Abstract

Purpose
The paper presents a system for customized binaural audio delivery based on the extraction of relevant features from a 2-D representation of the listener's pinna.

Design/methodology/approach
The most significant pinna contours are extracted by means of multi-flash imaging, and they provide values for the parameters of a structural head-related transfer function (HRTF) model. The HRTF model spatializes a given sound file according to the listener's head orientation, tracked by sensor-equipped headphones, with respect to the virtual sound source.

Findings
A preliminary localization test shows that the model is able to statically render the elevation of a virtual sound source better than non-individual HRTFs.

Research limitations/implications
Results encourage a deeper analysis of the psychoacoustic impact that the individualized HRTF model has on perceived elevation of virtual sound sources.

Practical implications
The model has low complexity and is suitable for implementation on mobile devices. The resulting hardware/software package will hopefully allow an easy and low-tech fruition of custom spatial audio to any user.

Originality/value
We show that custom binaural audio can be successfully deployed without the need of cumbersome subjective measurements.

Keywords: 3D audio, spatial sound, binaural, HRTF, pinna, auditory localization, headphones

Article Classification: Research paper

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 body which acts as a further filter (Pulkki *et al.*, 2011). 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, as 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 (Cheng and Wakefield, 2001), or HRTF. 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 can be simulated.

Non-individual HRTF sets, typically recorded by using dummy heads, are known to produce evident sound localization errors (Møller *et al.*, 1996). On the other hand, obtaining personal HRTF data for a vast number of users is simply impractical because specific hardware, anechoic spaces, and long collection times are strictly required. Structural HRTF modeling (Brown and Duda, 1998) 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 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 (Spagnol *et al.*, 2013*a*; Geronazzo *et al.*, 2013*a*). In the 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 (Spagnol *et al.*, 2010*b*). 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 (Section 2), and delivered to the listener

(Section 3). Furthermore, static localization performances of the HRTF model are assessed through a preliminary subjective test (Section 4).

2 Extraction of pinna features

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 (Cheng and Wakefield, 2001). 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 (Hebrank and Wright, 1974), their relation with anthropometry has not been fully understood yet.

However, a previous work of ours (Geronazzo *et al.*, 2011) 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 (Spagnol *et al.*, 2013*a*), we exploited a simple ray-tracing law to show that in median-plane frontal HRTFs (with elevation ranging from $\varphi = -45°$ to $\varphi = 45°$) 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 opens the path for a very attractive approach to the parametrization or selection of HRTFs based on individual anthropometry: extrapolating the most relevant parameters that characterize the pinna's acoustic contribution just from one or more pictures of the user's external ear.

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 (Canny, 1986) is known to fail in low-contrast areas such as the pinna. This task can be instead achieved through a technique known as multi-flash imaging (Raskar *et al.*, 2004): 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.

(a) (b)

Figure 1*: The multi-flash camera prototype. (a) Electronic parts; (b) Full prototype.*

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 mobile device.

Figure 2*: Acquisition of the four pinna pictures.*

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 (Raskar *et al.*, 2004) 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 $-\varphi$ inclination are stored as distances in pixels. Then, a partial tracking algorithm (McAulay and Quatieri, 1986), 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 φ, where distance values take the role of partials (right side of **Fig. 3**);

- ***computation of pinna-related features***: the three distance tracks d_i, i = 1, 2, 3, are translated into notch frequency parameters through the following linear law (Spagnol *et al.*, 2010*a*):

$$f_n^i(\varphi) = \frac{c}{2d_i(\varphi)}$$

where c is the speed of sound, and approximated as fifth-order polynomial functions of the elevation angle F_n^i, i = 1, 2, 3.

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).*

The effectiveness of the multi-flash device in extracting pinna contours has been investigated in a previous work (Spagnol *et al.*, 2013*b*). Results on 30 subjects revealed that the ear canal point is always correctly identified, and that 93.3% of the edge tracks are correctly extracted. The device is currently being used for anthropometry-based HRTF selection tests (Geronazzo *et al.*, 2013*b*); in particular, HRTF selection procedures for static acoustic scenes (Geronazzo *et al.*, 2014) have been evaluated using the same technological platform.

3 The real-time system

In order to exploit the above discussed relation between anthropometry and HRTF features, we propose the modular system sketched in **Fig. 4**. The system includes the multi-flash camera described in the previous section, proposed as a stand-alone yet wearable device. In the realization of this system, the main technical challenge lies in the exchange of the relevant data between headphones and mobile device.

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.*

If we switch from a static to a dynamic environment where the source moves with respect to the listener and/or *vice versa*, both source direction and distance perception become much eased. The tendency to point towards the sound source in order to minimize interaural differences, even without visual aid, is commonly seen and openly disambiguates any front/back confusion (Wightman and Kistler, 1999). Active motion helps especially in azimuth estimation and to a lesser extent in elevation estimation (Thurlow and Runge, 1967). Furthermore, thanks to the *motion parallax* effect, slight translations of the listener's head on the horizontal plane can help discriminating source distance: if the source is near, its angular direction will drastically change after the translation, while for a distant source this will not happen.

A pair of common headphones augmented through motion sensors, as the one pictured in **Fig. 5** (AKG K240 MKII), can be used to control the parameters of the structural HRTF model. 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.

Figure 5*: Headphones augmented with a head pose tracker.*

3.1 Head tracking

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 represent the input to the binaural audio engine 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.

3.2 The model

The output of the pinna feature extraction algorithm implemented in Sec. 2 is fed to the structural HRTF model whose global view is sketched in **Fig. 6**. A fundamental assumption is introduced, i.e. elevation and azimuth cues are handled orthogonally and the corresponding contributions are thus separated. Vertical control is associated to the acoustic effects of the pinna and horizontal control is delegated to head diffraction. Indeed, median-plane HRTF patterns vary very slowly when azimuth increases, especially up to about 30° (Lopez-Poveda and Meddis, 1996).

Figure 6*: The structural HRTF model. The anthropometric extraction procedure on external ear pictures allows an elevation-dependent parametric tuning of the pinna filter block* $\boldsymbol{H_{pinna}}$*, composed of two peak filters and three notch filters. The subject's head radius allows instead an azimuth-dependent tuning of filter* $\boldsymbol{H_{head}}$*.*

As for azimuth effects, a simple spherical head model (Brown and Duda, 1998) is employed for filter $\boldsymbol{H_{head}}$, where the head radius parameter is defined by a weighted sum of the subject's head dimensions (Algazi *et al.*, 2001*a*). Pinna effects are instead

approximated by the more complex block $\boldsymbol{H_{pinna}}$ composed of two second-order peak filters $\boldsymbol{H_p^1}$, $\boldsymbol{H_p^2}$ and three second-order notch filters $\boldsymbol{H_n^1}$, $\boldsymbol{H_n^2}$, $\boldsymbol{H_n^3}$ (Dutilleux *et al.*, 2011) approximating the two main resonances and three main notches that can be commonly seen in typical PRTFs. The three input parameters of each filter describe, for each peak or notch, the desired center frequency ($F_{p/n}$), gain ($G_{p/n}$), and 3-dB bandwidth ($B_{p/n}$). Source elevation is the only independent parameter used by the pinna block and drives the evaluation of a number of polynomial functions each associated to a single filter parameter. Only the center frequency F_n of the three notches is currently mapped from the individual pinna picture as described in the previous section; all of the other parametric functions must be chosen *a priori*.

The model 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.

4 A preliminary localization test

In order to highlight the potential of the structural HRTF model, and in particular of its pinna block, we performed a preliminary localization test on 5 subjects (all male, age 23 to 41). All subjects had previous experience with binaural audio and reported normal hearing according to an adaptive maximum likelihood procedure (Green, 1993). Since our goal was to test the performance of the HRTF model independently of the other components of the system, in the following experiment we neither considered automatic extraction of pinna contours nor head tracking. The aim of the experiment was twofold: (1) to test whether the model can render the elevation of a sound source better than a non-individual HRTF set; and (2) to ascertain the importance of individual pinna contour extraction.

4.1 Stimuli

All stimuli used as sound source signal a train of three 300-ms uniform white noise bursts with 25-ms onset and offset ramps and 250 ms of silence between consecutive bursts. The average measured amplitude of the raw stimulus at the entrance of the ear canal was 60 dB(A). The signal was filtered through a headphone compensation filter (Lindau and Brinkmann, 2012) and applied to measured responses of a KEMAR mannequin without pinnae. It has to be highlighted that compensation was not individual; however, such kind of processing offers an effective equalization of the headphone up to 8−10 kHz on average.

Three distinct categories of experimental stimuli were then created, corresponding to the three experimental conditions of the listening test:

- condition $\boldsymbol{C_1}$: stimuli filtered through the structural HRTF model with individual pinna contour parametrization;
- condition $\boldsymbol{C_2}$: stimuli filtered through the structural HRTF model with non-individual pinna contour parametrization;
- condition $\boldsymbol{C_3}$: stimuli filtered through a pair of HRTFs measured on a KEMAR mannequin (Burkhard and Sachs, 1975).

The HRTFs used in condition $\boldsymbol{C_3}$ were taken from the CIPIC HRTF database (Algazi *et al.* 2001*b*). As for conditions $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$, parametric functions F_n^i, *i* = 1, 2, 3, directly derived from the manual extraction of pinna contours from an individual left pinna picture ($\boldsymbol{C_1}$) or from a picture of a KEMAR mannequin left pinna ($\boldsymbol{C_2}$). Constant functions G_n^i = −30 dB and B_n^i = 2 kHz, *i* = 1, 2, 3, were chosen to describe notch parameters not related to pinna anthropometry. The choice of the exact values was performed according to previous studies on the audibility and discrimination of spectral notches in the high-frequency range (Moore *et al.*, 1989; Poon and Brugge, 1993; Alves-Pinto and Lopez-Poveda, 2005). Parametric functions related to the two peaks F_p^i, G_p^i and B_p^i, *i* = 1, 2, were instead fitted to the magnitude spectrum of the pinna resonant component averaged on all 45 CIPIC subjects' left-ear responses (Geronazzo *et al.*, 2011).

Both conditions $\boldsymbol{C_2}$ and $\boldsymbol{C_3}$ were included as control conditions. As a matter of fact, the comparison between $\boldsymbol{C_1}$ and $\boldsymbol{C_3}$ is needed to test whether the model can render elevation better than a non-individual HRTF set, while results of the comparison between $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ reflect onto the psychoacoustic impact of having individual parameters.

4.2 Setup and protocol

Acquisition of pinna images was the first step. We created an ad-hoc capture environment in order to acquire a left side-face picture of the experimental subject. In a second phase, the picture was first rotated in order to horizontally align the tragus with the nose tip; then, the maximum protuberance of the tragus was chosen as the ear canal point. The three main pinna contours were manually traced and then used to calculate scaled distances from the ear canal point and consequently parametric functions F_n^i, *i* = 1, 2, 3, as previously described. The subject then entered a Sound Station Pro 45 silent booth and wore a pair of Sennheiser HDA 200 headphones plugged to a Roland Edirol AudioCapture UA-101 external audio card working at 44.1 kHz sampling rate.

Nine different stimuli per condition, each repeated six times during the test for a total of 9×3×6 = 162 trials, were presented to the experimental subject. These corresponded to 9 different elevation values from −45° to 45° in 11.25°-steps in the frontal half of the median plane. We chose to consider the median plane because, as already mentioned, relative azimuthal variations up to at least $\Delta\theta$ = 30° at fixed elevation cause very slight spectral changes in the HRTF, hence we expect pinna responses in this region to be elevation-dependent only. Elevations higher than 45° were instead discarded because of both the high degree of uncertainty in elevation judgments (Blauert, 1983) and the general lack of spectral notches in the corresponding HRTFs (Spagnol *et al.*, 2011).

The experiment was structured in six different blocks of trials each corresponding to one single condition and three repetitions of the nine stimuli, proposed in random order. The sequence of presentation of the blocks followed a latin-square design. In order to reduce fatigue of the subject, we added a 30-second pause between blocks. Subjects were instructed to enter the elevation judgment right after each sound presentation in a GUI designed in MATLAB (see **Fig. 7**) and inspired to the one appearing in a previous work (Begault *et al.*, 2001). Perceived elevation of the sound source was selected by placing a point in one of the two green portions of a circular ring surrounding a side view of a human head profile. Each portion, one in front and one on the back of the profile, spanned elevations from −45° to 45°. The back portion was added in order to take into account possible front/back reversals, that typically occur in localization tasks with non-individual HRTF sets (Wenzel *et al.*, 1993; Møller *et al.*, 1996).

***Figure 7**: The experimental GUI.*

4.3 Results

Fig. 8 reports the elevation results of the five subjects divided by experimental condition as scatterplots. Localization errors in elevation were analyzed with front/back reversals resolved. The average localization error for the three conditions, along with front/back and up/down reversal rates (the up/down reversal rate is calculated with a tolerance of 15° in elevation angle around the horizontal plane, and averaged over all target elevations except $\varphi = 0°$), are shown in **Table 1**.

***Figure 8**: Elevation scatterplots of the five subjects, with target and perceived elevation on x- and y-axes respectively. Black lines represent ideal response curves.*

Table 1: *Results of the localization test.*

Results show that for all subjects except one, condition C_1 scores the lowest elevation error. This represents an average improvement of 23.7% compared to C_2 and of 15.4% compared to C_3, with subjective peaks (subject SB) of 35.1% and 35.7% respectively. Such a result confirms that pinna contours have high significance for elevation cues, and that the proposed mapping for the structural HRTF model effectively succeeds in decreasing the elevation error with respect to non-individual HRTFs. Conversely, C_2 scores the highest elevation error in 4 cases out of 5, suggesting that the model cannot guarantee acceptable elevation performances when individual parametrization is not implemented.

The individually parameterized model also succeeds in remarkably improving the up/down reversal rates compared to the non-individual HRTF set: up/down reversals in C_1 are on average halved with respect to C_3. Front/back reversal rates deserve instead a separate analysis. If we exclude the two special cases of subjects SA and SB, who always perceive virtual sources in the front and in the back hemisphere respectively, the other subjects show different trends. A surprising result is that subject SD, who very often perceives stimuli as coming from the back in both conditions C_2 and C_3, almost eliminates front/back reversals when his individual contours are used.

5 Conclusions

The structural HRTF model as it currently is represents a notable extension of the only other customizable pinna model available in the literature (Satarzadeh *et al.*, 2007) 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 (Walker and Brewster, 2000). Further extensions of the model may include source projection behind, above, and below the listener.

We are currently working towards a solid and reliable implementation of the structural HRTF model. The localization test described in this paper serves as the first attempt to subjectively evaluate the degree of accuracy of the presented binaural audio system. Although referring to a preliminary test, these results encourage a deeper analysis of the psychoacoustic impact that the individualized HRTF model has on perceived elevation of virtual sound sources. The reliability of the structural HRTF model and its integration in the proposed real-time system will be assessed in the next stages of the evaluation process.

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. 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.

References

Algazi, V. R., Avendano, C. and Duda, R. O. (2001*a*), "Estimation of a spherical-head model from anthropometry", *Journal of the Audio Engineering Society*, Vol. 49 No. 6, pp. 472–479.

Algazi, V. R., Duda, R. O., Thompson, D. M. and Avendano, C. (2001*b*), "The CIPIC HRTF database", in *Proceedings of the 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, pp. 1–4.

Alves-Pinto, A. and Lopez-Poveda, E. A. (2005), 'Detection of high-frequency spectral notches as a function of level', *Journal of the Acoustical Society of America*, Vol. 118 No. 4, pp. 2458–2469.

Begault, D. R., Wenzel, E. M. and Anderson, M. R. (2001), "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*, Vol. 49 No. 10, pp. 904–916.

Blauert, J. (1983), *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press, Cambridge, MA, USA.

Brown, C. P. and Duda, R. O. (1998), "A structural model for binaural sound synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 6 No. 5, pp. 476–488.

Burkhard, M. D. and Sachs, R. M. (1975), "Anthropometric manikin for acoustic research", *Journal of the Acoustical Society of America*, Vol. 58, No. 1, pp. 214–222.

Canny, J. (1986), "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 8 No. 6, pp. 679–698.

Cheng, C. I. and Wakefield, G. H. (2001), "Introduction to head-related transfer functions (HRTFs): Representations of HRTFs in time, frequency, and space", *Journal of the Audio Engineering Society*, Vol. 49 No. 4, pp. 231–249.

Dutilleux, P., Holters, M., Disch, S. and Zölzer, U. (2011), "Filters and delays", in Zölzer, U. (Ed.), *DAFX: Digital Audio Effects*, John Wiley and Sons Ltd., Chichester, UK, pp. 47–81.

Geronazzo, M., Spagnol, S. and Avanzini, F. (2011), "A head-related transfer function model for real-time customized 3-D sound rendering", in *Proceedings of the 7th International Conference on Signal Image Technology & Internet Based Systems*, Dijon, France, pp. 174–179.

Geronazzo, M., Spagnol, S. and Avanzini, F. (2013*a*), "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery", in *Proceedings of the18th International Conference on Digital Signal Processing*, Santorini, Greece.

Geronazzo, M., Spagnol, S. and Avanzini, F. (2013*b*), "A modular framework for the analysis and synthesis of head-related transfer functions", in *Proceedings of the 134th Audio Engineering Society Convention*, Rome, Italy, paper no. 8882.

Geronazzo, M., Spagnol, S., Bedin, A. and Avanzini, F. (2014), "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions", in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, Firenze, Italy, pp. 4496-4500.

Green, D. M. (1993), "A maximum-likelihood method for estimating thresholds in a yes-no task", *Journal of the Acoustical Society of America*, Vol. 93 No. 4, pp. 2096–2105.

Hebrank, J. and Wright, D. (1974), "Spectral cues used in the localization of sound sources on the median plane", *Journal of the Acoustical Society of America*, Vol. 56 No. 6, pp. 1829–1834.

Lindau, A. and Brinkmann, F. (2012), "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings", *Journal of the Audio Engineering Society*, Vol. 60 No. 1/2, pp. 54–62.

Lopez-Poveda, E. A. and Meddis, R. (1996), "A physical model of sound diffraction and reflections in the human concha", *Journal of the Acoustical Society of America*, Vol. 100 No. 5, pp. 3248–3259.

McAulay, R. J. and Quatieri, T. F. (1986), "Speech analysis/synthesis based on a sinusoidal representation", *IEEE Transactions on Acoustics., Speech, and Signal Processing*, Vol. 34 No. 4, pp. 744–754.

Møller, H., Sørensen, M. F., Jensen, C. B. and Hammershøi, D. (1996), "Binaural technique: Do we need individual recordings?", *Journal of the Audio Engineering Society*, Vol. 44 No. 6, pp. 451–469.

Moore, B. C. J., Oldfield, S. R. and Dooley, G. J. (1989), "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz", *Journal of the Acoustical Society of America*, Vol. 85 No. 2, pp. 820–836.

Poon, P. W. F. and Brugge, J. F. (1993), "Sensitivity of auditory nerve fibers to spectral notches", *Journal of Neurophysiology*, Vol. 70 No. 2, pp. 655–666.

Pulkki, V., Lokki, T. and Rocchesso, D. (2011), "Spatial effects", in Zölzer, U. (Ed.), *DAFX: Digital Audio Effects*, John Wiley and Sons Ltd., Chichester, UK, pp. 139–183.

Raskar, R., Tan, K.-H., Feris, R. S., Yu, J. and Turk, M. (2004), "Nonphotorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging", *ACM Transactions on Graphics*, Vol. 23 No. 3, pp. 679–688.

Satarzadeh, P., Algazi, R. V. and Duda, R. O. (2007), “Physical and filter pinna models based on anthropometry”, in *Proceedings of the 122nd Audio Engineering Society Convention*, Vienna, Austria, pp. 718–737.

Spagnol, S., Geronazzo, M. and Avanzini, F. (2010*a*), “Fitting pinna-related transfer functions to anthropometry for binaural sound rendering”, in *Proceedings of the 2010 IEEE International Workshop on Multimedia Signal Processing*, Saint-Malo, France, pp. 194–199.

Spagnol, S., Geronazzo, M. and Avanzini, F. (2010*b*), “Structural modeling of pinna-related transfer functions”, in *Proceedings of the 7th International Conference on Sound and Music Computing*, Barcelona, Spain, pp. 422–428.

Spagnol, S., Geronazzo, M. and Avanzini, F. (2013*a*), “On the relation between pinna reflection patterns and head-related transfer function features”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21 No. 3, pp. 508–519.

Spagnol, S., Hiipakka, M. and Pulkki, V. (2011), “A single-azimuth pinna-related transfer function database”, in *Proceedings of the 14th International Conference on Digital Audio Effects*, Paris, France, pp. 209–212.

Spagnol, S., Rocchesso, D., Geronazzo, M. and Avanzini, F. (2013*b*), “Automatic extraction of pinna edges for binaural audio customization”, in *Proceedings of the 2013 IEEE International Workshop on Multimedia Signal Processing*, Pula, Italy, pp. 301–306.

Thurlow, W. R. and Runge, P. S. (1967), “Effect of induced head movements on localization of direction of sounds”, *Journal of the Acoustical Society of America*, Vol. 42 No. 2, pp. 480–488.

Walker, A. and Brewster, S. (2000), “Spatial audio in small screen device displays”, *Personal Technologies*, Vol. 4 No. 2, pp. 144–154.

Wenzel, E. M., Arruda, M., Kistler, D. J. and Wightman, F. L. (1993), “Localization using nonindividualized head-related transfer functions”, *Journal of the Acoustical Society of America*, Vol. 94 No. 1, pp. 111–123.

Wightman, F. L. and Kistler, D. J. (1999), “Resolution of front-back ambiguity in spatial hearing by listener and source movement”, *Journal of the Acoustical Society of America*, Vol. 105 No. 5, pp. 2841–2853.

Figure 1a

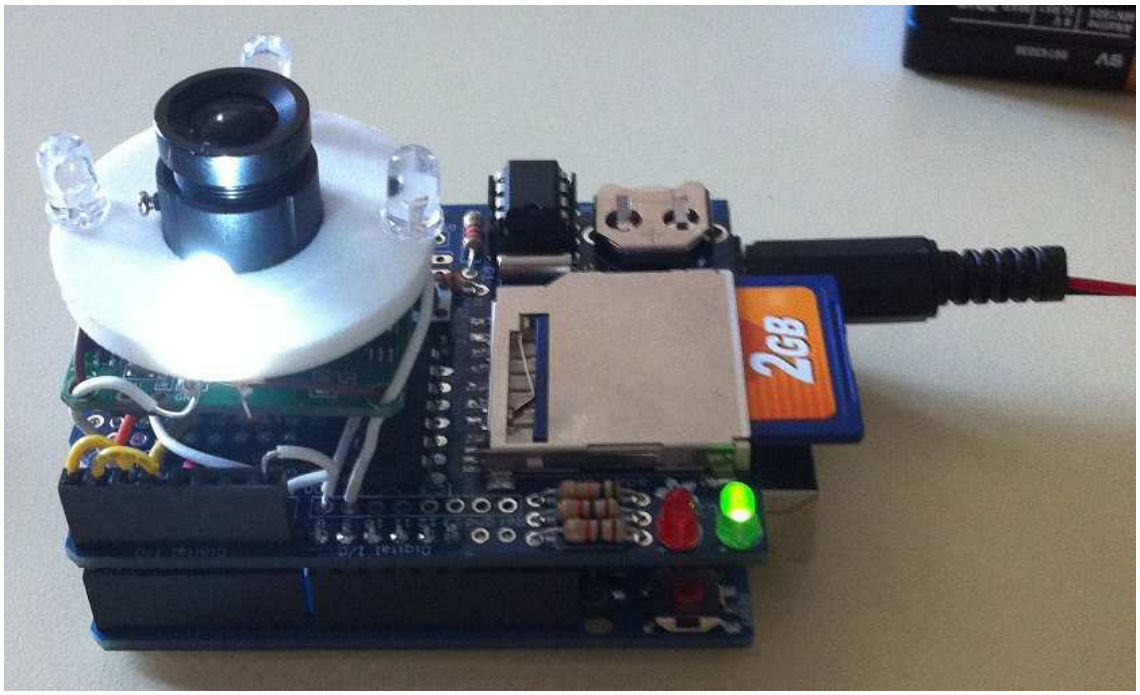

Figure 1b

Figure 2

Figure 3

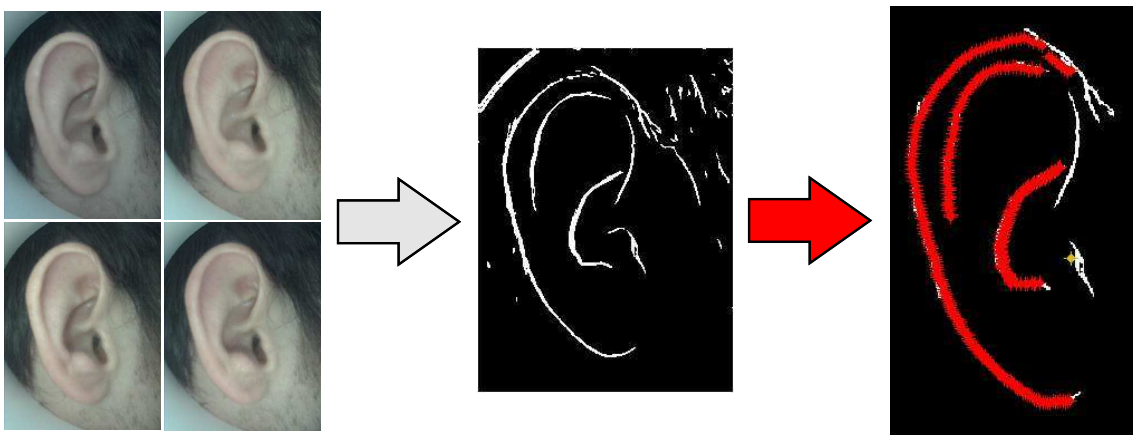

Figure 4

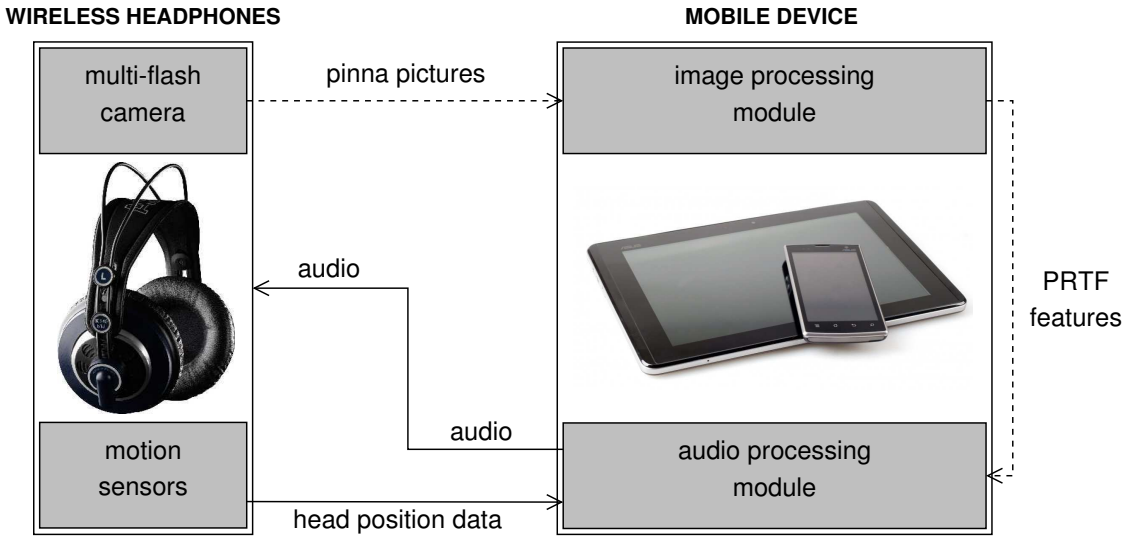

Figure 5

Figure 6

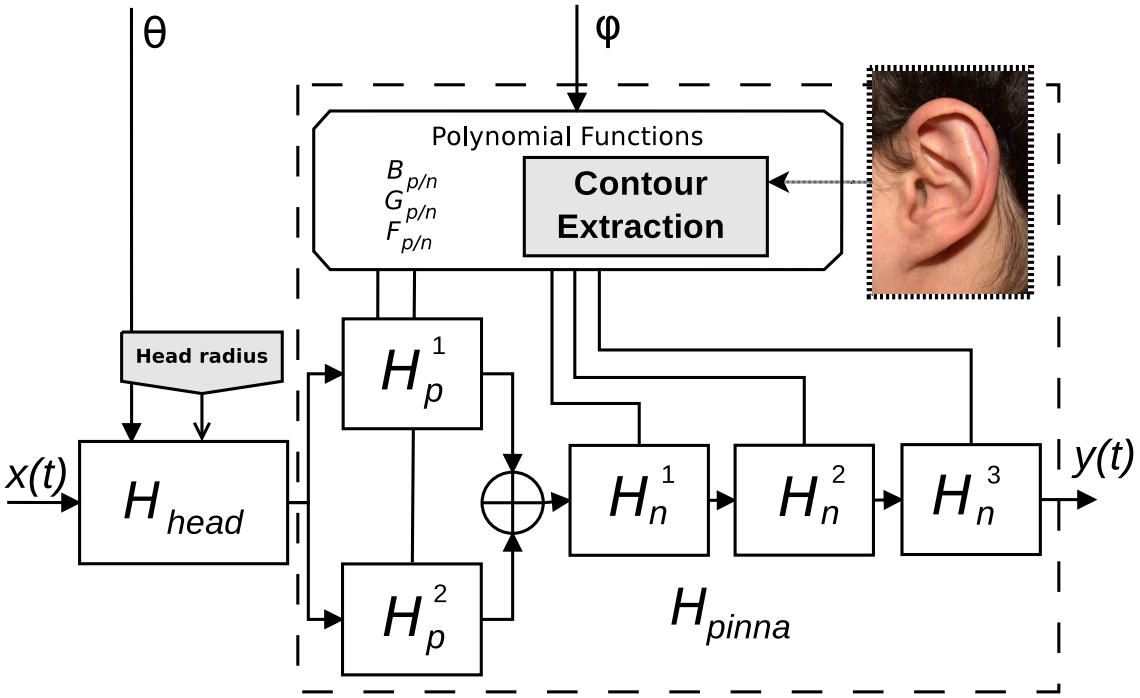

Figure 7

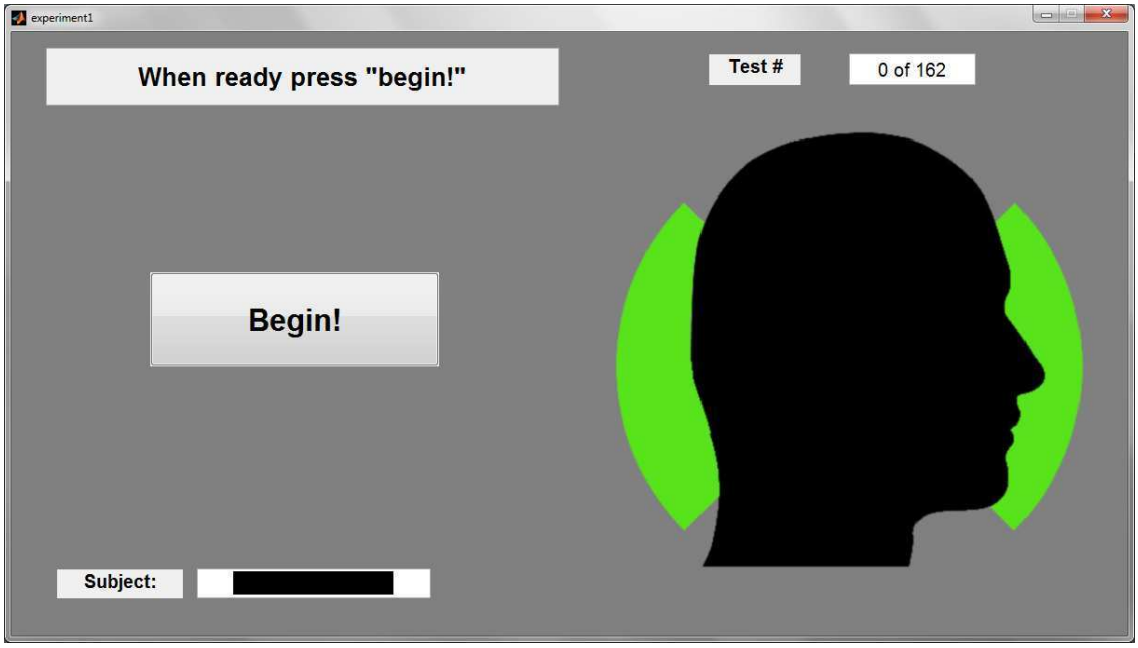

Figure 8

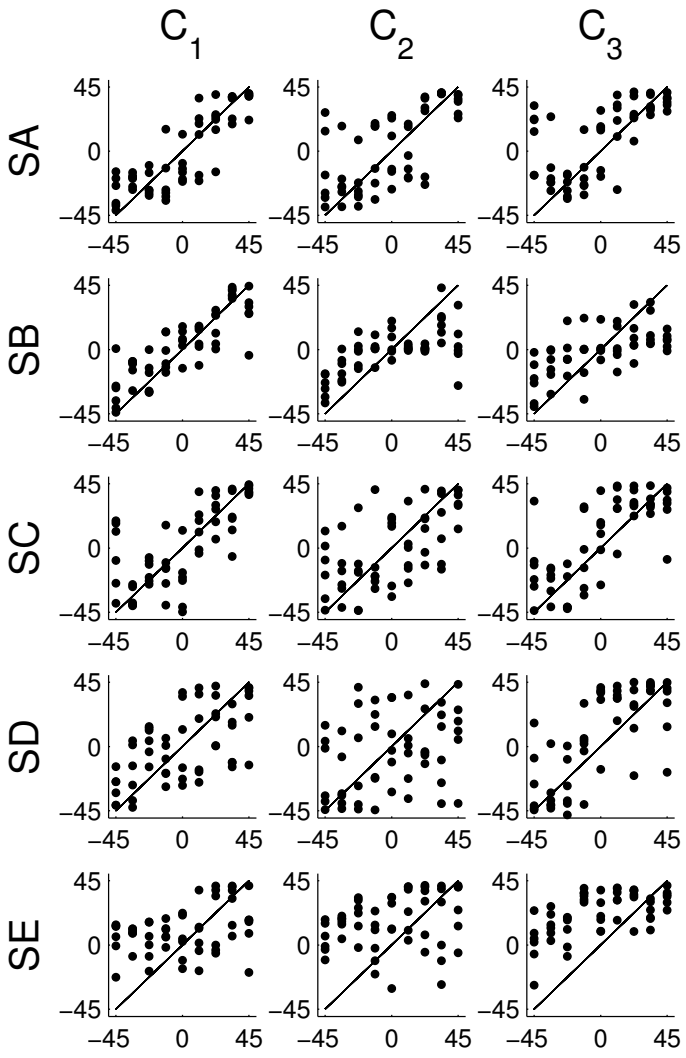

Table 1: *Results of the localization test.*

Criterion	Subject ID	C_1	C_2	C_3
Mean elevation error	SA	12.6°	17.5°	17.5°
	SB	12.4°	19.1°	19.3°
	SC	15.4°	19.9°	17.3°
	SD	20.3°	26.4°	19.0°
	SE	24.8°	29.3°	27.9°
	Mean	17.1°	22.4°	20.2°
Up/down reversal rate	SA	6.2%	18.7%	12.5%
	SB	0.0%	2.1%	6.2%
	SC	6.2%	8.3%	6.2%
	SD	6.2%	18.7%	8.3%
	SE	10.4%	25.0%	25.0%
	Mean	5.8%	14.6%	11.6%
Front/back reversal rate	SA	0.0%	0.0%	0.0%
	SB	100%	100%	100%
	SC	77.8%	55.6%	42.6%
	SD	3.7%	90.7%	79.6%
	SE	27.8%	35.2%	48.1%
	Mean	41.9%	56.3%	54.1%

Acknowledgments (if applicable): This work was partially funded by the research project PADVA (Personal Auditory Displays for Virtual Acoustics), grant no.CPDA135702, of the University of Padova. The authors wish to thank Mr. Sandro Scaiella and Mr. Giacomo Sorato for their help in the implementation and management of the localization test.

Biographical Details (if applicable):

Simone Spagnol received the BS degree in Computer Engineering in 2006 and the MS degree in Computer Engineering in 2008 from the University of Padova, Italy. He was Visiting Scholar at the Laboratory of Acoustics and Audio Signal Processing, Aalto University, Finland in 2010. He received the Ph.D. degree in Information Engineering (Curriculum in Information and Communication Technology) at the University of Padova in April 2012. He is currently a Postdoctoral Fellow at the University of Padova. His research interests include binaural sound localization and synthesis and sonic interaction design.

Michele Geronazzo received the BS degree in Computer Engineering in 2006 and the MS degree in Computer Engineering in 2009 from the University of Padova, Italy. He received the Ph.D. degree in Information Engineering (Curriculum in Information and Communication Technology) at the University of Padova in April 2014. He is currently a Postdoctoral Fellow at the University of Padova. His research interests include binaural technologies and multimodal interaction in virtual environments.

Davide Rocchesso received the Ph.D. degree from the University of Padova, Italy in 1996. He is associate professor at the Iuav University of Venice, Italy. He has been the coordinator of EU project *SOb* (the Sounding Object), and local coordinator of the EU project *CLOSED* (Closing the Loop Of Sound Evaluation and Design) and of the Coordination Action *S2S*^{2} (Sound-to-Sense; Sense-to-Sound). He has been chairing the COST Action IC-0601 *SID* (Sonic Interaction Design), and he is now coordinating the EU project *SkAT-VG* (Sketching Audio Technologies using Vocalizations and Gestures).

Federico Avanzini received the Ph.D. degree in Information Engineering from the University of Padova, Italy, in 2001. Since 2002 he has been with the Sound and Music Computing group at the Department of Information Engineering of the University of Padova, where he is currently Assistant Professor. His main research interests are in the area of sound synthesis and processing. He has been key researcher in numerous European projects (FP5, FP6) and national projects, and PI of the EU project *DREAM* (Culture2007) and of industry-funded projects. He was the General Chair of the 2011 International Sound and Music Computing Conference.