Application of binaural audio techniques for immersive fruition of cultural heritage

A Lapini[1], G Calamai[2], F Argenti[2] and M Carfagni[1]

[1] Department of Industrial Engineering, University of Florence, via S. Marta 3, Florence 50139, IT

[2] Department of Information Engineering, University of Florence, via S. Marta 3, Florence 50139, IT

E-mail: {alessandro.lapini@, giulio.calamai@stud., fabrizio.argenti@, monica.carfagni@}unifi.it

Abstract. In this paper we address the issue of enriching the fruition of museums and art shows for the visitors. We preliminary consider the design of an immersive audio environment by means of 3D audio rendering, aiming to provide each user a deeper connection with each exhibited artwork. By exploiting a real-time binaural audio system, artworks become virtual sources of the audio guide voice, making the user perceive they personally speak to her/him. As an advantage, the implementation of the proposed framework relies on basic audio and signal processing techniques, that is, no expensive or personalized equipment for the visitors is required. Furthermore, since the proposed method is independent of the recording stage, existing audio tracks or even real-time speech can be used.

1. Introduction

Immersive technologies are nowadays spreading across several fields and disciplines, improving the human perception in augmented reality or, more generally, in mixed reality environments. Remarkable engineering studies and applications in this broad context involve (but are not limited to) biomedicine, manufacturing processes, entertainment, gaming, military operations and also cultural heritage.

The efforts of technological progress in computer science have been historically focused firstly on improvements of the user' visual experience. Incidentally, this is nowadays witnessed by the wide spreading of stereoscopic displays [1, 2, 3, 4]. Nevertheless, human perception is particularly influenced by sound, too. Stereophonic audio (which has widespread much earlier than stereo vision) and the subsequent surround techniques are still far from providing a realistic experience. In the last years, the hardware and software improvements, especially in consumer equipment, have paved the way to introduce immersive audio technologies [5]. Spatial rendering techniques that allow the human auditory system to perceive the spatial localization of virtual sound sources are currently crossing the line between theory and practical implementations. Among them, binaural synthesis aims to control the sound directly in the human ears [6, 7].

The interest towards the application of binaural techniques to cultural heritage has grown both in institutions and private companies, as the EU financed project BINCI [8] demonstrates. The aim of this work is to describe a signal processing method for fast real–time binaural synthesis, whose main target application is, among others, the fruition of cultural heritage. The practical implementation of the system is beyond the scope of this work. The proposed method relies on the same logical approach proposed in [9], but, to the best of authors' knowledge, our framework presents some novelties.

2. Application scenario

A description of the application scenario based on binaural techniques is provided according to the following test case, whose sketch is reported in Figure 1.

We consider a visitor of a museum approaching to a specific artwork. We assume that he/she is provided with an audio–guide device and headphones or earphones; the audio–guide device must support the following capabilities:

(i) real–time processing of a monaural audio track that describes the artwork;
(ii) reproduction of a stereo signal through the headphones/earphones that are plugged into;
(iii) indoor positioning and head orienting system, briefly indicated as *positioning system*.

The visitor is required to select the audio track pertaining to the targeted artwork and the reproduction starts. A suitable binaural processing of the monaural audio track is then applied and the output signal is reproduced in the visitor's ears, in order to make him perceive the audio coming from the artwork itself. The visitor is allowed to freely walk around the artwork, getting closer or moving away, while the audio source accordingly moves with respect to him. This application can be straightforwardly generalized to two or more sources that can be located on the same artwork or even on different ones, obtaining, for instance, the effect of multiple voices incoming from different directions. Thanks to the earphones and to individual audio–guide, more visitors are allowed to target the artwork at the same time independently from each other.

A schematic view of the proposed system for the case of a single virtual source is depicted in Figure 2. For sake of generality, the visitor and the artwork are indicated as listener and the (virtual) source, respectively. At a given time instant n, the positioning system computes the source–listener reciprocal position and the head orientation of the listener with respect to the source, being the latter fixed or, in general, known a–priori. Research on indoor positioning systems is currently ongoing and the reader interested in indoor tracking might find useful to see [10, 11]. The information about position and orientation is then exploited to query and retrieve the corresponding couple of binaural filters, which are stored in a specific database. Such filters are applied to the monaural audio track in order to determine the left and right binaural audio signals until the time instant $n + M$, where M is the number of sampling instants between each position update. Finally, the signals are transmitted to the headphones/earphones of the listener. The algorithm then iterates to the following position update until the end of the audio track.

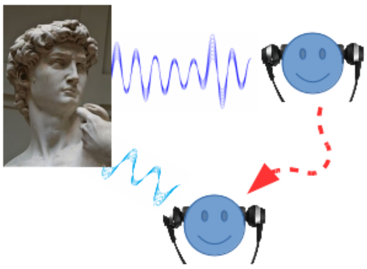

Figure 1. Sketch of the a possible application scenario for the proposed method.

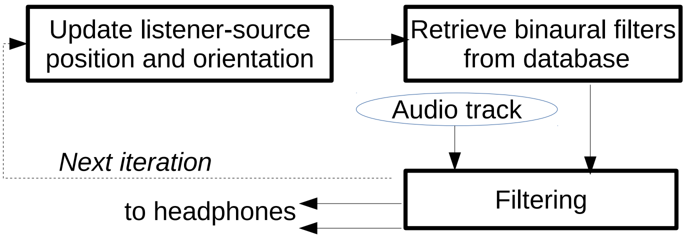
Update listener-source position and orientation
Retrieve binaural filters from database
Audio track
Next iteration
Filtering
to headphones

Figure 2. The proposed system for a real-time binaural processing considering a moving listener and a source located in a known position.

3. Time–varying binaural synthesis

We start this Section briefly introducing the modeling for human sound localization; the proposed method is subsequently described.

3.1. Characterization of auditory system

The modeling of human auditory system is a long term goal of several disciplines, being audio engineering among them. Nowadays it is known that human perception of sound spatiality is based on the decoding of the *binaural directional cues*, that is, the differences of sound between the ear channels. Time and level differences of the sound traveling into the listener's ears are due not only to the reciprocal position of the source with respect to the listener' head, but also to the interaction of the acoustic field with the listener's torso, head, and pinnae. These effects can be roughly thought as the incoming acoustic signal undergoing a *body filtering* stage before being captured by human auditory system.

The concept of *head–related impulse response* (HRIR) is commonly adopted to describe the relation between the source signal, namely x, and the signals in the listener's ears, namely y^L and y^R. We indicate the source–listener distance and the orientation of listener head with respect to the line passing trough the source and the listener with symbols r and $\boldsymbol{\phi}$, respectively. Thus, the output signals at the discrete time instants $n \geq 0$ are given by

$$y^{L,R}[n] = \sum_{m=0}^{\min(N-1,n)} h^{L,R}_{r,\boldsymbol{\phi}}[m]x[n-m] = \sum_{m=\max(n-N+1,0)}^{n} h^{L,R}_{r,\boldsymbol{\phi}}[n-m]x[m] \tag{1}$$

where $h^L_{r,\boldsymbol{\phi}}$ (respectively, $h^R_{r,\boldsymbol{\phi}}$) is the left (right) ear HRIR, and N is its length in terms of number of samples. Equation (1) represents the input–output relation of a linear time-independent (LTI) system, where the HRIR is the impulse response of the associated LTI filter.

Given a known test signal x, HRIR can be estimated by recording the received signals y^L and y^R by means of ear microphones plugged in real human ears or in realistic mannequin heads. By moving the source and fixing the listener (or vice versa), HRIRs corresponding at different azimuths, elevations and distances are gathered. Incidentally, a number of research paper linked to free HRIR databases exists, e.g., [12, 13, 14].

Binaural synthesis techniques, instead, aim to control directly the sound in the ear canals to match a recorded real case or with a simulated virtual case, by using measured or modeled HRIRs [15]. Our scenario belongs to the latter case. Specifically, given a monaural signal x and a couple of filters $(h^L_{r,\boldsymbol{\phi}}, h^R_{r,\boldsymbol{\phi}})$, a virtual sound source, positioned at distance r and oriented at $-\boldsymbol{\phi}$ with respect to the listener's head, is synthesized; the output signals y^L and y^R can be eventually reproduced by headphones/earphones.

3.2. Time–varying binaural modeling

Unfortunately, the HRIR defined in (1) is spatial dependent but not time dependent, that is, it is valid for constant r and $\boldsymbol{\phi}$ across time n. Hence, it cannot be directly applied for a moving source or listener, as it would be required in our scenario. By means of a simple example, we initially show that extending (1) to a time–varying scenario is not trivial nor unambiguous; then we propose a solution in a constructive manner.

Let's assume that the source is fixed and the listener moves. Without loss of generality, we consider that the position updating occurs at each n (i.e., $M = 1$) and and we indicate the time–varying r and $\boldsymbol{\phi}$ with r_n and $\boldsymbol{\phi}_n$, respectively. The output signal at $n = 0$ is then

$$y[0] = h_0[0]x[0],$$

where, for sake of brevity, the superscripts L, R have been omitted and the symbol h_0 is a short version of h_{r_0,ϕ_0}. At instant $n = 1$ one may think to extend (1) as

$$y[1] = h_1[0]x[1] + h_1[1]x[0]$$

or more generally

$$y[n] = \sum_{m=0}^{\min(N-1,n)} h_n[m]x[n-m] = \sum_{m=\max(n-N+1,0)}^{n} h_n[n-m]x[m]. \tag{2}$$

According to this model, at time n, there is no difference between the signal received by the moving listener and the signal that would be received by a fixed listener at $(r_n, \boldsymbol{\phi}_n)$. Since $x[n_0]$ is somehow associated to an incident acoustic wavefront, this approach is equivalent to state that the acoustic scattering contribution of the human body at n_0 vanishes for $n > n_0$. We believe that the physical validity of such model is questionable and should be preliminarily validated.

In this paper we propose an alternative model to (2). The received signal is given instead by

$$y[n] = \sum_{m=0}^{\min(N-1,n)} h_{n-m}[m]x[n-m] = \sum_{m=\max(n-N+1,0)}^{n} h_m[n-m]x[m]. \tag{3}$$

Because $x[n]$ is weighted by h_n, in this case each incoming wavefront is affected by a time–invariant scattering contribution of the human body, that is, the listener is fixed with respect to each incoming wavefront. Accordingly, this model is equivalent to the case of a moving source and a fixed listener. It is our opinion that this assumption is more valid with respect to (2) and entails an approximation error that is acceptable for a listener moving much slower than the speed of sound in the air.

3.3. Linear real-time interpolation of HRIR

In a real scenario the updating rate of HRIRs is usually slower than to the sampling rate, that is, $M \gg 1$. This is mainly due by the time required by the positioning system update and the database query/retrieval process. By defining

$$p(n) \triangleq M \lfloor n/M \rfloor , \tag{4}$$

where $\lfloor \cdot \rfloor$ is the floor operator, (3) becomes

$$y[n] = \sum_{m=0}^{\min(N-1,n)} h_{p(n-m)}[m]x[n-m] = \sum_{m=\max(n-N+1,0)}^{n} h_{p(m)}[n-m]x[m]. \tag{5}$$

For $M = 1$, (5) suitably coincides with (3). Practical implementation of (5) is quite simple;however, it can introduce audio impairments due to abrupt head orientation changes whenever the listener sufficiently moves during the position updating period M.

We overcome this problem by considering the following modified version of (3):

$$y[n] = \sum_{m=0}^{\min(N-1,n)} \hat{h}_{n-m}[m]x[n-m] = \sum_{m=\max(n-N+1,0)}^{n} \hat{h}_m[n-m]x[m], \tag{6}$$

being $\hat{h}_m$ the *time–varying* HRIR defined as linear interpolation of the retrieved HRIRs, that is,

$$\hat{h}_m[n] \triangleq h_{p(m-M)}[n] + \frac{m - p(m)}{M}\left(h_{p(m)}[n] - h_{p(m-M)}[n]\right) \tag{7}$$

and assuming $h_{-M} = h_0$. According to (7), the computation of interpolated HRIRs conveniently demands to retrieve only one HRIR every M samples. It has to be noted that $\hat{h}_{p(m)} = h_{p(m-M)}$, i.e., the interpolated HRIR at index $m = p(m)$ is equal to the penultimate retrieved HRIR. This choice preserves causality at the expense of a time delay of M samples in the position of the virtual source. However, according to our tests, it has a lower perceptual impact than the reproduction of audio glitches.

3.4. Fast real–time block processing

Let's indicate $n = kM + l$, for $k \in \mathbb{N}_0$ and $0 \le l < M$, and define $x_k[l] \triangleq x[kM + l]$ as the k-th input block. According to the linearity and the causality of (6), the contribution of $x_k[l]$ to the output signal y is given by

$$y_k[n] \triangleq \begin{cases} \sum_{m=\max(n-M+1,0)}^{\min(N-1,n)} \hat{h}_{n-m+kM}[m]x_k[n-m], & \text{for } 0 \le n < M+N, \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

so that

$$y[n] = y[kM + l] = \sum_{p=0}^{k-1} y_p[l + (k-p)M] + y_k[l]. \tag{9}$$

The summation in (9) accounts for the contributions of input blocks prior to k, while the last term represents the update contribution of the k-th block. Substituting (7) in (8) and considering only the non–trivial case ($0 \le n < M + N$) yields

$$y_k[n] = \sum_{m=\max(n-M+1,0)}^{\min(N-1,n)} \left[h_{(k-1)M}[m] + \frac{n-m}{M} \left(h_{kM}[m] - h_{(k-1)M}[m] \right) \right] x_k[n-m] \tag{10}$$

where, since $0 \le n - m < M$, the identity $p(n - m + kM) = kM$ has been used. By defining the following quantities,

$$g_{kM}[n] \triangleq \frac{h_{kM}[n] - h_{(k-1)M}[n]}{M} \quad \text{and} \quad g'_{kM}[n] \triangleq n g_k[n], \tag{11}$$

and dropping the extrema of summations for sake of brevity, (10) is rewritten as

$$y_k[n] = \sum_m \left(h_{(k-1)M}[m] - g'_{kM}[m] \right) x_k[n-m] + n \sum_m g_{kM}[m] x_k[n-m]. \tag{12}$$

Both the summations in (12) are linear convolutions and can be potentially computed by means of Fast Fourier Transform (FFT) techniques to reduce the computational costs [16]. Since the monaural signal x_k is known a–priori[1] of and only the HRIRs at the time instants $(k-1)M$ and kM are required, all the quantities in (12) are available at time instant kM. Hence, the proposed method can take advantage by processing a block of M input samples every M time instants, without introducing further processing delay. The procedure is reported in Algorithm 1.

4. Simulation results

The proposed method has been tested by means of computer simulations that aim to reproduce the conditions of the real application scenario.

[1] In the case of real–time audio track, it sufficient to introduce a delay of M samples in x.

Algorithm 1 Fast real–time binaural synthesis. Steps (*) must be repeated for each source.

```
Require: $x$, $M \in \mathbb{N}$
  set $k = 0$ and $K = \lfloor (\text{length}(x) - 1) / M \rfloor$
  set $y^{L,R}[n] = 0$, for $0 \le n \le \text{length}(x) - 1$
  repeat
    extract $x_k$ from $x$ (*)
    acquire $r_k$ and $\boldsymbol{\phi}_k$ by positioning system
    for all channels $L, R$ do
      retrieve $h_k$ in the HRIRs database
      compute $g_{kM}$ and $g'_{kM}$ using eq.(11)
      compute $y_k[n]$, for $0 \le n < M - N$, using eq.(12) (*)
      assign $y[n] \leftarrow y[n] + y_k[n]$, for $0 \le n < M - N$ (*)
    end for
    assign $k \leftarrow k + 1$
  until $k > K$
```

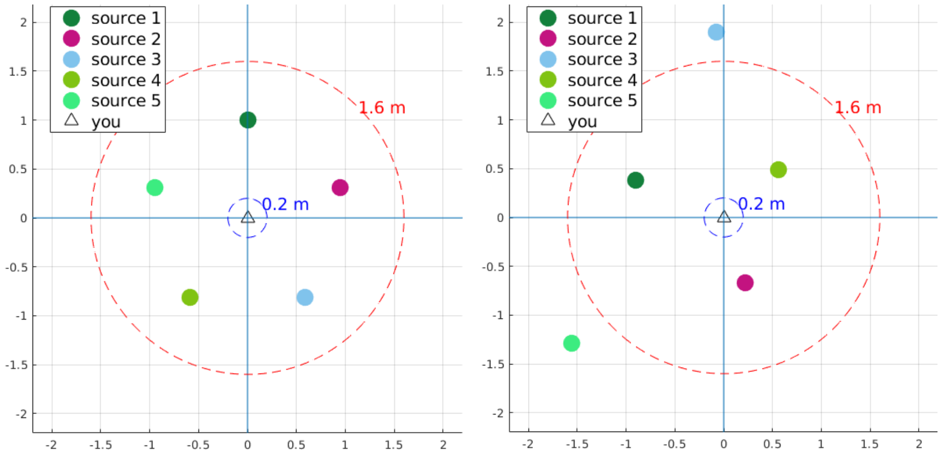

Figure 3. Two snapshots of the MATLAB GUI developed to test the binaural synthesis.

4.1. Setup

The real–time binaural synthesis has been implemented in MATLAB according to Algorithm 1 considering one or more source. After setting the sampling rate and position update rate, the audio track(s) is(are) preliminary loaded in the program, along with the HRIRs database. In our tests we have adopted the PKU-IOA HRTF database [13], which approximately covers elevation angles from -40 o to 90°, azimuth angles from 0 to 360°, and distances from 20 cm to 160 cm. For distances beyond 160 cm, HRIRs related to 160 cm are retrieved and a suitable extra delay is introduced. Since azimuth, elevation and distances are sampled in their respective ranges, the retrieval process requires an interpolation; incidentally, we have used a tetrahedral interpolation method described in [17] that provides an acceptable performances tradeoff in our case.

A 2D graphical user interface (GUI) has been developed to simulate the reciprocal movement on a plane between source(s) and listener, as shown in the snapshots reported in Figure 3. For sake of simplicity, the listener, which is the GUI user, is supposed to be fixed with respect to coordinate system. Source movements are implemented by means of mouse drag–and–drop and are real–time acquired by the GUI in order to simulate the positioning system.

4.2. Performance measurement

The proposed method has been tested on an system GNU/Linux 64–bit, kernel Linux4.4.0–112–generic SMP, mounting one Intel(R) Core(TM) i5–4200U CPU @ 1.60GHz. In the case of a single virtual source, the real–time system successfully works at a sampling rate up to $f_s = 65536$ Hz. For multiple virtual sources the system is capable to synthesize up to 5 source at $f_s = 32768$ Hz. As to the memory occupation, excluding the sources' audio tracks, it is mainly dictated by the HRIRs database, whose size scales approximately with f_s and it is about 26 MB at $f_s = 32768$ if stored in single precision (32–bit).

The audio quality has been objectively evaluated by considering the synthesis of a single tone at frequency f_0 emitted by a source moving with period $T = 8$ s on a circular trajectory around the listener. The performances have been evaluated in terms of Signal-to-noise and distortion ratio (SINAD) by means of the homologous function included in the MATLAB Signal Processing Toolbox. The SINAD is defined as

$$SINAD = (P_s + P_N + P_d)/(P_N + P_D),$$

being P_s, P_N and P_D the powers of the tone, of the background noise and of the distortion introduced by the the proposed method, in that order. The results obtained setting $f_s = 16384$ Hz for different f_0 (columns) and different position updating frequency f_{pos} (rows) are reported in Table 1. The *fixed* row refers to the SINAD for a fixed source, which is affected only by finite arithmetic noise. Interestingly, as a general rule, SINAD values increase from low to high frequency tones. Considering each single tone, we observe SINAD values for lower tones attain their best with lower values of f_{pos}, while for higher tones the best performance is for intermediate f_{pos}.

Table 1. Signal-to-noise and distortion ratio (dB) measured for the implemented method.

	$f_0 = 62.5$ Hz	125 Hz	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz
fixed	137.2	137.1	136.5	136.1	136.2	136.5	136.4
$f_{pos} = 4$ Hz	15.4	17.1	18.1	21.9	31.8	28.0	24.5
8 Hz	10.1	12.0	14.4	17.1	26.2	36.1	34.9
16 Hz	8.8	11.4	13.8	16.8	25.4	34.8	34.8
32 Hz	8.2	10.6	13.2	16.1	24.6	33.6	33.5
64 Hz	8.0	10.4	13.0	16.1	24.4	33.3	33.0

5. Conclusions

In this paper, a real–time binaural method for an immersive fruition of the cultural heritage has been proposed. The realistic scenario of a visitor moving inside a museum hall has been considered as the case study. The proposed approach successfully exploits a database of statically recorder head–related impulse response (HRIR), as well as the position and orientation of the visitor, to synthesize one ore more virtual audio sources reciprocally moving with respect to a him/her. By using the real–time linear interpolation of consecutive HRIRs, a formulation of a fast block processing algorithm has been derived. The experimental results have shown that the system achieves a good audio quality at a computational cost that is affordable by consumer devices. Further optimization of the processing stage, possibly targeted to specific portable hardware, as well as the synthesis of environmental effects, e.g. room reverberations, would be worthwhile for future works. Furthermore, a psychoacoustic evaluation by means of a selected auditory panel should be carried out in order to detect and mitigate eventual undesirable sound artifacts.

6. References

[1] IJsselsteijn W A, de Ridder H and Vliegen J 2000 *IEEE Transactions on Circuits and Systems for Video Technology* **10** 225–233 ISSN 1051-8215

[2] Lin Y H and Wu J L 2014 *IEEE Transactions on Image Processing* **23** 1527–1542 ISSN 1057-7149

[3] Shao F, Lin W, Gu S, Jiang G and Srikanthan T 2013 *IEEE Transactions on Image Processing* **22** 1940–1953 ISSN 1057-7149

[4] Tam W J, Speranza F, Yano S, Shimono K and Ono H 2011 *IEEE Transactions on Broadcasting* **57** 335–346 ISSN 0018-9316

[5] Hacihabiboglu H, Sena E D, Cvetkovic Z, Johnston J and III J O S 2017 *IEEE Signal Processing Magazine* **34** 36–54 ISSN 1053-5888

[6] Baumgarte F and Faller C 2003 *IEEE Transactions on Speech and Audio Processing* **11** 509–519 ISSN 1063-6676

[7] Brown C P and Duda R O 1998 *IEEE Transactions on Speech and Audio Processing* **6** 476–488 ISSN 1063-6676

[8] URL `http://www.binci.eu`

[9] Ahnert W, Feistel S, Lentz T, Moldrzyk C and Weinzierl S 2004 *Audio Engineering Society Convention 117* URL `http://www.aes.org/e-lib/browse.cfm?elib=12932`

[10] Liu H, Darabi H, Banerjee P and Liu J 2007 *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **37** 1067–1080 ISSN 1094-6977

[11] Martinelli A, Gao H, Groves P D and Morosi S 2018 *IEEE Sensors Journal* **18** 1600–1611 ISSN 1530-437X

[12] Algazi V R, Duda R O, Thompson D M and Avendano C 2001 *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)* pp 99–102

[13] Qu T, Xiao Z, Gong M, Huang Y, Li X and Wu X 2009 *IEEE Transactions on Audio, Speech, and Language Processing* **17** 1124–1132 ISSN 1558-7916

[14] BolaÃśos J G and Pulkki V 2012 *Audio Engineering Society Convention 133* URL `http://www.aes.org/e-lib/browse.cfm?elib=16501`

[15] Zölzer U, Amatriain X, Arfib D, Bonada J, De Poli G, Dutilleux P, Evangelista G, Keiler F, Loscos A, Rocchesso D *et al.* 2002 *DAFX - Digital Audio Effects* (John Wiley & Sons) ISBN 9780471490784 URL `https://books.google.it/books?id=h90HIV0uwVsC`

[16] Proakis J G and Manolakis D K 2006 *Digital Signal Processing (4th Edition)* (Upper Saddle River, NJ, USA: Prentice-Hall, Inc.) ISBN 0131873741

[17] Hugeng H, Anggara J and Gunawan D 2017 *2017 International Conference on Signals and Systems (ICSigSys)* pp 35–39