Binaural Spatialization: Comparing Head Related Transfer Function Models for Use in Virtual and Augmented Reality Applications

Simone Angelucci, Fabio Franchi, Fabio Graziosi, and Claudia Rinaldi(✉)

Department of Information Engineering, Computer Science and Mathematics (DISIM), Università degli Studi dell'Aquila, via Vetoio 1, 67100 L'Aquila, Italy
simone.angelucci@student.univaq.it,
{fabio.franchi,fabio.graziosi,claudia.rinaldi}@univaq.it
https://www.univaq.it/

Abstract. This work is focused on the binaural spatialization, presenting an analysis of the most common solutions to understand and classify their advantages and drawbacks and to find the one that results in a better virtual/augmented experience on the basis of subjective tests. This work is a preliminary step toward the implementation of an augmented reality system for cultural heritage enjoyment exploiting spatial audio through low-cost devices. Two different models are implemented in order to avoid the use of non individualized HRTFs and results show promising opportunities that must be further exploited.

Keywords: Audio Signal Processing · Binaural spatialization · Head Related Transfer Function · Augmented Reality (AR) · Virtual Reality (VR)

1 Introduction

The recent widespreading of multimedia applications for learning and entertainment, even pushed by the recent 5G paradigm, has also to take into account the proper treating of audio signals. In this context many research topics are involved as efficient coding solutions, innovative effects development, artificial reverberation advancements, audio synthesis improvements and audio spatialization issues. The latter aspect is particularly important while conceiving with different kind of Augmented and Virtual Reality (AR/VR) applications, [8,13].

This paper focuses on the exploitation of audio spatialization for improving AR/VR experiences for Cultural Heritage (CH). Indeed, the progress of digital information has significantly affected the evolution of CH dissemination [23], offering new technological possibilities for developing, e.g., the market of tourist

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 601–613, 2022.
https://doi.org/10.1007/978-3-031-18123-8_47

services [18] and CH organisations have to address new users needs by creating innovative applications [1], i.e. AR/VR based. Indeed, AR technology gives a different perception of reality, as it enriches reality with a computer-generated layer containing visual, audio, and tactile information while using a "virtual" representation of a classic museum allows access to aspects of the artwork that may otherwise be hidden [19].

However, in order to build a complete and complex cultural representation via digital heritage technologies, developers must also understand how users interact with the system or interface [29]. When designing AR applications, it is important to choose the best combination of techniques for presenting the appropriate digital information to the users. During the past years, the main aim of AR/VR applications for CH changed from a mere virtual recreation of object to display, to the development of an entire virtual environment able to disseminate and teach culture. The idea is the opposite of a "dead museum": users don't have to see an accumulation of 3D heritage objects, but feel and understand another culture through those items. An important aspect is the relation between AR/VR and education. This new way to present Culture enhances the learning process, encouraging students and researchers through stimulating methods of presentation of archival materials and historical events. Users can therefore travel through space and time without moving from their home [27]. Numerous AR/VR applications exist for CH or tourists enjoyment of places with a rich past, allowing a realistic navigation of environments that no longer exist or that may be inaccessible, [2,6,11,14,25].

The most of the previously cited experiences do not take into account the advantages that may arise by a proper exploitation of sounds together with visual effects, except for audio content presentation purposes. On the contrary, an acoustic guide, properly placed in the virtual space, may drive the user toward a certain direction or the acoustic landscape of a specific historical period could be reproduced to improve the virtual experience. To the authors' knowledge only a few experimentations have been carried on in this context. One of the few exceptions is given by the work in [17], where authors present a signal processing method for fast real-time binaural synthesis, whose main target application is the fruition of cultural heritage, and the work in [8] where a smart headphones set is presented that remotely takes the orientation of the listener's head and properly generates an audio output to attract the tourists'attention toward specific points of interest in the 3D space. An interesting analysis of hardware and software requirements for this purpose is presented in [20] without references to real applications.

In this paper we present a preliminary study on the possible models for binaural spatialization, already available in literature, that could be exploited for AR/VR applications for CH, taking into account the kind of devices involved in this context which do not offer a high definition audio experience. We are aware of existing solutions offered by Google (i.e. Resonance Audio, [10]) or the Oculus Unity Spatializer, [21] by Facebook and the Steam Audio Unity Plugin [28], but in this initial part of our research we decided to build our own solution in order

to precisely control all the involved parameters given the specific application and devices we are targeting. The paper is organized as in the following. The analyzed methods are summarized in Sect. 2, Sect. 3 reports the implementation of the previously described models, subsequently tests and results are discussed in Sect. 4, while conclusions are drawn in Sect. 5.

2 Analyzed Models

This section briefly describes two models that have been analyzed and tested to achieve binaural spatialization for AR/VR applications. The two models have been chosen due to their different characteristics to achieve the spatialization. The first one, i.e. the anthropometric model, is able to achieve the spatialization by adapting some parameters directly to the listener, so it is able to take into account the differences between human parts. The second one, the minimum phase representation model which is basically an approximation of the HRTFs, is useful to investigate which characteristics of the HRTFs are most relevant to locate a sound source in the space.

2.1 Anthropometric Models

During the previous years researches have focused on how human body parts can affect the acoustic experience and which of them are the most influencing ones. Models coming from these studies are called *anthropometric models*. They are based on several geometric approximations, e.g. a sphere for the head, and they also exploit particular manikins, known as KEMAR (Knowles Electronic Manikin for Acoustic Research), because they allow the use of in-ear microphones to measure the Head Related Transfer Functions (HRTF), i.e. the representation of the ways in which the positions of the head, trunk, and ears filter sounds, altering the way they are perceived. In this paper we mainly focus on the analysis and implementation of models related to the head and the pinna, namely the outer part of the ear.

It is well known that the variation in magnitude of the audio signal due to the head can be represented by a low-pass filter, as mentioned by Lord Rayleigh in [26].

One of the parameter that the human auditory system (HAS) uses to locate sounds in the space is the Interaural Time Difference (ITD). This parameter is given by the difference of the arrival times of a sound wave to the left and the right ear. By simply manipulating this parameter, it is possible to place a sound source in a virtual environment.

However, such parameter allows to locate sound sources only on a horizontal plane.

The ITD is included in the developed anthropometric model by using an all-pass filter with group delay given by the Woodworth-Schlosberg formula:

$$\tau_h(\theta) = \begin{cases} -\frac{a}{c}\cos\theta & \text{if} \quad 0 \leq |\theta| < \pi/2 \\ \frac{a}{c}\left(|\theta| - \pi/2\right) & \text{if} \quad \pi/2 \leq |\theta| < \pi \end{cases} \tag{1}$$

where θ represents the azimuth coordinate, a the head radius and c the speed of sound.

Considering human ears as two "observation points", the formula has to be used once for the right ear and once for the left one in order to take into account the Interaural Time Difference (ITD).

As previously said, the ITD allows sound sources positioning on a horizontal plane. It has been shown that the influence of the pinnas affects the positioning of sound sources on a vertical plane.

The pinna is mainly responsible for multiple reflections of the incoming wave [31] and, according to the procedure described in [4], the pinna effects can be modeled with five reflections by means of a tapped delay line.

2.2 Minimum Phase Representation

Several studies have focused on relevant characteristics of the HRTF for the localization task, [16,30].

In this paper we refer to the work in [16], regarding the sensitivity of humans to the variations of the phase spectra of the HRTFs where HRTFs were approximated by a minimum phase transfer function including also a factor capable of incorporating the ITD.

This approximation is justified from the fact that every system can be decomposed in a product of a minimum-phase system and an all-pass system [22]:

$$H(e^{j\omega}) = H_{min}(e^{j\omega})H_{ap}(e^{j\omega}) \tag{2}$$

The HRTF can thus be expressed as follows:

$$H(e^{j\omega}) = |H_{min}(e^{j\omega})|e^{j\phi(\omega)}e^{-j\tau} \tag{3}$$

where $|H_{min}(e^{j\omega})|$ is the magnitude of the HRTF, $e^{j\phi(\omega)}$ represents the phase response of the minimum phase transfer function (TF) and $e^{-j\tau}$ is the all-pass function able to model the ITD, as explained below. The ITD is thus a delay line, since it is an all-pass function, to be applied to the lagging ear, [22].

Since in every causal system the real and the imaginary parts of the frequency response are related to each other by the Hilbert transform, for minimum-phase systems, it is possible to state that the *cepstrum*, the sequence associated to the complex logarithm of the frequency response, is causal, and so the real and the imaginary part of the logarithm of the frequency response, which correspond to the magnitude and the phase respectively, are related to each other [22].

Due to the minimum-phase representation, the ITD cue has to be reintroduced as a constant delay. In this study the ITD has been extracted from a set of HRTFs taken by the ARI (Acoustics Research Institute) database through the Interaural Cross-Correlation (IACC) method, [9,15], where the ITD is interpreted as the delay which maximizes the similitude between the Head Related Impulse Responses (HRIRs) of the right and the left ears, i.e. the delay at which the cross-correlation is maximum.

3 Models Implementation

In this section details on the implementation of the previously described models are given.

3.1 Anthropometric Model

In order to model the low-pass effect of the head, we referred to the single pole-single zero function modelled by [4]. The transfer function of the IIR filter representing the head behaviour and taking into account both the angle of incidence θ of the acoustic wave and the head radius (a), is given in [31]:

$$H_{\mathrm{HS}}(z) = \frac{(\omega_0 + F_s\alpha) + (\omega_0 - F_s\alpha)z^{-1}}{(\omega_0 + F_s) + (\omega_0 - F_s)z^{-1}} \tag{4}$$

where $\omega_0 = c/a$, c is the speed of sound (343 m/s circa), α is a function of the azimuth angle given by, [4]:

$$\alpha(\theta) = \left(1 + \frac{\alpha_{min}}{2}\right) + \left(1 - \frac{\alpha_{min}}{2}\right)\cos\left(180° \frac{\theta}{\theta_{min}}\right) \tag{5}$$

The ITD is represented as a constant delay on all frequencies through an allpass filter given in 6, where τ is the group delay defined in Eq. 1, [31]:

$$a = \frac{1 - \tau}{1 + \tau} \tag{6}$$

Concerning the model of the pinna, the choice of coefficients has been done empirically. Indeed from previous studies, pinna effects can be found in the first 0.7 ms of the HRIRs, [4], thus, using a sampling frequency of 44.1 KHz, in the first 32 samples. This implies that a 32 taps FIR filter is enough for pinna modeling. Moreover, as previously said, five reflections are enough representative of the pinna effect, therefore parameters describing these reflections have to be properly chosen. First of all the intensity of the reflections is assumed to be independent on the source direction, thus the corresponding parameters are constant. For each reflection the characterizing parameters are reported in Table 1 where ρ is the reflection coefficient which is different for each reflection and A_n, B_n, D_n are experimentally derived for 3 different persons as described in [31].

It is worth noting that two different columns characterize the parameter D_n since it allows the adjustment of the model to the individual characteristics of the pinna and it indeed results to be different for one person of the 3 involved for the derivation of this table. The delay of each path is instead dependent on the audio source direction as follows, [4]:

$$\tau_{pn}(\phi, \theta) = A_n \cos(\theta/2) \sin[D_n(90 - \phi)] + B_n \tag{7}$$

where coefficients values can be obtained from Table 1.

Table 1. Coefficients table for Eq. 7

n	ρ_{pn}	A_n	B_n	D_n for PB & NH	D_n for RD
2	0.5	1	2	1	0.85
3	−1	5	4	0.5	0.35
4	0.5	5	7	0.5	0.35
5	−0.25	5	11	0.5	0.35
6	0.25	5	13	0.5	0.35

Table 2. Values of τ_{pn} for every event of the pinna with θ equals to $0°, 15°, 30°, 45°$ and $60°$. The generic value it has been calculated using the values of Table 1

n	0°	15°	30°	45°	60°
2	3	7.53	10.53	14.53	16.53
3	2.99	7.50	10.50	14.50	16.50
4	2.96	7.41	10.41	14.41	16.41
5	2.92	7.26	10.26	14.26	16.26
6	2.86	7.06	10.06	14.06	16.06

In the current paper only one column for D_n has been taken into account because there are no information related to the "original" listeners and the persons involved in the experimentation reported in Sect. 4 are physically very different.

The last aspect to be discussed is the relation between the FIR filter coefficients and the coefficients of the pinna anthropometric model. By the analysis of τ_{pn} it has been shown that the delay associated to the different pinna events were almost constant as the incident angle of the horizontal plane was varying, as it is possible to see in Table 2. It is thus reasonable to assume a mean value for each delay associated to a generic event. The delays obtained with this assumption were approximated to the upper and lower integer numbers of samples.

The values of the reflection coefficients were also interpolated since the number of reflections was less than the number of delays. Finally, due to these assumptions, also the extreme values of the reflection coefficients were adjusted since they were not defined.

3.2 Minimum Phase Representation

For the derivation of the minimum phase response, the real cepstrum solution has been exploited, as described in [24]. Then, in order to reintroduce the ITD, the cross correlation method has been used. This method returns the number of samples by which the sequence, representing the audio signal, must be shifted in order to take into account the ITD.

4 Tests and Results

The previously discussed models have been tested only on 4 adult individuals with different anthropometric characteristics. The subjects were 2 males and 2 females with no experience in listening tests and with no hearing problems. The age varied within 25 up to 70 years.

The audio source for both administered tests was a 4000 ms periodic pseudo-random sequence. The generation of the sequence and all the processings have been done by using MATLAB. All the listeners have used a pair of Marshall Major headphones to hear the processed sequences through a PC with a Realtek High Definition Audio on-board. The type of administered tests has required a differentiation, as described below.

4.1 Anthropometric Model Test

The anthropometric model has been tested over 27 different audio source positions on each person. In order to compare the performances of the tested model, the same tests were performed by using also 2 sets of non-individualized HRTFs per person, taken from the ARI database.

So, the tested positions were the ones in which the HRTFs were measured. It is worth noting that the original discretization between audio source positions was equal to 15° but preliminary tests showed that users were not sensitive to this tight distances, so a distance of 30° was chosen. The positions involved are reported in Fig. 1.

Since the anthropometric model can be adapted to anthropometric features of the listeners, for each of them the model was tested by exploiting both an average head radius of 8 cm and a personalized head radius (approximated on each listener's head). This is due to the fact that one of the main purposes of this experimentation is to understand how much a personalized model influences the correct perception of the source location when medium quality devices are used.

Each listener were asked to point on a grid pattern the perceived audio source position. This procedure was repeated for each position testing both the two HRTFs and the two anthropometric models. Results are discussed in Sect. 4.3.

4.2 Minimum Phase Model Test

Concerning the minimum phase model, the Four Interval - Two Alternative Forced Choice (4I-2AFC) has been chosen, [3], which consists of administrating to each individual 4 sequences, within which only 1 is different from the others and asking them to point out the different sequence [16].

In our case, the sequences were the audio source, a pseudo-random noise, filtered by a HRTF taken from the ARI database or its minimum-phase approximation. In particular the different sequence was the one filtered with the approximation, while the other 3 identical sequences were given by filtering the noise through a non personalized HRTF.

With this approach, the experimentation is successful if the 50% circa of the responses is correct because this would mean that the two models are actually indistinguishable and thus the minimum phase solution is properly approximating the HRTF.

4.3 Anthropometric Models Results

Results for the anthropometric model are reported in Table 3 where in a single cell the average percentage of error for the model involved and for a specific direction is reported, as it is evaluated only for the azimuth coordinate. Having a look at these results there is not a particular model that behaves better than others.

Results show a not optimal outcome since the typical problems associated to the use of non personalized HRTF, i.e. front/back reversal, arise also in the case of the anthropometric model. The most of the problems arise when θ is equal to 0°, 30° and 60° that are often perceived as 180°, 150° and 120° degrees respectively.

The less problematic directions were 90° and 270°. Indeed these are those directions for which the binaural parameters do not show ambiguities [7], thus rarely reporting "big" mistakes. The elevation perception which is not here reported, has shown very good results. A particular trend that is worth to discuss is that performance improve as the elevation increases. This result show a proper behaviour of this model for localization of the height of an audio source. While this behaviour of the elevation perception is in line with the results reported in [4], the same cannot be stated for the horizontal perception, indeed the authors in [4] state that the horizontal angle perception did not show any problem thus avoiding to report results. It is possible that these errors are due to the non personalized coefficients for the pinna model (i.e. D_n in Eq. 7), future works will be devoted also to understand this aspect.

Other data can be derived from Figs. 2 and 3, where the abscissa represents the horizontal angle of the emitted audio source, while the ordinate axis reports the angle perceived by the listener. Each set of plots is referred to the antropometric model with a minimum radius for the head and the anthropometric model with personalized head radius. The angles on each plot are parametrized with respect to a certain elevation angle, pointed as ϕ. For some elevation values a few points have been tested given the points chosen for the HRTFs. This point of view allows to notice that there is not a particular model that behaves better than others in terms of perceived localization, moreover, the diversity of choices made by the listeners is even more evident.

4.4 Minimum Phase Representation Results

As hinted before, these tests were based on the 4I-2AFC paradigm. Results have reported that equally perceived sequences were the ones generated by the non personalized HRTF, while the one perceived to be different was generated through minimum phase HRTF with external ITD. Results have thus shown a

100% percentage of correct responses, which is the opposite of the 50% expected for stating the correctness of the solutions. This is not a failure because the differences perceived were mainly related to the quality of the sound and no significant differences in the sources localization were reported.

Both the previously used HRTFs and the 27 positions have been exploited to exploit the minimum phase model through the 4I-2AFC paradigm. Each person was asked to explain which sound was different and why. The most of the reasons were related to a slightly different perceived position. In particular the sound filtered through the minimum phase representation appeared to be higher than the one reproduced through the reference HRTF. A reversal was perceived in only one case i.e. when the correct position was placed at $\theta = 90°$ and $\phi = 80°$ and was instead perceived at $\theta = 240°$. Nevertheless, with respect to the previous results it is possible to state that the minimum phase model introduces improvements.

Table 3. Test results for the anthropometric model. The first column represents the position tested indicated as (ϕ,θ), where ϕ is the elevation and θ is the azimuth.

Positions	HRTF n1	HRTF n2	Average radius	Adapted radius
(0, 0)	58%	31%	46%	48%
(0, 30)	19%	15%	23%	23%
(0, 60)	21%	6%	8%	19%
(0, 90)	4%	4%	8%	10%
(0, 120)	4%	10%	17%	8%
(0, 150)	0%	27%	27%	23%
(0, 180)	6%	21%	17%	17%
(0, 210)	10%	8%	8%	6%
(0, 240)	6%	2%	2%	2%
(0, 270)	6%	6%	8%	10%
(0, 300)	10%	15%	6%	8%
(0, 330)	12%	19%	17%	8%
(30, 0)	13%	25%	35%	65%
(30, 30)	17%	13%	15%	19%
(30, 90)	8%	10%	4%	4%
(30, 150)	17%	15%	4%	15%
(30, 180)	6%	38%	23%	25%
(30, 270)	8%	6%	19%	15%
(60, 0)	25%	23%	50%	27%
(60, 30)	17%	21%	10%	8%
(60, 90)	10%	8%	4%	13%
(60, 180)	15%	17%	0%	4%
(60, 270)	15%	8%	6%	8%
(80, 30)	19%	23%	25%	13%
(80, 90)	10%	8%	4%	6%
(80, 180)	13%	27%	0%	2%
(80, 270)	8%	8%	2%	8%

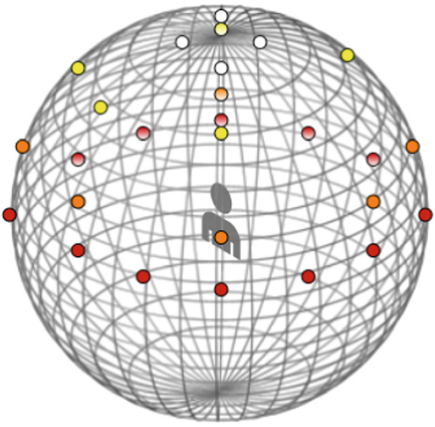

Fig. 1. Each colour represents a specific elevation, red is for 0, orange is for 30, yellow is for 60 and white is for 80. (Color figure online)

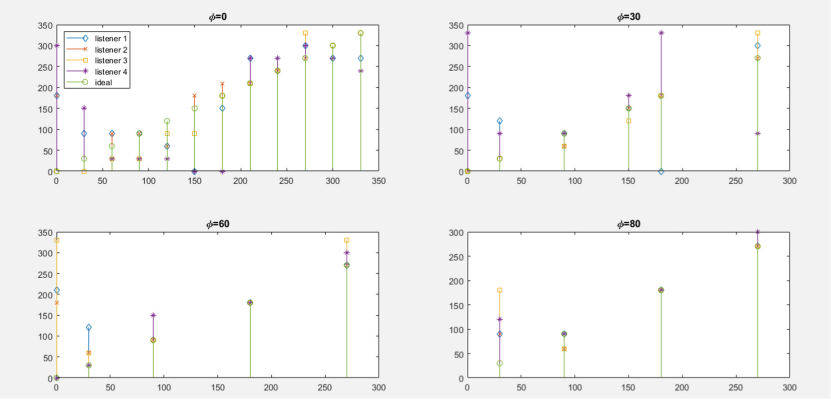

Fig. 2. Results related to the anthropometric model with average head radius

The validity of such approximation was already tested by Kulkarni et al., [16], in a bigger work regarding the sensitivity of humans to variations of the phase spectra of HRTFs, that is why the 4I-2AFC paradigm has been taken into consideration in this work. They obtained the minimum phase reconstruction by a set of non individualized HRTFs, just as presented here, so it seems reasonable to conclude that this model is not valid over these HRTFs. The problems related to the use of non individualized HRTFs are well-known so it would be interesting to repeat this study with individualized ones.

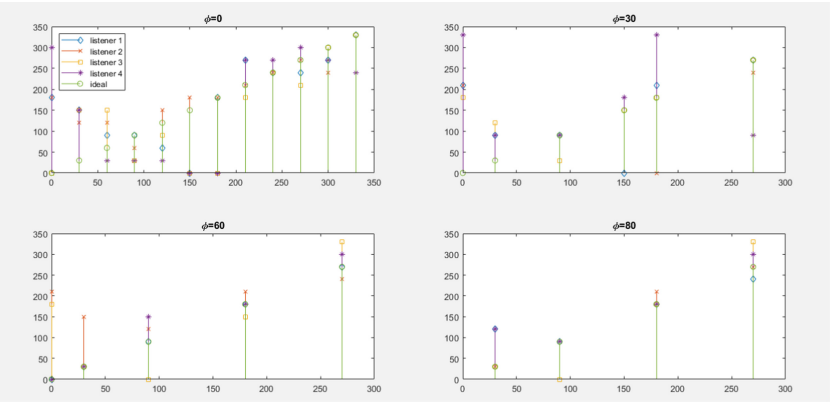

Fig. 3. Results related to the anthropometric model with listener adapted head radius

5 Conclusions and Future Works

This work dealt with the analysis of binaural spatialization in order to support AR/VR applications with particular reference to Cultural Heritage Enhancement services. This work presented the techniques able to support those kind of applications in terms of audio experience, taking into account the need of adapting to low-cost, low-complexity devices that may be exploited in the described scenario. Two models have been studied, implemented and tested in order to avoid the use of non individualized HRTFs in order to obtain binaural spatialization. Results obtained for the antropometric model demonstrate a proper behaviour of the solution when conceiving with the elevation perception, while the azimuth information perception suffered from the typical problems associated to the exploitation of non personalized HRTF such has front/back reversal. This is in contrast with previous studies, thus suggesting a possible change of used coefficients to be adapted to the listener.

Concerning the minimum phase solution, listeners were always able to distinguish the different sound source, but only in terms of its timbre, while the spatial position of the source was not significant for discrimination. Reference parameters properly derived by a set of HRTF directly obtained from the listener, would help in obtaining more reliable solutions and this is part of future works.

Future works are also devoted to solve the problem of front/back reversal [12], and to take into account the issue of the externalization. It is well-known that the only use of the HRTFs lacks on information about the external environment, so the sound is perceived as inside the head. The use of some reverberation in this cases is very helpful.

We have only considered sound sources at a fixed distance without varying it, so the problem of sources closer than 1 m, where the binaural parameters have a fundamental role [5], has to be taken into account as well. A sample demon-

stration system is under development in order to setup a validation scenario for audio spatialization supporting AR/VR applications using Google's Resonance Audio open-source library [10].

References

1. Addison, A.C.: Emerging trends in virtual heritage. IEEE Multimedia **7**(2), 22–25 (2000)
2. Bernardini, F., Rushmeier, H., Martin, I.M., Mittleman, J., Taubin, G.: Building a digital model of Michelangelo's Florentine Pieta. IEEE Comput. Graphics Appl. **22**(1), 59–67 (2002)
3. Bi, J., Kuesten, C.: The four-interval, two-alternative forced-choice (4I2AFC): a powerful sensory discrimination method to detect small, directional changes particularly suitable for visual or manual evaluations. Food Qual. Prefer. **73**, 202–209 (2019)
4. Brown, C.P., Duda, R.O.: A structural model for binaural sound synthesis. IEEE Trans. Speech Audio Process. **6**(5), 476–488 (1998)
5. Brungart, D.S., Rabinowitz, W.M.: Auditory localization of nearby sources. Head-related transfer functions. J. Acoust. Soc. Am. **106**(3), 1465–1479 (1999)
6. Brusaporci, S., Graziosi, F., Franchi, F., Maiezza, P., Tata, A.: Mixed reality experiences for the historical storytelling of cultural heritage. In: Bolognesi, C., Villa, D. (eds.) From Building Information Modelling to Mixed Reality. STCE, pp. 33–46. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-49278-6_3
7. Carlile, S.: The physical and psychophysical basis of sound localization. In: Virtual Auditory Space: Generation and Applications, pp. 27–78. Springer (1996). https://doi.org/10.1007/978-3-662-22594-3_2
8. D'Auria, D., Di Mauro, D., Calandra, D.M., Cutugno, F.: A 3D audio augmented reality system for a cultural heritage management and fruition. J. Digit. Inf. Manage. **13**(4) (2015)
9. Estrella, J.: On the extraction of interaural time differences from binaural room impulse responses. Master's thesis, Technische Universität Berlin (2010)
10. Gorzel, M., et al.: Efficient encoding and decoding of binaural sound with resonance audio. In: Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio. Audio Engineering Society (2019)
11. Guidi, G., et al.: 3D digitization of a large model of imperial Rome. In: Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM 2005), pp. 565–572. IEEE (2005)
12. Gupta, N., Barreto, A., Ordonez, C.: Improving sound spatialization by modifying head-related transfer functions to emulate protruding pinnae. In: Proceedings IEEE SoutheastCon 2002 (Cat. No. 02CH37283), pp. 446–450. IEEE (2002)
13. Härmä, A., et al.: Augmented reality audio for mobile and wearable appliances. J. Audio Eng. Soc. **52**(6), 618–639 (2004)
14. Ikeuchi, K., Nakazawa, A., Hasegawa, K., Oishi, T.: The great buddha project: modeling cultural heritage for VR systems through observation. In: ISMAR, vol. 3, pp. 7–16 (2003)
15. Katz, B.F., Noisternig, M.: A comparative study of interaural time delay estimation methods. J. Acoust. Soc. Am. **135**(6), 3530–3540 (2014)
16. Kulkarni, A., Isabelle, S., Colburn, H.: Sensitivity of human subjects to head-related transfer-function phase spectra. J. Acoust. Soc. Am. **105**(5), 2821–2840 (1999)

17. Lapini, A., Calamai, G., Argenti, F., Carfagni, M.: Application of binaural audio techniques for immersive fruition of cultural heritage. IOP Conf. Ser. Mater. Sci. Eng. **364**, 012099 (2018). https://doi.org/10.1088/1757-899X/364/1/012099
18. Madirov, E., Absalyamova, S.: The influence of information technologies on the availability of cultural heritage. Procedia. Soc. Behav. Sci. **188**, 255–258 (2015)
19. Malpas, J.: Cultural heritage in the age of new media. In: Kalay, Y.E., Kvan, T., Affleck, J. (eds.) New Heritage: New Media and Cultural Heritage, Abingdon (2008)
20. Murphy, D., Neff, F.: Spatial sound for computer games and virtual reality. In: Game Sound Technology and Player Interaction: Concepts and Developments, pp. 287–312. IGI Global (2011)
21. Oculus: Explore the oculus unity spatializer with the sample scene. https://developer.oculus.com/documentation/unity/audio-osp-unity-scene/. Accessed 15 Jan 2021
22. Oppenheim, A.V., Buck, J.R., Schafer, R.W.: Discrete-Time Signal Processing, vol. 2. Prentice Hall, Upper Saddle River (2001)
23. Palombini, A.: Storytelling and telling history. Towards a grammar of narratives for cultural heritage dissemination in the digital era. J. Cult. Heritage **24**, 134–139 (2017)
24. Pei, S.C., Lin, H.S.: Minimum-phase FIR filter design using real cepstrum. IEEE Trans. Circuits Syst. II Express Briefs **53**(10), 1113–1117 (2006)
25. Pietroni, E., Pagano, A., Rufa, C.: The Etruscanning project: gesture-based interaction and user experience in the virtual reconstruction of the Regolini-Galassi tomb. In: 2013 Digital Heritage International Congress (DigitalHeritage), vol. 2, pp. 653–660. IEEE (2013)
26. Rayleigh, L., Lodge, A.: Iv. on the acoustic shadow of a sphere. Philos. Trans. R. Soc. London Ser. A, Contain. Papers Math. Phys. Charact. **203**(359–371), 87–110 (1904)
27. Roussou, M., Efraimoglou, D.: High-end interactive media in the museum. In: International Conference on Computer Graphics and Interactive Techniques: ACM SIGGRAPH 1999 Conference Abstracts and Applications, vol. 8, pp. 59–62 (1999)
28. Software, V.: Steam audio unity plugin. https://valvesoftware.github.io/steam-audio/doc/phonon_unity.html. Accessed 15 Jan 2021
29. Thwaites, H.: Digital heritage: what happens when we digitize everything? In: Ch'ng, E., Gaffney, V., Chapman, H. (eds.) Visual Heritage in the Digital Age. SSCC, pp. 327–348. Springer, London (2013). https://doi.org/10.1007/978-1-4471-5535-5_17
30. Zagala, F., Noisternig, M., Katz, B.F.: Comparison of direct and indirect perceptual head-related transfer function selection methods. J. Acoust. Soc. Am. **147**(5), 3376–3389 (2020)
31. Zölzer, U.: DAFX: Digital Audio Effects. Wiley, New York (2011)