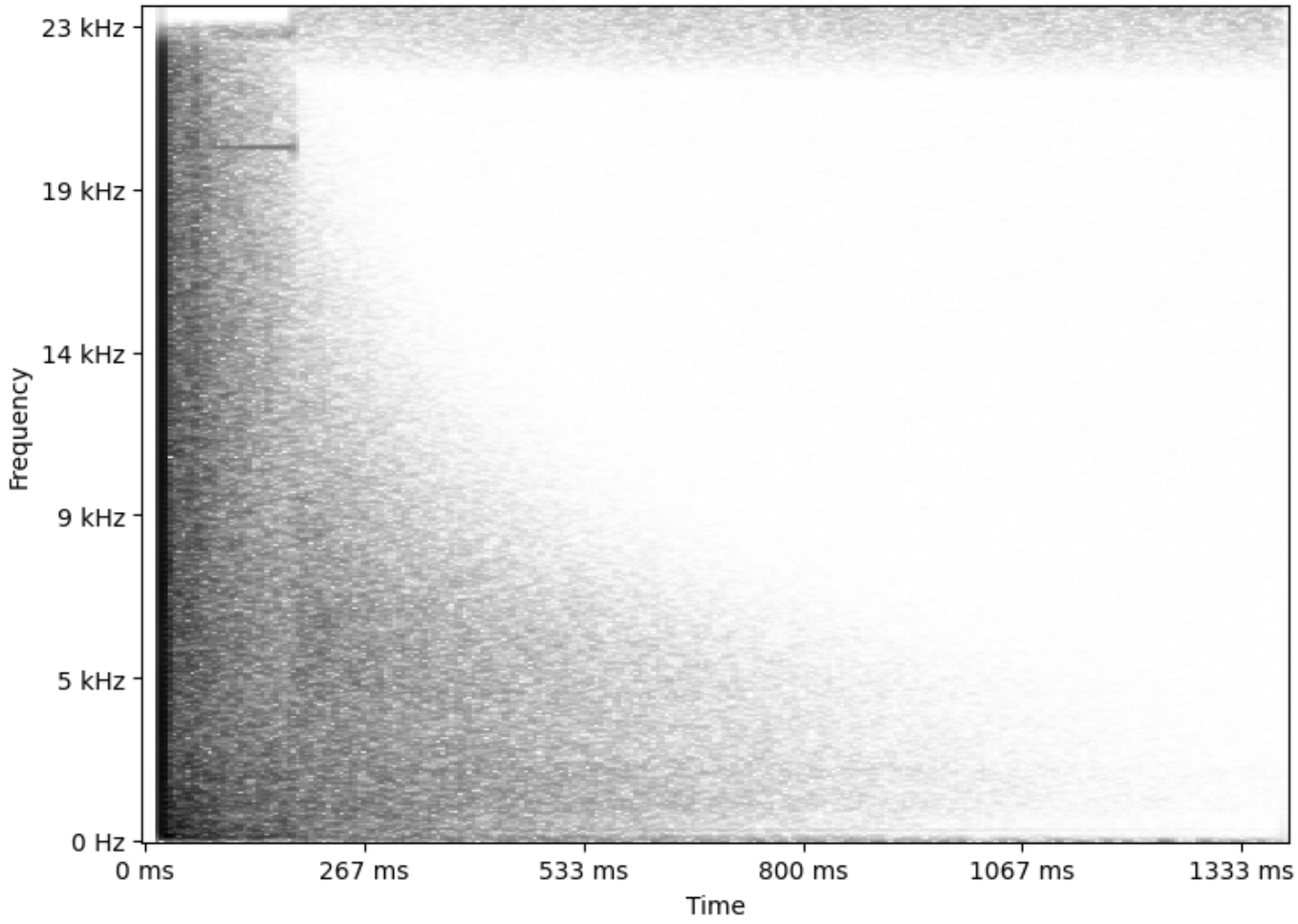

Figure 1: Magnitude spectrogram for blending the on-site RIR from Bella Center Copenhagen with an RIR from a room in Ilmenau (H2505). The blending time is 160 ms with a short crossfade duration of 0.3 ms. Note that there is a slight timbral difference at the crossfading time.

The question for the listeners was to observe at which blending step the illusion breaks down. The experiment was conducted using a single virtual sound source. This raises the question of how much of the target room's reflections and reverb can be included in the summed RIR without compromising the plausible and externalized sound event within the listening room. As a result, it can be stated that this approach worked quite well down to a certain blending time. The times used have been 160 ms, 80 ms, 40 ms (still quite good) and could be shortened to 20 ms, 10 ms and 6 ms where most listeners still found the rendering to be externalized, but clearly not as plausible as with the larger times. A relatively high amount of the target room's RIR can be embedded in the summed RIR before the virtual sound source is no longer plausible and externalized inside the listening room. Concrete times need to be evaluated in formal listening tests. The times also depend on how different the listening room and the target room are. Further, it depends on the type of audio signal; speech seems to be more critical than music.

General Observations

Below please find some observations from the demo and anecdotal evidence. Again, more research and more testing is necessary:

- As a well-known effect, making a dry room more reverberant is easier than vice versa. This has been observed in the context of loudspeaker based immersive systems (like Wave Field Synthesis), too.
- The late reverberation tail seems to have a strong effect on perceived room characteristics, like the perceived size of the room.
- If the DRR stays constant, localizability and externalization can still be kept.
- The late reverberation makes up little in terms of total energy. It can be modified without much influence on DRR.
- "Stretching" a room's late reverb tail by an average (target) RT60 mostly retains timbre.
- Bandwise "stretching" to a target RT60 changes timbre.
- Changing early reflections for small changes (moving a virtual wall) can still retain externalization.
- The early reflections are critical for the perception of externalization and localization of a virtual sound source. Thus, too much change of the early reflections in the model can destroy plausibility.

Summary and Outlook

The first informal listening tests have shown that in an acoustically relatively dry environment modified RIRs lead to the following changes:

- Given not too extreme changes, the RIR from a much more reverberant room still leads to virtual sources being perceived as externalized.
- The more the modified RIRs belong to a more reverberant or larger room, the more the sound sources start to sound diffuse, like from a much larger room. This effect is wanted, but it may destroy plausibility at some point.

An acoustically dry listening environment helps to overlay the acoustic impression of a larger room. There are still a number of open questions in this area needing further research:

- What qualities do audio professionals require from the simulation of rooms (both studio and recording environment)?
- Which auditive cues can be altered to change room acoustics?
- What about room divergence in VR compared to the AR situation observed here? In VR there could be a full audio-visual override.
- How much does the ability to explore the scene (6DOF) help to get plausible immersion?

The final question is: When does the illusion break down, leading to the well known in-head-localization?

Overall, the scenarios rendered with the described RIR blending approach work better than expected by the authors - despite the obvious room divergence.

References

[1] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 8th Int. Conf. on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal, 2016.

[2] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual matching of room acoustics for auditory augmented reality in small rooms - literature review and theoretical framework," *Trends in Hearing*, vol. 26, 2022. DOI: 10.1177/23312165221092919.

[3] K. Brandenburg, C. Fascella, N. Merten, *et al.*, "Implementation of and application scenarios for plausible immersive audio via headphones," in *Proc. of the 155th Conv. of the Audio Eng. Soc.*, New York, US, 2023.

[4] C. Pörschmann, P. Stade, and J. M. Arend, "Binauralization of omnidirectional room impulse responses - algorithm and technical evaluation," in *Proc. of the 20th Int. Conf. on Digital Audio Effects (DAFx-17)*, Edinburgh, UK, 2017.

[5] S. Werner, F. Klein, A. Neidhardt, *et al.*, "Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation," *Applied Sciences*, vol. 11, no. 3, p. 1150, Jan. 2021.

[6] K. Brandenburg, F. Klein, A. Neidhardt, *et al.*, "Creating auditory illusions with binaural technology," in *The Technology of Binaural Understanding*, J. Blauert and J. Braasch, Eds., Cham: Springer International Publishing, 2020, pp. 623–663, ISBN: 978-3-030-00386-9. DOI: 10.1007/978-3-030-00386-9_21. [Online]. Available: https://doi.org/10.1007/978-3-030-00386-9_21.

[7] A. Neidhardt, "On the plausibility of simplied acoustic room representations for listener translation in dynamic binaural auralizations," Ph.D. dissertation, Ilmenau, 2023.

[8] S. D. Ewert, N. Gößling, O. Buttler, *et al.*, "Computationally-efficient and perceptually-motivated rendering of diffuse reflections in room acoustics simulation," *arXiv preprint arXiv:2306.16696*, 2023.

[9] J. Herre and S. Disch, "MPEG-I immersive audio–reference model for the virtual/augmented reality audio standard," *JAES*, vol. 71, no. 5, pp. 229–240, 2023.

[10] U. Sloma, N. Merten, T. Thron, *et al.*, "Proof of concept of a binaural renderer with increased plausibility," in *49th Annu. Meet. for Acoustics, DAGA*, Hamburg, 2023.

[11] U. Sloma, T. Thron, N. Merten, and K. Brandenburg, "Multichannel immersive audio demo with reduced number of RIR measurements," in *50th Annu. Meet. for Acoustics, DAGA*, Hannover, 2024.

[12] C. Schneiderwind, M. Richter, N. Merten, and A. Neidhardt, "Effects of modified late reverberation on audio-visual plausibility and externalization in AR," in *2023 Int. Conf. on Immersive and 3D Audio (I3DA)*, Bologna, Italy, 2023.