Extraction of Pinna Features for Customized Binaural Audio Delivery on Mobile Devices

Simone Spagnol,
Michele Geronazzo
Department of Information Engineering
University of Padova
Padova, Italy
{spagnols,geronazzo}@dei.unipd.it

Davide Rocchesso
Iuav University of Venice
Venice, Italy
roc@iuav.it

Federico Avanzini
Department of Information Engineering
University of Padova
Padova, Italy
avanzini@dei.unipd.it

ABSTRACT

The paper presents a system for customized binaural audio delivery based on the extraction of the relevant features from a 2-D representation of the listener's pinna. A procedure based on multi-flash imaging for recognizing the main contours of the pinna and their position with respect to the ear canal entrance is detailed. The resulting contours drive the parametrization of a structural head-related transfer function model that performs in real time the spatialization of a desired sound file according to the listener's position with respect to the virtual sound source, tracked by sensor-equipped headphones. The low complexity of the model allows smooth implementation and delivery on any mobile device. The purpose of the desired system is to provide low-tech custom binaural audio to any user without the need of tedious and cumbersome subjective measurements.

Categories and Subject Descriptors

H.5.5 [**Information Interfaces and Presentation (e.g., HCI)**]: Sound and Music Computing—*Modeling, Signal analysis, synthesis, and processing*; H.5.1 [**Information Interfaces and Presentation (e.g., HCI)**]: Multimedia Information Systems—*Artificial, augmented, and virtual realities*

General Terms

Algorithms, Design

Keywords

spatial audio, binaural, HRTF, pinna

1. WHY CUSTOMIZE SPATIAL AUDIO?

In recent years spatial sound has become increasingly important in a plethora of application domains. Spatial audio models are able to provide information about the relationship between the sound source and the surrounding environment including the listener and his/her body which acts as a further filter, an information which is hardly imitable by visual or tactile displays. Among the diverse possibilities offered by spatial audio technologies, binaural (i.e., headphone-based) reproduction systems - if properly designed - allow tailoring immersive and realistic auditory scenes to any user without the need of loudspeaker-based systems. These technologies go well with mobile devices in particular:

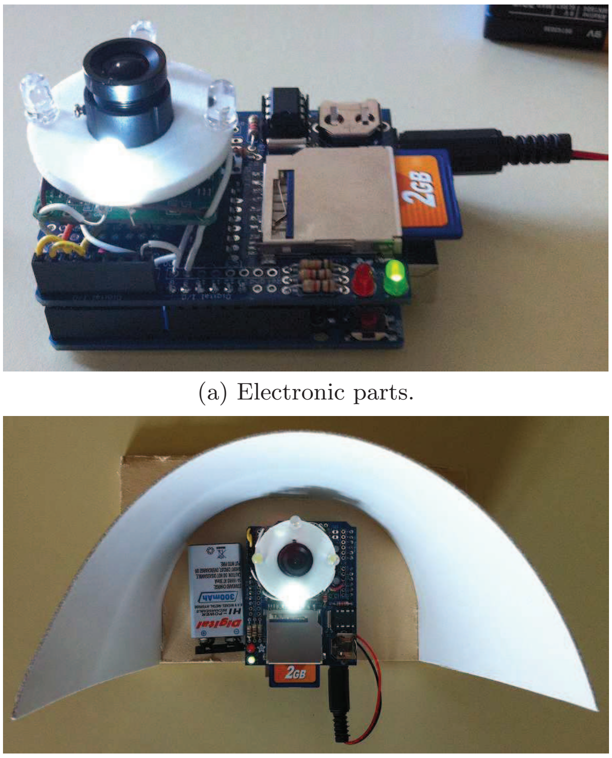
1. no integrated loudspeaker can guarantee the same audio quality as a pair of common earbuds or headphones;
2. headphones are typically used on the move without the need to hold anything in hand.

Binaural audio rendering approaches are typically based on the concept of head-related transfer function (HRTF) [3]. HRTFs capture the transformations undergone by a sound wave in its path from the source to the eardrum, and in particular those caused by diffraction, reflection, and resonance effects onto the torso, head, shoulders and pinnae of the listener. Such characterization allows virtual positioning of a number of sound sources in the surrounding space by filtering the corresponding signals through a pair of HRTFs, thus creating left and right ear signals to be delivered by headphones. In this way, three-dimensional sound fields with a high immersion sense can be simulated.

Non-individual HRTF sets, typically recorded by using dummy heads, are known to produce evident sound localization errors [9]. On the other hand, obtaining personal HRTF data for a vast number of users is simply unpracticable because specific hardware, anechoic spaces, and long collection times are strictly required. Structural HRTF modeling [1] represents an attractive solution to all of these shortcomings. By isolating the effects of different components (head, pinnae, ear canals, shoulders, torso), and modeling each one of them with a corresponding filtering element, the global HRTF is approximated through a proper combination of all the considered effects. Moreover, by relating the temporal/spectral features of each component to corresponding anthropometric quantities, one can in principle obtain a HRTF representation that is both computationally economical and customizable.

A complete structural filter model of the HRTF is currently being studied by the authors [14, 13, 4, 5]. In the

MoMM2013, 2-4 December, 2013, Vienna, Austria
Copyright 2013 ACM 978-1-4503-2106-8/13/12 ...$15.00.



(a) Electronic parts.

(b) Full prototype.

Figure 1: The multi-flash camera prototype.

model, special care is reserved to the contribution of the external ear to the HRTF: data and results collected to date allow in particular the development of a parametric HRTF model customizable according to individual anthropometric data, which in turn can be automatically estimated through straightforward image analysis. This means that by feeding such model with a suitable characterization of the listener's anthropometry, and by rendering the resulting audio stream through motion tracked headphones, low-tech custom binaural audio can be delivered in real time on a standard mobile device. This paper explains in particular how custom audio streams can be derived from a set of distinctive pictures of the listener's pinnae, with a special focus on the extraction of the relevant parameters for HRTF customization.

2. EXTRACTION OF PINNA FEATURES

There is no doubt that, if we fix 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 [3]. 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. Unfortunately, although we know that these peaks and notches are responsible for vertical localization ability [7], their relation with anthropometry has not been fully understood yet.

However, a previous work of ours [12] highlighted that while the resonant component of the pinna-related counterpart of the HRTF (known as PRTF) is similar among different subjects, the reflective component of the PRTF comes along critically subject-dependent. In the same context, and more rigorously in a following work [14], we exploited a simple ray-tracing law to show that in median-plane frontal HRTFs (with elevation ranging from $\phi = -45^\circ$ to $\phi = 45^\circ$) 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 opens the path for a very attractive approach to the parametrization of the HRTF based on individual anthropometry: extrapolating the most relevant parameters that characterize the PRTF just from one or more pictures of the user's pinna.

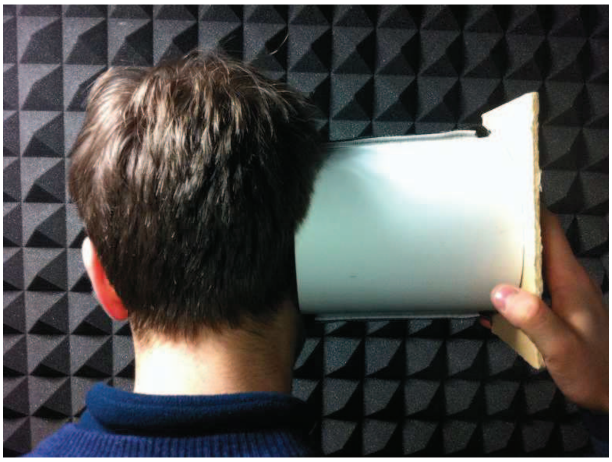


Figure 2: Acquisition of the four pinna pictures.

Clearly, extracting the relevant features from a 2-D representation of the pinna implies a mandatory image processing step. The clearest contours of the pinna as well as the ear canal entrance must be recognized in order to calculate distances between reflection and observation points and translate them into notch frequencies.

Intensity edge detection techniques applied to a single picture of the pinna are hardly exploitable. In particular, the Canny method [2] is known to fail in low-contrast areas such as the pinna, especially in those cases where shadows are not projected below the considered edge. This task can be instead achieved through a technique known as multi-flash imaging [10]: by using a camera with multiple flashes strategically positioned to cast shadows along depth discontinuities in the scene, the projective-geometric relationship of the camera-flash setup can be exploited to detect depth discontinuities (in our case, pinna contours) and distinguish them from intensity edges due to color discontinuities.

In order to investigate the potential of such technique, a multi-flash camera prototype was custom built by the authors. The device, pictured in Fig. 1, is composed of a TTL serial JPEG camera connected to a battery-powered Arduino UNO microcontroller board equipped with a data logging shield. Four Super Bright White LEDs are symmetrically positioned around the camera eye and can be turned on independently. The electronic components are secured to a rigid board and enclosed in a hemi-cylindrical PVC foil mimicking the pinna helix shape (see Fig. 1(b)). Finally, because a dark environment is desirable when shooting such kind of pictures, the open side of the arc can be closed by a black silk cut with Velcro fastening strips.

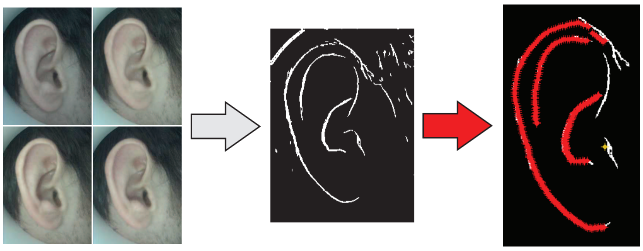


Figure 3: Contour extraction procedure. Left: the four pinna pictures. Middle: the depth edge map. Right: the extracted contour points (red) and the ear canal point (orange).

Acquisition of the required pinna pictures is managed as follows. As depicted in Fig. 2, the subject presses the open top side of the device right around the left or right pinna trying to align the hemi-cylinder with the helix. The shape of the device affords correct orientation as referred to the outer ear. An Arduino sketch takes a set of four pictures, each synchronized with a different light flash. Because of the required storage time this basic procedure takes approximately 30 seconds, during which the subject should try to keep the device as still as possible with respect to his/her pinna. The four pictures, stored in a wireless SD card as 320×240 pixel .jpg files, are ready to be transmitted with low latency to a smart device.

An image processing algorithm is then able to recognize and separate the contours which are of interest to us and straightforwardly calculate the associated notch frequencies as functions of the elevation angle of the sound source. In order to do so, the four pictures (left side of Fig. 3) need to be fed to a collection of scripts automatically performing the following steps:

- *motion correction*: in order to compensate for subject motion, the four pictures are first rotated and then translated for the best possible relative alignment according to a standard 2-D correlation function;
- *depth edge detection*: based on the four pictures and their relative differences in shadow and lighting, a *depth edge map* is computed through the algorithm proposed in [10] as a binary matrix whose white pixels represent the most prominent depth discontinuities (middle panel of Fig. 3);
- *map refinement*: only the connected components containing at least 100 pixels are kept in the depth edge map;
- *ear canal detection*: the connected component corresponding to the tragus edge is isolated; the ear canal entrance is taken as the darkest point of one of the initial pictures falling in the tragus edge's bounding box;
- *contour tracking*: for each desired elevation angle ϕ, all the 0→1 transitions in the depth edge map along the ray originating from the ear canal point and heading towards the pinna with $-\phi$ inclination are stored as distances in pixels. Then, a partial tracking algorithm [8] (originally used to group sinusoidal partials along consecutive temporal windows according to their spectral location) is exploited to track the three longest contours (concha wall, helix inner border, and helix outer border in increasing order of distance) along elevation, where distance values take the role of partials (right side of Fig. 3);
- *computation of pinna-related features*: the three distance tracks are translated into notch frequency parameters through a simple linear law and approximated as functions of the elevation angle.

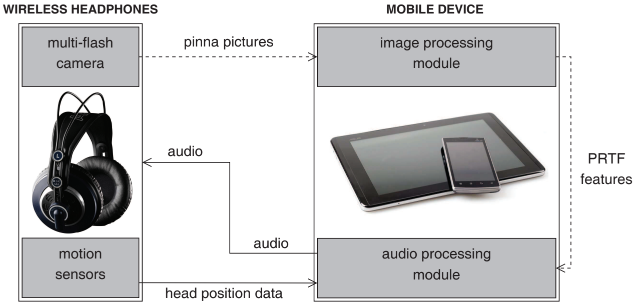


Figure 4: A simplified scheme of the system's architecture and software. Broken line arrows refer to offline data exchange, solid line arrows to real-time data exchange.

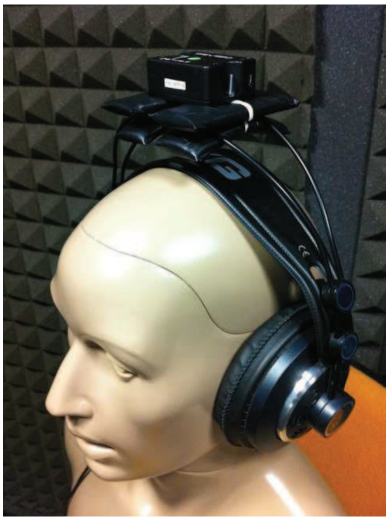


Figure 5: Headphones augmented with a head pose tracker.

3. THE REAL-TIME SYSTEM

The subsequent steps are schematized in Fig. 4. Once the notch frequency parameters are available, they are fed to our structural HRTF model. The model, whose details can be found in [14, 4], is 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 audio processing environment such as Pure Data. Two appropriately synchronized instances (one per ear) of the model allow for real-time binaural rendering.

In order to fully exploit the potential of such a 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 is needed. A pair of common headphones augmented through motion sensors, as the one pictured in Fig. 5 (AKG K240 MKII), easily fits to such a goal. The Trivisio Colibri wireless motion tracker installed on top of the headphones incorporates indeed a number of sensors (a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis digital compass) able to track the 3-D orientation of the user's head thanks to the 6-DoF motion processing they convey.

Data from the motion sensors (pitch, roll, and yaw rotations of the head) are sent in real time by radio transmission to the audio processing module and translated into a couple of polar coordinates (θ, ϕ) of a fixed or moving sound source. These coordinates finally represent the input to the structural HRTF model that performs the convolution between a desired sound file and 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 ideally perceive the position of the virtual sound source as being independent from his or her movement.

4. FUTURE DEVELOPMENTS

The structural model as it currently is represents a notable extension of the only other customizable pinna model available in the literature and described in [11] as it includes a large portion of the frontal hemispace, and is thus suitable for real-time control of virtual sources in a number of applications involving frontal auditory displays, such as a sonified mobile screen [15]. Further extensions of the model, such as to include source positions behind, above, and below the listener, will be objects of future research.

The presented real-time system is however currently lacking of a solid implementation of the theorized components of the structural HRTF model, onto which we are currently working. Once this fundamental component is integrated, extensive listening sessions will attest the degree of accuracy and realism of the presented 3-D audio scenes. Still, the multi-flash device and head tracker are currently being used for anthropometry-based HRTF selection tests [6].

Many improvements can be done at design level. First, we will switch to a completely wireless system through the use of a pair of wireless headphones. Fast shooting is also desired to reduce the duration of the picture acquisition routine down to a few seconds and make motion correction become much less critical. We will also consider whether a smaller and more compact version of the multi-flash camera device can be slotted inside one of the two headphones' cups if space (both inside the cup and between the lens and the ear of the user wearing the headphones) permits. A reasonable compromise would be proposing the multi-flash as a separate yet wearable device. Ultimately, in-place sensing and processing on a mobile device having all the required motion and image sensors represents in perspective the optimal solution in terms of cost and flexibility. The resulting hardware/software package will hopefully allow an easy and low-tech fruition of custom spatial audio to any user.

5. REFERENCES

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

[2] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-8(6):679–698, November 1986.

[3] 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.

[4] 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.*, pages 174–179, Dijon, France, November-December 2011.

[5] M. Geronazzo, S. Spagnol, and F. Avanzini. Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, Santorini, Greece, July 2013.

[6] M. Geronazzo, S. Spagnol, and F. Avanzini. A modular framework for the analysis and synthesis of head-related transfer functions. In *Proc. 134th Conv. Audio Eng. Soc.*, number 8882, Rome, Italy, May 2013.

[7] J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.*, 56(6):1829–1834, December 1974.

[8] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Process.*, 34(4):744–754, August 1986.

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

[10] R. Raskar, K.-H. Tan, R. S. Feris, J. Yu, and M. Turk. Non-photorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 23(3):679–688, August 2004.

[11] P. Satarzadeh, R. V. Algazi, and R. O. Duda. Physical and filter pinna models based on anthropometry. In *Proc. 122nd Conv. Audio Eng. Soc.*, pages 718–737, Vienna, Austria, May 2007.

[12] 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)*, pages 194–199, Saint-Malo, France, October 2010.

[13] S. Spagnol, M. Geronazzo, and F. Avanzini. Structural modeling of pinna-related transfer functions. In *Proc. 7th Int. Conf. Sound and Music Computing (SMC 2010)*, pages 422–428, Barcelona, Spain, July 2010.

[14] S. Spagnol, M. Geronazzo, and F. Avanzini. On the relation between pinna reflection patterns and head-related transfer function features. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(3):508–519, March 2013.

[15] A. Walker and S. Brewster. Spatial audio in small screen device displays. *Pers. Technol.*, 4(2):144–154, June 2000.