Personalized 3D sound rendering for content creation, delivery, and presentation

Federico Avanzini[1], Luca Mion[2], Simone Spagnol[1]

[1]Dep. of Information Engineering, University of Padova, Italy; [2]TasLab - Informatica Trentina, Trento, Italy

E-mail: federico.avanzini@dei.unipd.it, luca.mion@infotn.it, simone.spagnol@dei.unipd.it

Abstract: **Advanced models for 3D audio rendering are increasingly needed in the networked electronic media world, and play a central role within the strategic research objectives identified in the NEM research agenda. This paper presents a model for sound spatialization which includes additional features with respect to existing systems, being parametrized according to anthropometric information of the user, and being based on audio processing with low-order filters, thus allowing for significant reduction of the computational costs. This technology can offer a transversal contribution to the NEM research objectives, with respect to content creation and adaptation, intelligent delivery and augmented media presentation, by improving the quality of the immersive experience in a number of contexts where realistic spatialization and personalised sound reproduction is a key requirement, in particular in mobile contexts with headphone-based rendering.**

Keywords: 3D sound, multimodal interaction, virtual auditory space, augmented reality.

1 INTRODUCTION

In the networked electronic media world, strategies for innovation and development have increasingly focused on applications that require spatial representation and real-time interaction with/within 3D media environments. One of the major challenges that such applications have to address is user-centricity, reflecting e.g. on developing complexity-hiding services so that people can personalise their own delivery of services. In these terms, multimodal interfaces represent a key factor for enabling an inclusive use of new technology by all. To achieve this, multi-modal realistic models to describe our environment are needed, and in particular models that accurately describe the acoustics of the environment and the communication through the auditory modality.

Models for spatial audio can provide accurate information about the relation between the sound source and the surrounding environment, including the listener and his/her body which acts as an additional filter. This information can not be substituted by any other modality (e.g., visual or tactile). However, today's spatial representation of audio tend to be simplistic and with poor interaction capabilities, being multimedia systems currently focused on graphics processing mostly, and integrated with simple stereo or surround-sound. We can identify three important reasons why many media components lack such realistic audio rendering. First, the lack of personalization of services and content, since current content delivery systems do not exploit information about the environment in which they are working, and no adaptation on user is provided except for profiling at the metadata level. Second, the increasing need for bandwidth and high computational costs which easily overload the resources available both on the channel and on the terminal, especially when concerning mobile devices. Third, current auralization technologies rely on invasive and/or expensive reproduction devices (e.g., HMDs, loudspeakers), which cause to the user a perceived non-integrated experience due to an unbridged gap between the real and the virtual world.

With reference to the NEM strategic agenda [1], these three points are directly linked to the research fields of Content creation, Delivery, and Media presentation. Hence the need for advanced models for 3D audio rendering emerges transversally from the strategic changes identified in the agenda.

Stereo is the simplest system involving "spatial" sound, but a correct spatial image can only be rendered along the central line separating the loudspeakers (the 'sweet spot'). Surround systems based on multichannel reproduction, such as 5.1 or 10.2 systems [2], or ambisonics [3], also suffer from similar "crosstalk" problems (i.e., the sound emitted by one loudspeaker is always heard by both ears). Crosstalk cancellation techniques commonly employed are effective only in a very limited listening region.

Wave-Field Synthesis is a currently active line of research. This method, initially proposed in [4], uses arrays of small and individually driven loudspeakers to reproduce a faithful replica of a desired spatial sound field. As a result, the spatial image is correct in the whole half-space at the receiver side of the array. Research in this direction is progressing rapidly, however wave-field methods require expensive and cumbersome reproduction hardware, which makes them suitable only for specific application scenarios (e.g., digital cinema [5]).

On a different level lie 3D audio rendering approaches based on headphone reproduction. In this paper we focus on this latter family of approaches, and present a model for 3D audio rendering that can be employed for immersive sound reproduction. The proposed approach allows for an interesting form of content adaptation and personalization, since it includes parameters related to

Corresponding author: Federico Avanzini, University of Padova, Italy +3904982777856, avanzini@dei.unipd.it

user anthropometry in addition to those related to the sound sources and the environment. Our approach has also implications in terms of delivery, since it operates by processing a monophonic signal exclusively at the receiver side (e.g., on terminal or mobile device) by means of low-order filters, which allows for reduced computational costs. Thanks to the low complexity, the model can be used to render scenes with multiple audio-visual objects in a number of contexts such as computer games, cinema, edutainment, and any other context where realistic sound spatialization, personal, and personalised sound reproduction is a major requirement, in particular in mobile contexts with headphone rendering.

The remainder of the paper is organized as follows: Sec. 2 reviews the main concepts of 3D binaural sound reproduction; Sec. 3 presents our recent and current research in this field; finally, Sec. 4 discusses the relevance of this work in relation with the main research challenges of the NEM agenda.

2 3D BINAURAL SOUND REPRODUCTION

Possible disadvantages of headphone-based systems (e.g., invasiveness, non-flat frequency responses, lack of compensation for listener motion unless a tracking system is used), are counterbalanced by a number of desirable features. They eliminate reverberation and other acoustic effects of the real listening space, they reduce background noise, and they provide personal audio display, which are all relevant aspects especially in mobile contexts. On a more technical note, headphone based systems allow to deliver distinct signals to each ear, which greatly simplifies the design of 3D sound rendering techniques.

2.1 Head-related transfer functions

A sound source can be virtually positioned in space by filtering the corresponding (monophonic) source signal with so-called *head related transfer functions (HRTFs)*, thus creating left and right ear signals that are subsequently delivered by headphones as shown in Fig. 1 [6]. The HRTFs depend on the relative position between listener and sound source. For a given position, they capture the transformations experienced by a sound wave in its path from the source to the tympani, which are caused by diffraction and reflections by the torso, head, shoulders and pinnae of the listener. Consequently the HRTFs exhibit a great person-to-person variability.

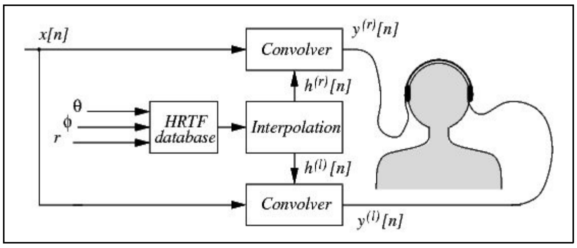

Figure 1: A simplified 3D audio reproduction system based on headphones and HRTFs

The rendering scheme of Fig. 1 assumes the availability of a database of measured HRTFs. Acoustic measurement of individual HRTFs for a single subject is an expensive and cumbersome procedure, which has to be conducted in an anechoic chamber, using in-ear microphones, specialized hardware, and so on. Therefore individual HRTFs cannot be used in most real-word applications. Alternatively, generalized HRTFs are typically measured on so-called “dummy heads”, i.e. mannequins constructed from averaged anthropometric measures, representing standardized heads with average pinnae and torso. However, this limits to some extent the realism of the rendering: in fact one dummy head might sound more natural to a particular set of users than another, depending on anthropometric measures and also on technicalities in the measurement procedure.

A second problem is that HRTF measurements can only be made at a finite set of locations, and when a sound source at an intermediate location must be rendered, the HRTF must be interpolated. If interpolation is not applied (e.g., if a nearest neighbour approach is used) audible artefacts like clicks and noise are generated in the sound spectrum when the source position changes. Clearly this problem becomes even more severe in interactive settings, where both the listener and the sound sources are moving in the environment and the rendering must be dynamically updated.

2.2 Structural models

As opposed to the rendering approach based on measured HRTFs, the structural modeling approach [7] attempts to simulate the separate filtering effects of the torso, head, and pinna. These filtering blocks, each accounting for the contribution of one anatomical structure, are then combined to form a model of the HRTF.

The head causes both time and level differences between sound waves reaching the two ears, which occur because sound has to travel an extra distance in order to reach the farthest ear, and is acoustically “shadowed” by the presence of the head. Correspondingly, head effects are simulated using delay lines and low/high-pass filters [7]. The external ear acts as both a sound reflector and a resonator: acoustic rays are reflected on the “bass relief” form of the pinna, and moreover the cavities of the external ear add sound coloration with their resonances. Correspondingly, pinna effects are simulated using resonant and notch filters [8], and are especially relevant in rendering sound source location in the vertical direction. The torso contributes additional sound reflections.

Finally, room effects can also be incorporated into the rendering scheme: in particular early reflections from the environment can be convolved with the external ear (pinna) model, depending on their incoming direction. A synthetic block scheme of a generic structural model is given in Fig.2.

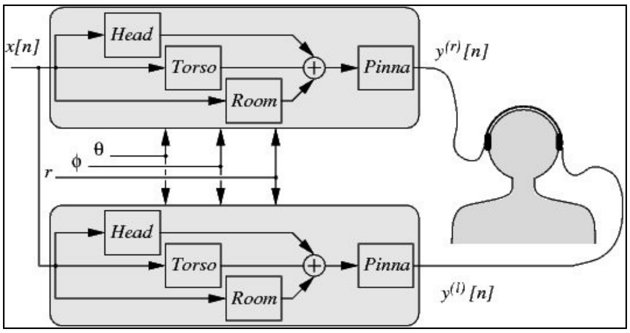

Figure 2: A simplified 3D audio reproduction system based on structural HRTF modeling.

3 CURRENT RESEARCH

3.1 Low-order structural models

We have recently proposed an approach to derive low-order filtering structures for a structural HRTF model [9]. The main results are summarized in the remainder of this section.

Similarly to previous literature [7], the diffraction effects of the human head are approximated with those of a sphere, which are known analytically. Given such a "spherical HRTF", represented by the transfer function *H(μ,θ,ρ)* (where *μ* is a normalized frequency, *θ* is the angle of incidence, and *ρ* is the source-head distance), we apply principal component analysis (PCA) to obtain a series expansion of this transfer function on a suitable basis of vectors. As a result, *H* is expressed as follows:

$$H(\mu,\theta,\rho)=\sum_{i=1}^{p}H_i(\mu)a_i(\theta,\rho)$$

where H_i are the frequency-dependent basis vectors, a_i are a set of coefficients that depend on spatial variables only, and *p* is the number of principal components used.

This representation has two main advantages. First, the basis vectors H_i have relatively simple responses, and can be approximated with low-order filters. Second, the decoupling of frequency and spatial variables implies that when a set of *N* sound sources (i.e., N monophonic signals X_k, *k=1...N*) has to be rendered, the rendering is achieved through the following equation:

$$Y(\mu)=\sum_{i=1}^{p}H_i(\mu)\sum_{k=1}^{N}a_i(\theta_k,\rho_k)X_k(\mu)$$

where *Y* is the signal produced after the diffraction of *all* the signals X_k on the spherical HRTF. This means that the rendering is achieved by linearly combining all the source signals with the coefficients a_i and then processing the resulting signal through the basis vectors H_i. The advantage is that there are always only *p* filtering operations regardless of the number *N* of sources to be processed.

Figure 3 shows an example of the results of the proposed approach. The top panel depicts the spherical HRTF for a fixed distance *ρ* and for various incidence angles *θ*: note the transition from a high-pass to low-pass characteristics as the angle of incidence is varied. The bottom panel shows the approximation obtained with the PCA approach described above: in this case only *p=3* components have been used, nonetheless the corresponding approximation is already quite accurate.

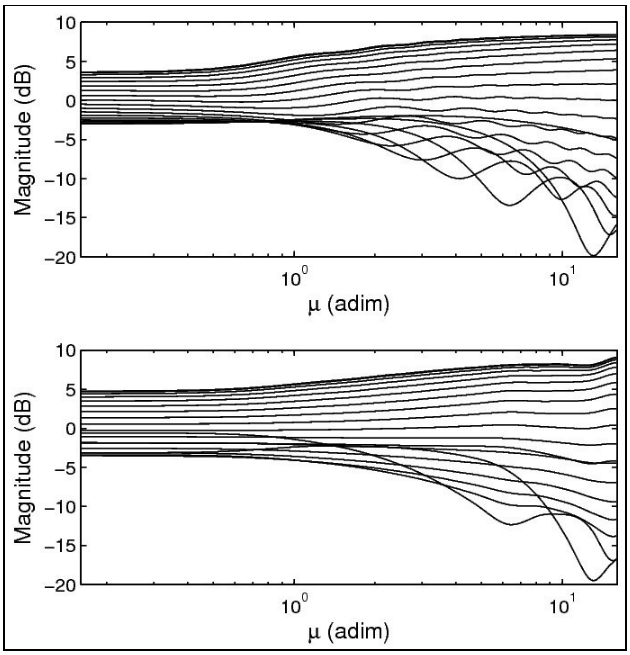

Figure 3: Example of analytical (top panel) and approximated (bottom panel) spherical HRTF magnitude curves. Curves are computed for a fixed head-source distance *ρ*, and are parametrized through the angle of incidence *θ*.

3.2 Experimental validation

We have developed a real-time implementation of the structural model, with the aim of experimentally validating the proposed approach through listening tests in interactive settings [10].

All the tests that have been conducted are based on similar set-ups in which virtual audio-visual objects are placed in given spatial locations, and users are free to move in the virtual environment. Head position and orientation are captured by a marker-based motion tracking system, and these data are used to drive the graphic and audio rendering, displayed by means of a head-mounted display (HMD) and insulated headphones, respectively. 3D audio rendering uses the spherical HRTF model described above, as well as a simplified pinna model that simulates the first frequency notch introduced by sound reflections on the external ear.

In an experiment on the perception of sound source angular position, subjects were asked to judge the incoming direction of acoustic stimuli produced by virtual sources on a sphere centered at the listener's head and with radius of 1 m. Reverberation was also added to simulate the characteristics of a real small-sized room. Stimuli were presented through headphones with markers applied to track head movements. No visual feedback was provided in this case.

We used two experimental conditions: passive playback and active movement. In the first condition, subjects were asked to mark on a grid the perceived direction of the sound source without moving, while in the second one they had to move their head to face the virtual source. Interestingly, subjects proved to be much more confident on their judgement in dynamic conditions (see Fig. 4), confirming that the relatively simple structural model used in this work is effective in rendering spatial sound especially in interactive settings where the user is free to move in the scene.

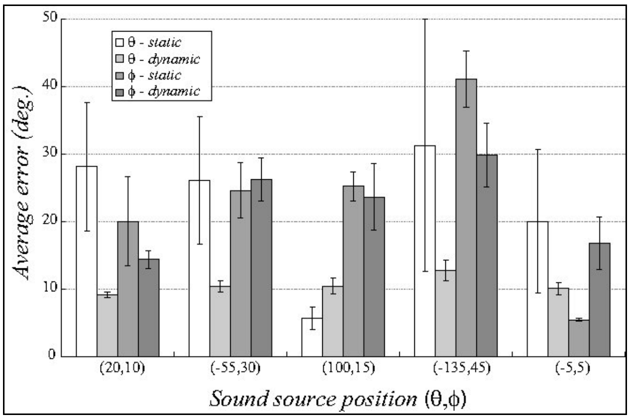

Figure 4: Average (across subjects) absolute error estimation for azimuth θ and elevation Φ, in static and dynamic conditions, for five sound source locations.

It is known that rendering of sound source distance is a more challenging task than rendering of angular direction. In an experiment on sound source distance perception [10], subjects were asked to judge verbally (yes/no) whether a simulated audio-visual object was within reach. The set-up was similar to the previous experiment. Three experimental conditions were used (video-only, audio-only, audio-video). Participants were allowed to explore the scene freely (by moving their head and/or torso) prior to giving their judgements, thus influencing the 3D rendering (Fig. 5 shows the case of sound rendering).

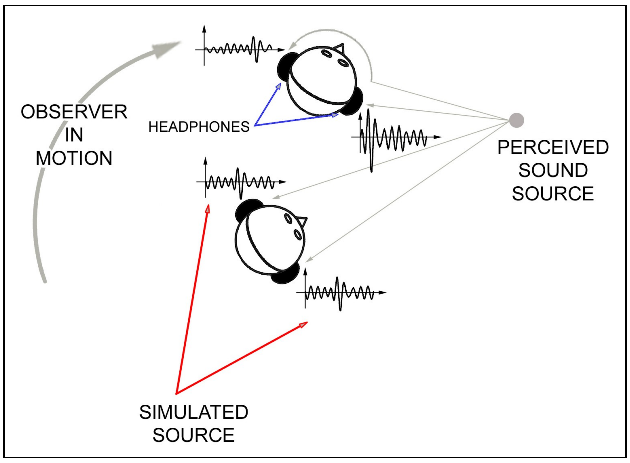

Figure 5: Simulating a virtual sound source for a moving observer.

Results showed that the precision in participants' judgements when the target was only audible was very similar to those obtained when the target was only visible or when it was both visible and audible. Again, these results support the conclusion that structural HRTF models are effective in rendering spatial sound especially in interactive settings.

4 DISCUSSION

Current research on 3D binaural sound reproduction is relevant for the NEM research agenda [1] at many levels. In particular it has implications transversally on the three main research “pillars” identified in the NEM vision, namely Content creation, Delivery, and Media presentation.

4.1 Content creation and manipulation

In this context, the main topics for research include auralization technologies that are able to create realistic 3D sound scenes. It is recognized that content formats must include sound, moreover auralization tools have to be adapted to the type of content to be created (e.g., game, music, video, TV, rich media), and have to allow for interactivity, realism, immersion, customization, adaptation to the terminal and the equipment available at the user's location (e.g., stereo headphones or loudspeaker setups). In particular auralization technologies are expected to become more and more used in games. The game market will help auralization to enter the multimedia content market (e.g., music, video, TV, and rich media contents) [1].

The technologies presented in the previous section are relevant to this vision in many points. First, using headphone-based reproduction, in conjunction with head tracking devices, allows for interactivity, realism, and immersion that are not achievable yet with multichannel systems or wavefield synthesis, due to limitations in the user workspace and to acoustic effects of the real listening space.

Second, the techniques outlined in Sec. 3 allow for an interesting form of content adaptation, i.e. adaptation to users' anthropometry: in fact the parameters of the rendering blocks sketched in Fig. 2 can be related to anthropometric measures (e.g., the interaural distance, or the diameter of the cavum conchae), with the advantage that a generic structural HRTF model can be adapted to a specific listener, thus further increasing the realism of the sound scene and quality of experience.

4.2 Delivery

The technologies discussed in this paper fit well into an *object oriented* sound reproduction approach [5]. The general idea behind this definition is that, when transmitting data of a 3D sound scene, the *sound sources* are transmitted. Each sound source is an audio signal together with additional meta-data describing the spatial

Corresponding author: Federico Avanzini, University of Padova, Italy +3904982777856, avanzini@dei.unipd.it

source position and other relevant properties. This means that, in order to convey a sound scene composed on N sound sources, only the corresponding N audio streams plus meta-data need to be transmitted. The audio scene can then be rendered on arbitrary reproduction setups, the final rendering depends on the reproduction system, and is left to the intelligence of the terminal.

In the context of 3D binaural sound reproduction considered in this paper, the scene will be rendered by (a) processing each individual sound source signal through a pair of HRTFs, estimated using the associated meta-data, and (b) summing all the left- and right-signals to obtain the final stereo signal to be delivered at the earphones.

This architecture can also allow for effective scalability depending on the network resources. Sound sources can be prioritized based e.g. on psychoacoustic criteria (i.e., priority depends upon audibility of the source). In case of limited bandwidth, the number N of sound sources delivered to the terminal can be reduced accordingly (i.e., the least perceivable sources are removed from the scene). This would allow for graceful degradation of the rendering depending on the available bandwidth, and would result in satisfactory quality of experience even in cases of limited quality of service.

4.3 Media presentation

In this area of research, it is recognized that multimodal user interfaces can increase the naturalness and the effectiveness of the interaction in a transparent way. A relevant example in the context of this paper is augmentation of visual scenes by acoustic descriptions when important details are offscreen. It is emphasized that authentic, true-to-original media reproduction requires novel displays to offer realistic and immersive reproduction especially in the context of video (holographic eyeglasses, wearable organic light-emitting diodes (OLEDs) and similar).

One advantage of the techniques discussed here is that they have minimal hardware requirements with respect to those implied by realistic video reproduction, and with respect to other technologies for immersive sound reproduction (multichannel systems and wavefield synthesis).

A second advantage is that computational requirements at the terminal side are also low. Rendering a sound source in a virtual location in space simply requires filtering a monophonic audio signal through two low-order filters. On a more technical note, the modeling approach outlined in Sec. 3.1 implies that rendering of multiple sources can be achieved by (a) summing all the corresponding signals, weighted by their location-dependent principal components, and (b) filtering this compound signal through the same direction-independent filters. This means that the computational load is almost constant with respect to the number of sources to be rendered (including possible phantom sources that simulate reflections from the environment).

The above remarks are particularly relevant for mobile applications. Whereas video mobile display is currently limited by the technology available (both in terms of reproduction devices and computational resources), the techniques considered in this paper allow for 3D binaural audio display without the above limitations, and are therefore suited for mobile applications (particularly mobile virtual/augmented reality).

5 CONCLUSION

This paper has presented our current work on the development of a structural model of HRTFs, which is based on low-order filter structures that simulate the acoustic effects of head and ears on an incoming, spatially located sound. Listening experiments with users have shown that such simple models are effective in rendering spatial sound, especially in interactive settings where the user is free to move in the scene.

One of the main advantages of such a structural modeling approach with respect to measured HRTFs is that the model can be parametrized according to anthropometric information of the user, thus allowing for an interesting form of content adaptation, i.e. adaptation to users' anthropometry.

Subsequent discussion has emphasized that this approach to 3D sound rendering can offer a transversal contribution to the NEM research objectives, with respect to content creation and adaptation, intelligent delivery and augmented media presentation, by improving the quality of the immersive experience in a number of contexts where realistic spatialization and personalised sound reproduction is a key requirement.

Acknowledgements

We wish to thank Benoit Bardy and Bruno Mantel (Lab. of Motor Efficiency and Deficiency, Montpellier-1 University) for insightful discussions and fruitful collaboration in the experiments described in Sec. 3.2.

References

[1] Strategic Research Agenda. “Networked and Electronic Media” European Technology Platform, Sep. 2008. www.nem-initiative.org

[2] T. Holman. 5.1 Surround Sound: Up and Running. Focal Press, 2000.

[3] M.A. Gerzon. Ambisonic in multichannel broadcasting and video”, J. Audio Eng. Soc. 33:859-871 (1985).

[4] A.J. Berkhout. A holographic approach to acoustic control”, J. Audio Eng. Soc. 36:977-955 (1988).

[5] G. Gatzsche, B. Michel, J. Delvaux, and L. Altmann. Beyond DCI: The integration of object oriented 3D sound into the Digital Cinema. In Proc. 2008 NEM Summit, pages 247-251. Saint-Malo, Oct. 2008.

[6] C. I. Cheng and G. H. Wakefield. Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in time, frequency, and space. J. Audio Eng. Soc., 49(4):231-249, Apr. 2001.

[7] C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. IEEE Trans. Speech Audio Process., 6(5):476–488, Sep. 1998.

[8] P. Satarzadeh, V. R. Algazi, and R. O. Duda, 2007. Physical and Filter Pinna Models Based on Anthropometry. In *Proc. 122nd AES Convention*, Vienna, 2007.

[9] S. Spagnol and F. Avanzini. Real-time binaural audio rendering in the near field. Accepted for publication in *Proc. Int. Conf. on Sound and Music Computing (SMC09)*, pages 201-206, Porto, July 2009.

[10] L. Mion, F. Avanzini, B. Mantel, B. Bardy, and T. A. Stoffregen. Real-time auditory-visual distance rendering for a virtual reaching task. In *Proc. ACM Int. Symposium on Virtual Reality Software and Technology (VRST07)*, pages 179-182, Newport Beach, CA, Nov. 2007.

Corresponding author: Federico Avanzini, University of Padova, Italy +3904982777856, avanzini@dei.unipd.it