Model-based customized binaural reproduction through headphones

Michele Geronazzo
University of Padova

Simone Spagnol, Davide Rocchesso
IUAV - University of Venice

Federico Avanzini
University of Padova

ABSTRACT

Generalized head-related transfer functions (HRTFs) represent a cheap and straightforward mean of providing 3D rendering in headphone reproduction. However, they are known to produce evident sound localization errors, including incorrect perception of elevation, front-back reversals, and lack of externalization, especially when head tracking is not utilized in the reproduction . Therefore, individual anthropometric features have a key role in characterizing HRTFs. On the other hand, HRTF measurements on a significant number of subjects are both expensive and inconvenient. This short paper briefly presents a structural HRTF model that, if properly rendered through a proposed hardware (wireless headphones augmented with motion and vision sensors), can be used for an efficient and immersive sound reproduction. Special care is reserved to the contribution of the external ear to the HRTF: data and results collected to date by the authors allow parametrization of the model according to individual anthropometric data, which in turn can be automatically estimated through straightforward image analysis. The proposed hardware and software can be used to render scenes with multiple audiovisual objects in a number of contexts such as computer games, cinema, edutainment, and many others.

1. A CUSTOMIZABLE HRTF MODEL

There is no doubt that, if we set the direction of the sound source with respect to the listener, the greatest dissimilarities among different people's HRTFs are due to the massive subject-to-subject pinna shape variation. The external ear plays an important role by introducing peaks and notches in the high-frequency spectrum of the HRTF, whose center frequency, amplitude, and bandwidth greatly depend on the elevation angle of the sound source.The relative importance of major peaks and notches in typical HRTFs in elevation perception has been disputed over the past years; in general, both seem to play an important function in vertical localization of a sound source.

However, in a previous work [1] we highlighted that while the resonant component of the pinna-related counterpart of the HRTF (known as PRTF) is in broad terms similar among different subjects, the reflective component of the PRTF comes along critically subject-dependent. In the same work, we exploited a simple ray-tracing law to show that in median-plane frontal HRTFs (with elevation ranging from $\phi = -45°$ to $\phi = 45°$) the frequency of the spectral notches, each assumed to be caused by its own reflection path, is related to the shape of the concha, helix, and antihelix on the frontal side of the median plane at least. This finding opens the door for a very attractive approach to the parametrization of the HRTF based on individual anthropometry, that is, extrapolating the most relevant parameters that characterize the PRTF just from a 2-D representation of the user's pinna.

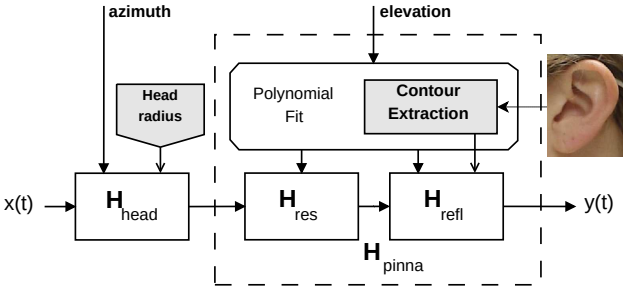

Figure 1. The structural HRTF model. H_{head} is the head filter; H_{res} and H_{refl} are the resonant and reflective components of the pinna model H_{pinna}, respectively.

Following this important result, a complete structural HRTF model that takes into account the user's anthropometry was proposed in [2] and is schematically depicted in Figure 1. In the model, elevation and azimuth cues are handled orthogonally: vertical control is associated with the acoustic effects of the pinna (H_{pinna}) while the horizontal one is delegated to head diffraction (H_{head}). The model is designed so as to avoid expensive computational and temporal steps, allowing implementation and evaluation in a real-time audio processing environment. Two instances of such model, appropriately synchronized through interaural time delay estimation methods, allow for real-time binaural rendering.

The core of the above structure is the pinna model: here two second-order peak filters (filter structure H_{res}) and three second-order notch filters (filter structure H_{refl}) synthesize two resonance modes and three pinna reflections respectively, with the associated parameters either derived from the subject's anthropometry or taken from average measurements on a group of subjects. Obviously, extracting the relevant features from a picture implies a mandatory image processing step. The clearest contours of the pinna and the ear canal entrance must be recognized in order to calculate distances between reflection and observation points and convert them to notch frequencies.

Intensity edge detection techniques applied to a single

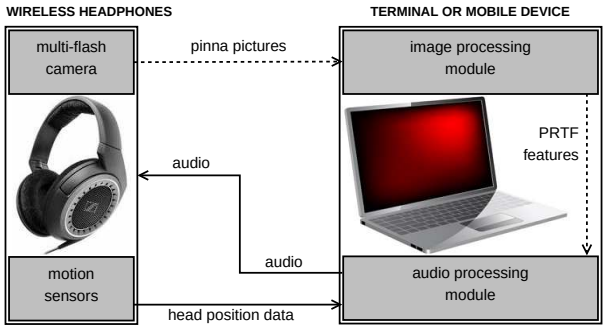

Figure 2. The system's architecture and software.

picture of the pinna are hardly applicable. This task can be instead achieved through a technique known as multi-flash imaging [3]: by using a camera with multiple flashes strategically positioned to cast shadows along depth discontinuities in the scene, the projective-geometric relationship of the setup can be exploited to detect depth discontinuities (in our case, pinna contours) and distinguish them from intensity edges due to color discontinuities.

2. SYSTEM ARCHITECTURE

In order to fully exploit the potential of a customizable HRTF model in both static and dynamic listening scenarios, an appropriate audio device equipped with sensors able to detect the relevant parameters needed to fine tune the model both before and during listening has to be designed. Our idea is that a couple of common wireless headphones augmented through motion sensors and possibly a multi-flash camera could easily fit the goal. Figure 2 illustrates the architecture of the wireless system we are currently realizing, including all of the data exchanges between the headphones and a master device running both the image and audio processing software.

The headphones incorporate a number of sensors (a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis digital compass) able to continuously track the user's head pose in the 3D space thanks to the full 6-DoF motion processing they convey [4]. Moreover, a digital camera equipped with four bright LEDs acting as flash lights and positioned around the camera eye can be slotted inside one of the two cups of the headphones depending on the available space, both inside the cup and between the lens and the ear of the user wearing the headphones. Alternatively, the multi-flash camera shall be proposed as a separate wearable or interchangeable device. Synchronization between each of the flash lights and the related shot is managed by a simple microcontroller.

Storage of the resulting pictures and transmission to the master device is managed through a wireless SD card. The pictures will be received by an image processing program performing the following steps:

- *depth edge detection*: based on the available pictures and their relative differences in shadow and lighting, a depth edge map is computed through the algorithm proposed in [3];
- *pinna contour recognition*: the most prominent contours are extracted among the available depth edges based on both their shape and length and the consistency between their relative positions, and stored as a specifically designed data format;
- *ear canal detection*: the ear canal entrance is approximated by one specific point of the tragus edge;
- *computation of pinna-related features*: distances between the extracted contours and the ear canal entrance are translated into notch frequency parameters through straightforward trigonometric computation, approximated as functions of the elevation angle, and fed to the audio processing module.

This last module, implemented in a real-time environment, systematically receives at each time frame the data from the motion sensors (pitch, roll, and yaw rotations of the head) through radio transmission and translates it into a couple of polar coordinates (azimuth, elevation) of a fixed or moving sound source with respect to the user. The couple of coordinates finally represents the input to the structural HRTF model that performs the spatialization of a desired sound file through the user's customized synthetic HRTFs. This way, provided that the center of rotation of the head does not excessively translate during the rotation (distance between the user and the sound source cannot indeed be tracked in real time by the available sensors), the user will perceive the position of the virtual sound source as being independent from his or her movement.

The HRTF model currently includes a large portion of the frontal hemispace and the proposed architecture could thus be suitable for real-time control of virtual sources in a number of applications involving frontal auditory displays, such as a sonified screen. Further extensions of the HRTF model, capable of including source positions behind, above, and below the listener, may be obtained in different ways, and will be objects of future research.

3. REFERENCES

[1] S. Spagnol, M. Geronazzo, and F. Avanzini, "Fitting pinna-related transfer functions to anthropometry for binaural sound rendering," in *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP'10)*, (Saint-Malo, France), pp. 194–199, October 2010.

[2] M. Geronazzo, S. Spagnol, and F. Avanzini, "A head-related transfer function model for real-time customized 3-D sound rendering," in *Proc. INTERPRET Work., SITIS 2011 Conf.*, (Dijon, France), pp. 174–179, November-December 2011.

[3] R. Raskar, K.-H. Tan, R. Feris, J. Yu, and M. Turk, "Non-photorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging," *ACM Trans. Graphics (Proc. SIGGRAPH)*, vol. 23, no. 3, pp. 679–688, 2004.

[4] S. Nasiri, S.-H. Lin, D. Sachs, and J. Jiang, "Motion processing: the next breakthrough function in handsets," 2010.