Exposé

Master Thesis in Computer Science

An iOS Implementation of Motion-Tracked Binaural Rendering

Author
Aki Saksala

Supervisors
Prof. Dr. Stefan Weinzierl
David Ackermann

Berlin, 2 March 2021
Abstract

The motion-tracked binaural method enables the recording of arbitrarily complex sound scenes and their subsequent immersive, pseudo-binaural reproduction on headphones. As one of the applications, this technique serves as a potential alternative for current ways of recording and livestreaming orchestra performances, providing a sound image that accurately reproduces the concert hall environment. The planned thesis explores the implementation of the MTB renderer and a corresponding motion tracking system on consumer-available hardware.

Contents

1 Introduction 2
2 State of the art 2
3 Methods 5
4 Time schedule 6
1 Introduction

The motion-tracked binaural (MTB) technique [1] provides a method for the recording of a sound environment and its immersive reproduction on headphones. Utilizing an array of microphones arranged on a sphere and replicating the positioning of the listener’s ears, it preserves spatial characteristics such as interaural time differences and interaural level differences, which are vital cues for the auditory system to determine the direction of sound sources. Another important factor that greatly increases the realism of the reproduced scene is tracking the listener’s head movements and simulating the independence of sound source locations from the orientation of the head. This improves both the ability to locate sounds as well as the perceived externalization of the scene (the impression that sound sources are placed further away around the head, rather than inside the head).

Livestreams and recordings of classical music performances, such as the ones offered in the Digital Concert Hall of the Berlin Philharmonic Orchestra [2], are one area where this technique could be employed. While such online media oftentimes comprises audio content of decidedly high quality, it is typically recorded and reproduced in a manner that resembles traditional record productions and does not attempt to present an exact replica of the concert hall environment. Possibilities of creating an immersive reproduction of the concert hall that accurately preserves spatial sound characteristics are not yet widely being leveraged. Regarding both livestreams and recordings of concerts, a sound image that convincingly mimics listening to the performance as an audience member in the hall could contribute to a lively and engaging experience. The accurate directional perception and head tracking would be especially well-suited for orchestral pieces in which instruments play offstage or behind the audience. Similarly, circumstantial sounds in the concert hall that usually play a lesser role in a recording context, such as applause or incidental noise before the beginning of the performance, also benefit from a pseudo-binaural rendering and could further augment the impression of being situated in the concert hall.

The planned thesis explores the application-focused realization of the MTB approach. The goal is the implementation of the MTB renderer and a corresponding motion tracking system that can be deployed on consumer hardware without the need to set up any custom equipment.

2 State of the art

In the MTB method, sound environments are recorded using a solid spherical microphone with the diameter matching an average human head. The equator of the sphere is equipped with an array of sixteen equidistant electret condenser capsules placed flush with the surface. These capture sounds representative of discrete ear positions, depending on the horizontal rotation of the head. During reproduction, the MTB rendering software is supplied with the signals of all capsules and simultaneously receives the current orientation of the listener’s
Figure 1: MTB microphone: a 3D model (a), the interior view (b) and the assembled microphone from the side with visible microphone capsules (c). Figure adapted from [3].

head from a motion tracking device. The signals of the capsules are then interpolated in order to transform the discrete spatial samples of the sound field into a continuous space. The resulting pseudo-binaural sound reflects changes in horizontal rotation of the head as well as arbitrary movements of sound sources in the environment. Figure 1 shows the structure of an MTB microphone as it was constructed at the department.

The microphone positioning in traditional binaural dummy head recordings is decided at the time of recording and cannot be retroactively changed. These recording methods do not offer the possibility of reacting to the listener's head orientation at the time of playback and ensuring that the locations of sound sources are perceived to be independent of head movements. In methods that do provide this form of head tracking, the perceived positional invariance increases the accuracy of localizing sound sources [4], reduces problems of front-back ambiguity due to small interaural time and level differences [5] and substantially enhances the listener's perception of externalization of sounds [6]. The MTB method benefits from comprising a fairly simple concept and thus poses a good alternative to other methods of binaural synthesis that allow for the incorporation of head movements. Dataset-based methods commonly referred to as dynamic binaural synthesis require not only the acquisition and storage of a large number of binaural room impulse responses (one for each combination of head orientation and sound source position) but also knowledge about the location of all sound sources and a reverberation-free recording—conditions that cannot be feasibly established in a concert hall environment. Techniques involv-
Figure 2: Overview of the signal transmission in the MTB method. The raw audio channels recorded on the left can be streamed live or stored for later playback. The movement tracking and signal interpolation is performed at the location of the recipient. Figure from [7].

ing the binaural encoding, transmission and recomposition of the sound field, such as dynamic binaural recording, require a spherical microphone array with a significantly higher number of capsules to produce a sufficient sound quality [7].

The sound image produced by the MTB approach has been shown to provide a good externalization, localizability and plausibility of sounds as well as a high degree of naturalness of the scene, resulting in only minor tone colorations and degradation in spatial characteristics, when compared directly to a dynamic binaural synthesis [3]. It was also shown that only using a subset of eight of the sixteen microphone channels resulted in no significant decrease in reproduction quality. With only eight audio channels needing to be transmitted, the required bandwidth for an online broadcast is comparable to a traditional surround sound transmission. If all movement-related rendering can be executed on the client hardware, all playback devices can receive the same source material, which makes the system well-suited for livestreaming applications and also enables storing the raw recording to be played back at a later time, without the need to store any additional data. Figure 2 illustrates the transmission of the signal for livestreams as well as stored recordings. Since it is important to retain control over the sound in a music recording context, the possibility to integrate additional spot microphones as a way to influence the presence and balance of specific instruments is currently being explored at the department.

Headphones have, slowly over time, been receiving increased support for motion tracking features. Bose Corporation launched a beta version of their Bose AR platform [8] in 2019, together with an SDK that enabled app developers
to process positional information from compatible headphone models sold by the company. The project has, however, been discontinued in June 2020 [9, 10]. In September 2020, Apple Inc. has released a head-tracked spatial audio feature for compatible headphones as part of the iOS 14 operating system release. The two currently supported models are the AirPods Pro headphones that have already been available since October 2019 and the AirPods Max headphones that were launched in December 2020 [11]. While the spatial audio feature is intended to place virtual external speakers around the listener’s head by processing 5.1-channel, 7.1-channel or Dolby Atmos content, a headphone motion API was also added to the platform’s Core Motion framework as part of the release [12]. This API provides access to raw movement data from the compatible headphone models, which would eliminate the need for an external head tracking device in the way it was used in the initial implementation.

3 Methods

The MTB rendering system will be implemented on the iOS platform, chosen for its prevalence and the expected continued support for the headphone motion framework mentioned in section 2. It will be developed natively in the Swift programming language [13] and will support headphones that are compatible with the framework. If possible, all movement information will be obtained from the integrated motion sensors, thus replacing the external motion tracking device that was previously used. A modular structure of software components will be aimed for, to facilitate the modification and substitution of parts. After completion of the project, this software architecture should, for instance, enable the integration of the renderer into existing applications as well as the replacement of the MTB renderer with a renderer using a different technique. The architecture and source code will be clearly documented to allow for use in future projects.

The developed application is intended to serve as a functioning prototype that can be compared to the setup used for quality evaluation [3]. It will be assessed if the rendering performance on smartphones and tablets is sufficient, especially with regards to the input latency and the reliability of tracking the head orientation. Additional testing with the existing MTB microphone in one of the concert halls of the Berliner Philharmonie is also a possibility, involving the recording of a concert and an informal sound evaluation.
4 Time schedule

The following table contains an overview of the required tasks and the approximate time period in which they will be completed.

<table>
<thead>
<tr>
<th>Task</th>
<th>Timeframe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literature research, project planning</td>
<td>February-March</td>
</tr>
<tr>
<td>Study of current renderer implementation</td>
<td>March-June</td>
</tr>
<tr>
<td>Draft of software requirements</td>
<td>March-April</td>
</tr>
<tr>
<td>Draft of software architecture</td>
<td>April-May</td>
</tr>
<tr>
<td>Software implementation</td>
<td>May-July</td>
</tr>
<tr>
<td>Testing</td>
<td>June-July</td>
</tr>
<tr>
<td>Performance evaluation</td>
<td>July-August</td>
</tr>
<tr>
<td>Writing the thesis</td>
<td>August-September</td>
</tr>
</tbody>
</table>

References


