POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

Master of Science in
Computer Engineering

Selection of Head-Related Transfer Function through Ear Contour Matching for Personalized Binaural Rendering

Candidate

Marco Dalena
Student Id. number 783527

Thesis Supervisor	Assistant Supervisor
Prof. Augusto Sarti	Ing. Lucio Bianchi

Academic Year 2012/2013

POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

Laurea Magistrale in
Ingegneria Informatica

Selezione della Head-Related Transfer Function tramite Matching dei Contorni dell'Orecchio per Resa Binaurale Personalizzata

Candidato

Marco Dalena
Matricola 783527

Relatore

Prof. Augusto Sarti

Correlatore

Ing. Lucio Bianchi

Anno Accademico 2012/2013

Selection of Head-Related Transfer Function through Ear Contour Matching for Personalized Binaural Rendering
Master thesis. Politecnico di Milano

This thesis has been typeset by LaTeX and the smcthesis class.

Author's email: marco.dalena@mail.polimi.it

To my family

Sommario

In questa tesi viene proposto una nuova metodologia per eseguire la resa binaurale tramite Head-Related Transfer Function, estraendo caratteristiche significative da una fotografia dell'orecchio. Dimostreremo che rispetto alle usuali tecniche di scelta della Head-Related Transfer Function migliore, la nostra metodologia porta agli stessi risultati con ordine di grandezza dei secondi anzichè dei minuti. Noi proponiamo una metodologia basata sull'estrazione dei contorni per l'estrazione delle caratteristiche salienti basata sulla trasformata di Hough generalizzata per effettuare il confronto. Queste due tecniche hanno il pregio di essere largamente usate e di essere ben documentate.

Dal momento che una metodologia basata su estrazione di caratteristiche salienti solitamente richiede l'uso di un database, e dal momento che il nostro obiettivo finale è di fornire una Head-Related Transfer Function, abbiamo scelto uno tra i database più usati per l'area di ricerca riguardante le HRTF: il database CIPIC.

La nostra metodologia riguarda aspetti di percezione del suono da parte degli umani, dunque è stata progettata anche una metodologia per testare il nostro approccio e validare la bontà della nostra teoria.

Abstract

In this work we porpose a novel approach to personalized binaural rendering through Head-Related Transfer Function based on feature extraction and feature matching of an ear picture. We show that with respect to usual Head-Related Transfer Function retrieving procedure, our methodology give the same results performing more then one order of magnitude faster. We propose an approach based on edge extraction for feature extraction and on Generalized Hough Transform for matching step. This two technique are widely used and largely documented.

Since feature extraction approaches usually involves databases, and since our final goal was to retrieve an Head-Related Transfer Function, we choose one of the most used databases for HRTF reasearch purpose: the CIPIC database.

Since our methodology involves human perception, a test methodoogy has been designed and perception tests has been performed to validate our theory.

Contents

List of Figures

Chapter 1

Introduction

This work of thesis is a result of experimental research aimed at provide a personalized binaural rendering using image matching techniques.

This work merges two very different context, which share only the idea behind the techniqus they use. On one hand we have the image processing world, which focuses on all the operation that can be done on an image to extract information from it and focuses on the operation that can be applied to images to compress, uncompress and send them through whatever medium.

On the other hand there is the Head-Related Transfer Function (HRTF) world, which is the a reasearh context that studies the relationship between sound perception and antropometry of the part of the body involved. It may sound quite strange but the sound that reaches the ear canal have been processed by other part of the body: for example it can bounce on the head and on a shoulder before reaching the ear and it is easy to imagine that they may have some impacts on the perception. Is it well known from physics that a sound from a certain distance may be considered as plane wave. When such a wave reaches a shoulder, for example, a diffraction phenomenon can be elicited; when the wave reaches the ear, it is no more a plane wave but a superposition of waves coming from different direction. All this waves mix togheter, arriving to the ear in different time instants: the brain is able to decode this mix and is able to localize the source, since it knows all the parameters that play role in the path. The brain basically knows the model of all the part of the body involved in the perception. HRTF can be seen as a filter between environment and ear canal: once the parameters of such a filter are knwon, it easy to switch from the inside (brain) to the outside (environment). Derivation of such a parameters is one of the current big research area of HRTF. The other area is related to the measurement as much accurate as possible of such a function. HRTF is the Fourier Transform of the Head-Related Impulse Response (HRIR), which is a measurable response of the head to the environment that need ad-hoc places where to measure and expensive ad-hoc structures. This approach, if correctly implemented, gives the best results in term of rendering, while modeling approach is still far from commercial purposes. It is straightforward to think that this two big area will collide when the relathionship between antropometry and perception will be fully understood and it will be possible to compare the results of one technique with the ones of the other.

It could appear confusing trying to merge such a different worlds, but actually is not. What all the image analysis techniques shares between each other is the final goal: given a pattern, find the exact or the best match in an image, but most of these techniques apply also in other fields, where a recurring pattern must be found whitin an ensamble. Just to make an example, let's think of a piece of music: it is composed by notes played by different instruments that overlap each other creating the melody: one can think to some indicators for evaluating portion of the melody, and try to find the same indicators in another piece of music.

The problem of finding a measured HRIR (and its Fourier Transform) whitin a database of HRIR could be easily seen as a pattern matching: one can think to search relevant feautures from the HRIR in other HRIRs. Unfortunally this approach is time- and power-consuming, since it easily reach over 1200 HRIRs for a single ear of a single person: searching even few features for every HRIR of a subject in a database of 100 ear (each one with 1200 directional HRIRs) leads to an explosion of the complexity of the algorithm.

Our approach will show that extracting feautures of the ear, which is a relevant responsible of the filtering effect, and using them to search whithin the same features what is the best approximation can save time, money and can be extended for commercial purposes.

1.1 Outline

This work present the following structure: in Section 2 we will cover the state of the art for both HRTF and image processing, focusing on the relevant techniques that fit our purposes. Section 3 will present our methodology for solving the problem, giving the knoledge to understand the implementation of the methodology, presented in Section 4. Finally we present in Section 5 a test methodology and the results of the testing phase to evaluate our work. We conclude with the future developments of this work in Section 6.

Chapter 2

Background

The aim of this work is to recognize a proper Head-Related Transfer Function (HRTF) within a database by using 2D ear features matching, thus mixing two different world into the same environment: the HRTFs world, with specific address to the HRTF measurement, and the image processing world, with specific address to the image matching.

The HRTF is the Fourier Transform of the Head-Related Impulse Response (HRIR), that is a response that characterizes how an ear receives a sound from a point in space; a pair of HRTFs for two ears can be used to synthesize a binaural sound that seems to come from a particular point in space. It is a transfer function, describing how a sound from a specific point will arrive at the ear (generally at the outer end of the auditory canal). The estimation of the direction is made possible by the cues derived from one ear (monaural cues) and comparing cues received at both ears (binaural cues). The monaural cues comes from the interaction between the sound source and the anatomy of the hearing system, in which the source sound is modified before it enters the ear canal and it is processed by the middle and inner ear. Such monaural cues are intensity of the perceived sound (related to the distance of the source), ratio between direct and reverberant sound and frequency filtering effect of the outer ear that allows to distinguish the direction on a vertical plane. The binaural cues are the Interaural Time Difference (ITD), that is the time the sound has to travel to reach both ears, from the nearest ear to the farthest one, Figure 2.1, and the Interaural Level Difference (ILD), that is the difference of perceived sound intensity between the ears due also to the shadowing effect of the head, Figure 2.2.

Lately the ear has been considered a unique antropometric human characteristic, especially into the shape of the auricle (or pinna), as fingerprints are from many years. This unicity reflect in a fundamental way in the shaping of HRTFs. It introduces peaks and notches in the high-frequency spectrum, 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; still, both seem to play an important function in vertical localization of a sound source. For this reason finding a way to map the shape of the ear to the paramaters that affect the perception is important and can lead to approaches of HRTF parametrization. This work is not focused on parametrization, nevertheless we cannot underestimate the importance of the shape

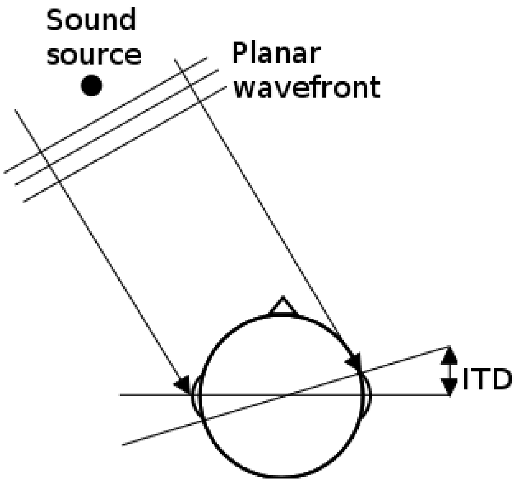

Figure 2.1. Interaural Time Difference: the sound that reach the left ear must travel a small distance to reach the right one.

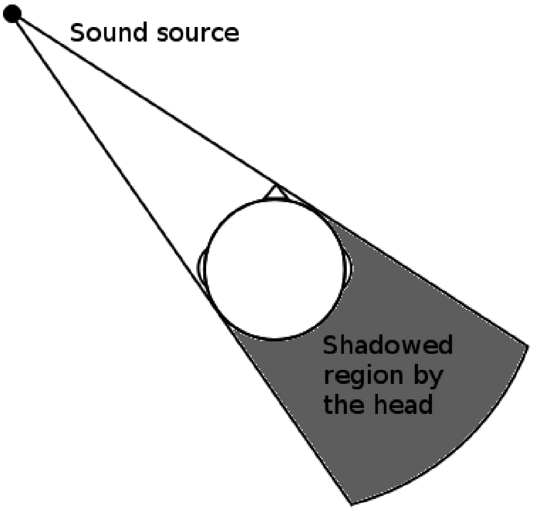

Figure 2.2. Interaural Level Distance: the intensity of the sound at the left ear is higher then the one at the right ear because the head shadows the farthest ear.

of the ear, which still plays a fundamental role in this thesis. We will focus on the extraction of the features from an ear picture, trying to find a relation between them and the associated HRTF.

The feature extraction and matching step is part of the image processing world, which is an area investigated since 1960 that had seen a fast growing from the 2000s due to the increase of computational power. For these reasons, literature about image processing is very large. This work will focus on the most suitable technique for our purpose, that is that is the ear recognition process.

In the following we will proceed in describing the state of the art for both HRTF-measured apporach and ear recognition system.

2.1 Head-Related Transfer Functions

Up to now the usual procedures to retrieve a personalized HRTF are catergorizable in two main groups: measure the HRTF and synthesize the HRTF by modeling some exernal ear features.

The measure of the HRTF is usually done by placing the subject in an anechoic room, with a microphone in its ears. A source is placed at a distance such that the wave coming from the source to the ear can be approximated as a plane wave. Next step is to sample tha space to cover as many directions as possible around the subject, ince HRTF is a function of direction of arrival: this can be done by forcing the source to move along an arc above the subject and rotating the subject or keeping the subject fixed in its position and moving the source in the space. Both methods require ad hoc structures on which loudspeaker are mounted and some time to take the desired number of measures: according to [27, 29], using a step of 5 ° in azimuth and a step of 5 ° in elevation, the needed position are more then 1200 and, using a signal 1.5 s long to obtain a Signal-to-Noise Ratio of 70 dB [26], the time needed to to complete the whole measurement is about 30 minutes.

The second family of techniques aims at parametrizing an HRTF based on a model of the system made by ears, head, shoulder and torso: these techniques are mainly based on the system proposed, by Brown and Duda [11], which employs a combination of infinite impulse response (IIR) filter to model head-shadow effect, a finite impulse response (FIR) filter to model pinna-echo effect and a FIR filter to model shoulder-echo effect.

The method proposed by Brown and Duda state that the components of the model have a one-to-one correspondence with the shoulders, head and pinnae, with each component accounting for a different temporal feature of the impulse response, meaning that the method works in the time domain. Figure 2.3 shows the model as stated by the authors; they state also that the model can be further simplified or complicated, for example in [14] by using a 3 classes model of the pinna (resonator, diffractive and reflective). In their model ρ denotes the reflection coefficient, while T and τ denote time delays.

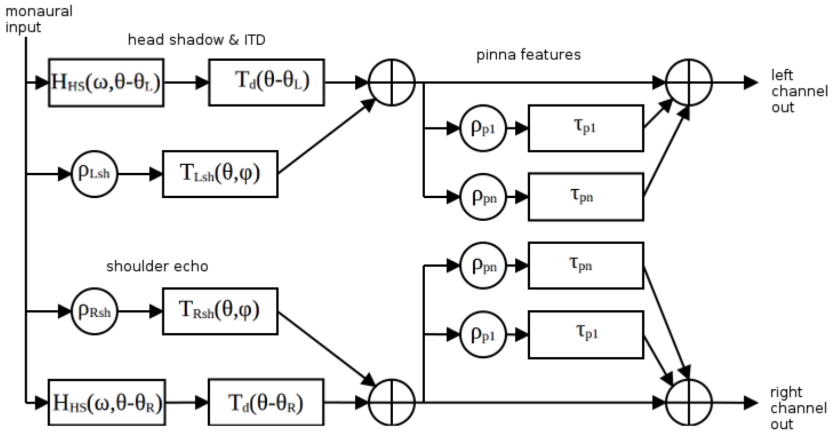

Figure 2.3. Structural model proposed by Brown and Duda

The work of Spagnol et al. [36] is based on the modeling approach and had inspired this thesis work. In [36] the authors try to construct a multi-notch filter suitable for anthropometric parametrization as a replacement to the simpler comb filter used by Algazi et al. in [5]. To design the notch filters with the appropriate paramaters, the Pinna-Related Transfer Function (PRTF) needs to be analyzed. Since the database used is the CIPIC database [4], which provides only the Head-Related Impulse Responses (HRIR), the time-domain Fourier transform of the HRTF, a step of PRTF extraction from HRIR is needed.

The authors of [36] state that knowing that pinna reflection delays usually range between 100 and 300 μs in the median plane, as stated by Batteau in [8], one can shorten the HRIR by applying a 1-ms Hann window starting from the HRIR onset [32]. In this way spectral effects due to reflections caused by shoulders and torso are removed from the response, while those due to the pinna are preserved. Concerning head diffraction compensation, if one virtually treat the pinnaless head as a sphere, the ear canal lies around azimuth $\theta = \pm 90\,^\circ - 100\,^\circ$, since human ears typically lie slightly behind and below the x axis [38], the source-ear angular distance is certainly greater than $90\,^\circ$ for sources between elevation $\phi = 0\,^\circ$ and $\phi = 45\,^\circ$ at least. It can be noticed [17] that the response due to scattering of the sphere for a source in the frontal side of the median plane at 1 meter (where CIPIC HRTF measurements were taken) presents an approximatly flat frequency spectrum.

From the obtained PRTF one can proceed extracting the feature that will be used to fill the filter paramters. To do this a separation algorithm is used, whose details are in [19], and then frequency notch is extracted by using a picking algorithm [35].

To relate the reflections and the anthropometry of the ear, a ray-tracing approach is used instead of a wave-based one. This seems a crude approximation of the wave equation, but the authors of [36] state that this approximation can be considered valid under the assumption that the wavelength of the sound must be small compared to the dimension of the reflective surface, which is the case of audible spectrum's high frequencies where spectral notches due to pinna reflections appear.

The matching procedure that allows to switch from anthropometry to filters parameters is described in [36] as follow: the basic assumption that drives the analysis procedure is that each notch track, an algorithm capable of individualizing nothces, is associated with a distinct reflective surface on the subject's pinna. Since the available data for each subject is a side-view of head showing the left or right pinna, extraction of the "candidate" reflection surfaces must be reduced in a two-dimensional basis. The choice has been to investigate as possible reflective surfaces a set of three contours directly identifiable from the pinna photograph, together with two hidden surfaces approximating the real inner back walls of the concha and helix, as shown in Figure 2.4.

2.2 Ear Recognition History

The idea of using ears as recognition element, such as fingerprint, is due to the French criminologist Alphonse Bertillon, who in the early 1890s [9] wrote

> The ear, thanks to these multiple small valleys and hills which furrow across it, is the most significant factor from the point of view of identifi-

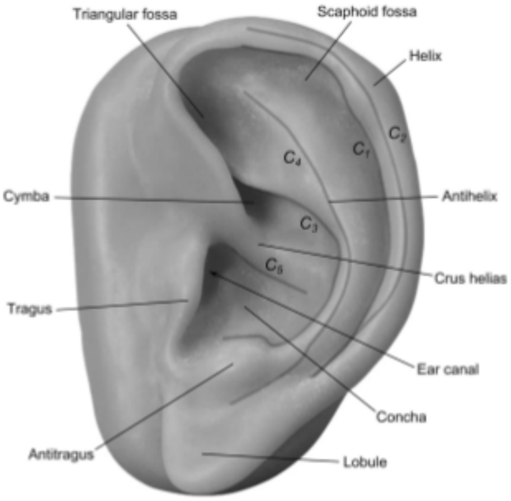

Figure 2.4. Pinna anatomy and the five possible contours that can be chosen for the matching procedure. C1: helix border; C2: helix wall; C3: concha border; C4: antihelix and concha wall; C5: crus helias, [34].

> cation. Immutable in its form since birth, resistant to the influences of environment and education, this organ remains, during the entire life, like the intangible legacy of heredity and of the intrauterine life.

The first ear recognition system is Iannarelli's system [22] which was originally developed in 1949: it is a manual system based upon twelve measurements as illustrated in Figure 2.5. Each photograph of the ear is aligned such that the lower tip of a standardized vertical guide on the development easel touches the upper flesh line of the concha area (12), while the upper tip touches the outline of the antitragus(9). Then the crus of helix is detected and used as a center point. Vertical (11), horizontal (7), diagonal (6), and anti-diagonal lines are drawn from that center point to intersect the internal and external curves on the surface of the pinna (1,2,3,4). The twelve measurements are derived from these intersections and used to represent the ear.

Abaza et al. [1] in 2013 collect into a survey the most used techniques of automatic ear recognition; they state also that currently no commercial system to automatically identify or verify the identity of individuals by way of their ear biometric exists.

Nevertheless efforts in this direction has been made since 1999, by Moreno et al. in [28], where they describe the first fully automated system for ear recognition. They used multiple features and combined the results of several neural classifiers. Their feature vector included outer ear points, ear shape, and wrinkles, as well as macro-features extracted by a compression network. To test that system, two sets of images were acquired.

Later, Burge and Burger [13] presented one of the most cited ear biometric methods in the literature. They located the ear by using deformable contours on a Gaussian pyramid representation of the image gradient. Then they construct a graph model from edges and curves within the ear, and use this graph to run the

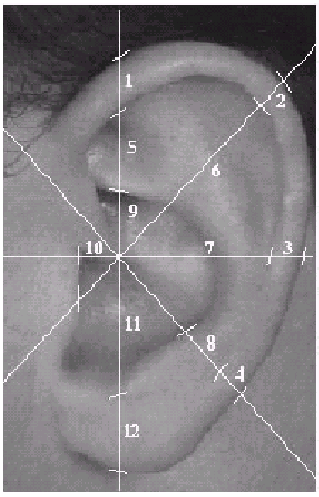

Figure 2.5. The locations of the anthropometric measurements used in the "Iannarelli System", [12].

matching algorithm for authentication.

In 2005 Mu et al. [30] extended the work of Moreno et al. [28] representing the ear feature vector as a combination of the outer ear shape and inner ear structure.

In the following sections we will present a summary of the most used techniques, following the taxonomy of the survey of Abaza et al. [1].

2.3 Ear Biometric System

An ear biometric system may be viewed as a typical pattern recognition system: an input image is reduced to a set of features that is subsequently used to compare against the feature sets of other images.

Ear recognition can be accomplished using 2D images of the ear or 3D point clouds that capture the three-dimensional details of the ear surface. In the 2D case every color or grayscale image can be used, it is evaluated the variation of the luminance gradient of the image (from dark to light and viceversa) and a point on the contour is represented as a couple of cartesian coordinates (x, y). 3D techniques use a set of discrete 3D homogenous vertices to identify the contours: the third dimension cannot be simply evaluated by associating it to the intensity of the pixel at position (x, y) since intensity is heavely affected by lighting and shadowing; usually range sensor are used to collect geometric information of the ear because they not suffer from the already mentioned problems. In the following we will assume to be working in a 2D domain, unless differently specified.

The usual processing in literature, according to [1], follows the scheme in Figure 2.6.

We will now procede in presenting the most used techniques suitable to accomplish the task for each block.

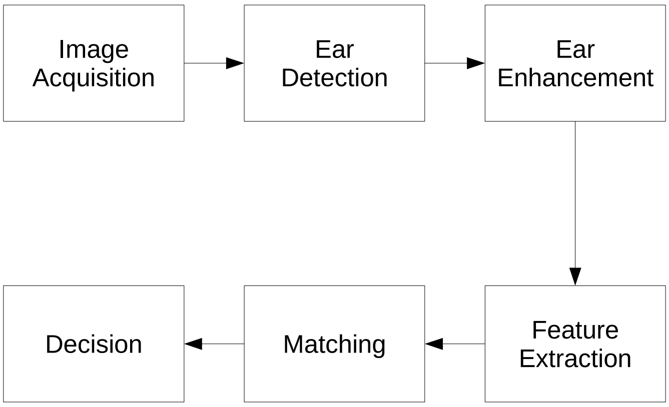

Figure 2.6. Classic ear recognition system block diagram

2.3.1 Ear Detection

Ear detection is an essential task for automated ear recognition system, though in many publications this is achieved manually. There are several approaches aimed to provided fully automated ear detection. We now briefly describe the main categories, and then we explain what are the suitable techniques for our purpose and why.

- Computer-assisted ear segmentation: semi-automated method requiring user-defined landmarks on the image, meaning that it is asked to the user to select points on the ear that will be used to perform the segmentation in an automated fashion.

- Template matching: techniques in this category uses an ear template (with reference to Figure 2.4, it could be the outer helix border) and search it in the image after edge extraction. Techniques for 2D and 3D has been developed.

- Shape-based: similar to template matching, but what is searche din the edged image are geometrci shapes or parametric curves. Techniques for 2D and 3D has been developed.

- Morphological operators: techniques based on Mathematical Morphology (MM), a theory and technique for analysis and processing of geometrical that, once applied to digital image, make use of the classic computer vision concepts (such as dilation, erosion, opening, closing, etc.) and new operators such as morphological gradients (difference between dilation and erosion of a given image), top-hat trasform (difference between image and its opening or closing), etc.

- Hybrid techniques: these techniques combine two or more standard techniques to get ear detection that could be, for example, Principal Component Analysis (PCA), Gabor filters, template matching, histograms, graphs, ray-tracing, etc.

This work focuses more on the recognition part, thus the detection part has been done manually, after image edge extraction.

Nevertheless we will explain a bit more in detail the template matching technique, since this will be part of the ear recognition step.

Burge and Burger [13] located the ear using deformable contours on a Gaussian pyramid representation of the image gradient. Then edges are computed using the Canny operator, and edge relaxation is used to form larger curve segments, after which the remaining small curve segments are removed.

Ansari and Gupta [6] used outer helix curves (which can be seen in Figure 2.4) of ears. Using the Canny edge detector, edges are extracted from the whole image and subsequently segmented into convex and concave curves, Figure 2.7. From these segmented edges, expected outer helix edges are determined: outer curve is the most likely to be a part of outer helix curve so all the identified curves are considered to be possible outer ear curves which are parallel and the average distance of two curves and the standard deviation of their distance are less than a certain threshold value. For calculating distance a path in the direction toward concave side of each edge is traced and if it intersects any other edge, distance is calculated between the points from which the path was started to the point at which the path intersects any other edge. The curves which satisfiy the following properties are identified as potential ear curves: 1. a curve is running parallel to it in the direction of concave side; 2. the parallel curve is at a certain range of average distance from the outer curve; 3. the standard deviation of distances between the points of the curves is less than a threshold.

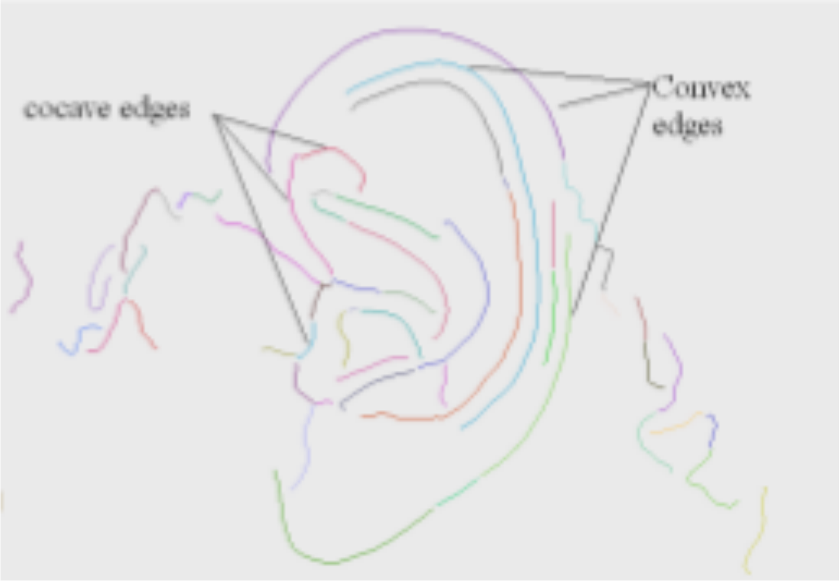

Figure 2.7. Concave and convex curves of the ear.

AbdelMottaleb and Zhou [2] segmented the ear from a face profile based on template matching, where they modeled the ear by its external curve.

What these tecnique share is the edge detection using canny, which is also our approach to the problem.

2.3.2 Ear Recognition System

In this section we analyze the most used ear recognition algorithm in literature, grouped by feature extraction scheme used to represent the ear biometric.

- Intensity-Based - uses the intensity in the 2D black and white image to capture the 3D features of the ear: the more the light is present (pixel tends to white) the more the point is higher with respect to the plane tangent the face to the ear canal; the more the light is absent (pixel tends to black) the more the point lies on the tangent plane.

- Force Field - the image is treated as an array of Gaussian attractors that act as the source of the force field (as shown in Figure 2.8). The directional properties of that force field are exploited to locate a small number of potential energy wells and channels that are used during the matching stage.

- 2D Curves Geometry - extraction of the ear contours and centroid from the ear image, and then construction of concentric circles using the centroid; feature vectors can be defined based on the interest points between the various contours of the ear and the concentric circles or based on the angle of the contour representation and the geometrical parameters of the curves.

- Generic Fourier Descriptors (GFD) - rotation-invariant descriptors to extract meaningful features from ear images. This descriptor is quite robust to both ear rotations and illumination changes. It is a Polar Fourier Tranform (PFT) followed by a classic Fourier Descriptor (FD) extraction.

- Scale-Invariant Features Transform - uses the original algorithm [25] to find point on the image that are invariant with respect to the scaling operation.

Our work take as reference the curves pointed out in the work of Spagnol et al. [36], thus our needs were quite different with respect to the “classic” ear recognition system: with reference to Figure 2.4, we were intersted in outer and inner helix border, anti-helix border and concha border, visible in Figure 2.9. For this reason we cannot use, for example, a force field technique: as one can clearly see in Figure 2.8, the extracted contour is not sufficient for our purposes.

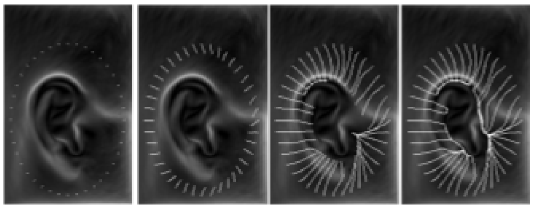

Figure 2.8. Force Field Transform, [21]

An approach that seemed to be promising too was based on feature extraction related to the geometry and morphology of the ears, but at the moment is not clear how they are related to HRTFs. For this reason we had to choice another approach. Our choice to perform ear recognition has been template matching in combination with Hough Transform. Template matching as been already discussed in Section 2.3.1 while Hough Transform will be discussed in Section 3.

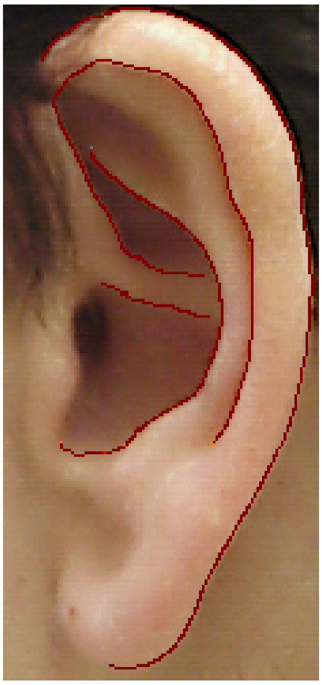

Figure 2.9. Outer and inner helix border, anti-helix border and concha border marked in red.

2.4 Outline

To the best effort of the author of this work, the approach explained in the following has never been tried. The usual procedure to retrieve a personalized HRTF using measured ones is to perform perception tests on several subjects (number vary from 14 to 50). This approach is the easiest one, since do not require measurements that take some time, as already explained in Section 2, but requires to the subject to spend time in exploring all the HRTFs in the database, or at least a subset of them, to find the right one. For example Seeber at al. in [33] aimed to develop a fast and effortless method to select the best HRTF for a certain subject on the basis of questionnaire that evaluate objective localization criteria such as externalization of the source, minimization of the fron-back confusion, match of the presented-perceived directions and minimization of the perceived source width. For their work they used the AUDIS-catalogue [10].

Their test was made of two step, resumed as follow:

- preliminary test - five pulses of white noise are presented to 17 subjects, asking them to evaluate the overall directional impression; this test lead to the conclusion that it is difficult to evaluate a predetermined HRTF because subjects adapt quicly to the given HRTF, a direct comparison of similarly reproducing HRTFs is hampered because of the specific order of HRTFs and within the test, the assessment factors change;

- selection procedure for non-individual HRTFs - the chance to randomly choose within a set of HRTFs is presented to the 46 subjects who can select and listen

> each sound processed with each HRTF, play whichever sound how many times they need and take notes about each HRTF. To reduce the selection time, a two-step procedure is used: first step consists of extracting five out of twelve HRTFs that greatly enhance spatial perception of the frontal area, then second step consist of choosing one of the five selected HRTFs that best matches the criteria of sound beeing inside an azimuth range of $-40\,^{\circ}$ left to $+40\,^{\circ}$ right, moving horizontally in equally-spaced steps always with constant elevation from frontal plane at constant (far away) distance.

Tame et al. in [37] uses a small subset of five HRTFs "maximally different", performing a k-means clustering on the HRTFs measure in the Acoustic Research Institute (ARI) database [3] to obtain a subset of them. They devise two test, one in a controlled environment and the other via web: the purpose of the test is to give to the subject an audio cue which might occur at either of the two key positions but whose intended location was easily confused due to the effects of front-back confusion. The material was selected such that the distance from the microphone to the audio source was of a similar distance to that of the ARI measurement positions from their test subjects.

Our approach aims to give to the subject the best HRTF match only by taking a photo of its ear and searching in a database the ear that best match the given one.

Section 3 explains the methodology we used to reach our goal; Section 4 gives details on the implementation of our methodology; Section **??** gives information on the methodology of tests in literature and explain our choose and finally Section 5 shows the results of our test.

Chapter 3

Methodology

In this chapter we will present the approach we have chosen, trying to explain the difficulties and problems we have faced while proceding in the development of the algorithm.

With reference to Figure 2.6, we will present our (reduced) block diagram (Figure 3.1), describing for each block the procedure we used.

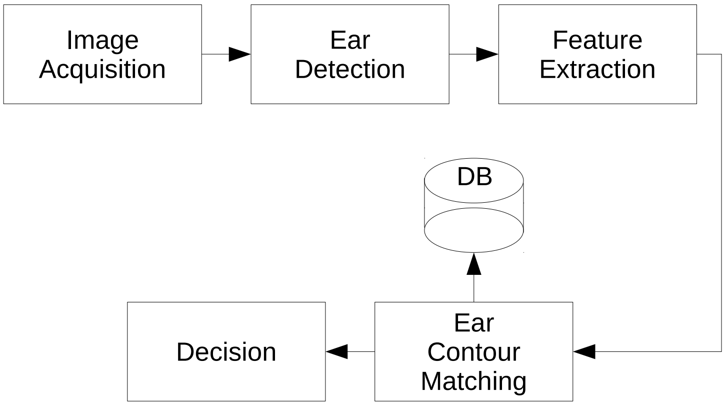

Figure 3.1. The reduced block diagram: the enhancement phase has been removed, since manual detection is the method that ensure the most reliable results in terms of ear detection; in the matching step a database is involved.

3.1 Image acquisition

Image acquisition could become a quite critical operation, depending on what follow in the cascaded blocks and depending on what is the final goal of the application.

Lighting and shooting distance in general heavely affect the acquired image, thus one should carefully decide the condition in which to operate.

For what concern light condition, the presence of too much light affect the edge detection (see Section 2 for more information). A related problem to lighting is the presence of shadows, most of all due to the morphology of the ear.

Decide the position of the camera with respect to the ear is another critical task; perspective distortion in acquisition can greatly modify the shape of the ear even if it is possible to retrieve the undistorted image having, for example, two or more circles with known dimension or 4 coplanar points in the image ([15, 24]).

Lastly we had to choose a reference unit in the image, since we wanted to have the images scaled at the same dimension (more details in Section 4, when we will explain the matching algorithm).

3.2 Ear Detection

With reference to Section 2, ear detection is a fundamental step that must be much reliable as possible to avoid severe errors in the phases that follow it.

We have used a manual tecnique to localize the exact position of the ear, throwing away the portion of the image we do not need. Manual detection is the method that ensure the most reliable results in terms of ear detection, for this reason the enhancement step is not needed.

3.3 Feature Extraction

In our methodology, feature extraction consist in an edge extraction from a database of images; the results of this operation sholud be as clean as possible, meaning having only the edge we were interested in (see Section 2 for further information). If needed, an additional manual cleaning is performed.

During the edge extraction and cleaning procedure, we have also stored (by manual selection) the coordinates of the ear canal, that will be useful in our implementation of the algorithm (details in Section 4).

For what concerns the user images, the acquired images have been processed in order to get the edges; the edged images have been used in the algorithm for edge matching.

3.4 Ear Contour Matching

Our approach to the matching problem is based on the Generalized Hough Transform. We will now proceed with a brief presentation of the Hough transform, both in its Standard and Generalized version.

3.4.1 Hough Transform

Standard Hough Transform (SHT) is a very popular voting algorithm for line and circle recognition inside images [16, 20].

Ballard [7] had generalized this algorithm to detect arbitrary shapes into images (from now on we refer to this generalization as Generalized Hough Transform, GHT).

Standard Hough Transform

Suppose we want to find circular boundaries into an image, whose cartesian equation is

$$(x-a)^2 + (y-b)^2 = 1. \tag{3.1}$$

We will work on binary images, that means we transform an RGB image into a black and white image, where white pixels are the ones where the magnitude of the colored image is above a certain threshold. We will call these white pixels *edge pixels*. For every edge pixel, if it lays on the circle, we can state that the locus for the parameters of that pixel is a circular cone like the one in Figure 3.2; this can be seen from equation 3.1 by fixing x and y and let varying a, b and r. The intersting aspect of the Hough Transform is that every edge point on the circle has the same value for the triplet parameter a, b, r; thus many of such circular cone will have a common point where they intersect.

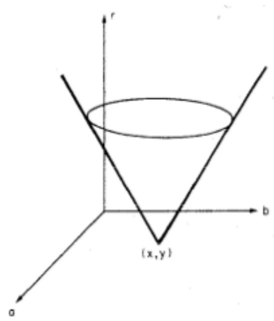

Figure 3.2. Circular cone: locus of the edge points laying on a circle [7].

In its original version [20] the SHT does not use gradient information; in [7] it is introduced the use of the gradient of the edges. A change in gradient is intended to be a point where there is a change of intensity (from black to white and viceversa).

Formally the gradient is the derivative of a function $f(x_1, x_2, ...x_n)$ along all the directions (say dimension) of the considered space

$$\nabla f = \frac{\partial f}{\partial x_1}\mathbf{e_1} + \frac{\partial f}{\partial x_2}\mathbf{e_2} + ... + \frac{\partial f}{\partial x_n}\mathbf{e_n} \tag{3.2}$$

where $\mathbf{e_i}$ are the orthogonal unit vectors pointing in the coordinate directions.

In two dimension, we have

$$\nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} \tag{3.3}$$

where $\mathbf{i}, \mathbf{j}$ are the standard unit vectors. If we consider a curve in its analytic form $f(\mathbf{x}, \mathbf{a}) = 0$, where $\mathbf{x}$ is an image point and $\mathbf{a}$ is a parameter vector, we get

$$\frac{\partial f}{\partial x}(\mathbf{x}, \mathbf{a}) = 0 \tag{3.4}$$

that introduce a term $\frac{dy}{dx}$, whose value is known and it is

$$\frac{dy}{dx} = \tan\left[\phi(\mathbf{x}) - \frac{\pi}{2}\right] \tag{3.5}$$

where $\phi(\mathbf{x})$ is the direction of the gradient. This way, the parameters locus reduces to a line (Figure 3.3), reducing the numebr of paramaters to manage in the Hough Space. The most important step in the SHT is the construction of the so called **accumulator array A**, a vector that keeps track of the number of entries for a certain triplet value **a,b,r** or, if using gradient direction information, the number of entries for the values of vector **a**. The steps to increment the value of **A** are the following:

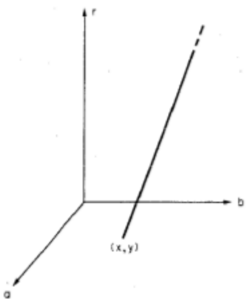

Figure 3.3. Locus of the edge points laying on a circle, using gradient direction information [7].

- for a specific curve $f(\mathbf{x}, \mathbf{a}) = 0$ with parameter vector **a**, form the accumulator $\mathbf{A}(\mathbf{a})$;
- initialize $\mathbf{A}(\mathbf{a}) = 0$;
- for each edge pixel **x** compute all **a** such that $f(\mathbf{x}, a) = 0$ and $\frac{\partial f}{\partial x}(\mathbf{x}, \mathbf{a}) = 0$;
- increment the corresponding accumulator entries $\mathbf{A}(\mathbf{a}) = \mathbf{A}(\mathbf{a}) + 1$.

After each pixel **x** has been considered, local maxima in **A** correspond to curves of f in the image.

Generelized Hough Transform

Now we proceed in explaining the generalization of the SHT for generic shapes, as stated in [7].

We can define the following parameters

$$\mathbf{a} = (\mathbf{c}, \mathbf{s}, \theta) \tag{3.6}$$

where $\mathbf{c} = (x_c, y_c)$ is the reference origin of the shape, θ is its orientation and $\mathbf{s} = (s_x, s_y)$ describe a scale factor. Figure 3.4 shows the parameter for a generic shape S.

The algorithm for finding the best set of parameters of **a** is almost the same as the SHT, the difference is that for GHT the distance between centre of the image and edge points is no more the same for all edge points.

The key to generalize the algorithm to arbitrary shape stands in the usage of directional information: they make the algorithm faster (reducing the number of parameters to compute in the Hough Space) and they improve the algorithm accuracy.

For each gradient point $\mathbf{x}$ we need to compute modulus and argument function of $\mathbf{r}$:

$$|\mathbf{r}| = \mathbf{a} - \mathbf{x} \tag{3.7}$$

$$\angle \mathbf{r} = \phi(\mathbf{x}) \tag{3.8}$$

where now $\mathbf{r}$ will vary in magnitude and direction.

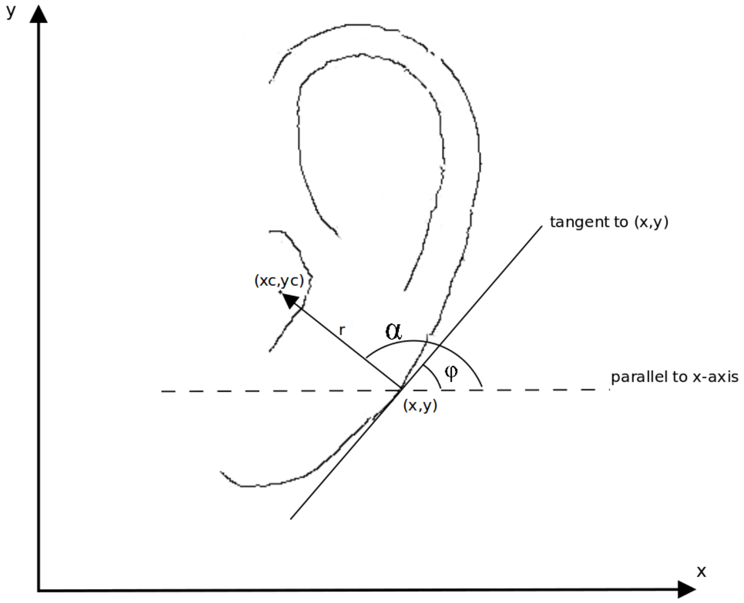

Figure 3.4. Parameters for a generic shape S: r and α are the polar coordinates of the reference point of shape, ϕ is the angle that the tangent to an edge point form with the x axis.

To keep track of a shape S, the so-called **R-table** is constructed as follows:

- choose a reference point $\mathbf{c} = (x_c, y_c)$ for the image
- for each edge pixel $\mathbf{x}$ compute gradient direction $\phi(\mathbf{x})$ and radius $\mathbf{r} = \mathbf{c} - \mathbf{x}$
- store $\mathbf{r}$ as funtcion of ϕ.

Referring to Figure 3.4, every entry of the R-table consist in a couple (r, α) for every value of the gradient ϕ. Before describing how the R-table is used to detect instances of the shape S, we define a scaling operator $T_s[R(\phi)] = s \cdot R(\phi)$ such that all vectors in the R-table are scaled by s. Similarly we define a rotation operator $T_\theta[R(\phi)] = Rot\{R[(\phi - \theta)\, mod\, 2\pi], \theta\}$ such that all indeces are incremented by $-\theta\, modulo\, 2\pi$, the right $\mathbf{r}$ vectors are found and rotated by θ.

Figure 3.5 may help in understanding better: an edge pixel with orientation ϕ can be considered as corresponding to the boundary point $\mathbf{x_A}$ with reference point $\mathbf{y_A}$ or can be considered as $\mathbf{x_B}$ on a rotated instance of the shape with reference point $\mathbf{y_B}$ that is $\mathbf{r_A}$ translated to $\mathbf{x_B}$ and rotated by $+\Delta\theta$.

The R-table is then used to detect istances of the shape S following these steps:

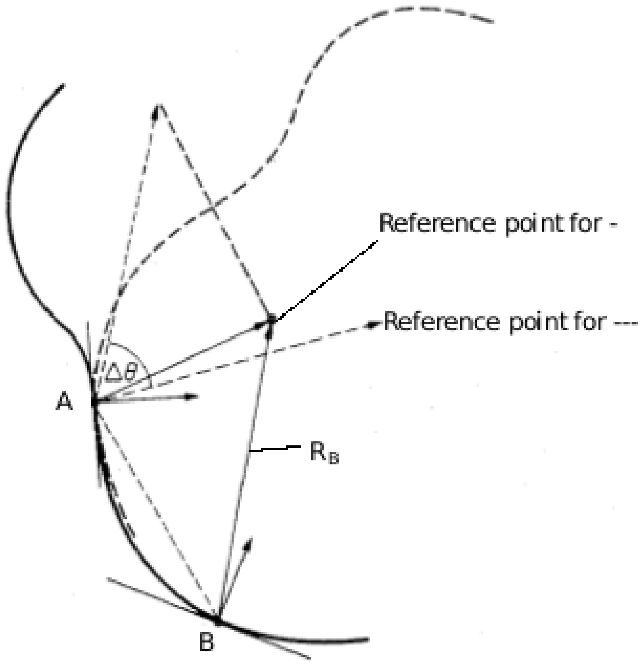

Figure 3.5. Transform of the R-table by a $\Delta\theta$ rotation. Point A can be viewed as on the continous-line shape or as point B on the dashed-line shape, rotated by $\Delta\theta$. In this last case the appropriate $\mathbf{R}$ is obtained by translating $\mathbf{R_B}$ to A and rotating it by $\Delta\theta$.

- for each edge pixel $\mathbf{x}$ in the image, increment all the corresponding points $\mathbf{x} + \mathbf{r}(\phi)$ in the accumulator array $\mathbf{A}$, where $\mathbf{r}(\phi)$ is an R-table entry indexed by ϕ;
- apply $T_s[R(\phi)]$ and $T_\theta[R(\phi)]$ for every ϕ to each R-table related vector;
- maxima in $\mathbf{A}$ correspond to possible instances of the shape S.

3.5 Decision

The decision step is inherently part of the Ear Contour Matching step; in this section we want to motivate the choice to take the maxima returned from the GHT procedure.

When the GHT procedure ends, the accumulator $\mathbf{A}$ contains, for each row, the possible locations of the reference point of the reference image S (previoulsy called generic shape S) in the database images: this is the voting mechanism on which the GHT relies. The maxima in the accumulator return the position of the reference point, and thus the correct scale and rotation of the database ear. We collect, for each ear in database, the maxima and then we chose among them the maximum: this will be the best match between the acquired ear and the ears in database.

Once the ear has been obtained, the associated HRTF is known and the procedure is over.

Chapter 4

Implementation

In this chapter we will explain and motivate our solutions for the methodology described in Section 3.

4.1 Image acquisition

Acquisition of the image has been done with a simple mobile phone camera, which nowadays have enough resolution for our purpose, in a controlled environment for what concern light exposure, angle and distance of acquisition. On this picture has been performed edge extraction.

For what concern light, test with different ligthing has been performed resulting in wrong or too incomplete edges when image is too dark or there was too much light. A good way to proceed is avoid direct intense light and prefer a more diffuse one (as reference, a daylight lighting is a good compromise). We have used light of three neon placed under the ceiling in a C-like figure that allowed a very diffuse and distributed light, avoiding shadow zones except for the one created by the ear itself.

The choice of the acquiring angle and distance has been made on a simulation of what an user should do to get a self-picture of its ear: this is because the future development of this work will give the user the chance to get its personalized HRTFs using just a picture of its ear.

Acquisition angle is 90 ° on the right with respect to the front direction. The choice of moving to the right instead of moving to the left is arbitrary: it would make no difference choose one with respect to the other since the algorithm works both with left and right ears.

Acquisition distance is about 40 cm, that is the distance between the shoulder and the elbow: the reason is still trying to simulate a user behaviour. This distance could greatly affect the step after image acquisition in the process: by taking a reference measure of the ear this problem can be avoid, as we will explain in the following.

To get a measurement reference in the image, we have analyzed three approach: using a coin with concentric circles (1 eur), using a card with fixed dimension (a driver license) or get the measure manually. One of the reason for the first two object is,as already said, that a user must be able to take a picture of its ear in every moment, thus using a common object that can be found easily. The other reason is that with at least two circle, with known dimension, or with 4 coplanar points (like the corner of the driver license) in an image is possible to remove perspective distortion ([15, 24]) that a user cannot avoid, especially if the picture is taken by himself.

We have faced great problems in using a coin, due to its thickness that affected the individuation of the correct edges. We have tried to get also the edges from the driver license without any relevant result. Since this is a pre-engineered work, we have decided to get the ear measure manually, by placing a rule near the ear. Reference measure have a great importance since help in removing prespective distortion due to the acquiring distance. Taking a picture nearer the ear will result in a bigger image of it while taking a picture from a higher distance will result in a smaller ear: having the true dimension of the ear one is capable to scale the image to the correct size.

4.2 Ear Detection

We decided to perform this step in a manual fashion. This technique is the most reliable (in terms of accuracy in finding edges) and it is the fastest to implement.

We decided to crop manually the the user image to get the portion of it in which we were interested in, by using an image manipulation program (GIMP).

Since manual detection gave us the exact portion of the image we were intersted in, we have decided to not perform image enhancement, removing this step from the chain.

4.3 Feature Extraction

Feature extraction step consist in a simple edge extraction.

A gaussian smoothing filter on the image has been applied prior to perform a canny edge extraction to reduce the noise, that are hair in our case. After edge extaction, a threshold on the resulting image has been applied, to reduce the number of false edges. Then a non-maximal-suppresion step has been performed on the edges, to thin them, meaning removing unwanted spurious points on the edges.

An image manipulation program (GIMP) has been used on database images and to clean everything except the ear wanted ear contour (see Section 2). A manual cleaning of the edges has been applied also on too noisy user image to remove what was outside the ear contours, leaving unaltered everything else.

An example of the result of this chain can be seen in Figure 4.1.

4.4 Feature Matching

Before starting to explain our implementation of the Generalized Hough Transform algorithm, we want to spend few word on the choice of the database.

Ear recognition, as other biometric-based recognition, is dependent on the database upon which it is performed: there are a multitude of databases available for research purposes, but this work is related to ear biometrics and HRTFs, so the

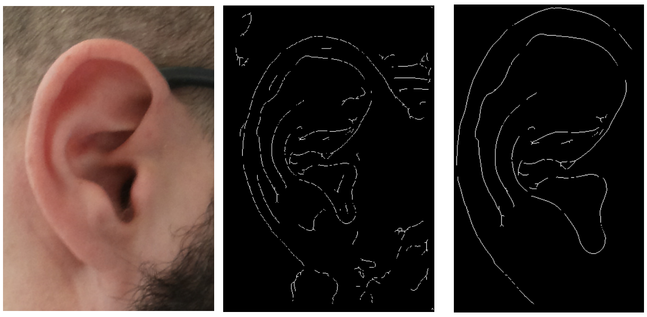

Figure 4.1. From left: user image, edge extraction, edges after cleaning.

database on which this work rely is the CIPIC database [4], with 160 subject and HRTFs and ear photograph for each one.

The choice of this database has several reasons: it contains, for every subject, HRTF and photos of the ear, it has been used by Spagol et al. [36], that is the article that inspired this work, and it is probably the most used database for HRTF testing.

Nevertheless, this database as some drawbacks: the lighting in the ear images vary, making difficult to perform processing on them; the resolution of the images is too low, affecting again the processing.

4.4.1 Hough Transform

We decided to use two different implementation of the GHT algorithm, and use each one of them for evaluating the other in terms of efficiency and error, even if computing the error in our case is possible only by performing some subjective test (results can be read in Section 5).

First implementation uses polar coordinate. We decompose it in two parts; the first follows this steps:

1. Find the barycenter of the database image
2. Find the coordinates of the contour points of the database image
3. Construct the R-table as follow
   (a) given A(x, y) contour point of the database image, find ρ and α polar coordinate of A(x, y) with respect to the barycenter
   (b) store x, y, ρ and α in a row of the table

The R-table allows to use the contour edge points and gradient angle to recompute the location of the reference point.

The second part of the algorithm concerns the detection of the shape:

1. quantize the parameter space $H[x_{c,min}, x_{c,max}][y_{c,min}, y_{c,max}]$
2. find the coordinates (x, y) of the countour points of the user image, for each one

(a) using gradient angle ϕ, retrieve from the R-table all the values of (ρ, α) for every ϕ_i

(b) for each value of (ρ, α), compute the candidate reference points as

$$x_c = x + \rho \cdot \cos(\theta) \tag{4.1}$$

$$y_c = y + \rho \cdot \sin(\theta) \tag{4.2}$$

where S is a scaling factor and θ keep into account possible rotations of the shape.

3. increase the counter (voting mechanism) $H = H(x_c, y_c)$

Possible locations of the barycenter, and thus position fo the contours, are given by local maxima in $H(x_c, y_c)$.

Now we want to expand the above algorithm for a generic rotation θ and a uniform scaling S $(x', y') \rightarrow (x'', y'')$:

$$x'' = (x' \cdot cos(\theta) - y' \cdot sin(\theta)) \cdot S; \tag{4.3}$$

$$y'' = (x' \cdot cos(\theta) + y' \cdot sin(\theta)) \cdot S \tag{4.4}$$

The coordinate of the new centre will be

$$x_c = x - x'' = x - (x' \cdot cos(\theta) - y' \cdot sin(\theta)) \cdot S; \tag{4.5}$$

$$y_c = y - y'' = y - (x' \cdot cos(\theta) + y' \cdot sin(\theta)) \cdot S \tag{4.6}$$

The algorithm become

1. quantize the parameter space $H[x_{c,min}, x_{c,max}][y_{c,min}, y_{c,max}][\theta_{min}, \theta_{max}][S_{min}, S_{max}]$

2. find the coordinates (x, y) of the countour points of the user image, for each one

   (a) using gradient angle ϕ, retrieve from the R-table all the values of (r, α) for every ϕ_i

   (b) for each value of (r, α), compute the candidate reference points as

$$x' = r \cdot \cos(\alpha) \tag{4.7}$$

$$y' = r \cdot \sin(\alpha) \tag{4.8}$$

$for(\theta = \theta_{min}, \theta \leq \theta_{max}, \theta ++)$
$for(S = S_{min}, S \leq S_{max}, S ++)$

$$x_c = x - (x' \cdot cos(\theta) - y' \cdot sin(\theta)) \cdot S; \tag{4.9}$$

$$y_c = y - (x' \cdot cos(\theta) + y' \cdot sin(\theta)) \cdot S \tag{4.10}$$

3. increase the counter (voting mechanism) $H = H(x_c, y_c, \theta, S)$

Again, possible locations of the barycenter are given by local maxima in $H(x_c, y_c, \theta, S)$.

In this implementation we consider as barycenter the coordinates of the ear canal, that we had acquired manually in the acquisition step.

The Hough Transform is known to be time- and memory-consuming, but it has some useful advantages:

- robust to partial or slightly deformed shapes;
- robust to the presence of other object in the image (lines, curves, etc.);
- robust to noise;
- can find multiple occurrences of the shape in the same processing pass

To speed up the algorithm, the scale factor is computed outside the GHT algorithm, using the height of the input ear (measured manually when acquiring the picture of the ear, as explained in Section 4.1) and the height of the database ear stored in a separate file (available within the CIPIC database download).

The rotation factor can be freely chosen, by default it is an array containing values from $-15\,^\circ$ to $+15\,^\circ$ with $5\,^\circ$ step: it seems a reasonable range of values by looking the possible rotation of the ears in database.

The second implementation is described in the following, uses cartesian coordinates and uses a different version of the R-table. Same reasoning for scale and rotation apply also to this version of the algorithm.

1. Find the coordinates of the contour of the user image
2. Find the coordinates of the contour of the database image
3. For every value of rotation
   (a) apply rotation to database image
   (b) resize user image to match (if needed) dimension of rotated image
   (c) sum user and database image
   (d) store result in R-table with rotation value

The most similar contour will be the one with the maximum number of countour points in common with user image.

4.5 Decision

With reference to Section 2, we choose the maximum between the maxima returned by both implementations. In the first case we get the maximum number that indicate the barycenter of tha database image that best matches the contours of the user image; in the second case we get the maximum number of common points between the edges of the user image and the database images.

We expect that the two algorithms give the same match, but we are aware that for very similar morphology of two ear in the database they can return two different results.

Chapter 5

Test Methodology and Results

5.1 Test Methodology

In this section we propose formal listening tests with the aim of providing a subjective validation of the previously introduced methodologies. In particular, we will discuss in detail the design phase of listening experiments, presenting the criteria that most suite the need of our work.

An important factor that we have considered during the design phase of our listening test is that the listeners must not be overloaded, in order to prevent a loss of accuracy in the judgments.

The usual procedure for subjective test involving HRTF and localization is the following: a short training phase is performed prior to the subjective test. This phase is necessary for several reasons: the first and most important is to give the subject the idea of how externalization is perceived in headphones.

Externalization is the opposite phenomenon of lateralization: anyone listening to the headphones has probably found the sound sometimes to be localized in the head. The sound field becames flat and lacking the sensation of dimensions.This phenomenon is often referred in literature as lateralization, meaning "in-the-head" localization. Lateralization occurs because the information in which the human auditory system relies when positioning the sound sources is missing or ambiguous. The problem is emphasized on the recording material, that is originally intended to be played via speaker systems.

Externalization means essentially "out-of-head" localization and aims to give the the perception thath the source position moves from the axis between the ears to outside the head.

The second reason to perform a training phase is to allow the subject to be more confortable with the request of pointing the direction of the perceived source.

A graphical aid is sometimes provided to assist the listener with the intended location of the externalized source test.

For what concern the listening test, the subjects usually are provided with a visual reference of the virtual scene on which indicate the direction of the source. Another popular solution is to use an head tracker attached to the headphones, asking the subject to look at the direction of the source after the sound stopped and track the movement of the head.

Our choice has been to sample the walls of a room with 204 points, subsampling the CIPIC measurement space (made by 1250 points), using azimuth positions from $-80\,^\circ$ to $80\,^\circ$ with $10\,^\circ$ step (where $0\,^\circ$ is the direction coming from the front with no elevation) and elevation positions from $-30\,^\circ$ to $210\,^\circ$ with $30\,^\circ$ step (where $0\,^\circ$ is the direction coming from the front with no azimuth). Since this is a preliminar test, we used a quite large step on elevation because we were interested in understanding if the subject was capable of discerning sources coming from upward or downward, coming from the front or the back and coming from left or right, thus requiring low precision in the exact source direction.

The subject has been equipped with a red-light laser, and we asked him to point with it the direction from which the source was perceived, trying to keep the light of the laser as a continuation of the line joining elbow and wirst.

For each stimulus we asked to the subject to give a number between one and five to indicate the confidence of the perceived direction, where one means totally sure and five means no externalization.

The sound used is a male speech from European Broadcasting Union (EBU) Sound Quality Assessment Material (SQAM) CD [18]. Such sound sample comes from easily accessible sources and has already been adopted for listening tests in the context of sound field rendering evaluation [23]. We choose a vocal signal because human voice is considered critical to evaluate the sense of audio quality and it is known that localization is most sensitive with speech [23]. Test sample has been first converted into monophonic signal, and then it has been filtered with the left and right HRTF(s) returned by the algorithms to obtain the virtual source for a given pair (azimuth,elevation).

As already explained in Section 4.5, we expected same result for both algorithm. Since the two algorithm could return different ear match, we designed a test suitable for both the possibility: this way we are also capable of evaluating the the goodness of our algorithms both in case of same or different match.

We choose to test ten position in our sampled space, having ten stimuli per match, for a total of twenty stimuli played as follows: for every position, first stimulus was the result of the filtering between test sound and HRTF associated to the ear match of the first algorithm and second stimulus was the result of the filtering between test sound and HRTF associated to the ear match of the second algorithm. After five position (meaning ten stimuli played, five for each match), we gave a one minute break to the subject to avoid its overload, thus reducing the chance of getting random answers. We ended the test by playing last ten stimuli in the already explained fashion.

If the algorithms gave the same result, we have used as second match an HRTF as different as possible with respect to the first match, chosing the minimun value between all the maxima, with reference to Section 4.4.

If the algorithms gave different results, we have followed the test procedure as explained.

In both cases we have been able to judge the matching: in first case we expected a number of correct guesses for first stimuli higher than the number of correct guesses for the second ones, while in second case we expect the same number of correct guesses for both first and second stimulus. This analysis will be done in Section 5.

In the design phase of such experiments we had need to ensure that uncontrolled factors will not deviate the results of the listening tests. For this reason we have proposed to each subject the same source positions for first and second stimulus but in a different and random order. In this way we can ensure that the judgments made by the subjects are independent from the actual sequence of stimuli.

5.2 Results

In this section we present the results of the listening test performed on 15 subjects; 11 of them are experienced in listening to sound in a critical way. In particular, the test subjects are staff members of the *Sound and Music Computing Lab* and students coming from the Masters' Degree in *Sound and Music Engineering* offered by Politecnico di Milano (Polo Regionale di Como).

We can state that the proposed methodology, even if in its first stage, had some potential to develop. Figure 5.1 shows that indeed some locations are best rendered with respect to others. The locations in front and on the back of the listener state that the "front-back confusion"effect is still an issue, which causes must be further invastigated. Also some lateral location have not been correctly located: this is probably due to another well know issue, the so.called "cone of confusion". As already said, the ability of resolving the location of a sound source is made possible by using interaural time differences and interaural level differences. However no such time or level differences exist for sounds originating along the circumference of circular conical slices, where the cone's axis lies along the line between the two ears. Consequently, sound waves originating at any point along a given circumference slant height, that is the distance from any point on the circle to the apex of the cone pointing to the ear canal, will have ambiguous perceptual coordinates. That is to say, the listener will be incapable of determining whether the sound originated from the back, front, top, bottom or anywhere else along the circumference at the base of a cone at any given distance from the ear. Of course, the importance of these ambiguities are vanishingly small for sound sources very close to or very far away from the subject, but it these intermediate distances that are most important in terms of fitness.

Figure 5.1 shows mean values of the test locations (red crosses) and their 95% confidence values (blue lines) for azimuth and elevation. Blue circles are position of the virtualized source.

We can notice that some red crosses are missing (for elevation values around 200 ° for stimulus 3 and 4): for this values subject were not able to correctly indentify if the source was coming from the back or from the front (front-back confusion, [31, 39]).

Also we can notice that some confidence interval are missing: this is because we have chosen a quite severe interval to evaluate the algorithm.

For certain values we appreciate a great displacement between real position and perceived position of the source: for example for azimuth 0 ° and third stimulus there is a misplacement of 30 °, even bigger for fifth stimulus. We expected great error in elevation, since this is a common problem for non-personalized HRTFs; the error in azimuth perception is probably due to some unreliable subject.

As already explained in previous paragraph, wrong externalization made impossible locating the direction of the sound outside the azimuth plane. Figure 5.2 shows

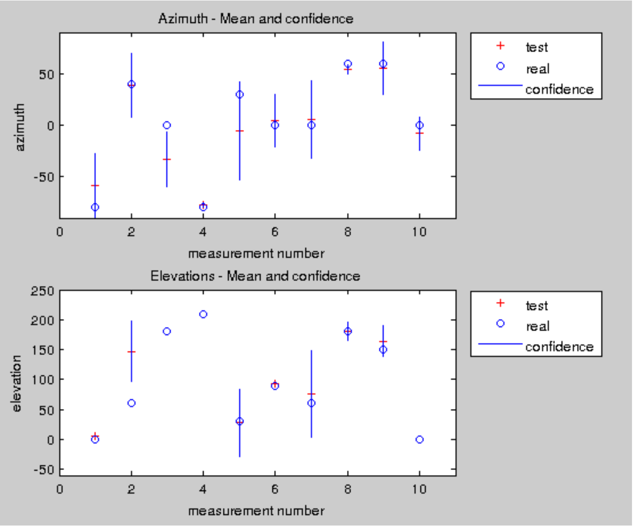

Figure 5.1. Values of mean and 95% confidence (blue lines) for azimuth and elevation. Blue circles are position of the source, red crosses are the mean of the position perceived by subjects.

the mean, for each stimulus, of inside head source: worst case is for stimulus seven with a 1.88% of the subjects reporting bad externalization.

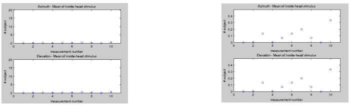

Figure 5.2. Left image - Lateralization effect: for each stimulus few subject perceived the source inside their head. Right image - Detailed y-axis scale: the error can be better appreciated.

With reference to the test methodology, we asked to the subjects to judge their assesment of the source localization by means of numerical values between 1 and 5 (where 1 means absolutely sure and 5 means impossible to judge). Figure 5.3 show an example of subject confidence for evaluation of source direction for stimulus eight: only two subjects rate their perception as unreliable (rate 4 and 5), while the majority were pretty sure of the direction.

An interesting result from this test has been that almost every subject had complained difficulty in pointing the direction not because this was not clear but because was not easy poiting without following with the rest of the body the direction of the sound.

As expected, our algorithms gave for the 78% of the subject the same result: this mean that only four subject out of fifteen presented two different result, which is a small number to use for cross-validating the algorithms. Nevertheless we can still state that our approach is valid, since two quite different implementation of the same algorithm gave the same results.

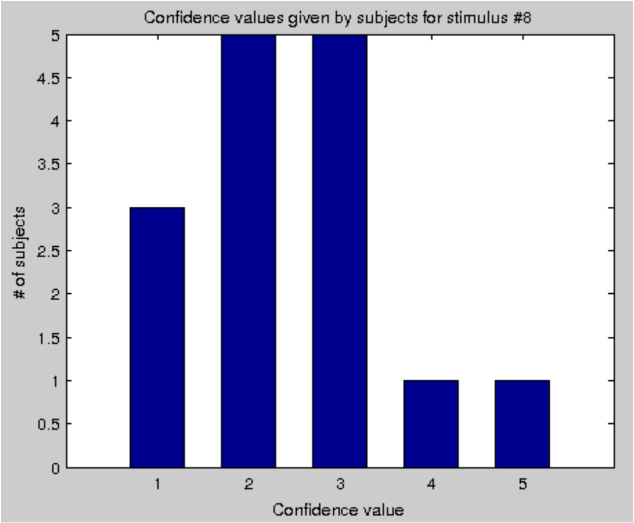

Figure 5.3. Example of subject confidence in judging the direction: the majority rate with 2 and 3, meaning they were pretty sure about the direction.

Chapter 6

Conclusions and Future Works

In this work we have presented a novel approach for selecting an Head-Related Transfer Function through ear contour matching for binaural rendering.

The usual procedure for retrieving an HRTF from a multitude is based on subjective approaches base on listening tests. Since the judgments of a human listener are involved, the results of these tests take human perception into account.

The dual procedure for personalized binaural rendering is measure the HRTF of the subject. However, both methods have high cost, making their use not possible in a large number of context.

In this work we have proposed an alternative approach using a 2D image of the ear that provides a psychoacoustic evaluation obtained with metrics that are representative of human perceptive experience.

We believe that using a larger database, with better picture quality, the results can only increase in terms of quality of best matching. We believe also that increasing the number of edges on which we perform the match can lead to slightly better results due to the few remaining edges not included in the algorithm.

6.0.1 Future works

This work may be inserted in a larger project that have the purpose of making possible for every user who wants to try the binaural experience at his home to take a picture of its ear and get, in few seconds, his personalized binaural rendering.

Some step must be followed since this will be possible: ear detection must be done automatically as the procedure of cleaning the results of the edge extraction; also the retrieving of the dimension of the ear should use one of the two proposed apporach (coin or driver license).

The precision of the algorithm can be increased by taking into consideration also luminance of the gray-scale picture, using it as third dimension (more light means much height) to use also the 3D shape of the ear, thus having another parameter that can be added to the Hough Space.

Ringraziamenti

Per primi vorrei ringraziare i miei genitori, grazie ai quali ho avuto chiara sin da subito la strada che avrei voluto seguire e che mi hanno permesso di percorrerla fino ad oggi.

Vorrei ringraziare anche il Prof. Sarti, grazie al quale ho potuto lavorare non solo a questa tesi ma ho potuto partecipare ad una serie di progetti collaterali a cui difficilmente si può avere accesso. Sento di non aver concluso questo lavoro nella maniera che avrei voluto e me ne dispiaccio.

Vorrei ringraziare tutto il gruppo del Laboratorio di Sound And Music Engineering, in particolare il Prof. Antonacci e l'Ing. Bianchi che mi hanno assistito in questo lavoro e che lo hanno reso più facile grazie ai loro preziosi consigli.

Un pensiero va anche a mia sorella, che credo sia più entusiasta di me per questo traguardo raggiunto, nonostante io creda di non essere ancora riuscito a spiegarle di preciso la mia figura di cosa è in grado di occuparsi.

Un ringraziamento a coloro i quali reputo i miei amici: per motivi simili abbiamo dovuto separarci ma nonostante questo sono conscio che il rapporto che abbiamo costruito negli anni di studio passati insieme è qualcosa che supera senza problemi tutte le barriere territoriali che ci dividono.

Ringrazio Nicola, più che un amico un fratello: 25 anni non sono pochi, dovremmo pensare alle nozze d'argento.

Ringrazio Vito perchè non ha mai smesso di non credere in me: oggi ho avuto la mia rinvicita.

Ringrazio infine Dario: abbiamo trovato il modo di essere vicini anche se lontani.

Li ringrazio anche per la breve ma intensa esperienza musicale che ci ha accomunati: probabilmente è un bene che sia finita ma è stata bella ugualmente.

Infine un ringraziamento a Francesca: mi ritengo fortunato, la tua personalità e il tuo appoggio incondizionato mi hanno aiutato in questi anni, questo traguardo è anche merito tuo.

Bibliography

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon. A survey on ear biometrics. *ACM Comput. Surv.*, 45(2):22:1–22:35, Mar. 2013.

[2] M. Abdel-Mottaleb and J. Zhou. Human ear recognition from face profile images. In D. Zhang and A. Jain, editors, *Advances in Biometrics*, volume 3832 of *Lecture Notes in Computer Science*, pages 786–792. Springer Berlin Heidelberg, 2005.

[3] Acoustic Research Institute. Ari database.

[4] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102, 2001.

[5] V. R. Algazi, R. O. Duda, and P. Satarzadeh. Physical and filter pinna models based on anthropometry. In *Audio Engineering Society Convention 122*, May 2007.

[6] S. Ansari and P. Gupta. Localization of ear using outer helix curve of the ear. In *Computing: Theory and Applications, 2007. ICCTA '07. International Conference on*, pages 688–692, 2007.

[7] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2):111–122, Jan. 1981.

[8] D. Batteau. The role of the pinna in human localization. In *Proc. R. Soc. Lond. B*, 1965.

[9] A. Bertillon. *La photographie judiciaire, avec un appendice sur la classification et l'identification anthropometriques.* Gauthier-Villars, 1980.

[10] J. Blauert, M. Brueggen, A. W. Bronkhorst, R. Drullman, G. Reynaud, L. Pellieux, W. Krebber, and R. Sottek. The audis catalog of human hrtfs. volume 103, pages ASA+, 1998.

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

[12] M. Burge and W. Burger. Ear biometrics. In *BIOMETRICS: Personal Identification in a Networked Society*, 1998.

[13] M. Burge and W. Burger. Ear biometrics in computer vision. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 822–826 vol.2, 2000.

[14] S. Carlile. The physical and psychophysical basis of sound localization. In *Virtual Auditory Space: Generation and Applications*, Neuroscience Intelligence Unit, pages 27–78. Springer Berlin Heidelberg, 1996.

[15] Y. Chen and Horace. Planar metric rectification by algebraically estimating the image of the absolute conic. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 88–91 Vol.4, Aug 2004.

[16] R. O. Duda and P. E. Hart. Use of the hough transformation to detect lines and curves in pictures. *Commun. ACM*, 15(1):11–15, Jan. 1972.

[17] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058, 1998.

[18] EBU Tech. 3253. Sound quality assessment material: recordings for subjective tests, 9 2008.

[19] M. Geronazzo, S. Spagnol, and F. Avanzini. Estimation and modeling of pinna-related transfer functions. In *Proc. 13th Int. Conf. Digital Audio Effects (DAFx-10)*, 2010.

[20] C. Hough. Method and means for recognizing complex patterns, december 1962. US Patent 3,069,654.

[21] D. J. Hurley, M. S. Nixon, and J. N. Carter. Force field feature extraction for ear biometrics. *Computer Vision and Image Understanding*, 98(3):491 – 512, 2005.

[22] A. Iannarelli. *Ear Identification, Forensic Identification Series.* Paramount Publishing Company, Fremont, CA., 1989.

[23] B. Klehs and T. Sporer. Wave field synthesis in the real world: Part 1 - in the living room. In *Audio Engineering Society Convention 114*, Mar 2003.

[24] M. I. A. Lourakis. Plane metric rectification from a single view of multiple coplanar circles. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 509–512, Nov 2009.

[25] D. Lowe. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157 vol.2, 1999.

[26] L. B. Majdak P., Balazs P. Multiple exponential sweep method for fast measurement of head related transfer functions. *Journal of the Audio Engineering Society*, 2007.

[27] J. C. Makous and J. C. Middlebrooks. Two dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

[28] B. Moreno, A. Sanchez, and J. Velez. On the use of outer ear images for personal identification in security applications. In *Security Technology, 1999. Proceedings. IEEE 33rd Annual 1999 International Carnahan Conference on*, pages 469–476, 1999.

[29] M. Morimoto and H. Aokata. Localization cues in the upper hemisphere. *J. Acoust. Soc. Jpn. (E), vol 5*, 1984.

[30] Z. Mu, L. Yuan, Z. Xu, D. Xi, and S. Qi. Shape and structural feature based ear recognition. In S. Li, J. Lai, T. Tan, G. Feng, and Y. Wang, editors, *Advances in Biometric Person Authentication*, volume 3338 of *Lecture Notes in Computer Science*, pages 663–670. Springer Berlin Heidelberg, 2005.

[31] 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, 1996.

[32] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, 118(1):364–374, 2005.

[33] B. U. Seeber, H. Fastl, A. T. Akustik, and T. München. Subjective selection of nonindividual head-related transfer functions. In *In Proceedings of the 2003 International Conference on Auditory Display*, pages 1–4, 2003.

[34] Spagnol. Pinnaanatomy.

[35] S. Spagnol, M. Geronazzo, and F. Avanzini. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, pages 194–199, Oct 2010.

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

[37] R. Tame, D. Barchiesi, and A. P. Klapuri. Headphone virtualisation: Improved localisation and externalisation of non-individualised HRTFs by cluster analysis. In *Proceedings of the Audio Engineering Society Convention*, 2012.

[38] C. A. V. R. Algazi and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 2001.

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