IMMERSIVE AUDIO-GUIDING

Nuno Carriço
DETI
University of Aveiro
Portugal
carrico@ua.pt

Guilherme Campos
DETI / IEETA
University of Aveiro
Portugal
guilherme.campos@ua.pt

José Vieira
DETI / Institute of Telecommunications
University of Aveiro
Portugal
jnvieira@ua.pt

ABSTRACT

An audio-guide prototype was developed which makes it possible to associate virtual sound sources to tourist route focal points. An augmented reality effect is created, as the (virtual) audio content presented through headphones seems to originate from the specified (real) points.

A route management application allows specification of source positions (GPS coordinates), audio content (monophonic files) and route points where playback should be triggered.

The binaural spatialisation effects depend on user pose relative to the focal points: position is detected by a GPS receiver; for head-tracking, an IMU is attached to the headphone strap.

The main application, developed in C++, streams the audio content through a real-time auralisation engine. HRTF filters are selected according to the azimuth and elevation of the path from the virtual source, continuously updated based on user pose.

Preliminary tests carried out with ten subjects confirmed the ability to provide the desired audio spatialisation effects and identified position detection accuracy as the main aspect to be improved in the future.

1. PROJECT MOTIVATION

Tourism and its economic impact have been growing markedly in recent decades [1][2]. The importance of enriching the visitor experience, promoting cultural tourism and adopting differentiation strategies are widely acknowledged [3], as well as the key role played in those efforts by digital information and communication technologies (ICT) [4][5].

Audio guides are increasingly popular in tourism applications (e.g. in museums, parks, historic sites and cities), both indoors and outdoors. A variety of systems are commercially available. Some are intended as aids to improve intelligibility by avoiding noise and interference (especially important in heritage sites under intense visitor pressure) in otherwise conventional guided tours [6][7][9]. Others are designed to operate autonomously (i.e. without live human guiding), delivering pre-recorded (often multilingual) interpretation content [6][7][8][10][11][12][13]. The diagram in Figure 1 covers both cases. Autonomous systems can be triggered manually by the user [10] or automatically based on route sensing (GPS, infra-red and radio-frequency ID sensors being among the most common).

For example, ‘hop-on hop-off’ urban tour buses, now commonplace even in middle-sized cities, are invariably equipped with audio-guiding systems.

Figure 1: Typical audio guiding system architecture

Typically, operation is autonomous, with pre-recorded audio contents triggered at certain positions detected by GPS along the bus route; visitors are given a pair of disposable headphones (relatively ‘low-fi’ and uncomfortable) to be plugged into audio terminal units placed by each seat, as illustrated in Figure 2.

Figure 2: Bus audio guide unit with language selection

This project aims at radically improving the visitor experience provided by this kind of systems, making it as immersive as possible. The idea is to create binaural audio augmented/mixed reality (AR/MR) effects by using geo-location and applying auralisation and source spatialisation techniques. While not new, these techniques have been explored mainly in the context of computer games. As these are increasingly geared towards mobile devices, AR and MR gain ground over VR (see, for example, [14]) and geo-location becomes an essential feature. Geo-located spatial audio systems have been proposed for various applications, including artistic soundscaping (e.g. the SoundDelta system [15]) and guidance systems for the visually impaired (e.g. the NAVIG system [16]). The applicability to tour guiding...
DAFx-258

is obvious [13][17]. However, to the best of the authors’ knowledge, there are no widespread commercial audio-guide models based on geo-location and incorporating spatialisation capabilities. The USOMO system [18] features binaural spatialisation, but is restricted to indoor usage.

2. SYSTEM OVERVIEW

A prototype was developed to address outdoor situations, taking the urban bus tour example mentioned above as the reference scenario. The goal is to turn focal points specified along the route (e.g. buildings, statues, trees…) into virtual sound sources, so that the interpretation content, delivered through headphones, be perceived by the visitor as originating from those focal points. This requires pre-recording appropriate content for each focal point, and processing this audio content in real time through filters capable of imprinting appropriate 3D directional cues according to listener pose (position and head orientation) relative to the corresponding source. Playback should be triggered when the vehicle enters route segments specified in the vicinity of the virtual source locations, as illustrated in Figure 3.

![Figure 3: Virtual audio source (S) locations and corresponding trigger point (TP) regions along a route](image)

Figure 4 represents the overall structure designed to achieve this goal. Its core element (playback block) relies on an auralisation engine, as the binaural spatialisation effect is obtained by convolving the anechoic input sound with head-related transfer function (HRTF) filter pairs (to generate left and right channel output). The filter pair applied at a given moment must be selected (from an HRTF database) according to the azimuth and elevation of the virtual source relative to the listener. For real-time operation, this information (and thus the HRTF filter pair) must be continuously updated based on:

- Listener position – given by a GPS receiver (GPS block);
- Listener head orientation – detected by an inertial head-tracking device attached to the headphone strap (IMU block);
- Source position – specified at the route definition stage (route manager block).

The following sections describe the implementation (based on C++ programming) and integration of these four blocks on a Windows environment.

3. PLAYBACK

3.1. Audio streaming and auralisation

The playback system was implemented with the help of the PortAudio [19] open-source library. As shown in Figure 5, it takes its input (44100Hz recordings of the virtual sound sources) from local memory files in 16-bit raw audio format and streams it through a real-time auralisation engine to generate output for binaural (i.e. headphone or earphone) presentation.

![Figure 5: Audio streaming through auralisation engine](image)
LibAAVE incorporates room acoustic modelling based on the mirror-image source (MIS) method. From the input data on 3D room configuration, primary source positions and listener location, the acoustic model works out the propagation paths reaching the listener considering wall reflections up to a user-defined order (this must be set low enough to allow real-time operation). The direction (azimuth and elevation) of each path relative to the listener head is also calculated using the input information on head orientation (pitch, yaw and roll angles).

The audio processing block can then determine the appropriate delay, attenuation and HRTF filtering to be applied to each audio component transmitted through each path and generate the resulting binaural output by adding together all those contributions. Different HRTF sets can be selected, taken from public-domain databases, namely the KEMAR-based MIT Media-Lab set [23] and CIPIC [24]. The system allows arbitrary movement of both sources and listener. Cross-fading between successive audio output blocks is applied to avoid audible HRTF transition glitches.

Only a fraction of LibAAVE’s capabilities are utilised in the outdoor scenario explored here, as it does not involve a room model – the engine is configured to process only direct sound (no reflections). Also, a single primary source is considered at a time. Under these conditions, real-time operation is comfortably achievable. In a future extension to indoor scenarios, LibAAVE could be configured to take into account the acoustic influence of the room – albeit through a simplified model – without compromising real-time operation.

### 3.2. Playback control

Two playback trigger modes were defined. In both, audio tracks, once triggered, are played through without interruption, regardless of listener position. However, while in mode 1 tracks can be played only once along a route (i.e. are never re-triggered), in mode 2 they will be replayed if the listener re-enters the respective trigger region.

A program thread is constantly checking the current listener position, received from the GPS block, against the route information to detect if the listener has entered the trigger region of a playable virtual source. In that case, streaming is activated; each time the playback thread extracts an audio block from the output circular buffer, the auralisation engine processes a new one to refill it.

The number of samples per audio block and the size of the output buffer are configurable. To minimise latency, it is desirable to keep them as low as possible.

### 4. POSITION DETECTION (GPS)

The Global Positioning System (GPS) block is responsible for tracking listener position (amounting to bus position in the reference scenario) and continuously feeding the playback block with updated values of latitude and longitude – GPS measured altitude is not taken into account in this application. The chosen GPS receiver was a XUCAI GD75 USB dongle – see Figure 7. Its main characteristics are listed in Table 1. Data is sent from the GPS dongle to the laptop in the ASCII format using RS232 emulation.

![Figure 7: GPS receiver for position detection](image)

### 5. HEAD-TRACKING (IMU)

For a given source position, sound perception depends not only on listener position but also on head orientation. This is normally specified by three rotation components:

- **Yaw**: around the vertical axis;
- **Pitch**: around the lateral (left-right) axis;
- **Roll**: around the longitudinal (back-front) axis.

If a virtual sound scene is to be recreated over headphones, head movements must be tracked and compensated for in real time. It is therefore necessary to use a head-tracking device capable of providing real-time pitch, yaw and roll angle data to the playback block. An inertial measurement unit (IMU) attached to the headphone strap is possibly the most appropriate choice for this purpose. An Intersense InertiaCube3 unit was employed – see Figure 8. Its main characteristics are listed in Table 2.
Table 2: IMU features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface</td>
<td>USB</td>
</tr>
<tr>
<td>Latency</td>
<td>4 ms (via USB)</td>
</tr>
<tr>
<td>Maximum refresh rate</td>
<td>180Hz</td>
</tr>
<tr>
<td>Degrees of freedom</td>
<td>3 axis (Yaw, Pitch, and Roll)</td>
</tr>
<tr>
<td>Angular range</td>
<td>360º (all axis)</td>
</tr>
<tr>
<td>Precision</td>
<td>Yaw: 1º; Pitch and Roll: 0.25º</td>
</tr>
<tr>
<td></td>
<td>(at the temperature of 25º C)</td>
</tr>
<tr>
<td>Maximum angular speed</td>
<td>1200 º per second</td>
</tr>
</tbody>
</table>

A software development kit is available to assist programmers using this device and provide examples regarding its operation, configuration and data acquisition.

To ensure correct integration, the IMU sensor was also tested with the help of a simple C++ application which displayed the received yaw, pitch, and roll values.

6. ROUTE MANAGER

A practical means of defining and configuring tourist routes is indispensable for efficient system operation. An application – whose user interface is presented in Figure 9 – was developed for this purpose. It allows the specification of a set of virtual sources, individually characterised in terms of (area 2 of Figure 9):

- Location (latitude and longitude);
- Height relative to a listener at the trigger region;
- Trigger region: centre point location (latitude and longitude) and radius;
- Corresponding anechoic audio file name.

This information is stored in a ‘route file’ (area 3 of Figure 9) under a very simple format (one text line per source) which is then passed to the playback block.

The latitude and longitude coordinates for the source and the trigger region centre can be entered manually (area 1 of Figure 9) but, as illustrated in Figure 4, there is also the option of acquiring them in-situ with the help of the GPS receiver.

7. VALIDATION

7.1. Test design and preparation

In order to obtain a preliminary assessment of system operation, a set of subjective tests was prepared on a short walking route with three virtual sound sources defined within the campus of the University of Aveiro, as depicted in Figure 10. Source locations are designated by ‘S’; their corresponding trigger regions (interior of the dashed circles, centred at points TP) are shown to scale. Table 3 lists the audio files used (44.1kHz, 16-bit mono speech recordings regarding the chosen campus locations). Audio streaming (recall Figure 5 and section 3.2) was set for 1024-sample blocks and a 5-block output buffer. This choice of settings had seemed to ensure smooth audio playback and avoid any noticeable latency effects.
The definition and configuration of this test route was itself an opportunity to validate an important part of the system – the route manager application described in the previous section.

A walking route was preferred to a driving route (the system’s reference usage scenario) because it simplified the logistics of the tests, seemingly without compromising their quality. In fact, as they involve shorter distances and less predictable user trajectories, walking routes would appear much more demanding in terms of position detection accuracy and precision.

Ten randomly chosen subjects (6 males and 4 females in the 20-35 age range, with no reported hearing problems) were invited to walk the route wearing the system. Figure 11 presents the equipment carried by the test subjects:

1. Head-tracker (Intersense InertiaCube 3).
2. GPS receiver (XUCAI GD75 USB dongle).
3. Headphones (Sony MDR-ZX110).
4. Processing unit (laptop).

The subjects were briefed on the purposes and design of the tests and informed on the characteristics of the route: chosen source focal points (see Table 3), radius specified for each trigger region (respectively 10, 12 and 15m) and respective centre point locations.

7.2. Test execution and results

The first set of tests were carried out using trigger mode 1 (no re-triggering – recall trigger modes described in 3.2). The subjects were asked to use a three-point discrete scale [from 1 (bad) to 3 (good)] to rate the experience regarding triggering (Q1: ‘does sound start at a seemingly correct distance?’) and spatialisation (Q2: ‘does sound appear to originate from the correct direction?’). The assessment – see Table 4 – was clearly positive in both regards for S2 and S3 and also positive for S1 regarding Q1, with no ‘bad’ ratings from any subject. However, the spatial effect of S1 was rated quite poorly; none of the subjects rated it ‘good’ and the majority considered it ‘bad’.

Table 4: User ratings (first test set)

<table>
<thead>
<tr>
<th>Source</th>
<th>Audio file</th>
<th>Duration (s)</th>
<th>Visual Cue</th>
<th>Height (m)</th>
<th>Q1 – triggering</th>
<th>Q2 – spatialisation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>Std. Dev.</td>
</tr>
<tr>
<td>S1</td>
<td>Welcome_speech.wav</td>
<td>38</td>
<td>Stone</td>
<td>0</td>
<td>2.4</td>
<td>0.52</td>
</tr>
<tr>
<td>S2</td>
<td>Library.wav</td>
<td>45</td>
<td>Corner</td>
<td>10</td>
<td>2.7</td>
<td>0.48</td>
</tr>
<tr>
<td>S3</td>
<td>Media_Centre.wav</td>
<td>19</td>
<td>Window</td>
<td>5</td>
<td>2.6</td>
<td>0.52</td>
</tr>
</tbody>
</table>

These bad results for S1 are not surprising, since shorter distances between source and listener are expected to amplify the ill effects of imprecise position detection. Unlike sources S2 and S3, placed well outside their respective trigger regions (see Figure 10), S1 was deliberately located at the centre of its trigger region to expose this effect. The spatialisation effect was very noticeably disrupted by the instability of GPS position readings (error up to 5m – see Table 1), causing abrupt changes in perceived source location. As expected, results for Q2 in S1 improved (from 1.4 to 2.4) when the subjects were asked to stop and make their assessment as soon as playback started (i.e. at the edge of the trigger region).

Under trigger mode 2, additional tests were conducted with the listeners asked to stand still for one minute inside the TP2 circle (15m radius) after the end of playback of source S2 in two situations: 1) more than 5m away from the trigger region limit and 2) less than 5m away from the trigger region limit. Obviously, playback re-triggering is not supposed to occur in either of them. However, it did in the second, again highlighting GPS position measurement errors. In this instance, they cause the listener to be occasionally detected outside the trigger region and subsequent position readings inside it are of course interpreted as a re-entry. In the first situation, re-triggering was never observed.

8. DISCUSSION AND FUTURE WORK

Whilst confirming the ability to provide the desired audio spatialisation effects, the preliminary tests identified lack of precision in GPS position detection as the main problem affecting the user experience. Although the impact of this problem may be significantly mitigated in the reference scenario (tour bus) for reasons pointed out in the previous section (higher predictability, larger distance to virtual source locations), solving it is essential for system versatility.
Simply applying moving-average filtering to the GPS output is not appropriate, as it would improve precision at the expense of responsiveness. Exploring sensor fusion techniques to combine IMU and GPS data is the most promising approach.

Work is under way to port the applications supporting the various blocks (playback, GPS, route manager…) to Android, as the system structure can be made simpler, lighter and more versatile by concentrating all the communication and processing functions on a smartphone or tablet – see Figure 12. As the figure suggests, operation would be completely autonomous, audio content being downloaded from the Internet according to the chosen route.

For this reason, establishing R&D partnerships with tour operators is also among the envisaged future work threads. The development of a demonstration route with excellent content design is key in this effort.

9. REFERENCES


DAFx-263