Externalization in data-based Binaural Synthesis: Effects of Impulse Response Length

Florian Völk
AG Technische Akustik, MMK, TU München, Germany
Email: florian.voelk@mytum.de

Introduction

Databases of head related impulse responses (HRIRs) for binaural synthesis can be measured either in anechoic or reflective environments. The time until energy decays to negligible values after impulsive excitation ranges from some milliseconds under anechoic conditions to some seconds in reflective environments. The computational load for dynamic binaural synthesis increases significantly with the impulse response length. For that reason short impulse responses would be preferable. If an authentic binaural simulation of an everyday listening situation is desired, an impulse response length in the order of the reverberation time or more would be necessary. A shortening of the impulse responses might lead to perceptual aberrations.

This paper considers the dependence of the perceived distance of auditory images (externalization) on the impulse response length. Impulse response databases of different lengths are used for dynamic binaural synthesis of a frontal virtual source. The distances of the resulting auditory images are measured in a psychoacoustic localization experiment. Statistical analysis of the resulting relative externalization differences suggests that shortening of the impulse responses generally reduces the externalization of auditory images.

Clarification of Terms

The term head related impulse response is usually used for impulse responses recorded in the ear canals of a subject under anechoic conditions. If the recording is carried out in a reflective environment, the resulting impulse responses are called binaural room impulse responses (BRIRs). To point out that the impulse responses contain information stemming from the room, the term room is included. Whenever in the following one single term for both groups of impulse responses (with and without reflections) without an explicit distinction between them is necessary, the term head related impulse response will be used.

We define the goal of a binaural synthesis as the creation of the auditory events that arise in the real scene to be synthesized in complete darkness (cf. Völk et al. [1]). In a straightforward manner, externalization is defined here as the perceived distance of an auditory event to the center of the head (following Kim and Choi [2]), but with the additional requirement of dark circumstances.

Aim of this Work

There are two common approaches to the collection of the HRIR library for a binaural synthesis: a model- and a data-driven method (for an overview cf. Vorländer [3]). The difference between these approaches is the procedure used for the room simulation. The first approach is based on a HRIR library measured under anechoic conditions. These HRIRs are convolved with a room impulse response that may be measured or - under certain conditions - rendered in real-time (cf. Vorländer [3]). The latter procedure allows maximum flexibility since changes of the room's acoustical properties during system operation and even simulations of non-existent rooms are possible. The second more traditional and restrictive approach relies on BRIR measurements in the room of which a simulation is desired. This room may be an anechoic chamber. Here, the data-driven and the model-driven approach without room simulation are identical. If there are reflections in the recording room, the data-based approach requires lots of measurements, making the synthesis of reflective environments a time consuming and resource-intensive task (e.g. memory requirements). In the data-based case, the computational burden of a dynamic binaural synthesis increases significantly with the length of the used impulse responses.

This paper deals with the question whether an artificial shortening of the impulse responses influences externalization. The results will be compared to existing psychoacoustical data and models for distance perception with binaural synthesis. In addition, possible influences of the impulse response length on height perception will be studied.

Procedure

The experiments reported in this paper were conducted under dark circumstances. Inputs to other modalities (e.g. the tactile sense) were neglected. It is assumed that they play an inferior role when subjects are seated in a dark room and sound levels remain in the range around 70 dB (A) for broadband stimuli as applied in the current work.

Unpaid voluntary subjects were seated in a laboratory room (reverberation time see Figure 1) after reading the instructions. They put on the headphones (STAX lambda pro new) and after that, the room was completely darkened.

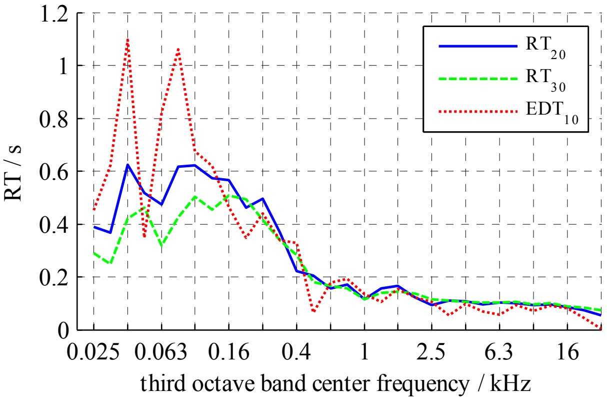

Figure 1: Reverberation time (RT) of the laboratory room ($V \approx 84$ m^3) where the experiments and the recording of the impulse responses took place. The reverberation times were computed from the right frontal HRIR of full length.

During the experiment, the subjects were allowed to move and rotate their heads and bodies freely (while remaining seated on the chair).

Subjects had to state the parameter of interest orally. Their responses were recorded and evaluated manually later in an offline analysis. The experimental method was selected according to the procedure of absolute magnitude estimation. Subjects were instructed to depict the perceived magnitudes on a scale they were familiar with, the metric length scale.

Stimuli

All sound stimuli were presented using a data based dynamic binaural synthesis system (cf. Völk et al. [4]), synthesizing a loudspeaker in front of the subjects.

HRIR Sets

For the binaural synthesis, a BRIR-database was used that has been recorded with a swept sine method (cf. Farina [5], Müller and Massarani [6], sweep duration 5 s, frequency range from 10 Hz to 22 kHz) in the blocked ear canals of an individual (male, 28 years, known as a so called good localizer, a person whose HRIRs have shown good localization results with a number of subjects in previous studies, cf. Møller et al. [7], Seeber and Fastl [8]).

The data set was recorded in a laboratory environment with a relatively low reverberation time between 100 and 200 ms for frequencies above about 300 Hz, but increasing up to 1s for frequencies below 100 Hz (see Figure 1). The sweep signals were played back during the recording at a sample rate of 44.1 kHz by a 24 bit D/A-converter (RME Fireface 400) and a loudspeaker (Klein & Hummel O200) at 80 dB SPL. For the recording, miniature microphones (Sennheiser KE 4-211-2) connected to the microphone amplifiers of the RME device, which was also used as 24 bit A/D converter, were utilized. All post processing was carried out in MATLAB with double precision word length (8 byte).

The resulting data were filtered with a linear phase FIR filter to equalize the undesired distortion of the headphones, microphones, amplifiers, and other playback as well as recording equipment.

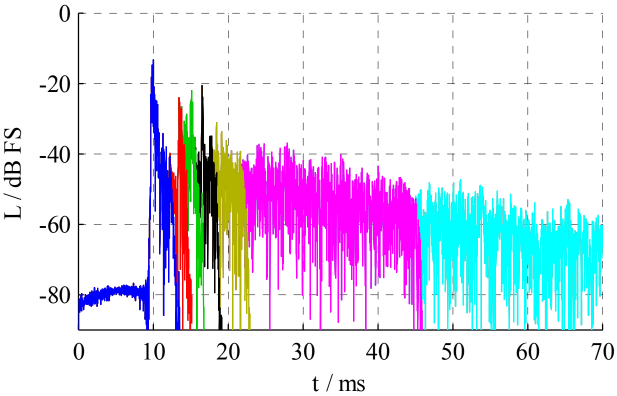

Figure 2: Frontal right impulse responses used for the listening experiments. From the original measurements, impulse response sets with lengths of about 13.6, 15.2, 17, 19.3, 23.2, 46.4, 92.9, and 185.8 ms were generated by Gaussian modulation (indicated by different colours).

From these data, eight HRIR sets were generated by Gaussian modulation of the original data (time between 90 % and 10 % of the amplitude: 0.7 ms) so that the impulse response lengths of the resulting sets were about 13.6, 15.2, 17, 19.3, 23.2, 46.4, 92.9, and 185.8 ms (see Figure 2). It was intended to separate each of the first order reflections. So, the first four impulse responses should contain direct sound, floor, ceiling, and wall reflections (in this order). Figure 3 shows the frequency response of the left and right frontal measurements, calculated from the longest part of the impulse responses used in this experiment (about 185.8 ms).

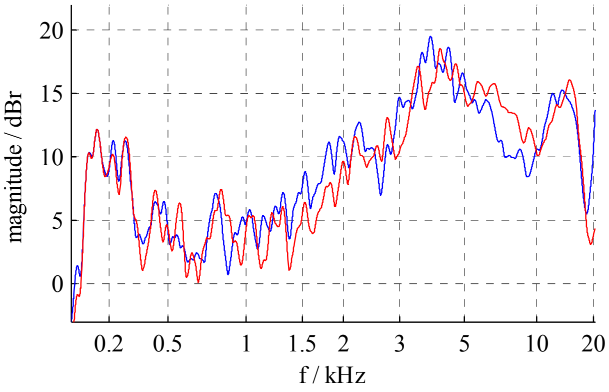

Figure 3: Magnitude of the HRTFs at 0°, mean of Fourier-t-Transformation (FTT, cf. Terhardt [9]) of the HRIRs of 185.8 ms duration.

Sound Stimuli

The experiments were conducted using two different stimuli: a male speech and a broadband noise signal. As the latter, pulsed broadband uniform exciting noise (UEN, cf. Fastl and Zwicker [10]) was used (see Figure 4).

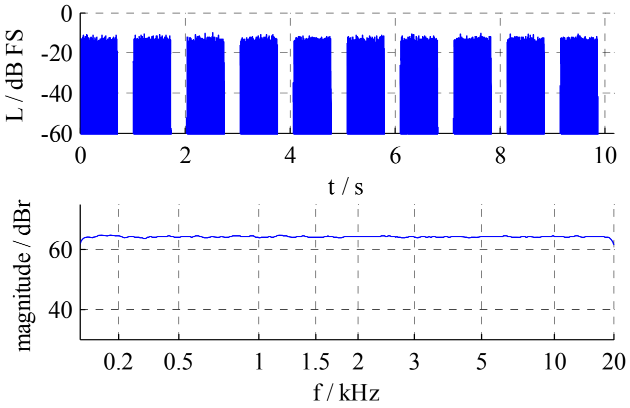

Figure 4: Time function and magnitude of the long term spectrum computed as mean of the Fourier-t-Transform. (FTT) of the used Uniform Exciting Noise (UEN).

This stimulus contains equal intensity in each critical band, thus providing all spectral cues contained in the HRIRs to the listener with the same perceptual weight. Therefore, all possible spectral information is available to the hearing system, but no influence of the sound stimulus on the auditory event should be present. To add temporal information besides the random temporal structure of the noise, the UEN was pulsed with 700 ms pulse duration and 300 ms pause duration in between the impulses. Following Blauert and

Braasch [11], 200 ms is the minimal duration allowing dynamic localization cues, so the latter should be available. The pulses were modulated with 20 ms Gaussian gating signals.

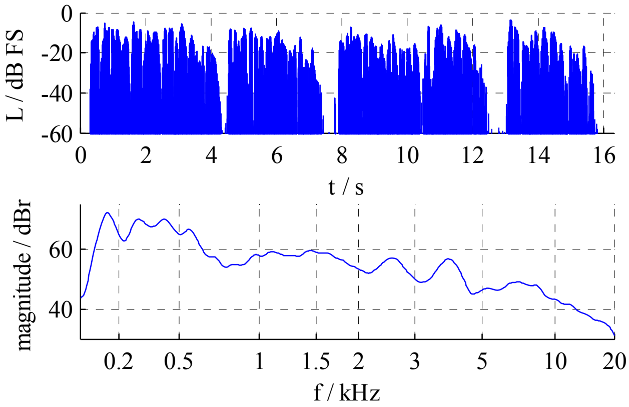

Figure 5: Time function and magnitude of the long term spectrum computed as mean of the Fourier-t-Transform. (FTT) of the used speech signal.

Figure 5 shows the other used sound signal, a recording of a German male speaker with 16 bit word length and 44.1 kHz sampling frequency.

Results

To prove the reliability of the subjects, for each experiment the median of the ranges containing all answers per stimulus was compared to a critical limit. The critical limits were set empirically to one meter for the denoted distances and half a meter for the height judgments. The results for each subject were computed as medians of the answers to each stimulus (intra-individual medians). From these values, the inter-individual medians and inter-quartile ranges were calculated. In addition, the data sets were checked for significant differences (ANOVA with post-hoc comparison according to Bonferroni). The resulting inter-quartile ranges are displayed as lines with markers at the quartiles.

Distance

15 normal hearing subjects (one female and 14 male) participated in the experiment regarding distance perception. Seven of the subjects had previous experience in listening tests, four of them were familiar with listening in virtual acoustical displays and experienced with localization experiments. The others had never participated in listening experiments before (naive subjects). The criterion rejected three of the naive subjects (two male and one female), so that the results are computed from the answers of 12 male subjects aged between 23 and 29 years (mean value: 25.4 years, four experts, three experienced, and five naive). The experiment was conducted as one run with a mean duration of 17.7 minutes (range: 9 to 41.7 minutes). Figure 6 shows the resulting distance judgments of the auditory events created by the different impulse response sets for a frontal virtual loudspeaker versus the different used impulse response lengths. The intended distance of the virtual loudspeaker is defined as the spacing between the loudspeaker and the subject at the recording time. For the current experiment, the intended distance was 2.1 m. Circles indicate medians for male speech, triangles for UEN as sound stimulus. Different colors for the medians are utilized to relate the results to Figure 2, where the corresponding impulse responses are shown.

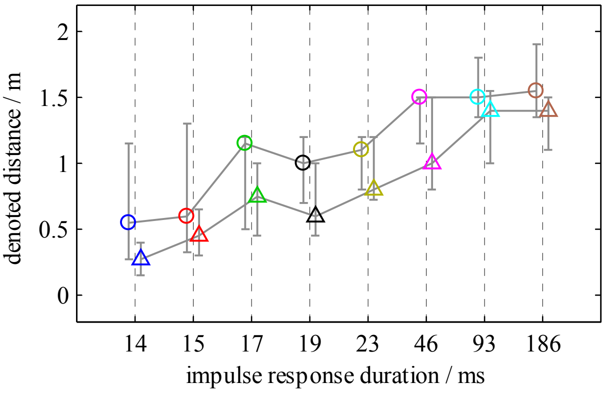

Figure 6: Distance of the auditory events (ordinate) created by dynamic binaural synthesis of a frontal virtual loudspeaker using HRIRs cut to different lengths (abscissa). Medians are indicated by circles for male speech and by triangles for uniform exciting noise as sound stimulus. The different colours relate the results to the different impulse responses shown in Figure 2.

ANOVA suggests that the results for impulse response lengths between 14 and 19 ms differ each highly significant (1 % significance level) from the results for each length between 46 and 186 ms. In addition, highly significant differences occur between the responses for 14 and 23 ms as well as for 23 and 93 ms (1 %). The results for 15 and 23 ms are significantly different (5 % significance level). The mean values of the inter-quartile ranges for the different HRIR-durations are 0.55 m for both sound stimuli (speech and UEN). ANOVA suggests a highly significant difference between the results for the two distinct sound stimuli.

Height

In the experiment concerning height perception 13 male subjects participated. Six of them had previous experience in listening tests, four of them were familiar with listening in virtual acoustical displays and experienced with localization experiments. The others had never participated in listening experiments before (naive subjects). The criterion rejected four of the naive subjects from the analysis, so the results are computed out of the answers of 9 male subjects aged between 23 and 29 years (mean: 25.8 years, four experts, two experienced and 3 naive). The experiment lasted for a mean duration of 16.5 minutes (7 to 34.9 minutes). Figure 7 illustrates the perceived distances over the different impulse response lengths. Again, the different colors link this illustration to the used impulse responses, shown in Figure 2. Circles depict the medians for speech signals, triangles for UEN. ANOVA indicates no significant differences at all for the perceived heights (nor between the impulse response lengths neither between the stimuli). The mean size of the inter-quartile ranges is 0.26 m for speech and 0.24 m for UEN as sound signal. For the speech stimulus, two groups are evident: the mean inter-quartile range for the leftmost five stimuli is 0.16 m, but 0.35 m for the rightmost three.

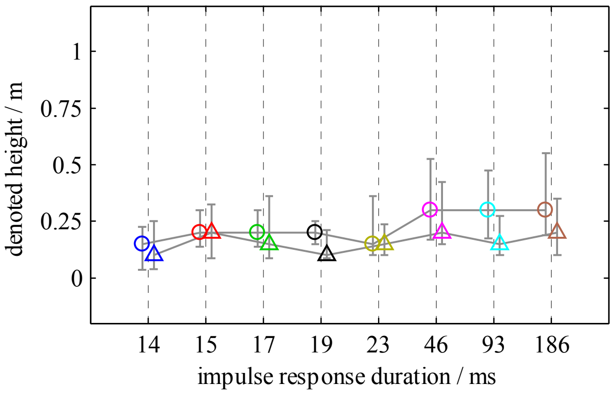

Figure 7: Height of the auditory events above the horizontal plane (ordinate) created by dynamic binaural synthesis of a frontal virtual loudspeaker using HRIRs cut to different lengths (abscissa). Circles indicate medians for male speech, triangles for uniform exciting noise as stimulus.

Discussion and Summary

This paper considers the denoted distance of auditory images (externalization) in dependence of the impulse response length using dynamic binaural synthesis. The results show that higher order reflections (included in impulse responses longer than about 25 ms) increase distance judgments substantially. In addition, the early reflection pattern can influence externalization judgments. In the presented study, reverberation after about 100 ms did not influence externalization judgments. In contrast, a highly significant difference between the used stimuli occurred in such a way that denoted distances of auditory events tend to be farther out with male speech than with a broadband noise stimulus. The height judgments of the auditory events were not significantly influenced by the impulse response length.

Comparison to Previous Work

Begault [12] found, similar to the study on hand, an externalization enhancement when adding synthetic reverberation to binaurally synthesized speech signals. At the same time, the number of errors in height localization increased. Figure 7 indicates similar mechanisms as the inter-quartile ranges tend to be larger for the longest impulse responses, especially for the speech stimulus. Bronkhorst and Houtgast presented in [13] a model for auditory distance perception based on the ratio of energies of direct and reflected sound. Figure 8 shows the results if this model is applied to the impulse responses used in the current work (model outputs are similar for left and right side, the algorithm is only applicable to the longer impulse responses). The model predicts the measurements for the speech stimulus nearly perfect, whereas the noise stimulus causes somewhat different judgments.

Acknowledgements

The author is indebted to Prof. Dr.-Ing. H. Fastl and Dr.-Ing. S. Kerber for valuable suggestions and critical comments on the draft material. Part of this work was supported by grant FA 140/4 of the Deutsche Forschungsgemeinschaft (DFG). The author gratefully acknowledges the support of Sennheiser electronic GmbH & Co. KG who provided the microphones for the HRIR measurements.

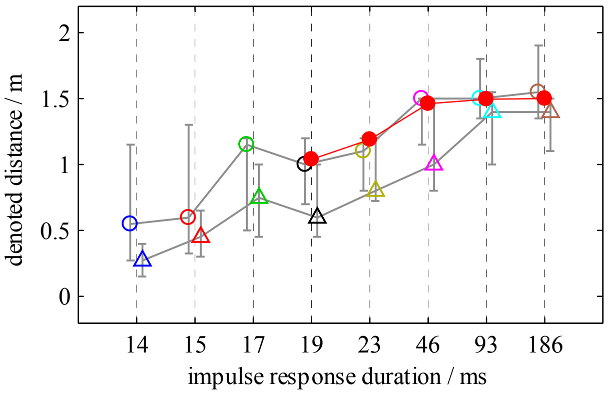

Figure 8: Comparison between model predictions (filled circles) computed according to Bronkhorst and Houtgast [13] and psychoacoustical data measured in the current work (unfilled symbols). Circles represent medians of auditory event judgements for speech, triangles for uniform exciting noise as sound stimulus, presented over a binaurally synthesized virtual loudspeaker.

References

[1] Völk, F.; Heinemann, F.; Fastl, H.: Externalization in binaural synthesis: effects of recording environment and measurement procedure. *Proc. Acoustics '08*, 6419-6424 (2008)

[2] Kim, S.; Choi, W.: On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach. J. Acoust. Soc. Am. **117**, 3657-3665 (2005)

[3] Vorländer, M.: *Auralization – Fundamentals of Acoustics, Modeling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, Berlin, Heidelberg (2008)

[4] Völk, F.; Kerber, S.; Fastl, H.; Reifinger, S.: Design und Realisierung von virtueller Akustik für ein Augmented-Reality-Labor, *Fortschritte der Akustik, DAGA '07*, DEGA e. V., Berlin (2007)

[5] Farina, A.: Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. *108th AES Convention*, Preprint 5093 (2000)

[6] Müller, S.; Massarani, P.: Transfer-Function Measurement with Sweeps, J. Audio Eng. Soc. **49**, 443-471 (2001)

[7] Møller, H.; Jensen, C. B.; Hammershøi, D.; Sørensen, M. F.: Selection of a typical human subject for binaural recording. Acta Acustica united with Acustica **82**, 215 (1996)

[8] Seeber, B. U.; Fastl, H.: Subjective Selection of Non-Individual Head-Related Transfer Functions. *Proc. of ICAD 2003* (2003)

[9] Terhardt, E.: Fourier transformation of time signals: Conceptual revision. ACUSTICA **57**, 242-256 (1985)

[10] Fastl, H.; Zwicker, E.: *Psychoacoustics – Facts and Models.* 3rd ed., Springer, Berlin, Heidelberg (2007)

[11] Blauert, J.; Braasch, J.: Räumliches Hören. Contribution for the *Handbuch der Audiotechnik (Chapter 3., Stefan Weinzierl, Ed..), Springer Verlag*, Berlin, Heidelberg (2007)

[12] Begault, D. R.: Perceptual effects of synthetic reverberation on three-dimensional audio systems, J. Audio Eng. Soc. **40**, 895-904 (1992)

[13] Bronkhorst, A. W.; Houtgast T.: Auditory distance perception in rooms, NATURE **397**, 517-520 (1999)