Customized 3D sound for innovative interaction design

Michele Geronazzo[*]
Department of Information Engineering
University of Padova
Via Gradenigo 6/A
35131 Padova, Italy

Simone Spagnol[†]
Department of Information Engineering
University of Padova
Via Gradenigo 6/A
35131 Padova, Italy

Federico Avanzini[‡]
Department of Information Engineering
University of Padova
Via Gradenigo 6/A
35131 Padova, Italy

ABSTRACT

This paper considers the impact of binaural 3D audio on several kinds of applications, classified according to their degree of body immersion and their own coordinate system deviation from a physical condition. A model for sound spatialization, which includes additional features with respect to existing systems, is introduced. A significant reduction of computational costs is allowed by model parametrization according to anthropometric information of the user and audio processing through low-order filters, thus resulting affordable for several kinds of devices. According to the following examination, this approach to 3D sound rendering can grant a transversal enrichment to the CHI research purposes, in reference to content creation and adaptation, resourceful delivery and augmented media presentation. In several contexts where personalized spatial sound reproduction is a central requirement, the quality of the immersive experience could only benefit from this sort of adaptable and modular system.

Categories and Subject Descriptors

H.5.5 [**INFORMATION INTERFACES AND PRESENTATION**]: Sound and Music Computing—*Modeling*; H.5.1 [**INFORMATION INTERFACES AND PRESENTATION**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*; H.5.2 [**INFORMATION INTERFACES AND PRESENTATION**]: User Interfaces—*Auditory (non-speech) feedback*

General Terms

Design, Human Factors

Keywords

3D sound, mixed reality, customized HRTF

[*]e-mail: geronazzo@dei.unipd.it

[†]e-mail: spagnols@dei.unipd.it

[‡]e-mail: avanzini@dei.unipd.it

1. INTRODUCTION

Mixed reality (MR) applications anchor rendering processes to the world's reference frame, rather than to the listener's reference frame, as is the case for pure virtual reality (VR). The degree of immersion and the definition of the spatial frame are qualities connected to the concept of *virtuality continuum* introduced in the literature by Milgram *et al.* [20] for visual displays. These notions can be adapted to virtual auditory displays (VAD) and augmented audio reality (AAR), including sonic effects and overlaying computer-generated sounds on top of real-time acquired audio signals [13]. Our paper faces the problem of creating a model that can be employed for immersive sound reproduction over the different degrees of virtuality. Here we focus on headphone-based systems for binaural audio rendering, taking into account that possible disadvantages (e.g. invasiveness, non-flat frequency responses) are counterbalanced by a number of desirable features. Indeed, these systems eliminate reverberation and other acoustic effects of the real listening space, reduce background noise, and provide adaptable audio displays, which are all relevant aspects especially in enhanced contexts. With this kind of system each ear receives distinct signals, greatly simplifying the design of 3D sound rendering techniques.

Nowadays most of MR systems are able to control fluently two dimensions of an auditory space, i.e. sound sources positioned in the horizontal plane referring to a head-centric coordinate system. Head-tracking technologies plus dummy-head frequency responses or adaptable models for horizontal localization [6] allow for accurate discrimination between sound sources placed around the user and into the defined subspace.

The third dimension, elevation or vertical control, requires a user specific characterization in order to simulate the effective perception in the vertical plane mainly due to the shape of the external ear (the pinna). The crucial tasks needed to find a suitable model for describing the pinna contribution together with extraction of the parameters related to anthropometric measurements are thorny challenges. Our research aims at understanding which possible approximations can be introduced in such pinna model so as to make vertical control available to a CHI designer.

The proposed approach allows for an interesting form of content adaptation and customization, since it incorporates both parameters related to the user's anthropometry and the spatial ones. In terms of delivery, our model works by processing a monophonic signal at the receiver side (e.g., on a terminal or mobile device) using low-order filters, thereby allowing a reduction of the computational costs. Its low-complexity nature can easily be used to represent scenes with multiple audiovisual objects in various situations such as computer games, cinema, edutainment. It can also be used in any scenario requiring highly realistic sound spatialization and personalized sound

reproduction. Furthermore, a customized Head-and-Torso (HAT) model [3], or other equivalent contributions, must be connected to our pinna model in order to achieve a complete structural representation and a full 3D experience.

Sec. 2 includes examples of real-time systems in mixed reality contexts where 3D sound enriches the immersion and interactivity in multiple scenarios, while in Sec. 3 we propose a basic overview of spatial sound rendering techniques and the motivations that brought us toward a structural modeling approach. Finally a complete description of our structural model together with a possible parametrization of pinna-related HRTF features based on anthropometry is sketched in Sec. 4.

2. CHI-RELATED APPLICATIONS

A 3D audio scene, created by binaural sound reproduction, will be estimated from each individual sound source signal using the associated meta-data, and then summing the left and the right signal to produce the final stereo signal sent to the earphones. This architecture can also allow for effective scalability depending on the available computational resources or bandwidth. Psychoacoustic criteria can define the sound sources' rendering priority and attributes, such as audibility of the source. Specifically, in relation to the amount of bandwidth available the least perceivable sources can be removed from the scene and this graceful degradation of the rendering scene would result in a satisfactory experience even in cases of limited quality of service.

In typical virtual audio applications the user's head is the central reference for audio objects rendering. In principle the location of the user's head establishes a virtual coordinate system and builds a map of the virtual auditory scene. In the case of a mixed environment sound objects are placed in the physical world around the user and hence, conceptually, positioned consistently with a physical coordinate system. Locating virtual audio objects into a mixed environment requires the superimposition of one coordinate system onto another.

Depending on the nature of the specific application, several settings can be used to characterize the coordinate system for virtual audio objects and the location of objects in the environment. A simple distinction is the choice to refer to a global positioning system, or a local coordinate system. An ideal classification can help the definition of the possible applications that use spatial audio technologies. In some cases it is necessary to make the two coordinate systems match in a way that virtual sound sources appear in specific locations into the physical environment, while in other applications virtual sources are floating somewhere around the user because the target lies on a disjoint conceptual level in user interaction.

In order to help the presentation, a visual scheme of two different applications is shown in Fig. 1. The characterization moves around a simplified two-dimensional space defined in terms of *degree of immersion* (DI) and *coordinate system deviation* (CSD). Our point of view is a simplification of the three-dimensional space proposed by Milgram *et al.* [20] consisting of Extent of World Knowledge (EWK), Reproduction Fidelity (RF) and Extent of Presence Metaphor (EPM). The correspondences are traced paying attention to the three entities involved: the real world, the MR engine and the listener.

The MR engine is the intermediary between reality and the representation perceived by the listener; in that sense *CSD* matches with EWK. A low *CSD* means a high EWK: the MR engine knows everything about the objects' position in reality and can render the synthetic acoustic scene as if the listener perceives a coherent world.

On the other hand the issue of realism concerns the technologies

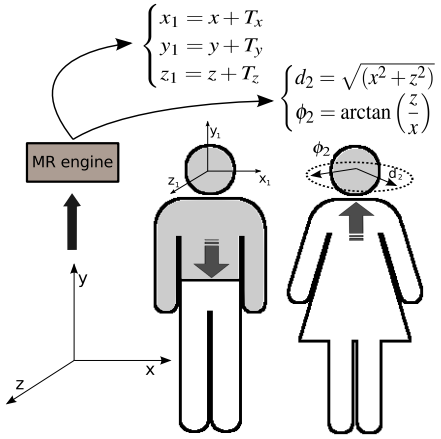


Figure 1: *Coordinates distinction in mixed reality.*

involved in the MR engine, and the complexity of the taxonomy for such system increases considerably. EWK and RF are not entirely orthogonal and our choice is to define DI according to the following idea: when a listener is surrounded by a real sound, all his/her body interacts with the acoustic waves, i.e. a technology with high realism rate is able to monitor the whole listener's body position and orientation (high DI).

Returning to Fig. 1, the subject on the left provides an example of high DI corresponding to a high percentage of the body being tracked into the virtual coordinate system (an almost fully gray-filled body), and it exhibits a low virtual coordinate deviation from the physical coordinate system, due to a simple translation. On the contrary the subject on the right exhibits a low DI and a high CSD, represented by a gray head and a listener-centered 2D virtual space.

Concrete examples of the two cases are the following. The female user is wearing a mobile device and is navigating throughout her many messages and appointments. The male user is in a dangerous telepresence operation, and his head and torso are tracked to immerse his body in a distant real place. The first scenario depicts a totally virtual world and in many cases not fully three-dimensional, the latter represents the exact superposition of the virtual and real worlds.

2.1 Virtual coordinates

The most common case of a floating virtual coordinate system is the one where the only anchor point relative to which the event is localized is the user's head. Usually, virtual sound sources are rendered to different directions and are purely virtual (minimum DI and maximum CSD).

As an example, information services such as news, e-mails, calendar events or other types of messages can be positioned in the virtual acoustic space around the user's head. Calendar events, in the form of speech messages, are rendered in different directions depending on the timetable of the user's agenda, so that noon appears in the front. In the case of a hierarchical menu structure presentation, as commonly found on mobile devices, spatial user-interface designs such as the one presented in [19] can be adopted.

Immersive virtual reality applications also use specific virtual coordinate systems, usually related to the geometry of a graphical virtual reality scene. In computer game applications that use spatial audio techniques, the virtual coordinate system is defined according to the game scene and sometimes combined with information on the physical location of a user (e.g. head tracking via webcam).

Telepresence is another case of a floating virtual coordinate system and is similar to virtual auditory display systems if focused on the immersive experience of the user. An interesting mixed reality case is the bidirectional augmented telepresence application where a binaural telepresence signal is combined with a pseudoacoustic environment [17]. The MR engine merges the local pseudoacoustic environment with a remote pseudoacoustic environment to acoustically produce the other person's environment. In this case the *CSD* related to the remote environment is very low.

In collaborative virtual environments, Benford *et al.* [8] have shown that spatial cues can combine with audiovisual cues in a natural way to aid communication. The well-known "cocktail-party" effect shows that people can easily monitor several spatialised audio streams at once, selectively focusing on those of interest. In multiparty teleconferencing, the positioning of each talker can be done freely in a virtual meeting room.

Researchers such as Walker and Brewster have explored the use of spatial audio on mobile devices, e.g. for addressing problems of visual clutter in mobile device interfaces [25]. This could provide help to disambiguate speakers in multi-party conferences, and affords further exploration for gesture-based spatial audio augmentation in mobile multi-party calling scenarios.

2.2 Physical coordinates

When placing virtual audio objects in established locations of the physical world around a user, the coordinate system used for rendering virtual sounds must match with a map of the physical environment. It would be ideally possible to put a virtual audio object near any physical object in the real world. Localized audio messages close to an artwork exposed at the museum as well as introductions to an exhibition are examples of audio guide systems [5]: an acoustic Post-it is binded to a physical coordinate system. A recorded message is played to a visitor when he/she is in a certain location of the museum. The user location and orientation are kept updated and the acoustic features of the building are kept monitored as well, resulting in an very high *DI*, thus an auralized dynamic soundscape and different spoken messages are played through his/her wireless headphones.

The above remarks are particularly relevant for mobile applications and eyes-free mobile control. Navigation aid systems for the visually impaired represent a socially strong use of these technologies. In these applications the map of the physical space can be global or local (the specific room). Two final examples for local physical coordinate systems are virtual auditory displays for air crews on simulated mission flights and collision alarm systems for flight pilots. In these latter applications the associated physical coordinate system is moving with the airplane and in both cases a low *CSD* is obtained, that is, the matching between virtual and physical coordinate systems is the critical task.

3. BINAURAL SOUND REPRODUCTION

Techniques for sound source localization in space follow different approaches [18]. A first distinction regards the sound reproduction method, i.e. the use of loudspeakers opposed to headphone-based systems. Binaural techniques lie between the two groups (binaural reproduction can be obtained either with loudspeakers or headphones [15]) and enables authentic auditory experiences if the eardrums are stimulated by sound signals bearing roughly the same pressure as in real life conditions [10]. Two other approaches belonging to the loudspeaker-only reproduction category are (i) the attempt to recreate the full sound field over a larger listening area (e.g. Wavefield Synthesis technique [9]) and (ii) the intent to introduce just the elements that the auditory system needs in order to perceptually determine the location of the sound (e.g. Ambisonics [14]).

Nevertheless, the use of headphone-based reproduction - onto which this paper focuses on - in conjunction with head tracking devices grants a degree of interactivity, realism, and immersion that is not easily achievable with multichannel systems or wavefield synthesis, due to limitations in the user workspace and to acoustic effects of the real listening space.

3.1 Head-related transfer functions

Head-Related Transfer Functions (HRTFs) capture the transformations undergone by a sound wave in its path from the source to the eardrum, typically due to diffraction and reflections on the head, pinnae, torso and shoulders of the listener. Such characterization allows virtual positioning of sound sources in the surrounding space: consistently with its relative position to the listener's head, the signal is filtered through the corresponding pair of HRTFs creating left and right ear signals to be delivered by headphones [12]. In this way, three-dimensional sound fields with a high immersion sense can be simulated and integrated in mixed reality contexts.

However, recording individual HRTFs of a specific listener requires specific facilities, expensive equipment, and delicate audio treatment processes. These elements make it difficult to use customized HRTFs in virtual environments, considering the high cost of other immersive components (motion tracker, head mounted display and haptic devices). For these reasons generalized HRTFs (i.e dummy-head HRTFs), also called non-individualized HRTFs, are used in some applications resulting, as tolerable drawbacks, in evident sound localization errors such as incorrect perception of the source elevation, front-back reversals, and lack of externalization [21].

A series of experiments were conducted by Wenzel *et al.* [26] in order to evaluate the effectiveness of non-individualized HRTF for virtual acoustic displaying. A very similar perceived horizontal angular accuracy in both real conditions and with 3D sound rendering is obtained by employing generalized HRTFs; however the experiments show that the use of generalized functions increases the rate of reversal errors. Also in this direction, Begault *et al.* [7] compared the effect of generalized and individualized HRTF, with head-tracking and reverberation applied to a speech sound. Their results showed that head tracking is crucial to reduce angle errors and particularly to avoid reversals, while azimuth perception in generic HRTF listening is marginally deteriorated if compared to the individualized one and is balanced by the introduction of artificial reverberation.

To sum up, while non-individualized HRTFs represent a cheap and straightforward mean of providing 3D perception in headphone reproduction, listening to non-individualized spatialized sounds is likely to result in sound localization errors that cannot be fully counterbalanced by additional spectral cues, especially in static conditions. In particular, elevation cues cannot be characterized through generalized spectral features. In conclusion, alongside critical dependence on the relative position between listener and sound source, anthropometric features of the human body have a key role in HRTF characterization.

3.2 Structural models

As one possible alternative to the rendering approach based on directly measured HRTFs or less accurate generic ones, the use of structural models represents an attractive solution to synthesize an individual HRTF or build an enhanced generalized HRTF. The contributions of the listener's head, pinnae, ear canals, shoulders and torso are isolated and arranged in different HRTF subcomponents

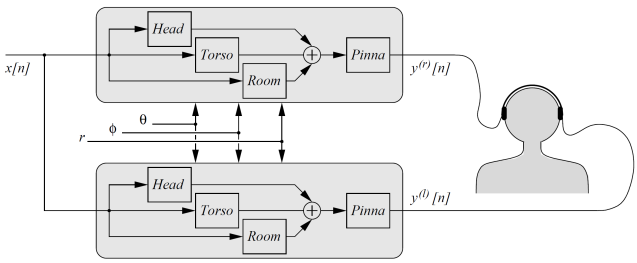


Figure 2: *Generalized 3D audio reproduction system based on a structural HRTF model.*

each accounting for some well-defined physical phenomenon; the linearity of these contributions allows the reconstruction of the global HRTF from a proper combination of all the considered effects. Relating each subcomponent's temporal and/or spectral features (in the form of digital filter parameters) to the corresponding anthropometric quantities would then yield a HRTF model which is both economical and personalizable [11].

Furthermore, 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. The choice of the room model is flexible to the specific application and not only directed at reproducing a realistic room behaviour, but also at introducing some externalization. A synthetic block scheme of a generic structural model is given in Fig. 2.

It is important to point out that the techniques we discuss have minimal hardware requirements as prerequisite with respect to those implied by realistic video reproduction, and in comparison with other technologies adopted to manage immersive sound reproduction such as multichannel systems and wavefield synthesis. A second advantage of the model is the opportunity 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., interaural distances, or pinna shapes) so that a generic structural HRTF model can be adapted to a specific listener, allowing further increase of the quality of audio experience thanks to an enhanced realism of the sound scene.

4. A CUSTOMIZED 3D SOUND SYSTEM

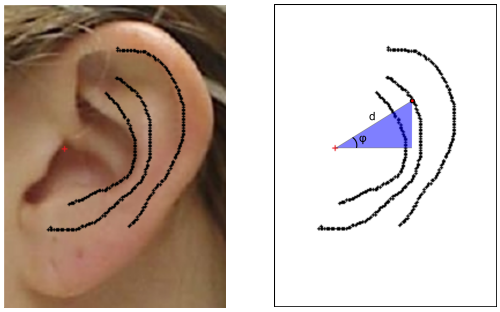
A complete structural approach requires the customization of all the components introduced in Section 3. Our work's focus is on the pinna block and a comprehensive definition of acoustic phenomena is required to understand its relevant characteristics.

Sound waves coming towards a subject's head have to travel an extra distance in order to reach the farthest ear and become acoustically "shadowed" by the presence of the head itself; time and level differences between the two sound signals reaching left and right ears are known as the binaural quantities *ITD* (Interaural Time Difference) and *ILD* (Interaural Level Difference). Therefore, head acoustic effects are modelled using delay lines and low/high-pass filters [11] accordingly with the dimension of the subject's head.

Before entering the ear canal, the sound waves undergo other spectral modifications due the external ear acting as both a sound reflector and a resonator:

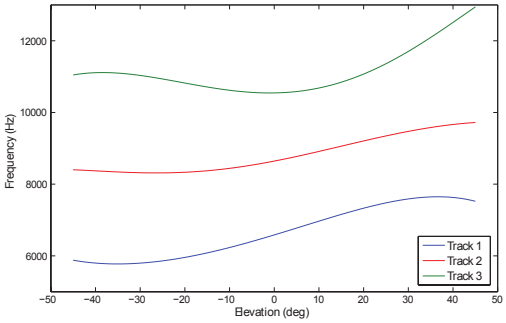
1. reflections over pinna edges. According to Batteau [4], sound waves are typically reflected by the outer ear, as long as their wavelength is small enough compared to the pinna dimensions. The interference between the direct and reflected waves causes sharp notches to appear in the high-frequency domain of the received signal's spectrum;

2. resonant modes in pinna cavities. As Shaw claimed [22], since the concha acts as a resonator some frequency bands of both the direct and reflected sound waves are significantly enhanced. The amplification is correlated to the elevation of the source.



(a) *Contour tracing*

(b) *Mapping to polar coordinates*



(c) *Notch track extraction and approximation*

Figure 3: *Notch frequency extraction from a picture of the pinna (CIPIC Subject* 048*).*

Taking into account these two behaviours we propose a first-stage method to extract the relevant psychoacoustic features from a specific pinna image. Our goal is to define a model that can be easily merged with the several solutions proposed in literature regarding the head, torso, shoulders, and room blocks.

4.1 Pinna-based customization

Fixing the sound source direction with respect to the listener's orientation, the greatest dissimilarities among different people's HRTFs are due to the massive subject-to-subject pinna shape variation.

The pinna's contribution (commonly known as Pinna-Related Transfer Function, PRTF [1]), extrapolated from the HRTF, exhibits a sequence of peaks and notches in its magnitude. In order to conduct a separate analysis of the spectral modifications due to reflections from those due to resonances, we implemented an algorithm (details of which can be found in [16]) that iteratively compensates the PRTF magnitude spectrum with an approximate multi-notch filter until no significant notches are left. Once convergence is reached at iteration n, the PRTF spectrum contains the estimated resonant component, while a combination of the n multi-notch filters provides the reflective component.

Having access to a public or institutional HRTF or PRTF database, the algorithm can be tested on median-plane data in order to conduct a statistical analysis of the peak and notch characteristics. The results shown in this paper are obtained on the CIPIC database [1] using a first order statistic and keeping the procedure simple to extract the relevant features. The detailed filter structure guiding the parameters' extraction is described in [24] and a high-level representation is depicted in Fig. 4.

Given that resonances have a similar behaviour in all of the analyzed PRTFs, customization of this component for the model may be overlooked. The mean magnitude spectrum was instead calculated and analyzed for resynthesis. More in detail, we applied a straightforward procedure that extracts for every available elevation angle the two maxima of the mean magnitude spectrum, which outputs the gain G_p^i, central frequency CF_p^i and the corresponding 3dB bandwidth, BW_p^i, of each resonance peak, $i = 1,2$. Then, a fifth-order polynomial (with the elevation ϕ as independent variable) was fitted to each of the former three parameters, yielding the functions that will be used in the model to continuously control the evolution of the resonant component when the sound source is moving along elevation.

Analysis of the reflective part revealed that while PRTFs generally exhibit poor notch structures when the source is above the head, as soon as elevation decreases the number, spectral location, and depth of frequency notches grows to an extent that differs from subject to subject, and that their evolution could be directly related to the location of reflection points over pinna surfaces. Assuming that the coefficient of all reflections occurring inside the pinna is negative [23], the extra distance travelled by the reflected wave with respect to the direct wave must be equal to half a wavelength in order for destructive interference (i.e. a notch) to occur, which translates into a notch frequency that is inversely proportional to such distance. Hence, under the simplification that the reflection surface is always perpendicular to the soundwave, we consider the mapping function

$$d(\phi) = \frac{c}{2CF_n}, \tag{1}$$

where $d(\phi)$ is the distance of the hypothetical reflection point from the ear canal at elevation ϕ, CF_n is the notch central frequency, and c is the speed of sound.

These results allow us to perform the procedure, sketched in Fig. 3, in order to extract notch frequencies from a representation of the pinna contours. The three most prominent and relevant contours of the pinna are manually traced with the help of a pen tablet. These are translated into a couple of polar coordinates (d,ϕ), with respect to the point where the microphone lied during the HRTF measurements, through simple trigonometric computations. Finally, the notch frequency (CF_n) is derived just by reversing Equation 1 and the sequence of points (CF_n^j,ϕ) for each of the three notch tracks, $j = 1,2,3$, is linearly approximated through a fifth-order polynomial $CF_n^j(\phi)$.

For what concerns the other two parameters defining a notch, i.e. gain G_n and 3dB bandwidth BW_n, there is still no evidence of correspondence with anthropometric quantities. A first-order statistical analysis, subdivided by notch track and elevation, among CIPIC subjects reveals a high variance within each track and elevation; only a slight decrease in notch depth and a slight increase in bandwidth as the elevation increases are reported. In absence of clear elevation-dependent patterns, the mean of both gains and bandwidths for all tracks and elevations ϕ among all subjects is computed, and again a fifth-order polynomial dependent on elevation is fitted to each of these sequences of points, yielding functions $G_n^j(\phi)$ and $BW_n^j(\phi)$, $j = 1,2,3$.

4.2 A novel approach

The proposed model was designed so as to avoid expensive computational and temporal steps such as HRTF interpolation on different spatial locations, best fitting non-individual HRTFs, or the addition of further artificial localization cues, allowing implementation and evaluation in a real-time environment.

A fundamental assumption is introduced, i.e. elevation and azimuth cues are handled orthogonally and the corresponding contributions are thus separated in two distinct parts. The vertical control is associated with the acoustic effects relative to the pinna and the horizontal one is delegated to head diffraction. Indeed, an informal inspection of different HRTF sets reveals that median-plane reflection and resonance patterns usually vary very slowly when the azimuth's absolute value is increased, especially up to about 30°. This approximation encourages us to define customized elevation and azimuth cues that maintain their average behaviour throughout

[1] http://interface.cipic.ucdavis.edu/sound/hrtf.html

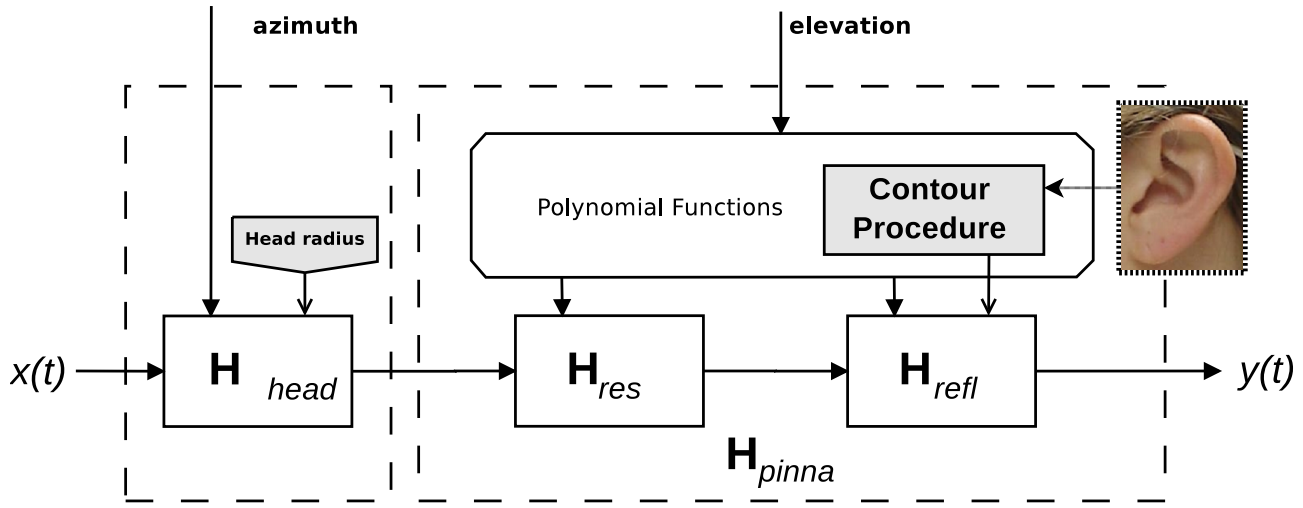


Figure 4: *The customized structural HRTF model.*

the front hemisphere.

The global view of our model is sketched in Fig. 4 and is provided with the input parameters obtained through the analysis procedure described in Section 4.1. A simple spherical model that approximates head shadowing and diffraction, described in [11], is employed, where the head radius parameter a is defined by a weighted sum of the subject's head dimensions considered in [2]. This is only a possibility among the several solutions employing a spherical or ellipsoidal model; even a KEMAR pinna-less response can be put in series before our pinna model. Finally, pinna effects are approximated by a "resonances-plus-reflections" block, following the previous observations and thus allowing elevation control.

Source elevation ϕ is the only independent parameter used by the pinna block and drives the evaluation of the polynomial functions yielding peak and notch spectral forms (parameters: center frequency, 3dB bandwidth and gain). Only the notch center frequency is customized on the individual pinna shape, hence the corresponding polynomial must be computed offline, immediately after taking a couple of photographs, previous to the rendering process.

5. CONCLUSIONS

In this paper we presented a customized structural model of the HRTF that can be used in real-time environments for 3D audio rendering. One of the main advantages of such approach, with respect to high resource investments to support measured HRTFs, is that the model can be parameterized according to anthropometric information of the user. Thus an interesting form of content adaptation represents a key role for innovative auditory experiences. We have presented several application domains and scenarios where these technologies can be applied to HCI research.

Future works on the pinna model are oriented at improving vertical control through the analysis of a 3D representation of the pinna that allows to investigate its horizontal section. The simplified equation 1, on reflection distance calculation, should embed the displacement caused by the flare angle of the pinna, because the pinna structure does not lie on a parallel plane in relation to the head's median plane. This improvement is crucial especially in subjects with protruding ears.

The extensions required to have a full, surrounding binaural experience are leading our research towards a model for source positions behind, above, and below the listener. The increase of body immersion requires the inclusion of the shoulders and torso contribution, adding further reflection patterns and shadowing effects to the overall model, especially when the source is below the listener. However, this preliminary stage can still introduce a real 3D control of a sound source in a number of "frontal" applications, e.g. a sonified screen.

6. REFERENCES

[1] R. V. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson. Structural composition and decomposition of HRTFs. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 103–106, New Paltz, New York, USA, 2001.

[2] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a spherical-head model from anthropometry. *J. Audio Eng. Soc.*, 49(6):472–479, 2001.

[3] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064, 2002.

[4] D. W. Batteau. The role of the pinna in human localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168(1011):158–180, August 1967.

[5] B. B. Bederson. Audio augmented reality: a prototype automated tour guide. In *Conference companion on Human factors in computing systems*, CHI '95, pages 210–211, New York, NY, USA, 1995. ACM.

[6] D. R. Begault. *3-D sound for virtual reality and multimedia*. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[7] D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society.Audio Engineering Society*, 49(10):904–916, Oct 2001.

[8] S. Benford and L. Fahlén. A spatial model of interaction in large virtual environments. In *Proceedings of the third conference on European Conference on Computer-Supported Cooperative Work*, pages 109–124, Norwell, MA, USA, 1993. Kluwer Academic Publishers.

[9] A. J. Berkhout, D. de Vries, and P. Vogel. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

[10] A. W. Bronkhorst. Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, 98(5):2542–2553, 1995.

[11] C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Transactions on Speech and Audio Processing*, 6(5):476–488, 1998.

[12] 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, April 2001.

[13] M. Cohen, S. Aoki, and N. Koizumi. Augmented audio reality: telepresence/vr hybrid acoustic environments. In *Robot and Human Communication, 1993. Proceedings., 2nd IEEE International Workshop on*, pages 361 –364, nov 1993.

[14] R. K. Furness. Ambisonics-an overview. In *Audio Engineering Society Conference: 8th International Conference: The Sound of Audio*, 5 1990.

[15] W. G. Gardner. *3-D audio using loudspeakers*. The Kluwer international series in engineering and computer science, Teil 444. Kluwer Acad. Publ., Boston u.a., 1998. Verfasserangabe: by William G. Gardner ; Quelldatenbank: HBZ ; Format:marcform: print ; Umfang: X, 154 S. : graph. Darst.

[16] M. Geronazzo, S. Spagnol, and F. Avanzini. Estimation and modeling of pinna-related transfer functions. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.

[17] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho. Augmented reality audio for mobile and wearable appliances. *J. Audio Eng. Soc*, 52(6):618–639, 2004.

[18] B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence: Teleoper. Virtual Environ.*, 17:527–549, December 2008.

[19] G. Lorho, J. Hiipakka, and J. Marila. Structured menu presentation using spatial sound separation. In *Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction*, Mobile HCI '02, pages

419–424, London, UK, 2002. Springer-Verlag.

[20] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. pages 282–292, 1994.

[21] H. Møller, M. Sørensen, J. Friis, B. Clemen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc*, 44(6):451–469, 1996.

[22] E. A. G. Shaw. *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter Acoustical features of human ear, pages 25–47. R. H. Gilkey and T. R. Anderson, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1997.

[23] S. Spagnol, M. Geronazzo, and F. Avanzini. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *IEEE International Workshop on Multimedia Signal Processing*, pages 194–199, Saint-Malo, France, October 2010.

[24] S. Spagnol, M. Geronazzo, and F. Avanzini. Structural modeling of pinna-related transfer functions. In *In Proc. Int. Conf. on Sound and Music Computing (SMC 2010)*, 2010.

[25] A. Walker and S. Brewster. Spatial audio in small screen device displays. *Personal and Ubiquitous Computing*, 4:144–154, 2000. 10.1007/BF01324121.

[26] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.