How binaural room impulse responses influence the externalization of speech

Florian Wendt[1], Robert Höldrich[1], Marton Marschall[2]
[1] *University of Music and Performing Arts Graz, Austria;*
Institute of Electronic Music and Acoustics. Email: {wendt, hoeldrich}@iem.at
[2] *Technical University of Denmark; Department of Health Technology. Email: mmars@dtu.dk*

Introduction

Binaural reproduction is becoming popular as people consume radio, TV, and music over headphones out and about. One of the challenges when listening with headphones is that the sound image appears inside the head. Binaural renderers are meant to overcome this by using binaural room impulse responses (BRIRs) to reproduce at the ears the exact sound waves that one would hear when listening to the real sound source.

Nevertheless, binaural reproduction often does no better than a standard stereo recording at getting the sound image out of the head. This is because our individual pinna, head, and torso are different to the one simulated by the renderer. Research has shown that when the features of individual HRIRs are accurately simulated with headphones, listeners report externalized images, e.g. [1]. If a dummy head is used instead, then the image may be externalized, but it is usually diffuse or localized close to the head; especially for the synthesis of sources that are directly in the front of the listener [2].

Another failing in binaural reproduction is the acoustical divergence of the synthesized room and the actual listening room, yielding negative influence on the externalization. Research has shown that congruent room impulse responses (RIRs) in terms of amount of reverberation and direct-to-reverberant energy ratio between synthesized and listening room are needed to generate an externalized image, e.g. [3, 4]. Similar to generic HRIRs, a divergent room yields diffuse sound images which are perceived either close or even inside the listeners head.

This study focuses on the relative contribution of individual HRIRs and the congruency of RIRs on externalization. Based on an externalization model we sketch a listening experiment where listeners are asked to rate the externalization of speech in a simulated environment. After this, we present the results and discuss them in the last section.

A conceptual model for externalization

The starting point for this work was an early study on externalization presented by Plenge in [5]. The study suggests that externalization includes top-down processing, where the resulting auditory impression depends on prior knowledge about the auditory event. Based on this assumption, Plenge introduces a conceptual model of the externalization process, consisting of two memory stages: The instrumental means of our localization ability, i.e. HRIRs, is stored in the long-term memory and the short-term memory is filled with information on the sound source and the room characteristics, i.e. RIR. In contrast to the learning and adaptation process to new HRIRs which is rather slow [6], the content in the short-term memory is volatile and adaptation to new RIRs happens each time we enter a new listening situation. According to the model, sound images are externalized if the information provided by ear signals is in compliance with the information in both memories. If there is any contradiction between the memory stages and what we hear, the sound image is localized in the head.

Experimental setup and conditions

To examine the interaction of HRIR and RIR on externalization we performed a listening experiment with simulated acoustics. The experiment took place in the audiovisual immersion lab (AVIL) at DTU, a $6\,\mathrm{m} \times 7\,\mathrm{m} \times 8\,\mathrm{m}$ large anechoic chamber equipped with a 64-channel spherical loudspeaker array at a distance of $r = 2.4\,\mathrm{m}$ from the central listening position. The simulation employs the LoRa toolbox [7] and ODEON room acoustic software [8], which is a hybrid of an image source model and a ray tracing model.

Conditions studying the long-term memory consisted of *individual* or *generic* HRIRs. For individual HRIRs the loudspeaker sphere was used as playback device, whereas conditions with generic HRIRs were played back over Sennheiser HD800 headphones. The headphones signal was created by convolving the loudspeaker signals with corresponding HRIRs of a Brüel & Kjær HATS. Figure 1 shows the processing scheme for the playback of signals with individual and generic HRIRs. Obviously, for loudspeaker conditions the open headphone alters the signals somewhat as the sound has to propagate through the ear cups. This attenuation was considered in the generic conditions by equipping the HATS with the same headphones during the HRIR measurement. Figure 2 shows the attenuation due to headphone isolation for different loudspeaker directions.

Familiarizing with the room is essential for externalization [4, 5] and conditions with *congruent* or *divergent* RIRs were used to study the short-term memory. Congruent conditions were established by a familiarization phase in the room before evaluation, and for divergent conditions training and evaluation room differed. Training on a room was done with individual HRIRs (loudspeaker playback) and lasted 7 min. Listeners could move their heads freely during the training and four talkers appeared at different azimuth angles at $r = 2.4\,\mathrm{m}$ around the listener.

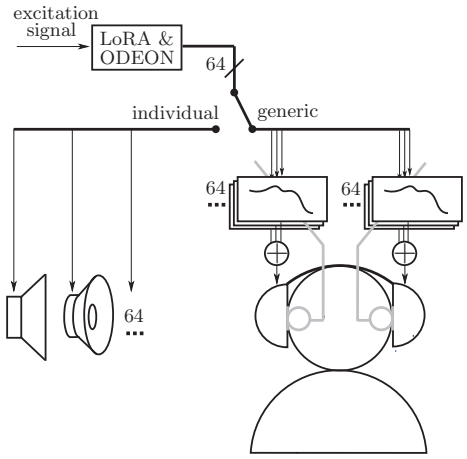

Figure 1: Processing scheme for individual and generic HRIRs. The room acoustics is simulated with LoRA and ODEON for the given 64 channel loudspeaker sphere. Individual HRIRs are tested using loudspeaker playback. Generic HRIRs are tested using open headphones. The headphones signal is created by convolving the loudspeaker signals with corresponding HRIR measurements of a dummy head with headphones on (grey).

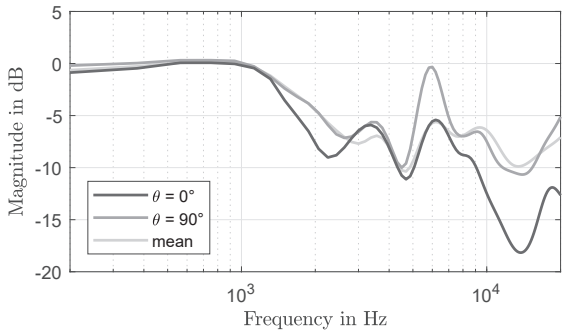

Figure 2: Measured attenuation of the external sound vs. frequency of the Sennheiser HD800 headphone for two horizontal directions at azimuth θ and the mean attenuation over all 64 loudspeaker directions. Magnitudes are third-octave smoothed and normalized to 1 kHz.

After training, listeners were instructed to face the loudspeaker directly in front of them and rate the degree of externalization of the condition under test. The condition consisted of a 2.5 s-long unknown speech sample, simulated at the position of the frontal loudspeaker. Ratings were input using a computer keyboard with a rating scale similar to that in [1]:

0 The source is in my head.

1 The source is not well externalized. It is at my ear, or at my skull.

2 The source is externalized. It is either before or behind the loudspeaker.

3 The source is well externalized, compact, and located at the loudspeaker.

Conditions could be either with individual or generic HRIRs (loudspeaker or headphone playback) and with congruent or divergent RIRs (trained on congruent or divergent room). During evaluation, the position of the listeners head was tracked using an OptiTrack system.

Table 1: Description and characteristics of the simulated rooms as used in the experiment.

room	T_{60}	*DRR*	size in m	description
*R*0	0.0 s	∞ dB		free field
*R*1	0.5 s	7.5 dB	9.5×7.6×3.0	classroom
*R*2	1.2 s	2.8 dB	14.2×9.0×5.5	auditorium

The playback of the condition under test started automatically when the listener was looking at the frontal direction and visual feedback was given whenever the direction deviated more than 2°.

Three different room simulations were studied, cf. Table 1. All listeners started with the training on *R*0. Conditions rated after this training were the congruent condition (simulated with *R*0 and denoted as *R*0⁄0) and a divergent condition simulated with *R*1 (denoted as *R*1⁄0). Both conditions were tested four times, with individual/generic HRIRs and with two repetitions. After each rating of a condition, the training on room *R*0 was continued for another 10 s before the next condition was tested. In this way the information in the short-term memory is maintained.
Then room *R*1 was trained and conditions *R*0⁄1, *R*1⁄1, and *R*2⁄1 were tested similarly. Lastly, room *R*2 was trained and conditions *R*1⁄2 and *R*2⁄2 were tested. This sequence of training was the same for all listeners, whereas the conditions presented after each training were an individual random permutation yielding $(2+3+2)$ conditions $\times 2$ HRIRs $\times 2$ repetitions $= 28$ rating tasks for each listener.

Seven listeners (3 female and 4 male, age 28-45 years) participated in the experiment including one of the authors. Except for the author, none of them was experienced in terms of binaural perceptual experiments.

Experimental results

The results are given in Figure 3 as median values by filled symbols for individual HRIRs and open symbols for the generic HRIRs with corresponding 95% confidence intervals. Conditions are grouped according to the training room.

As might be expected, externalization ratings for congruent conditions *R*0⁄0, *R*1⁄1, and *R*2⁄2 with individual HRIRs are the highest. Generic HRIRs on the other hand are rated lower.
A statistical analysis reveals the rated room to be a significant parameter (Friedman test $p < 0.01$) for generic HRIRs and by increasing the reverberation time T_{60} (and therefore decreasing the direct-to-reverberant energy ratio *DRR*), the difference between generic and individual HRIRs diminishes. For *R*1⁄1 and *R*2⁄2 no significance between individual and generic HRIR is obtained ($p > 0.3$). The significance of the rated room applies also for divergent conditions and ratings of *R*0⁄1 are significantly lower then those of *R*1⁄0, *R*1⁄2, and *R*2⁄1 (Wilcoxon rank sum test $p \ll 0.01$).

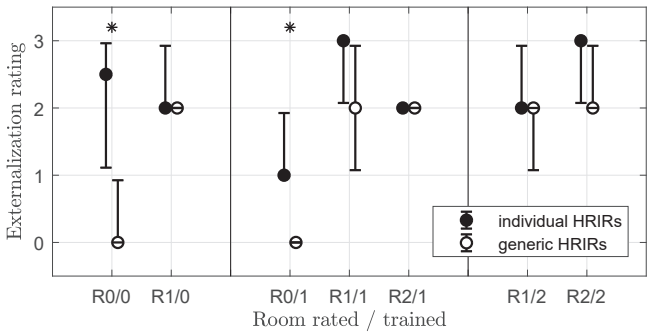

Figure 3: Medians and 95% confidence intervals of the externalization rating for different room simulations after training. Playback employed loudspeakers (individual HRIRs) and headphones (generic HRIRs) and significant differences ($p < 0.01$) are marked with an asterisk.

Table 2: Response frequency in % for well externalized percepts (rating 3) of tested conditions with individual and generic HRIRs.

cond.	*R*0⁄0	*R*0⁄1	*R*1⁄0	*R*1⁄1	*R*1⁄2	*R*2⁄1	*R*2⁄1
gen.	7	0	7	43	21	0	36
ind.	50	7	29	64	43	7	57

The influence of the training with individual HRIRs can be seen by comparing congruent and divergent conditions of the same room. Considering room $R0$, the congruent condition $R0⁄0$ with individual HRIRs is rated significantly higher then the divergent condition $R0⁄1$ ($p_{\text{ind}} = 0.04$). For generic HRIRs, on the other hand the training does not yield any improvement and externalization remains poor. Similarly, for room $R2$ the externalization of $R2⁄1$ significantly improves after training ($R2⁄2$) with individual HRIRs ($p_{\text{ind}} = 0.01$), although the ratings are generally higher compared to $R0$. And even with generic HRIRs a weak improvement of externalization is seen ($p_{\text{gen}} = 0.08$). Surprisingly, for room $R1$ the training is not a significant factor and does not yield any improvement ($R1⁄\langle 0, 2\rangle$ vs. $R1⁄1$); neither for the individual nor for generic HRIRs ($p \geq 0.15$). If we consider the response frequency of highest ratings, listed in Table 2 and grouped according the rated room, we see that conditions $R1⁄\langle . \rangle$ achieved higher percentages compared to conditions $R0⁄\langle . \rangle$ and $R2⁄\langle . \rangle$, especially for divergent conditions with individual HRIRs.

Conclusion and Outlook

The influence of BRIRs on the externalization of speech was considered in this study. Based on a conceptual model of externalization, we presented a listening experiment to examine contributions of individual/generic HRIRs and congruent/divergent RIR on the externalization.
We could show that the advantage of individual HRIRs compared to generic HRIRs on externalization diminishes with increasing reverberation energy. This is independent of any prior knowledge of the room and applies for congruent conditions (preceding training with congruent RIR) as well as for divergent conditions (preceding training with divergent RIR).

Agreeing with the findings in [4], training with congruent RIRs can improve externalization and corresponding conditions with individual HRIRs are rated highest. It remains unclear if the same applies for generic HRIRs, as the training was done only with individual HRIRs. However, in [4] the influence of training was found to be less distinctive for generic HRIRs.
Interestingly, externalization in room $R1$ behaves different and training was not found to be significant: divergent conditions were rated similarly high as congruent conditions. Compared to the other rooms, reverberation time and size of $R1$ is typical for domestic environments. Returning to the externalization model, this could suggest that the characteristics of such a *conventional* room are stored in the long-term memory. By contrast, as vision can affect externalization [9], an alternative hypothesis is that listeners expect the anechoic chamber to sound like $R1$ due to the similar ground size. Further research with more listeners is needed to answer this questions.

Overall we could show that even with generic HRIR or an unknown RIR, the binaural reproduction of speech can be perceived as externalized. In [4] the authors conjecture that individual HRIRs are more important than convergent RIRs. This is only partially confirmed by our results and we have shown that the need for individual HRIRs and/or congruent RIR depends on the listening room.

Acknowledgments

The authors thank all listeners taking part in the experiment. Special thanks are due to Torsten Dau for the opportunity to conduct the study at DTU and to Kasper Duemose Lund for helping with the setup.

References

[1] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3678–3688, 1996.

[2] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *Journal of the Acoustical Society of America*, vol. 94, no. 111, pp. 111–123, 1993.

[3] D. R. Begault, E. M. Wenzel, A. S. Lee, 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," in *Proceedings of the 108th Convention of the Audio Engineering Society*, vol. 5133, (Paris), 2000.

[4] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *8th International Conference on Quality of Multimedia Experience, QoMEX 2016*, 2016.

[5] G. Plenge, "Über das Problem der Im-Kopf-Lokalisation," *Acustica*, vol. 26, no. 5, pp. 241–252, 1972.

[6] F. Klein and S. Werner, "HRTF adaptation and pattern learning," in *Proceedings of ISAAR 2013: Auditory Plasticity - Listening with the Brain* (T. Dau, ed.), (Nyborg, DK), 2013.

[7] S. Favrot and J. M. Buchholz, "LoRA: A Loudspeaker-Based Room Auralization System," *Acta Acustica united with Acustica*, vol. 96, no. 2, pp. 364–375, 2010.

[8] G. Naylor, "ODEON - Another hybrid room acoustical model," *Applied Acoustics*, vol. 38, no. 2-4, pp. 131–143, 1993.

[9] J. Udesen, T. Piechowiak, and F. Gran, "Vision Affects Sound Externalization," in *AES Conference: 55th International Conference: Spatial Audio*, pp. 27–30, 2014.