Audio Engineering Society

Convention e-Brief 591

Presented at the 148th Convention
2020 June 2–5, Online

Real-time auralization while having prepared in advance for possible head movements

Mantas Tamulionis

Vilnius Gediminas Technical University, Faculty of Electronics, Vilnius, 03227, Lithuania

Correspondence should be addressed to Mantas Tamulionis (mantas.tamulionis@vgtu.lt)

ABSTRACT

In real-time binaural rendering (or auralization), it is important to ensure that no artifacts that may result from CPU overload are heard by the listener. The research is based on the work of A. Lindau. The author states that human cannot detect a difference between signals processed with different HRTFs that represent less than 3 degrees of head position change. The method proposed performs pre-filtering and prepares three variants of the auralized signal: one corresponding to the current position of the listener's head and other two required when the head rotates more than 3 degrees to either side. The right signal can be played immediately. This method allows reducing the size of HRTF database, computation time and saving CPU labor.

1 Introduction

Auralization is the process of creating realistic spatial sound that can strongly impact and immerse the user in virtual space. In order to perform auralization, we must first model the room virtually reproducing its acoustic properties. We need to know its dimensions and materials that are used to cover its surfaces (walls, floor, ceiling, etc.). Once we have a virtual room model, we can place sound sources and listeners. The filter we apply to anechoic sound that can simulate specific locations of the listener and sound source in a room is called *Room Impulse Response (RIR).* Each combination of position between the listener and the sound source in the room has a separate RIR filter and all these filters form the RIR database [1]. One of the most popular acoustics simulation techniques, also used in our study, is geometric acoustics and *Image Source Method (ISM).* Its implementation requires relatively small computing resources and is therefore popular when it comes to real-time tasks [2]. When analyzing the ISM method, it is important to understand the propagation of sound waves in any room. The listener will always be reached first by the direct sound, then the early reflections from the nearest surfaces and finally the late reflections which can also be called the *Diffuse Field.* These are chaotic reflections and their amount and propagation time are almost independent of the listener's position in the room. In the ISM method, each reflection is treated as a new direct sound wave and is called a new virtual sound source. Their intensity, direction, and the amount of additional virtual sound sources created depend on the surface absorption and scattering coefficients.

The next step in auralization, after applying the RIR filter, is modeling the listener's head position, or more precisely, simulating the audio signal coming into both eardrums of the listener. At this stage, we need to apply a so-called *Head Related Impulse Response (HRIR)* to the signal. This filter simulates the effect of the acoustic shadow of the human head, the shape of the outer ear and the length of the ear canal [3]. In most cases, we select a particular HRTF database for auralization, which ideally has filters that match every possible position of the human head in the horizontal (azimuth) and vertical (elevation) axes. From this database, we need to select the most

suitable filter for the listener's head position that currently exists. If there is a limited number of filters in the database (low spatial resolution) and a filter that does not exist needs to be applied, we need to interpolate two adjacent values and calculate the required filter. The combined effect of RIR and HRIR filters on an audio signal is usually attributed to a single filter, the *Binaural Room Impulse Response (BRIR),* which is a combination of the two filters discussed above [1]. The selected BRIR filter is suitable for one situation only, which corresponds to the specific positions of the sound source, the listener and his head. If at least one of these variable changes, we need to calculate a new version of the BRIR filter.

Figure 1. The process of making audio recordings in the university classroom.

Good auralization results should be as realistic as possible, so it is common to compare them with recordings made in the same room. In fact, the process starts when we first choose a real room where audio recordings can be made before modeling a virtual one. Then, knowing the dimensions of this real room and the materials covering its walls, we create its digital version. During audio recording, we play anechoic audio files in the room, and then apply auralization filters to the same files [4]. We can compare a real record and its computer simulated counterpart visually by looking at the signal spectra. We can also make comparisons with listening tests, allowing people to evaluate. Such listening tests are defined by clear assessment criteria [5]. It has been noticed that listeners' judgment differs depending on whether the test is conducted in an anechoic or reverberated environment. The results also differ when listeners perform their assessment using speakers or headphones.

To ensure a smooth real-time auralization, its results must be updated quickly enough. If the calculation takes too long, the user will notice it as a sound defect [6]. Total system latency (TSL) defines the time required for the system to complete all processes from the occurrence of a user motion to the delivery of the results to the user. This time period may not exceed the maximum allowable time, which is based on the user's perception threshold. A. Lindau found that a person does not notice a delay of up to 52 ms when listening to a speech signal in a reverberant room and up to 64 ms in an anechoic room.

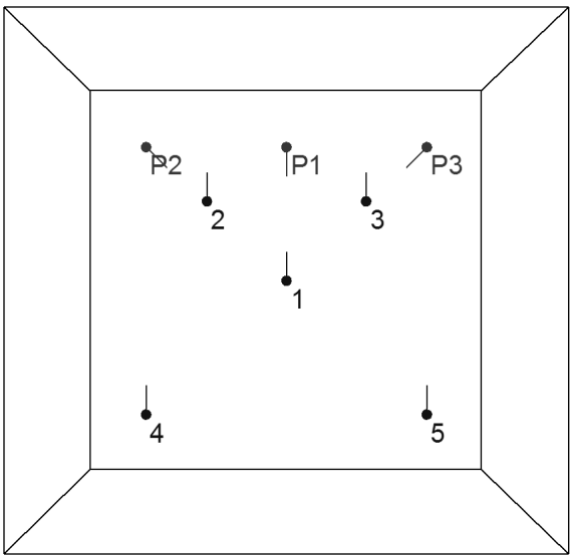

Figure 2. Positions of sound sources and receivers in a virtual room modeled with the acoustic simulation program “Odeon”. P1-P3 are speakers positions and points 1-5 are listeners positions.

2 Methods

In this study, we tested a novel algorithm that can produce high-quality real-time auralization with low computational resources. In our previous work, we have already explained the operation of the algorithm we developed when there is one sound source in a virtual room. This time we tested an algorithm where two audio sources are playing at the same time. The algorithm works rather commonly in the beginning – RIR filters that best suit the position of the speakers

and the listener in the room and the HRIR filter that matches the head position are selected, and the anechoic signal is processed. Two additional versions of the processed signal are prepared at the same time, which may be needed first when the user makes any movement. One version uses a HRIR filter that corresponds to the position of the head rotated by three degrees to the left, and the other version uses a HRIR filter that matches the head rotated by three degrees to the right. When the user's head position actually changes by more than three degrees, one of the ready-made versions of the processed signal is transmitted, and the algorithm calculates the new version that will be required when the user turns his head in the same direction for an additional three degrees. This idea of a three-degree head position change is thoroughly discussed in A. Lindau's work and defined as the minimum resolution required for the HRIR database [7]. The author has shown that human hearing cannot detect a change in head position of less than three degrees.

Before modeling a virtual room and developing an algorithm, we first chose a real room and made audio recordings there. The experiment was carried out in a university classroom (reverberant, almost empty, rectangular room) with a volume of about 115 m^3 (Fig. 1). We placed three sound sources in fixed positions (Genelec 8010 monitor speakers). We performed measurements in five different listener positions, playing the files at two different sound pressure levels (60dB and 80dB SPL). Measurements were made on two different stereo microphone systems – ORTF (two RODE NT5 small diaphragm condenser microphones) and Binaural (Sennheiser Ambeo Smart Headset). The second tool mentioned above is a headset with omni directional microphones on the outside. By inserting headphones into a person's ears and making a recording this way, we get a dummy head effect. The recordings were made by playing different audio files through the speakers – male voice, female voice, classical and electronic music. In the same experiment, we also captured room impulse responses that correspond to any possible combination between the three sound sources and the five positions of the listener. We used the previously discussed stereo microphone systems ORTF and Binaural, as well as an additional omni directional microphone "Sonarworks XREF 20", specialized for acoustic measurements with an almost linear frequency response. To create room impulses we used Matlab's "Impulse Response Measurer" tool and played exponential swept sine wave through different sound sources in the room. The program performed deconvolution and saved impulses to .wav files.

In order to check the quality of the RIR captured in a real room, we modeled a virtual room of the same dimensions and materials using the acoustic modeling program "Odeon". We selected the same positions of the sound sources and the listeners (Fig. 2), performed calculations and compared the results generated by "Odeon" and our algorithm (Fig. 3).

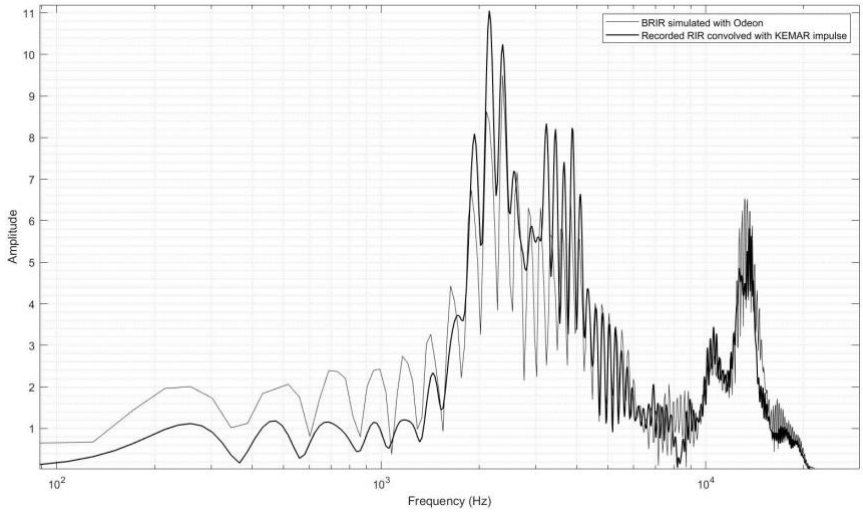

Figure 3. Comparison of our recorded RIR spectrum (after convolution with MIT-KEMAR HRTF) and the simulated impulse spectrum of the acoustic simulation program "Odeon".

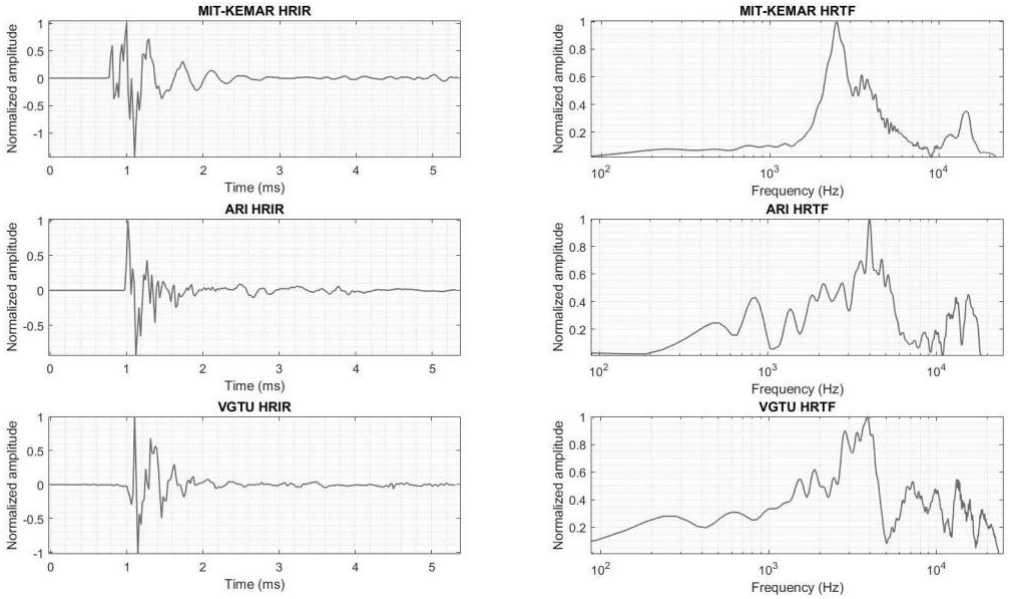

Figure 4. Impulses from three different HRTF databases with the sound source in front of the binaural head (0° azimuth and 0° elevation). Impulse length - 256 samples.

Both programs used the same HRTF database. The tests were performed first with MIT-KEMAR HRTF [8], then with Matlab default ARI HRTF, as well as the HRTF database produced in our own university anechoic room. This one was created using the same Sennheiser binaural headphones and in this work we call it VGTU HRTF database (Vilnius Gediminas Technical University).

We compared the performance of the algorithm when it uses different HRTF databases. We also compared the results when the HRTF impulse is truncated to 256 (Fig.4) or 128 samples (Fig.5). For this initial phase of the study, to simplify the experiment, we studied a head rotation in the azimuth (horizontal axis) only and left a fixed head elevation. Thus, the study would later be expanded to include more than 3 degrees change in head position on the vertical axis.

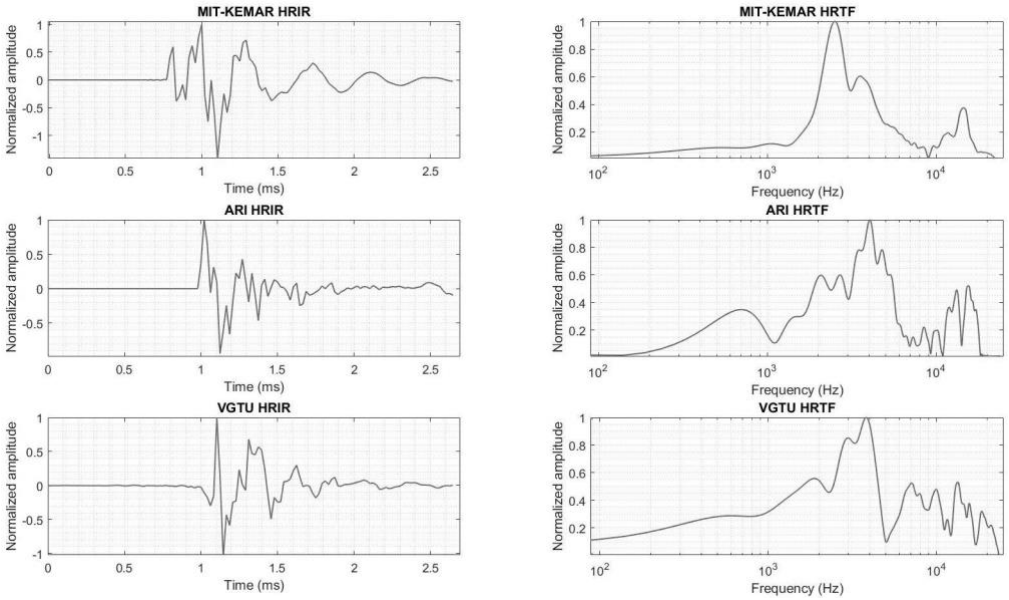

Figure 5. Impulses from three different HRTF databases with the sound source in front of the binaural head (0° azimuth and 0° elevation). Impulse length - 128 samples.

3 Conclusions

The essence of our proposed method is to always produce three parallel versions of a binaural signal, which we can switch between as soon as the user's head is turned more than three degrees to either side. The calculation of the new version required takes place within the delay limits. We measured the speed of a very fast human head rotation and found that a change of three degrees could probably occur only after 12 ms. In order to reduce the computational time of our algorithm, the length of the HRIR filter is shortened to 128 samples. In this way, a new version of the processed signal can be computed within 8 ms.

References

[1] S. Serafin, M. Geronazzo, C. Erkut, N. Nilsson and R. Nordahl, "Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions", *IEEE Computer Graphics and Applications,* vol. 38, no. 2, pp. 31-43 (2018).

[2] G. Koutsouris, J. Brunskog, C. Jeong, and F. Jacobsen, "A combination of the acoustic radiosity and the image source method", *J. Acoust. Soc. Am.* vol. 133 no. 6, pp. 3963–3974 (2013).

[3] L. Simon, N. Zacharov and B. Katz, "Perceptual attributes for the comparison of head-related transfer functions", *J. Acoust. Soc. Am.* vol. 140, no. 5, pp. 3623-3632 (2016).

[4] J. Ahrens, and C. Andersson, "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre", *J. Acoust. Soc. Am.* vol. 145, no. 4, pp. 2783-2794 (2019).

[5] A. Lindau, V. Erbes, S. Lepa, H. Maempel, F. Brinkman, and S. Weinzierl, "A spatial audio quality inventory (SAQI)," *Acta Acust. united with Acust.,* vol. 100, no. 5, pp. 984–994 (2014).

[6] A. Lindau, "The Perception of System Latency in Dynamic Binaural Synthesis" *Fortschritte der Akust. Tagungsband der 35. DAGA*, no. 1, pp. 1063–1066 (2009).

[7] A. Lindau, H. Maempel and S. Weinzierl, "Minimum BRIR grid resolution for dynamic binaural synthesis", *Acoustics'08 Paris*, pp. 3851-3856 (2008).

[8] W. G. Gardner and K. D. Martin, "HRTF Measurements of a KEMAR" *J. Acoust. Soc. Am.*, vol. 97, pp. 3907-3908. (1995).