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 in- 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.

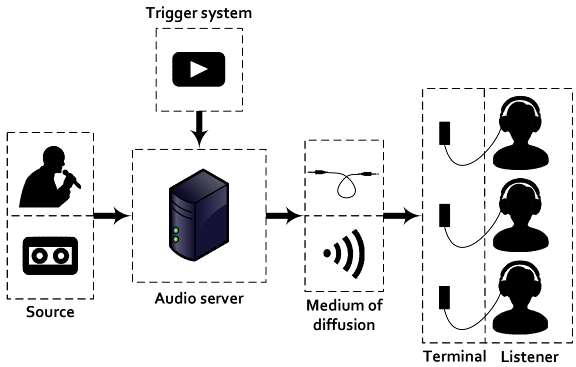

Figure 1: *Typical audio guiding system architecture*

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 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.

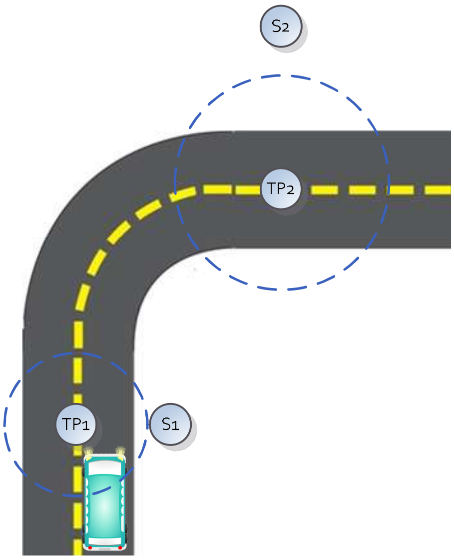

Figure 3: *Virtual audio source (S) locations and corresponding trigger point (TP) regions along a route*

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).

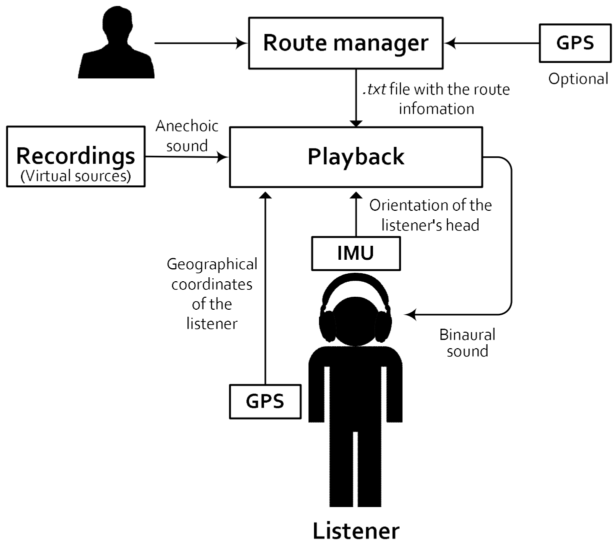

Figure 4: *System block diagram*

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.

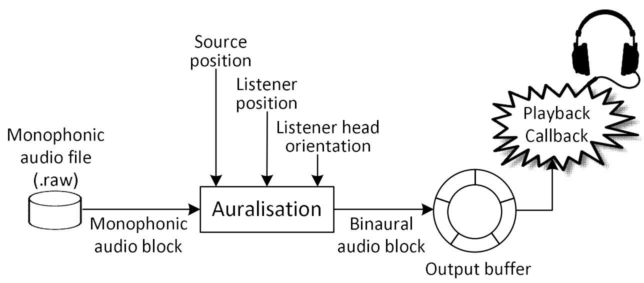

Figure 5: *Audio streaming through auralisation engine*

The auralisation engine was implemented using *LibAAVE*, a publicly available auralisation library [20] developed in a previous IEETA research project [21]. Its basic operation principles, described in [22], are illustrated in Figure 6.

.

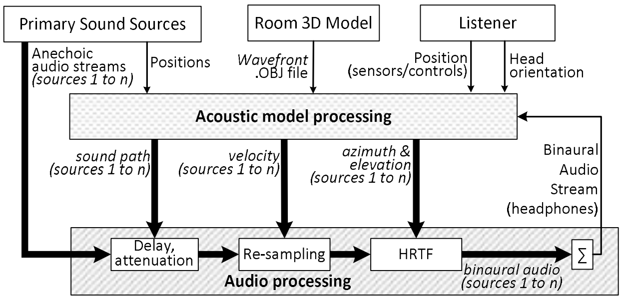

Figure 6: *LibAAVE operation structure* [21]

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 position, 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 considering 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 the 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 explore 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 achieved. 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 ASCII format using RS232 emulation.

Figure 7: *GPS receiver for position detection*

Table 1: *GPS receiver features*

Interface	USB
Communication protocol	NMEA 0183 (V3.0)
Maximum refresh rate	1Hz
Cold start time	<33s
Operating Temperature	-10º C a 70º C
Maximum error	5m (approx.)

To ensure correct integration, a simple C++ application was developed to test the device by displaying the received GPS position data. In addition to latitude, longitude and altitude, the application also extracted the number of satellites used by the receiver, since it is available from the same $GPGGA frames, constitutes an indicator of position measurement accuracy and may prove useful in scenarios to be explored in the future (e.g. transition to indoor situations).

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.

Figure 8: *IMU for head-tracking*

Table 2: *IMU features*

Interface	USB
Latency	4 ms (via USB)
Maximum refresh rate	180Hz
Degrees of freedom	3 axis (Yaw, Pitch, and Roll)
Angular range	360º (all axis)
Precision	Yaw: 1º; Pitch and Roll: 0.25º (at the temperature of 25º C)
Maximum angular speed	1200 º per second

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.

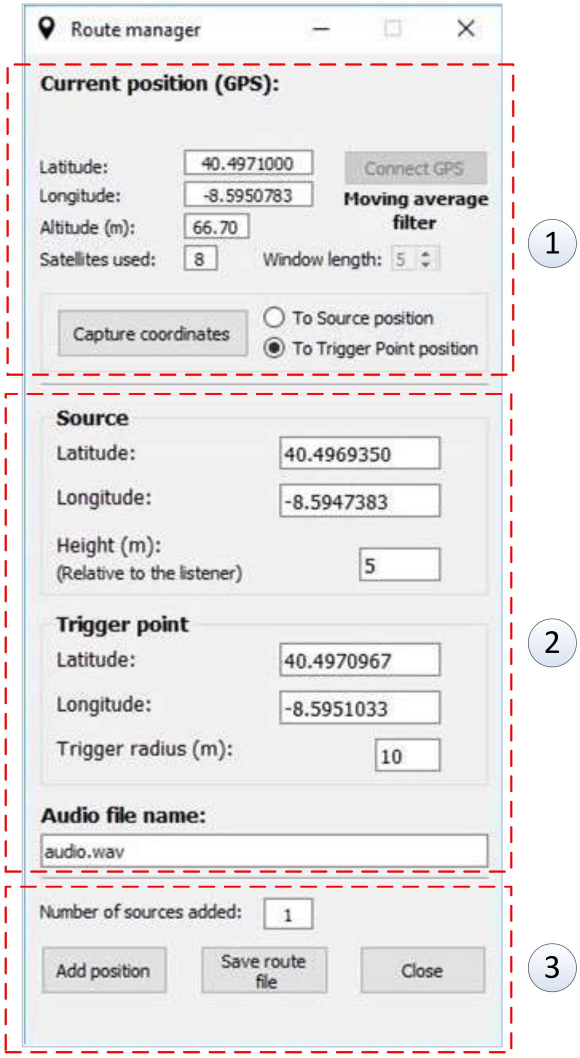

Figure 9: *Graphical user interface of the route manager*

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.

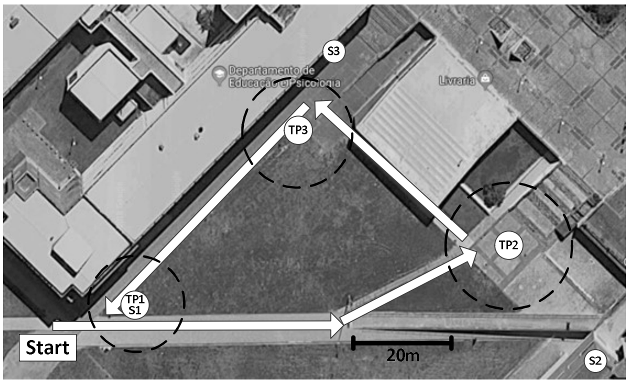

Figure 10: *Walking route for preliminary tests*

Table 3: Test route sources

Source	Audio file	Duration (s)	Visual Cue	Height (m)
S1	*Welcome_speech.wav*	38	Stone	0
S2	*Library.wav*	45	Corner	10
S3	*Media_Centre.wav*	19	Window	5

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).

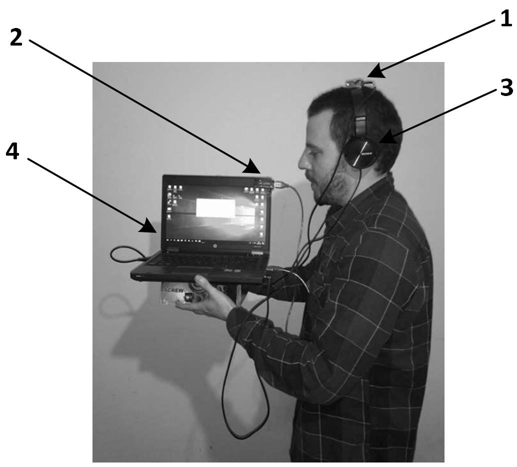

Figure 11: *Test equipment*

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)

	Q1 – triggering		Q2 – spatialisation	
	Mean	**Std. Dev.**	**Mean**	**Std. Dev.**
S1	2.4	0.52	1.4	0.52
S2	2.7	0.48	2.9	0.32
S3	2.6	0.52	2.7	0.48

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.

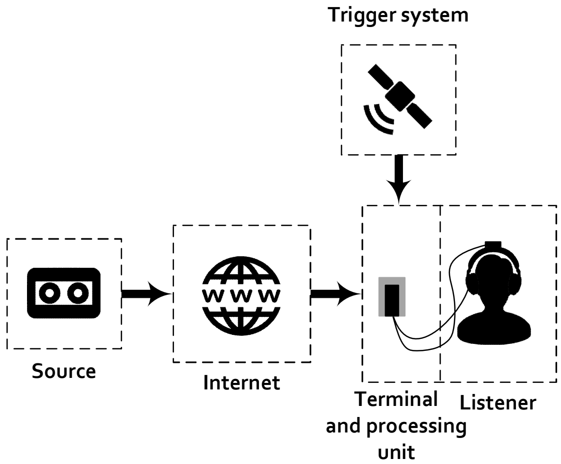

Figure 12: *Envisaged audio guiding system architecture*

Obviously, the IMU for head-tracking cannot be incorporated, as it must be attached to the headphone. The integration of a cost-effective, miniaturised head-tracking device, preferably with wireless connectivity, is another important future work front.

A wide variety of usage scenarios can be envisaged for a smartphone-based system. In the reference scenario, the triggering signal could be provided by a system installed on the bus, and hence constitute an added value of the ride. For indoor operation, position detection could no longer rely on GPS; alternative methods (e.g. based on radio-frequency ID tags, wi-fi sensors or ultrasonic beacons) would be required. By fully exploring *LibAAVE*'s capabilities, mentioned in section 3.1 (see Figure 6), the acoustic influence of the room could be modelled (early reflections and reverberation tail).

The perceived added value of the proposed audio AR effects in audio guides will no doubt be strongly influenced by other factors, namely:

• Quality of sound delivery – perfect-fit ear-enclosing high fidelity headphones are a must; comfortable, low-vibration bus seats would be desirable. Active noise cancellation systems may also prove indispensable.

• Content design – the added 3D audio dimension must be appropriately explored (e.g. for historic soundscape reconstruction). This requires expert story-telling based on appealing information and interpretation data brought to life by professional sound design/recording/editing.

To ensure the commercial success of audio guides incorporating the AR effects proposed here, these factors must be addressed simultaneously, which may impact on business models, possibly creating new (premium) market niche opportunities.

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

[1] WTTC "Travel & Tourism Economic Impact," London, 2018. Avail. at https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2017/world2017.pdf, Accessed April, 16, 2018

[2] UNWTO "Tourism Highlights," Madrid, 2017. Available at https://www.e-unwto.org/doi/pdf/10.18111/9789284419029, Accessed April, 16, 2018

[3] UNWTO "Tourism and Culture Synergies," Madrid, 2018. Available at https://doi.org/10.18111/9789284418978, Accessed April, 16, 2018.

[4] D. Buhalis and Z. Yovcheva, "The digital tourism Think Tank report: 10 best practices in tourism". Available at https://thinkdigital.travel/wp-content/uploads/2013/04/10-AR-Best-Practices-in-Tourism.pdf, Accessed April, 16, 2018.

[5] D. Han, T. Jung and A. Gibson, "Information and Communication Technologies in Tourism," chapter Dublin AR: Implementing Augmented Reality in Tourism, Z. Xiang and I. Tussyadiah Eds. Springer, Cham, 2014.

[6] Tamo GPS Multilingual Commentary System. Available at http://www.tamotec.com/Product//8259674739.html, Accessed April, 15, 2018.

[7] toGuide TriggerPoint Wireless. Available at http://www.toguide.pt/pt/hardware/hardware_show/scripts/core.htm?p=hardware&f=hardware_show&lang=pt&idcont=123, Accessed April, 15, 2018.

[8] Soolai Bus Audio Guide System. Available at https://soolai.en.made-in-china.com/product/LCOndajPETVW/China-Bus-Audio-Guide-System.html, Accessed April, 15, 2018.

[9] Takstar WTG-500 Tour Guide System. Available at http://www.takstar.com/en/product/detail-13-38-0-400, Accessed April, 15, 2018.

[10] Mix Tech Polska, ATGS02. Available at http://www.mixtechpolska.pl/en/tour-guide-system-ATGS02.htm, Accessed April, 15, 2018.

[11] Mix Tech Polska, ATGS03. Available at http://www.mixtechpolska.pl/en/tour-guide-system-ATGS03.htm, Accessed April, 15, 2018.

[12] Acoustiguide. Available at http://www.acoustiguide.com/smartphone-applications, Accessed April, 15, 2018.

[13] ECHOES – Geolocated experiences. Available at https://echoes.xyz/ Accessed July, 3, 2018.

[14] N. Paterson, K. Naliuka, S. Jensen, T. Carrigy, H. Haahr and F. Conway, "Design, implementation and evaluation of audio for a location based augmented reality game," in *Proceedings of ACM Fun and Games*, Leuven, Belgium, 15–17 Sept. 2010.

[15] N. Mariette and B. Katz, "SoundDelta - Largescale, multi-user audio augmented reality," in *EAA Symp. on Auralization*, Espoo, Finland, 2009, pp. 1–6.

[16] B. Katz, S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, M. Auvray, P. Truillet, M. Denis, S. Thorpe and C. Jouffrais, "NAVIG: augmented reality guidance system for the visually impaired," *Virtual Reality*, vol. 16, no. 4, pp. 253–269, 2012.

[17] N. Paterson, G. Kearney, K. Naliuka, T. Carrigy, H. Haahr, and F. Conway, "Viking ghost hunt: creating engaging sound design for location–aware applications," *Int. Journal of Arts and Technology*, 6(1), pp. 61-82, Jan 2013.

[18] Usomo – The Immersive Sound system. Available at http://usomo.de/en/, Accessed April, 15, 2018.

[19] PortAudio: Portable Cross-Platform Audio I/O. Available at *http://www.portaudio.com*, Accessed April, 16, 2018.

[20] AcousticAVE: Auralisation Models and Applications in Virtual Reality Environments.
Available at https://code.ua.pt/projects/acousticave,
Accessed April 16, 2018

[21] G. Campos, P. Dias, J. Vieira, J. Santos, C. Mendonça, J. P. Lamas, N. Silva and S. Lopes, "AcousticAVE: Auralisation Models and Applications in Virtual Reality Environments," in *Proc. 8th Iberian Congress of Acoustics (Tecniacústica)*, Murcia, Spain, Oct. 29-31, 2014.

[22] A. Oliveira, G. Campos, P. Dias, D. Murphy, J. Vieira, C. Mendonça and J. Santos, "Real-Time Dynamic Image-Source Implementation for Auralisation," in *Proc. Digital Audio Effects (DAFx'13)*, Maynooth, Ireland, Sept. 2013, pp. 368-372.

[23] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone," MIT MediaLab, 2000. http://sound.media.mit.edu/resources/KEMAR.html, Accessed April, 16, 2018.

[24] V. Algazi, R. Duda and D. M. Thompson, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, Oct. 2001, pp. 99-102.

[25] J. Jacoby and M. S. Matell, "Three point Likert scales are good enough," Journal of Marketing Research, 8(4), pp. 495-500, 1971.

[26] J. Moutinho, D. Freitas and R. E. Araújo, "Indoor Sound Based Localization: Research Questions and First Results," in L. M. Camarinha-Matos, S. Tomic and P. Graça (eds) *Technological Innovation for the Internet of Things*. DoCEIS 2013. IFIP Advances in Information and Communication Technology, vol 394. Springer, Berlin, Heidelberg.