Acoustic Selfies for Extraction of External Ear Features in Mobile Audio Augmented Reality

Michele Geronazzo*
Dept. of Neurosciences, Biomedicine and Movement Sciences
University of Verona

Jacopo Fantin, Giacomo Sorato, Guido Baldovino,
Federico Avanzini†
Dept. of Information Engineering
University of Padova

Abstract

Virtual and augmented realities are expected to become more and more important in everyday life in the next future; the role of spatial audio technologies over headphones will be pivotal for application scenarios which involve mobility. This paper introduces the SelfEar project, aimed at low-cost acquisition and personalization of Head-Related Transfer Functions (HRTFs) on mobile devices. This first version focuses on capturing individual spectral features which characterize external ear acoustics, through a self-adjustable procedure which guides users in collecting such information: their mobile device must be held with the stretched arm and positioned at several specific elevation points; acoustic data are acquired by an audio augmented reality headset which embeds a pair of microphones at listener ear-canals. A preliminary measurement session assesses the ability of the system to capture spectral features which are crucial for elevation perception. Moreover, a virtual experiment using a computational auditory model predicts clear vertical localization cues in the measured features.

Keywords: binaural audio, head-related transfer function, headphones, mobile augmented reality, computational auditory model

Concepts: •Human-centered computing → Interaction devices; •Computing methodologies → Mixed / augmented reality; Virtual reality; •Hardware → Signal processing systems; •Applied computing → *Sound and music computing;*

1 Introduction

Binaural audio technologies aim at reproducing sounds in the most natural way, as if listeners were surrounded by real sound sources for which our brain succeeds in perceiving the spatial qualities [Blauert 1983]. The rendering of virtual acoustic scenarios makes use of Binaural Room Impulse Responses (BRIR) that can be described as the combination of two main components: the first one represents characteristics of the environment, contained in the Room Impulse Response (RIR), while the second one is related to the characteristics of the listener, i.e. the Head-Related Impulse Response (HRIR) [Blauert 1983].

In particular, the HRIR (or their Laplace transform, the HRTF) describe the individual acoustic filtering of head, torso and ear of the listener. The HRTFs acquirement process requires special and expensive equipment (anechoic room, in-ear microphones, etc.), that is rarely available in real-world applications. HRTF measurement in a domestic environment is a challenging issue; recent trends exploit the availability of low-cost devices for acquisition of 3D meshes of the head [Gamper et al. 2015]), and algorithms for HRTF modeling and customization [Spagnol et al. 2013]. These solutions do not fully capture individual details of the external ear acoustics, due to the fine anthropometric structure of the pinna. Such information is collected in the so called pinna-related transfer function (PRTF) which contains salient localization cues especially for elevation perception (see [Spagnol et al. 2013] for a review), thus an accurate representation is mandatory in order to render the vertical dimension in virtual/augmented auditory displays where the use of non-individualized HRTFs is not acceptable [Wenzel et al. 1993].

This paper faces the issue of cost reduction in the HRTF measurement process, with particular focus on PRTF extrapolation for a mobile audio augmented reality (mAAR) system. Our final aim is to allow easy HRTF individualization, thus improving binaural spatial audio accuracy over non-individual HRTFs with novel personalization processes guided by such acquired PRTFs.

The proposed system includes headphones, provided with embedded external microphones for binaural capture of environmental sounds, as well as internal speakers for binaural audio reproduction. An attractive idea consists in using the embedded microphones in order to acquire HRTFs everywhere from sound stimuli played back by mobile device's speakers with the aim at building an acoustic self-portrait; the SelfEar project has the purpose of developing the signal processing algorithms and interaction with the device in order to obtain a self-adjust procedure. Furthermore, few studies have been conducted aiming to verify HRTF acquirement in non-anechoic environment with particular attention to directions in median plane [Ihlefeld and Shinn-Cunningham 2008] which are relevant for individual spectral content introduced in PRTFs.

In this contribution we present a series of measurements on the system, worn by a KEMAR dummy head [1] in a silent booth. As a preliminary analysis, we compared responses measured with the SelfEar system with those obtained with a professional equipment. Such a comparison uses a computational auditory model in order to predict localization performances for a virtual listener with a KEMAR-like acoustical behavior. In a subsequent more extensive analysis, we predicted its localization performances while listening with 41 non-individual HRTFs, in order to assess the saliency of the spectral features acquired by SelfEar.

2 Mobile audio augmented reality

In a mAAR system (see fig.1), the listener is able to enjoy a mix of real and virtual sound sources. Real sources are captured by headset microphones at the ear canal entrance and redirected to the headset speakers. In between, a correction filter compensates for errors introduced by different headphones and microphone positions

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

[†] e-mail:avanzini@dei.unipd.it

VRST '16,, November 02-04, 2016, Garching bei München, Germany
ISBN: 978-1-4503-4491-3/16/11
DOI: http://dx.doi.org/10.1145/2993369.2993376

[1] http://www.gras.dk/products/head-torso-simulators-kemar.html

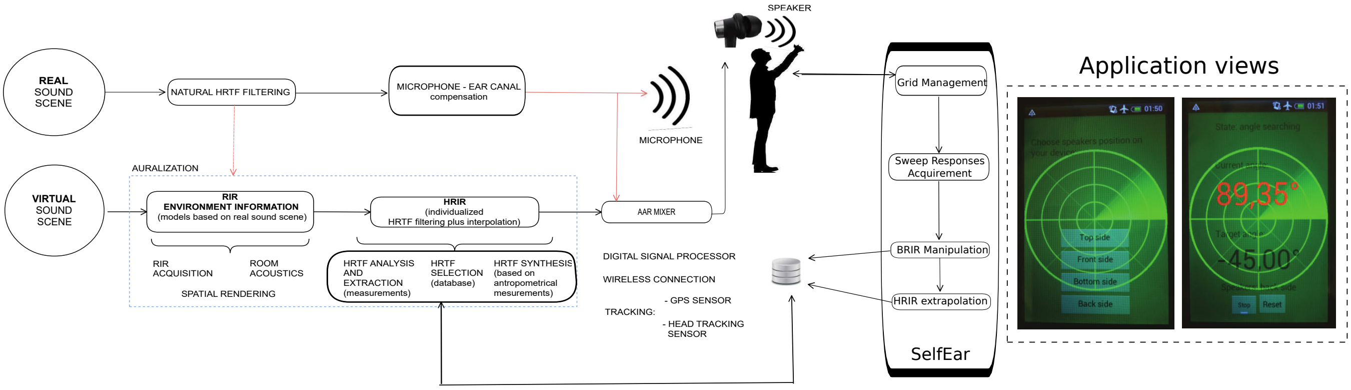


Figure 1: *Schematic view of SelfEar project for audio augmented reality. Screenshots of the two application views on the right side.*

compared to the unblocked entry point of the auditory channel, thus simulating natural listening condition. With such a setup, the headset becomes ideally transparent to real sound sources.

On the other hand, rendering virtual sources requires a dynamic and parametric auralization process in order to create a perfect superposition with reality. Auralization employs BRIRs, that must fit with the real placement environment. Producing realistic virtual and augmented acoustic scenarios over headphones, with particular attention to space properties and externalization issues, remains a major challenge due to the interplay of several components of a mAAR system [Loomis et al. 1999]:

- *ergonomic delivery system*: the ideal headphones should be acoustically transparent which means listeners are not aware of the sound emitted by transducers.
- *tracking*: tracking listener position in the environment is required to produce a common spatial representation between real and virtual scenes;
- *room acoustics knowledge*: spatial impression and perception of the acoustic space entail the knowledge of early reflections and reverberation of the real environment;
- *individual spectral cues*: head and pinna individually filter the incoming sound to listener ears and during playback.

3 The Selfear project

SelfEar is a mobile application for Android that measures user's personal HRIRs using sound stimuli played by the mobile device. The phone/tablet must be held with the stretched arm and positioned at several specific elevation angles along the subject's median plane. In-ear microphones capture the audio coming from the device speaker, thus recording the position-, listener-, and environment-specific BRIR. Data collected through the application can be employed at later processing stages to obtain an acoustic characterization of user's ear. After post-processing procedures that compensate acoustic effect of acquiring conditions and playback device, individualized PRTFs can finally be employed for spatial audio rendering. A promising technique involves HRTF selection based on psychoacoustic metrics and anthropometric similarities [Geronazzo et al. 2014] which might be parametrized by SelfEar-extracted acoustic features.

3.1 Source manager

The spatial grid management system of SelfEar guides the user through the BRIR measurement process by virtue of a self-adjusted procedure. In the launching view, the user is asked to select the device's speaker position (typically at the top, front, bottom or back side). This choice affects the device orientation during the sound stimulus playback in order to maximize speakers performance with respect to their directivity. The user then can press the "Start" button to begin the BRIR acquisition procedure. Before reaching the first target elevation, users must rise the device at eye-level exactly in front of their face, in order to create a proprioceptive reference for all target angles;[2] then, the procedure follows this logical flow:

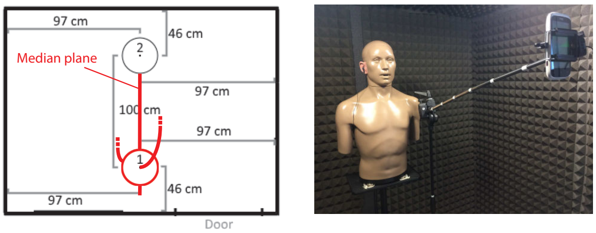
1. *Target reaching*: the current speaker orientation, corresponding to elevation in the user's median plane, is estimated using data from the accelerometer; the sequence of target elevations spans the range $[-40°, 40°]$ in ascending order with equal spacing (these values match those of the CIPIC HRTF database,[3] which will serve as a comparison to these measurements). An auxiliary beep signal sonifies the error between current and target position to aid the target reaching procedure, which would be particularly useful in case the display is not visible due to the speaker's position (e.g. in the back side). The pause between one beep and the following one is directly proportional to the difference between the current angle and the target. Target elevations have to be reached within a precision of $\pm 1°$.
2. *Position check*: once the target is reached, a 2 s timer starts; if the error from the target exceeds $\pm 2°$ three times before the timer ends, the procedure jumps back to the end of step 1.
3. *Sweep playback*: after position check, a sound stimulus (a sweep) is played from the speakers; if the error from the target exceeds $\pm 2°$ even once during the sweep playback, the entire procedure for the current elevation is reset.
4. *BRIR storing*: the recorded audio is locally stored (and the procedure returns to step 1 for next elevation, until all elevations are reached).

4 Acoustic measurements

A preliminary measurement session was performed in a non-anechoic environment using a KEMAR dummy head (in order to avoid errors due to active movement by the user). SelfEar acquired a set of BRIRs in the frontal median plane, allowing the estimation of individual elevation-dependent PRTF spectral features [Wenzel et al. 1993]: the two main peaks (resonances P1-omnidirectional mode, and P2-horizontal mode) and the three main notches (N1-3

[2] The device position can be controlled using head-pose estimation algorithms with the camera. We leave this improvement for future work.

[3] A public-domain database of high spatial resolution HRTFs. http://interface.cipic.ucdavis.edu/sound/hrtf.html



(a)

(b)

Figure 2: *Measurement setup. (a) Schematic top view: source (moving in the median plane, red line) and receiver positions. (b) SelfEar measument setup with selfie stick incorporated.*

generated by sound reflections on pinna reliefs).

4.1 Setup

All the measurement and experimental sessions were conducted inside a 2 × 2 m silent booth with a maximum acoustic isolation of 45 dB[4]. Figure 2a shows top-view schema of the measurement setup, identifying two positions: position #1 relative to the source which moves in the median plane, while position #2 to the receiver.

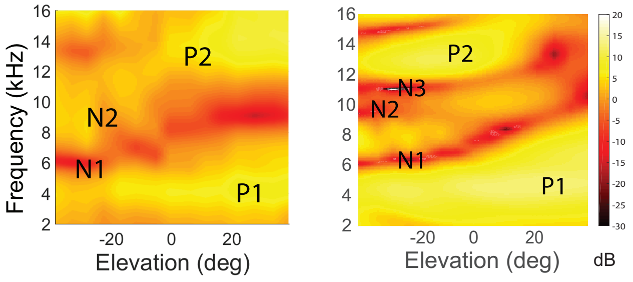
The playback device was an HTC Desire C smartphone supported by a self-produced boom arm with a selfie stick incorporated; the maximum SPL reached is 51 dB at the reference frequency of 500 Hz, at 1 m distance. The receiver was a pair of Roland CS-10EM in-ear headphones[5] with embedded microphones which were places at the entrance of KEMAR ear canals. The source signal was a one second logarithmic sine sweep signal from 20 Hz to 20 kHz.

Binaural microphones hanging from the booth ceiling with sound source in front of them at 1-m distance, captured the diffuse-field measurement that characterizes the environmental- and setup- specific acoustical features without KEMAR. The supporting structure consists of two pieces of iron wire that fall from the booth ceiling at the same positions of KEMAR ear canal entrances.

4.2 Acoustic data

A selfie stick held the smartphone which was placed inside the booth at position #1 of Fig. 2a and rotated on the median plane (see the red line in Fig. 2a) allowing a fine angular adjustment; the KEMAR dummy head wearing binaural microphones was placed at position #2 of Fig. 2a. The distance between smartphone and KEMAR was always one meter (we assume that PRTF spectral details for elevation perception are invariant with distance [Brungart and Rabinowitz 1999]). Measurements spanned 15 angles between -40° and $+40^\circ$ on the median plane.

For each measurement, the onset was detected by applying a cross-correlation function with the original sweep signal and the BRIR was then extracted by de-convolving sweep responses. Late reflections caused by the booth and the presence of equipment were removed by subtracting the corresponding diffuse field responses from BRIRs. Finally, PRTFs were estimated by windowing each impulse response with a 1 ms hanning window (48 samples) temporally-centered on the maximum peak and normalized on the maximum value in amplitude. All the normalized PRTFs were then band-pass filtered between 2 kHz and 15 kHz, ensuring the extraction of salient peaks and notches caused by pinna acoustics.



(a)

(b)

Figure 3: *PRTFs in the median plane with diffuse-field compensation. (a) SelfEar acquisition; (b) CIPIC KEMAR. The data are interpolated in order to have a smooth spatial transition. Identifiers of peaks, P_i, and notches, N_j, are also reported.*

Figure 3 provides a visual comparison between the results obtained using this procedure (with diffuse-field compensation) and the KEMAR measurements available in the CIPIC database. The latter (see Fig. 3b) contains P1 with central frequency at 4 kHz and P2 at 13 kHz; moreover N1 moves from 6 to 9 kHz, N3 from 11.5 to 14 kHz with increasing elevation; finally, N2 stars from 10 kHz and progressively disappears once reaching the frontal direction. SelfEar is capable of acquiring P1 and N1 effectively (see Fig. 3a). One can identify also P2 and, to a minor extent, N2. However, N3 is completely absent suggesting an acoustic interference introduced by headphones in pinna concha.

5 A virtual experiment

Using the predictions of an auditory model, we simulated a virtual experiment where a listener would be asked to provide an absolute localization judgment about spatialized auditory stimulus. We adopted a recent model [Baumgartner et al. 2013], that follows a *"template-based"* paradigm implementing a comparison between the internal representation of an incoming sound at the eardrum and a reference template. Spectral features from different HRTFs correlate with the direction of arrival, leading to a spectro-to-spatial mapping and a perceptual metric for elevation performances. The model is based on two processing phases: during peripheral processing, an internal representation of the incoming sound is created and the *target* sound (e.g. the PRTF acquired with SelfEar) is converted into a *directional transfer function* (DTF); in the second phase, the new representation is compared with a *template*, i.e. individual DTFs computed from individual PRTFs (from CIPIC database), thus simulating the localization process of our brain.

The virtual experiment was conducted simulating a listener with CIPIC KEMAR anthropometry. We predicted elevation performance for this virtual subject while listening with individual PRTFs, PRTFs from SelfEar acquisition, and 41 non-individual PRTFs from CIPIC. The precision for every j-th elevation response close to the target position is defined in the *polar error* (PE):

$$PE_j = \sqrt{\frac{\sum_{i\in A} (\phi_i - \varphi_j)^2 p_j[\phi_i]}{\sum_{i\in A} p_j[\phi_i]}}$$

where $A = \{i \in N : 1 \leq i \leq N_\phi, |\phi_i - \varphi_j| \, mod 180^\circ < 90^\circ\}$ defines local elevation responses within $\pm 90^\circ$ w.r.t. the local response ϕ_i and the target position φ_j, and $p_j[\phi_i]$ denotes the prediction, i.e. probability mass vector.

Experimental results are reported in Fig. 4. White areas denote high probability in correct localization responses, i.e. target angles equal to response angles, while progressively darkened areas denotes low

[4] This is neither an anechoic environment nor a reverberant room, thus it is a good compromise for a preliminary study in a controlled space.

[5] http://www.rolandus.com/products/cs-10em/

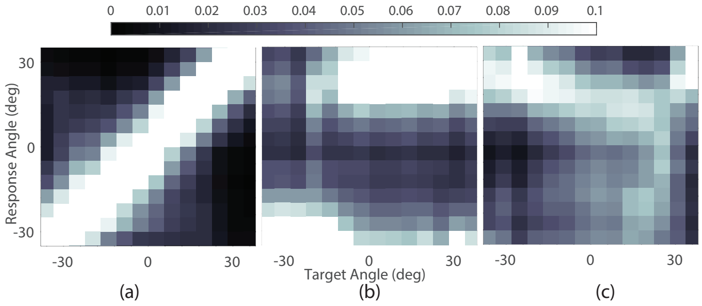


Figure 4: *Localization predictions. Response predictions for CIPIC KEMAR, while listening to sound sources located at target elevation angles with (a) individual, (b) SelfEar-acquired PRTFs, and (c) worst-case non-individual PRTF, respectively.*

probabilities in target and response agreements. The average PE provides an overall metric for localization predictions [Geronazzo et al. 2015]. Virtual listener had an average PE equal to 17.5° in the simulated ground truth condition, i.e. with individual PRTFs (see Fig. 4a). The same listener had localization judgments depicted in Fig. 4b while using PRTFs acquired with SelfEar, and a PE equal to 27.8°. Finally, the mean PE among 41 non-individual PRTFs is 25.8 ± 3.5 with 27% of them performing worse than SelfEar PRTF acquisition (see Fig. 4c for the worst-case listening scenario).

6 Discussion and conclusions

The SelfEar application allows low-cost HRTF acquisition in the frontal median plane capturing peculiar spectral cues of the listener's pinna. The application take advantage of a AAR technological framework for mobile devices. PRTFs acquired by SelfEar provided up-down localization cues according to a virtual simulated listening scenario; the two white areas of Fig. 4b in diagonal extremes clearly give insights for such an effect which might degrade while listening with anechoic non-individual PRTFs (see an inverted diagonal in Fig. 4c). However, SelfEar PRTFs were far from provide subtle elevation localization cues in proximity of the horizon. This prediction might be motivated by the visual comparison among peaks and notches in PRTFs: N2 and N3 did not clearly appear in SelfEar PRTFs leading to less spectral differences among elevations and less details for the virtual listener to be considered in localization judgments [Geronazzo et al. 2015]. We can consider the absence of concha reflections due to headset presence that dramatically reduces the concha volume, thus producing changes in resonant modes of the pinna structure [Prepeliţă et al. 2016]. The headset compensation will be take advantage from extra information such as listener's pinna anthropometry in order to artificially introduce N2-3 [Spagnol et al. 2013].

It is worthwhile to note that the most important notch and peak parameters for elevation perception, i.e. P1 e N1-2 [Iida et al. 2007], can be directly predicted from the PRTFs estimated through SelfEar. These spectral features can in turn be exploited for synthetic PRTF models, or for HRTF selection procedure.

The proposed system was tested without a human subject in a silent booth which is an acoustically treated environment. Further work is thus required in order to extend this PRTF-capture procedure to domestic environments, where the influence of background noise and random acoustic events must be taken into account, as well as subject movements during the acquisition. For such purpose, adaptive filtering approach can be able to extract PRTFs "'on-the-fly"' with random head movements [Ranjan et al. 2016].

Finally, it is indisputable that psychoacoustic evaluation with human subjects is mandatory in order to confirm the reliability of the SelfEar application providing effective individualized HRIRs in rendering virtual sound sources.

References

BAUMGARTNER, R., MAJDAK, P., AND LABACK, B. 2013. Assessment of Sagittal-Plane Sound Localization Performance in Spatial-Audio Applications. In *The Technology of Binaural Listening*, J. Blauert, Ed., Modern Acoustics and Signal Processing. Springer Berlin Heidelberg, Jan., 93–119.

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

BRUNGART, D. S., AND RABINOWITZ, W. M. 1999. Auditory localization of nearby sources. Head-related transfer functions. *J. Acoust. Soc. Am. 106*, 3, 1465–1479.

GAMPER, H., THOMAS, M. R. P., AND TASHEV, I. J. 2015. Anthropometric parameterisation of a spherical scatterer ITD model with arbitrary ear angles. In *2015 IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, 1–5.

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 *IEEE Int. Conf. on Acoust. Speech Signal Process. (ICASSP 2014)*, 4496–4500.

GERONAZZO, M., CARRARO, A., AND AVANZINI, F. 2015. Evaluating vertical localization performance of 3d sound rendering models with a perceptual metric. In *2015 IEEE 2nd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, IEEE Computer Society, Arles, France, 1–5.

IHLEFELD, A., AND SHINN-CUNNINGHAM, B. 2008. Disentangling the effects of spatial cues on selection and formation of auditory objects. *J. Acoust. Soc. Am. 124*, 4, 2224–2235.

IIDA, K., ITOH, M., ITAGAKI, A., AND MORIMOTO, M. 2007. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Applied Acoustics 68*, 8, 835 – 850.

LOOMIS, J., KLATZKY, R., AND GOLLEDGE, R. 1999. Auditory Distance Perception in Real, Virtual and Mixed Environments. In *Mixed Reality: Merging Real and Virtual Worlds*, Y. Ohta and H. Tamura, Eds. Springer.

PREPELIŢĂ, S., GERONAZZO, M., AVANZINI, F., AND SAVIOJA, L. 2016. Influence of Voxelization on Finite Difference Time Domain Simulations of Head-Related Transfer Functions. *J. Acoust. Soc. Am. 139*, 5 (May), 2489–2504.

RANJAN, R., HE, J., AND GAN, W.-S. 2016. Fast Continuous Acquisition of HRTF for Human Subjects with Unconstrained Random Head Movements in Azimuth and Elevation. In *2016 AES Int. Conf. on Headphone Tech.*, 1–8.

SPAGNOL, S., GERONAZZO, M., AND AVANZINI, F. 2013. On the Relation between Pinna Reflection Patterns and Head-Related Transfer Function Features. *IEEE Trans. Audio, Speech, Lang. Process. 21*, 3 (Mar.), 508–519.

WENZEL, E. M., ARRUDA, M., KISTLER, D. J., AND WIGHTMAN, F. L. 1993. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am. 94*, 1, 111–123. 00940.