POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Master of Science in
Computer Engineering

3D Models Extraction For Personalized Binaural Audio Applications

Candidate

Luca Bonacina
Student Id. number 795845

Thesis Supervisor

Prof. Augusto Sarti

Assistant Supervisor

Dr. Antonio Canclini
Dr. Marco Marcon

Academic Year 2014/2015

POLITECNICO DI MILANO

Scuola di Ingegneria dell'Informazione

POLO TERRITORIALE DI COMO

Laurea Magistrale in
Ingegneria Informatica

Estrazione Di Modelli 3D Per Applicazioni Di Audio Bineurale Personalizzato

Candidato

Luca Bonacina
Matricola 795845

Relatore	Correlatore
Prof. Augusto Sarti	Dr. Antonio Canclini Dr. Marco Marcon

Anno Accademico 2014/2015

3D Models Extraction For Personalized Binaural Audio Applications
Master thesis. Politecnico di Milano

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

Author’s email: Luca5.Bonacina@mail.polimi.it

Sommario

L'audio spaziale e la sintesi bineurale sono oggigiorno una componente fondamentale in numerosi sistemi commerciali di domotica e hi-fi. In questo mercato, i principali attori si affidano quasi esclusivamente a sistemi di altoparlanti per la riproduzione della scena acustica e fino ad ora le cuffie non sono state altrettanto impiegate. Un esteso utilizzo delle cuffie nella sintesi bineurale troverebbe applicazione in molteplici scenari dell'industria del multimedia. La capacita di localizzare nello spazio una sorgente sonora dipende fortemente dalla forma di orecchio,testa e torso di ciascun individuo. La loro anatomia determina specifiche caratteristiche di filtraggio che permettono di percepire il suono nell'ambiente. Queste caratteristiche sono codificate in una serie di filtri detti *Head Related Transfer Function* (HRTF). Tenendo conto dello stretto legame tra l'anatomia umana e le caratteristiche della HRTF, la necessita di una HRTF personalizzata è oggi riconosciuta. Le tradizionali tecniche di creazione di HRTF personalizzate si basano su acquisizioni acustiche complesse e costose; gli sviluppi delle tecniche numeriche, invece, permettono di calcolarle partendo da modelli 3D di testa, orecchie e torso. Tipicamente, questi modelli sono ottenuti tramite Laser Scanner o Risonanza Magnetica i quali forniscono risultati di alta qualità ma comportano alti costi e procedure complesse. La principale limitazione sta nella lunga durata delle acquisizioni. In questa tesi mi dedico alla generazione del modello 3D dell'orecchio utilizzando componenti commerciali a basso costo. L'obbiettivo principale è acquisire un modello 3D che possa trovare applicazione nel contesto dell'audio bineurale mantenendo il sistema di acquisizione il più semplice possibile. Il metodo proposto si basa su un approccio multi view che riduce notevolmente la complessità dell'acquisizione. Utilizzando un LeapMotion ed una Kinect si acquisiscono immagini dell'orecchio da diversi punti di vista che vengono poi processate per ottenere una *nuvola di punti 3D*. L'algoritmo di estrazione 3D sfrutta la *geometria epipolare* per associare coppie di punti nelle immagini che vengono poi triangolati per ottenere la nuvola di punti. Da ciascuna coppia di immagini si estrae una nuvola di punti, queste ultime sono poi fuse utilizzando la procedure di allineamento *iterative Closest Point*. La qualita dell'estrazione è valutata da un punto di vista sia geometrico che acustico. Nel primo caso la *Hausdorff Distance* è ultilizzata per confrontare il modello estratto (a bassa risoluzione) con un modello scansionato a Laser (ad alta risoluzione). Nel caso acustico, si calcola numericamente la HRTF da entrambi i modelli ad alta e bassa risoluzione. La metrica *Spectral Distortion* (SD) viene sfruttata per valutare le differenze di queste HRTF. I bassi valori di SD mostrano che la risoluzione del metodo proposto è sufficiente per ottenere HRTF personalizzate in un range frequenziale di [20Hz-5000Hz]. Inoltre, la relazione tra similarità acustica e geometrica di due modelli viene approfondita. La *Perceptually Weighted Hausdorff Distance* permette di includere considerazioni psicoacustiche attraverso un semplice sistema di pesatura: le parti del modello 3D percettivamente più rlevanti vengono pesate maggiormente delle altre. I risultati di questa analisi effetuata su 64 soggetti suggeriscono l'esistenza di una correlazione diretta tra somiglianza nel dominio geometrico e quello acustico.

Abstract

Spatial and binaural audio is nowadays a fundamental component of many domotic and hi-fi commercial systems. The main players in this market rely mainly on loudspeakers systems to reproduce the audio scene but so far the headphones haven't been largely exploited for this purpose. Moving the spatial audio rendering from loudspeakers to headphones would potentially enable a whole new set of scenarios for many fields of the media industry. The ability to localize a sound source in space strongly depends on the shape of each individual's ears, head and upper torso. Their anatomy determines specific filtering features that give us the perception of sound in the environment. Those features are encoded in a set of filters known as *Head Related Transfer Function* (HRTF). HRTF generation is of great interest in the research community; considering the strong link between specific human anatomy and associated HRTF features, the need of a *personalized* HRTF is nowadays undisputed. Traditional personalized HRTF generation relies on complex acoustical measurements that requires expensive hardware; advances in numerical techniques allow to predict the HRTFs starting from 3D models of the head, ear and torso. Typically those models are extracted by using a Laser Scanner or MRI. Even though they provide high resolution models, these solutions are extremely expensive and complex. The main limitation resides in the long duration of the scan session. In my thesis I focus on the generation of the 3D model of a person's ear using low-cost, off-the-shelf hardware. The main goal is to acquire 3D models that can find application in the context of binaural audio, while keeping the acquisition system simple and easy-to-use. The proposed 3D extraction method is based on a *Multi View* approach that greatly reduces the complexity of the 3D acquisition procedure. By using a Leap Motion and a Kinect device many images of the ear are captured from different viewpoints and then processed to obtain a 3D *point cloud*. The designed extraction algorithm works by exploiting the *Epipolar Geometry* to match couples of points in the acquired images; the matched points are then triangulated to extract the point cloud. One point cloud is extracted per each couples of captured images and then fused in a single 3D model by means of the *Iterative Closest Point* alignment procedure. The quality of the extraction procedure is then evaluated from both a geometrical and an acoustical standpoint. As far as the geometrical evaluation is concerned, the *Hausdorff Distance* is used to asses how an extracted, low-resolution ear model is similar to the correspondent ground-truth, Laser Scanned ear model. Regarding the acoustical evaluation, both the low and the high-resolution models are used to numerically compute the HRTFs. The *Spectral Distortion* metric is used to evaluate the similarity of those HRTFs; the high similarity values show that the resolution of the proposed method is high enough to effectively provide the personalized HRTF within a frequency range of [20Hz-5000Hz]. To further investigate how human antrophometry and HRTFs features are linked, the relation between geometrical and acoustical similarity is deeper studied. The *Perceptually Weighted Hausdorff Distance* is used to include psychoacoustical considerations in the geometric similarity computation. A simple weighting schema allows to give more importance to the psychoacoustically relevant parts of the 3D ear. The result of the analysis performed on a dataset of 64 subjects highlights the existence of a direct correlation between similarity in the geometric and acoustical domains.

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Spatial and binaural audio is nowadays a fundamental component of many domotic and hi-fi commercial systems. The main players in this market rely mainly on loudspeakers systems to reproduce the audio scene but so far the headphones haven't been largely exploited for this purpose, even though this problem is deeply studied in the literature. Moving the spatial audio rendering from loudspeakers to headphones would potentially enable a whole new set of scenarios for many fields of the media industry: a headphone based 3D audio rendering would be naturally applied to the VideoGame and the fast growing field of virtual reality, the Cinemas companies could cut the costs of the Hi Fidelity Loudspeakers and also the realism of the cochlear implants could be further engineered.

The ability to localize a sound source in space strongly depends on the shape of each individual's ears, head and upper torso. Their anatomy determines specific filtering features that give us the perception of sound in the environment. Therefore the knowledge of those filtering properties is fundamental for an effective headphone based spatial rendering.

Advanced signal processing techniques allow to filter a recorded sound in order to emulate its original position in the 2D or 3D space. These filters represent the frequency response of the outer ear that is named Pinna Related Transfer Function (PRTF) or the joint effects of Torso, Head and Ear frequency responses that, together, are called Head Related Transfer Function (HRTF).

Since the earliest studies [1] the need of taking into account the link between anatomy and filtering has been highlighted and the filters generation process has been tuned accordingly. Traditionally HRTF filters have been generated by acoustically measuring the ear transfer functions moving a loudspeaker (or a set of loudspeakers) around the listener's head and recording the signal at the eardrums with in-ear microphones. The task of numerically synthetize the HRTF/PRTF filters can be accomplished with two distinct approaches: one relies on generalized models of the human ear,torso and shoulders, in the other instead the generation procedure is personalized on the basis of the specific subject anatomy. The acoustic measures give the best result and represent the ground-truth for the software simulations. Regarding the former numerical approach, many solutions have been proposed in the literature: some of them based on structural or geometric models [2] [3], some on general anthropometric models [4] [5] [6] and some others on prediction representations[7]. The degree of personalization achieved by these methods is low and the generalization of the human sound localization process introduces high approximation. The latter numerical approach tries to reduce this approximation basing the whole HRTF synthesis on a 3D mesh model of the subject. To obtain the HRTF the 3D model is input to simulation software that can run different algorithms (FDTD [8], FEM[9], BEM [10], RAYTRACING[11][12]). The high personalization of the customized numerical simulations leads to a high fidelity HRTF synthesis.

In my thesis I mainly focus on the generation of the 3D model of a person's ear. This 3D model could be helpful in providing a better insight into the relation between HRTF and ear's geometry [13] and together with a good simulation algorithm could substitute the acoustical measurements.

Considering its importance, the extraction of the 3D model is a critical task. As it will be better explained, its quality directly affects the quality of the simulated HRTF: the higher the fidelity the lower the approximation error of the simulation algorithm. Among the many existing ways to acquire a 3D model of a scene (or a subject) Laser Scanners, Ultrasounds and MRI are the ones that have been used the most throughout the years. Even though MRI has been used for this purpose [10], these devices are extremely expensive and potentially harmful for the scanned subject. That's why some research centers prefer to acquire the model through a laser scanner, which represents a safer and cheaper solution. Laser scanners are not dangerous for the subject and they provide high resolution meshed models (a model in which the 3D shape is represented using triangles that connects the 3D coordinates of the model). On the other hand the high precision makes the whole extraction process sensible to very small displacements meaning that any little movement of the subject during the scan session would noticeably distort the results.

In my thesis I analyze the feasibility of some alternative methods and implement a framework for the extraction of the 3D model of the subject's ear based on infrared structured light and multi vision geometry. This thesis work is also motivated by the practical limitations of the above mentioned acquisition methods, in particular I focus on keeping the price of the experimental set-up low and on making the scanning procedure fast.

The scan time is a big concern, it can range from 10 minutes for MRI scans to hours for high quality Laser Scanners. In both cases the subject has to stay still during the acquisition and considering the duration of an average scan session this is not reasonable. That's why the procedure is broken down into many shorter scans that are post processed and combined. Even though shorter scans greatly reduce the sensibility to the subject's movement, the number of scans required to achieve a reasonable model quality increases a lot. Overall the acquisition takes long and it is often a complex task: for instance, in the case of laser scanners, the laser ray has to be orthogonally projected to the surface to acquire. In the case of complex surfaces like the ear's one this operation may take long time.

For these reasons alternative solutions have been proposed in the literature, as an example [14] obtained the 3D model laser scanning a mold of the ear: this approach is interesting because it makes the 3D model acquisition independent of the physical presence of the subject. This means that the mold can be scanned many times and the distortion due to the subject's movements are eliminated. Unfortunately the molding process could harm the subject's ear and extra material (silicon and syringe at least) is needed for each mold created. [15] includes an head tracking system to compensate the head movement and this effectively reduces the distortion of the mesh model but does not decrease the scan time.

In the chosen approach I exploit the computer vision theory of multi view geometry to extract the 3D model starting from two pictures taken from different viewpoints. With this method I capture many couples of images in few seconds with no need for the subject to stay still and I later process them in order to extract a point cloud. This procedure works better when the surface of the object in analysis is rich in details, for instance when it is textured, threaded or rough. Unfortunately the ear's skin is not well detailed, that's why I artificially enrich it projecting a structure light pattern. To capture the scene I use a LeapMotion device which is basically a system of two infrared cameras and three infrared LEDs (LEDs that illuminate light not visible by the human eye). To project the structured pattern instead I use a Kinect device. Please notice that thanks to its two cameras, the Leap Motion allows to capture the scene from two viewpoints at the same time. Also, since the cameras are sensible to infrared light only, the projected pattern is actually composed by infrared rays.

For each couple of images I obtain a partial 3D model in the form of point cloud: to obtain the total 3D model these parts have to be combined. To achieve this I use the *Iterative Closest Point* (ICP) alignment algorithm and I merge the resulting clouds.

The geometrical quality of the 3D extraction is evaluated by comparing the extracted clouds with ground-truth models; this is achieved by computing a shape similarity metric that allows to asses how much models are different. The generated models are also used to numerically estimate the HRTFs by means of suitable simulation software. These are compared to those predicted from the ground-truth models; an acoustical similarity metric is defined so to objectively determine the acoustical relevance of the 3D extraction procedure. As it will be shown, the proposed system provides clouds that can be effectively used to compute the personalized HRTFs without introducing noticeable error. Considering the binaural application context, a similarity metric that takes into account perceptual features is also defined. Using this measure it is possible to recognize, among many 3D models, the one whose computed HRTFs are the most similar to those generated from the tested model.

This thesis is organized as follows:

Chapter 2, the state of the art is reviewed, after a brief recap on basic spatial audio techniques the HRTF generation methods are analyzed: Acoustical, numerical and structural approaches are presented and compared. An analysis of the 3D extraction tools such as Laser Scanners, MRI and ultrasound is then provided.

Chapter 3 is devoted to the theoretical background. At first the camera related theory is presented: pin model, Radial Distortion, Multiple View and Epipolar Geometry are introduced. The second part is about 3D models, the ideas of 3D point cloud, Mesh and Cloud registration are explained together with the ICP algorithm. The last part describes the BEM technique and how to use it in order to estimate the HRTFs from a 3D model.

Chapter 4 provides an insight on the developed 3D extraction framework. After a brief introduction, the single stages of the workflow are discussed: the LeapMotion and the Kinect devices are described and all the practical details about the multi view approach and the infrared light projection are given. The practical steps to compute the HRTFs from a point cloud are also presented.

Chapter 5 reports the results of simulations and experiments conducted. The results of the calibration phase and examples of extracted clouds are presented. After having defined the quality metrics the results are evaluated from both an acoustical and geometrical standpoint. The relation between geometrical and acoustical similarity is also investigated.

Chapter 6 draws the conclusions and outlines possible future research directions.

Chapter 2

State Of The Art

In this chapter I provide an overview of the current Binaural Rendering technologies and 3D acquisition Systems. I start by introducing the main perceptual clues used by the human ear to localize sound in space. After a brief analysis of those, I give the definition of HRTF used throughout this work. Stated the fundamental role played by personalized HRTF in the localization process, I describe how to obtain it. The typical setup for acoustically measuring the HRTF is presented, together with its strength and limitations. I compare then two different approaches for HRTF synthesis: the Structural-Mathematical Models and the Numerical Predictions. In this case, the antrophometric information is extracted from 2D pictures of the subject. Even though they represent an appealing solution, their degree of personalization is limited by the lack of 3D details. Numerical predictions overcome this problem by taking advantage of 3D models of the ear,head and torso. They extract HRTF by simulating the sound scattering process, rather than approximating the HRTF structure itself. The obtainable personalization is higher; this, together with the advances in 3D acquisiton systems and computer vision, makes numerical predictions approaches more palatable for HRTF synhtesis. Considering the importance of 3D models, I present the main 3D acquisition systems. Range scanners controls the illumination of the obect to acquire by projecting light on it. They are comprised of a project and at least a sensor. They can work projecting light of different nature (Laser or diffuse) and measuring different quantities according to the type of object to aquire. Considering the application context (binaural rendering,in particular HRTF numerical prediction) laser scanner based on triangulation are preferred. Tomography imaging is another possible approach, it provides extremely high quality result but it's expensive.

This chapter is organized as follows: in section 2.1 I present the basic binaural cues and I define the HRTF. I also compare HRTF measurement systems with structural models and numerical predictions. In 2.2 I present the state of the art 3D acquisition systems.

2.1 Hedphone Based Binaural Rendering

The advances in mobile communication and IT technology (smartphones, high speed Internet, immersive gaming, 3D monitors, virtual reality...) call out loud for a more realistic and involving audio experience. Spatial audio rendering can provide a sense of being remotely immersed in the presence of people, musical instruments, and environmental sounds [16]. The immersive perception carried along with spatial audio technologies greatly improves the realism of the audio experience. Spatial sound technology has a long history. Until few years ago though, the majority of spatial sound systems were designed for loudspeaker reproduction only. Stereo and multi-channel are a familiar example. Also Ambisonics and Wave Field Synthesis use loudspeaker rendering and they provide great results. Though, there are many application scenarios where loudspeaker reproduction is not suitable at all. For this reason, the delivery of a high-quality spatial sound experience over headphones is of primary importance. Headphone based spatial rendering requires reproduction of the complex dynamic signals encountered in natural hearing. When both the earphones are used together, binaural techniques can be exploited. The idea behind binaural signal processing is to ad-hoc process the left and right signals separately so to emulate the phenomena typical of human hearing. Understanding any binaural technology requires knowledge of both the physics of sound propagation and the psychophysics of auditory perception.

Regarding sound perception, the human hearing system uses a number of cues to estimate the position of a sound source in a free field. The ones that hint localization the most are:

- the interaural time difference (ITD)
- the interaural level difference (ILD)
- monaural spectral cues that depend on the shape of the outer ear or pinna
- cues from torso reflection and diffraction
- the ratio of direct to reverberant energy
- cue changes induced by voluntary head motion
- familiarity with the sound source

In processing the left and right sound source signals, one has to use these cues to position the sound source in the desired virtual location. Although some of these cues are stronger than others, all of them should be consistently used. Indeed, as the human hearing system is complex, localization of a sound source can not be described completely by these separated cues. When a strong cue conflicts with a weak one, the strong cue will often dominate. The ITD and ILD are the primary cues for estimating the so-called lateral angle θ, that is the angle between the plane cutting the head vertically 2.1 and the line from the center of the head to the source position. The ITD describes the difference in arrival time when a wavefront hits the left and the right ear. The ILD tells about the energy what the ITD tells about the time: it provides information on the energy difference of the wave at the left and right ear. When the human anatomy is roughly approximated with simple geometry (as in the Snowman model [Approximating the head-related transfer function using simple geometric models of the head and torso]), the ILD and ITD can be easily computed. In his Duplex Theory Lord Raylegh [17] showed that the ITD provides

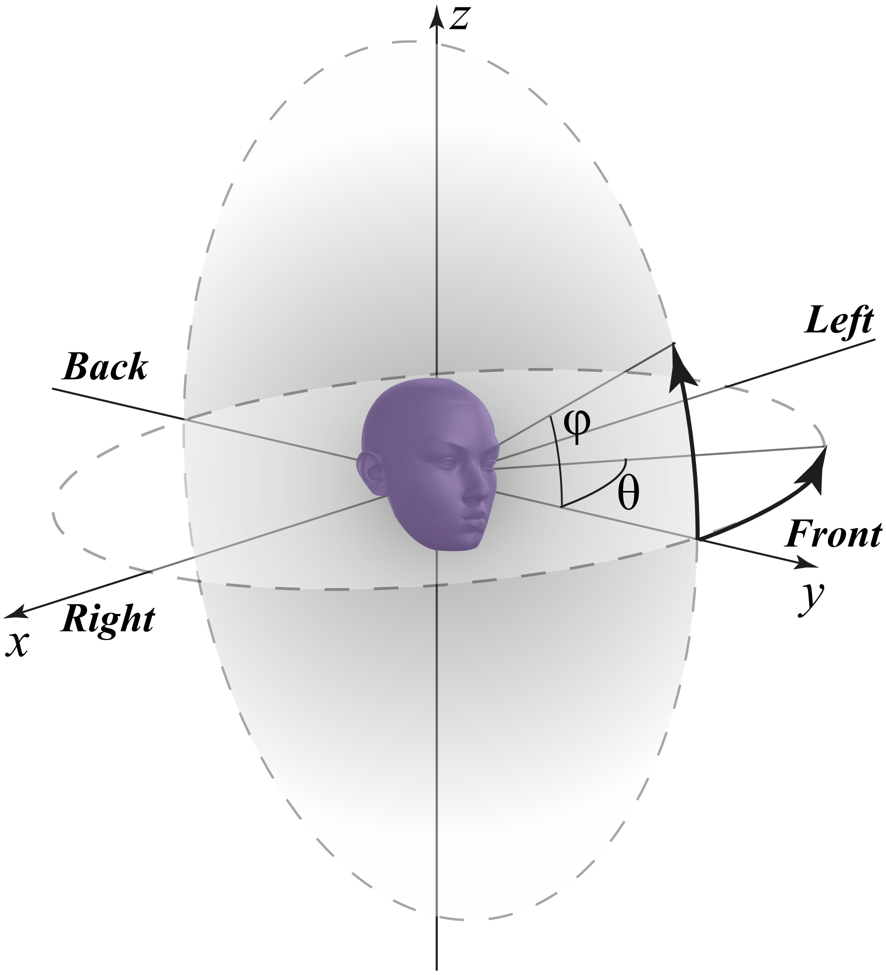


Figure 2.1. Coordinates System: the coordinate system used in this work is shown.

useful clues at low frequencies, where head shadowing is weak, and the ILD is more informative at higher frequencies, where the wavelength becomes comparable with the ear size. The crossover frequency between the two has been found to be around 1.5 kHz. While Duplex theory can effectively model the left-right displacement of a sound source, unfortunately it gives no elevation hints. Also there is an ambigous zone called "Cone of Confusion" where a pair of sources cannot be distinguished because carachterized by the same value of ITD [18]. This, together with the lack of elevation hints, are the main limitations of Duplex theory.

In general the cues for elevation are not as robust as those for the lateral angle. It has been found that the elevation perception is mainly due to changes in the signal's spectrum. The strongest spectrum changes occur when the wave length becomes comparable with the size of the pinna, this happens around 3 kHz. At this frequency the wave is scattered by the pinna and its spectrum is modified accordingly. The reflection and refraction of sound by the torso also provides elevation cues, even if weaker. They appear at lower frequencies, therefore they can be important for sources that have little high-frequency content LowFreqElev.

The three primary cues for range are the low-frequency ILD for close sources [19] and the ratio of direct to reverberant energy for distant sources [Auditory distance perception in rooms]. In particular, reverberant energy decorrelates the signals reaching the two ears [The decorrelation of audio signals and its impact on spatial imagery], and the differences between the timbre of direct and reverberant energy provides another localization cue, one that might be important for front/back discrimination as well. All of these cues contribute to externalization - the sense that the origin of the sound is outside of the head. Externalization is an important perceptual feature that is strongly degraded by a bad modeling of the propagation of sound from the source to the ear canal.

All of the cues described so far are static. However, dynamic cues that come form head motion plays a big role in the localization process. Considering that it's easier to recognize a known sound than a one never heard before, the familiarity with the source can improve the localization accuracy.

The acoustic cues for sound localization are a consequence of the physical processes of sound generation, propagation, diffraction, and scattering by objects in the environment, including the listener's own body. In principle, these processes can be analyzed by solving the wave equation subject to the appropriate boundary conditions. Fortunately, at typical sound pressure levels, the physical processes are essentially linear and time invariant (LTI). If we model the path of sound propagation from sound source to one ear as LTI system, this system is completely determined by its transfer function. We call this transfer function Head Related Transfer Function (HRTF) [20]. The HRTF is defined as the ratio of the Fourier transform of the sound pressure developed at the ear to the Fourier transform of the sound pressure developed at the location of the center of the listener's head with the listener absent. Formally:

$$\mathbf{HRTF}_{Left}(f,\theta,\phi,r) = \frac{\mathbf{X}_{Left}(f,\theta,\phi,r)}{\mathbf{X}_{FreeField}(f,\theta,\phi,r)}$$

$$\mathbf{HRTF}_{Right}(f,\theta,\phi,r) = \frac{\mathbf{X}_{Right}(f,\theta,\phi,r)}{\mathbf{X}_{FreeField}(f,\theta,\phi,r)}$$

Typically,the HRTF is analyzed in the far field, therefore the dependance on the distance r (from the axis origin) is often neglected. In this case the HRTF can be formalized as:

$$\mathbf{HRTF}_{Left}(f,\theta,\phi) = \frac{\mathbf{X}_{Left}(f,\theta,\phi)}{\mathbf{X}_{FreeField}(f,\theta,\phi)} \tag{2.1}$$

$$\mathbf{HRTF}_{Right}(f,\theta,\phi) = \frac{\mathbf{X}_{Right}(f,\theta,\phi)}{\mathbf{X}_{FreeField}(f,\theta,\phi)}$$

The inverse Fourier transform of the HRTF is the head-related impulse response (HRIR). Being directly dependent on physical processes, the HRTF varies with the different individual anatomy. Since the morphology greatly varies from one individual to another the use of non-individual HRTF in the spatial synthesis does not lead to convincing result. In [14] P.Guillon shows that non-individual binaural synthesis is imperfect and that the observed spatialization defects are not related to the technology itself, but to the use of inappropriate HRTF filters. The process of computing individualized HRTF can be carried out in three ways mainly:

- Acoustic Measurements
- Structural Models
- Numerical Prediction

2.1.1 Acoustic Measurements

Acoustic measurements entail no mathematical idealizations, are accurate over much of the audible frequency range, and can produce both HRTFs and head-related impulse responses HRIRs equally easily. However, room reflections make it difficult to measure the response at very low frequencies, physical constraints can make it difficult to position a loudspeaker at very low elevations, and measurement errors make it difficult to get a high degree of repeatability. The first step is to discretize an imaginary sphere of constant radius r, centered about the listener head's center. This step is needed to obtain a finite set of azimut and elevation values that represent the positions at which a sound source may be. For high-resolution spatial sampling schemes, a step of 5°, both for the azimut and the elevation, is typically used.

The typical measurement set up is comprised by at least one microphone, one loudspeaker and a device able to move one of the two. The need of changing position of either the microphone or the loudspeaker comes from the idea of discretizing the 3D sphere. There are two dual approaches to acoustically measure the HRTF: the direct and the reciprocal one. In the former an in-ear microphone is placed at the entrance of the blocked ear canal and the sphere is sampled moving the loudspeaker. In the latter the positions of microphone and louspeaker are switched.

In the direct method the sound source (loudspeaker) is moved in each sampled (θ, ϕ) position and a signal is played. The in-ear microphones (one for each ear) record the signal and the loudspeaker is moved to the next position. Even using fast positioning equipments, a system that uses only one loudspeaker (sparse array from now on) will always be slow. In this case the loudspeaker has to be moved to each measurement position to correctly discretize the sphere surface. If high spatial resolution is required, this is not the best choice because no parallelization in the measurement procedure is possible. Moreover, subjects have to wear a head tracking device to verify that their heads are always in the correct position, so to avoid misalignment in the measurements.

In principle, one may choose to use as many loudspeakers as sampling positions. However, this requires a large number of loudspeakers, amplifiers and digital-to-analog converters and this dramatically increases its cost.

Another approach is to position a group of loudspeakers on an arc of circumference, and to turn either the arc or the subject, so to cover the required positions. This solution represents a good compromise between speed and hardware complexity/cost and are therefore the most widely used. The well known CIPIC HRTF database [21] was measured with a hybrid system, moving an arc around the subject. Also [22] [23] [24] measured HRTF in the same way.

The main limitation of the direct approach is the duration of the measurements: results from long sessions are more likely to be misaligned due to the listers involuntary movements. To overcome this problem the measurements are speeded-up using ad-hoc driving signals (interleaved multi sine sweeps) [24] together with range extrapolation methods [25][26]. In particular range extrapolation methods predict the HRTF in the near field starting from its measurements in the far field. This allow to mount loudspeakers on an arc with fixed radius further cutting the costs of the system.

To reduce measurement time the alternative approach exploites the reciprocity principle, where source and receivers positions are exchanged. Reciprocal HRTF measurement was proposed by Zotkin et. al. [27], where they use a miniature sound source and microphones distributed in a spherical structure. This method is very fast, as the excitation signal has to be played only once for each ear. On the other hand the use of a miniature source delivers a considerably small signal-to-noise ratio (SNR) and restricts the measurement frequency range to frequencies above 1 kHz. If a high spatial resolution is desired, then many microphones are needed, increasing hardware costs. In general direct measurements is the preferred solution. The result are accurate over a reasonable frequency range and the price of the experimental set up is not too high. Even though interesting speed-up and simplification methods have been developed the acoustic measurement process is still long and complex.

2.1.2 Structural and Mathematical Models

For practical applications as well as theoretical understanding, it is often useful to be able to replace an experimentally measured HRTF by a mathematical model. By including only a small number of terms or a small number of coefficients, these models can often be simplified or smoothed to provide HRTF approximations. Many models have been proposed, including principal components models [28], spherical-harmonic models [25], neural network models [29], pole-zero models [30], and structural models [31].

A possible approach is to approximate the HRTF as a combination of basis functions of some kind. [28] Used PCA on a set of acoustically measured HRTFs to extract five basic spectral shapes (basis functions). A linear combination of them, accounts for approximately 90 percent of the variance in the original HRTF magnitude functions. [26] Represented the HRTFs as a weighted sum of surface spherical harmonics (SSHs) up to degree 17. The relatively small number of functions needed to reconstruct an HRTF make this approach interesting even though it gives little insight on the relation between ear's anatomy and HRTF. [6] employ an ANN to predict the HRTFs for a new subject based on his anthropometric parameters. The proposed approach uses nonlinear dimensionality reduction technique, Isomap, to improve the personalization quality. [32] predicts the main notches in the HRTFs using a linear prediction residual cepstrum and relates them to the main edges of the pinna.

Another recently developed approach [5] is to treat the HRTF synthesis problem as finding a sparse representation of the subject's anthropometric features. The idea is to treat the subject's antrophometric features as a linear superposition of the ones of a small subset of subjects from a training dataset. It is assumed that the HRTF data is in the same relation as these anthropometric features. The application of the same sparse vector directly on the HRTF tensor data leads to the synthesis of

the subject's HRTFs.
Structural HRTF modeling [31] represents another important approach to HRTF synthesis. By isolating the effects of different components (head, pinnae, ear canals, shoulders/torso), and modeling each one of them with a corresponding filtering element, the global HRTF is approximated through a proper combination of all the considered effects. Moreover, by relating the temporal/spectral features (or equivalently, the filter parameters) of each component to corresponding anthropometric quantities, one can in principle obtain a HRTF representation that is both computationally economical and customizable. In [33] the relation between pinna's contours and HRTF featrues is investigated. The basic assumption is that each HRTF notch track is associated with a distinct reflection surface on the subject's pinna. The developed structural model is based on ad-hoc filters tuned on the subject's anatomy. This work represents an interesting result even though the anatomical information comes from 2D images. A 3D analysis of the subject's anatomy could be beneficial. Generally speaking these approaches are computatinally light, therefore potentially suitable for real-time (or quasi real-time) applications. The main drawback is that they approximate the structure of the HRTF signal instead of the physical processes that generate it. As it will be explained in the next paragraph, latest advances in numerical theory enable to accurately simulate wave propagation and scatteirng processes. Even though structural and series expansion methods are an intersting approach they does not exploit this potenital. Consider also that the achievable degree of personalization is limited to the information carried along with 2D images. The described methods are not meant to expolit the rich anthropometric information that comes with 3D models. The chance of computing the HRTF starting from 3D shapes would provide new insght on the relation between the ear geometry and the spectral feature of HRTF. In the next paragraph I will decribe the main techniques to numerically predict the HRTF.

2.1.3 Numerical Prediction

Since the HRTF results from a scattering process, many computational acousticians have attempted to compute the HRTF by solving the wave equation subject to boundary conditions on the surface of the head [34] [35], [36][37]. The numerical HRTF is usually obtained by solving the Fourier transform of the wave equation (the Helmholtz equation) at wavenumbers corresponding to frequencies of interest, though direct simulation of the wave equation in the time domain has also been attempted [38]. When the solution is obtained using the direct configuration, the simulations must be repeated for each source position, similarly to the experimental setting. The solution can also be done in the reciprocal setting obtaining HRTF from all the sampled spatial position in a go. These numerical methods rely on a 3D model of the human body (Head, Torso And Ears) to get the needed information about the geometry and boundary conditions. These model are either 3D point clouds or their meshed version. Numerical simulation is attractive since it offers the possibility of extracting HRTFs without subjecting the user to measurement (beyond those needed to create the discretization of their body). Further, numerical simulation, if it were easy and accurate, offers the promise of allowing one to relate features in the HRTF with anatomical structure of the head, pinna and body. By manipulating the mesh and observing the resulting computed HRTF it might be possible to explore the sensitivity of the HRTF to particular features.

Different techniques can be used to numerically solve radiation and scattering problems, Boundary Elements Methods (BEM), Finite Elements Methods (FEM) and Finite Difference Time Domain Methods (FDTD) are the one that have been used the most throughout the literature. The boundary element method (BEM) is a numerical method for solving boundary-value or initial-value problems formulated by use of boundary integral equations (BIEs). In the BEM, only the boundaries (that is, surfaces for three-dimensional 3D problems) of a problem domain need to be discretized. In the context of HRTF computation the BEM allows to obtain a sound pressure value in any point of space solving the Kirchhoff–Helmholtz integral equation. The only needed inputs are the positions of the sound sources (loudspeakers), the sensors (microphones) and the Head, Torso and Ear 3D mesh model. The BEM can be applied in a direct or reverse (reciprocal) fashion. In the former the sound pressure is computed at the two ear positions when the head is illuminated by an acoustic source located at a given source position. The direct approach involves the solution of an acoustic scattering problem for each source direction and two receiver positions, meaning that as many runs are needed as the number of sources. The reverse approach makes use of acoustic reciprocity and computes the directional characteristics by expressing the acoustic pressure field on a discretized sphere surface. The sphere is around the head and the pressure field is due to two point source excitations located at the two ear positions. This approach allows to get the HRTFs form all the desired positions in only two runs (one per ear).

BEM methods have been proven to give convincing results and they represent the most used methods so far. The numerically computed HRTFs in the SYMARE database have been extracted using the BEM method. While the frequency domain Boundary Element Method (BEM) has so far been the dominant approach in this regard, a small number of studies have also adapted the Finite Difference Time Domain (FDTD) method to the task. FDTD simulation has the advantage of yielding a wideband frequency response in a single run, it does not explicitly require meshing of structural surfaces, and it is free of structural dependencies in algorithmic design; on the other hand, for 3D simulation it does require a rather large amount of computer memory, and the simulation landscape must be specified in terms of acoustic material properties at every cell (or voxel). In [11]is proposed a different approach based on acoustic ray tracing. It exploits the similarity between sound wave at high frequencies and light rays to geometrically solve the sound propagation/scattering problems. This approach could be interesting for real time application but its applicability is limited to source with little low frequency content.

Numerical prediction methods are promising especially because they can provide precise information about the relation between human anthropometric and HRTF. With these techniques is possible to modify the 3D shape (the ear for example) so to later recompute the HRTFs. Potentially, this operation allows to understand which parts of the ear are relevant the most to the HRTF generation process. 3D shapes can be acquired by 3D acquisition systems; exploiting different technologies one may extract models at different resolutions and fidelity. 3D models carry more anthropometric information than 2D images, the achievable degree of personalization is therefore higher. Considering these attractive features, I decided to adopt a BEM software to extract the HRTF from 3D models ad-hoc extracted. I will explain my approach in chapter **??**. Since 3D models are of primary importance in numerical predictions, in the next section I will present some important 3D acquisition systems.

2.2 3D Model Acquisition

Three-dimensional scanning has been widely used for many years for reverse engineering and part inspection [39]. The same techniques may be used to exatract 3D models in the context of binaural audio and HRTF numerical prediction. By 3D model, I refer to a numerical description of an object that can be used to render images of the object from arbitrary viewpoints and under arbitrary lighting conditions [40]. In particular the final outcome of the 3D extraction procedure will be a point cloud, i.e. a collection of point in a 3D space characterized by their cartesian coordinates. There exist many ways of acquiring a point cloud from a real shape and the choice of the technology to use strongly varies with the size, distance and the nature of the object itself. The most widely used technologies are summarized below:

RANGE SCANNER The main idea behind this family of devices is to control the illumination of the scene to acquire projecting special light patterns on it. This approach is also called active or structured light sensing. The fundamental component of these systems are a projector and, at least, a sensor: the former is used to illuminate the object's surface, the latter captures the scene. The goal of range scanner is to infer the distance of the illuminated surface's points by using the sensor; this information allows to extract the 3D cloud point. A first classification could be made on the basis on how this distance is computed, range scanners could be further grouped according to the projecting light's nature (laser, infrared, diffuse visible light) and to the number of sensors (single or multiple).
Laser scanners are probably the most widely used family of 3D scanners; these devices use either a laser line or single laser point to scan across an object. A sensor picks up the laser light that is reflected off the object and the system calculates the distance from the object to the scanner. How the distance is measured changes from one device to another. When the object to acquire is at a short distance (i.e. less than 2 m) the distance is computed by mean of trigonometric triangulation. The distance between the laser source and the sensor is known with high precision, as well as the angle between the laser and the sensor. As the laser light reflects off the scanned object, the system can discern what angle it is returning to the sensor at, and therefore the distance from the laser source to the object's surface. This technique is called triangulation because the laser dot, the camera and the laser emitter form a triangle. These three pieces of information fully determine the shape and size of the triangle and give the location of the laser dot corner 2.2. The triangulation method can work projecting a single laser point as well as a laser line or stripe; in all of the cases each image gives us a range profile (i.e. image of the

energy of the reflected light), and by sweeping the light over the surface of the object, it is possible to capture its shape.

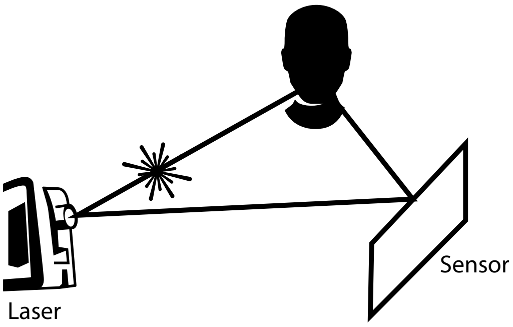


Figure 2.2. Laser Triangulation: the Laser, the sensor and the scanned object form a triangle.

Time of flight Laser scanners are based on a very simple concept: the speed of light is known exactly, so if it's known how long a laser takes to reach an object and reflect back to a sensor, it's known how far away that object is. This family of laser scanners works well for object that are far away (> 2 m) and, even though they are widely used for manufacturing and architectural measurements, they are not suitable for ear 3D extraction.
Another approach is to project a light pattern instead of a single point or line. This method triangulates the positions to get the distance from the object and it can work both with diffused and infrared light. Since there is no explicit need to use Lasers, this procedure can be cheaper than the ones previously described. Even tough laser based technology is more precise, projecting ad-hoc patterns (sinusoidal stripes, Gray Codes) can greatly improve the result's fidelity. The non-laser approach has been used for face recognition and biometrics purposes, it could represent an interesting approach for ear 3D model extraction too.
More than one sensor can be used to capture the scene. 3D scanners can take advantage of this capturing the scene form two (or more) different viewpoints and computing the distance from the object by means of geometrical considerations. The needed geometrical information can be derived from multi view theroy. The main idea behind multi view 3D scanners is to match points in the different images of the same scene and to infer depth (distance of the object) triangulating them with the unknown correspondent 3D point. The triangulation is usually possible after a calibration step in which the sensors system is "tested" to extract some geometric parameters. The multi view and projective geometry theory will be discussed in details in Chapter 3. [41] used a system of two cameras to capture ten couples of images of the ear, even though the pixel matching procedure has been carried out interactively, the obtained results make this approach palatable for further research.

TOMOGRAPHY 3D models can be extracted also from tomographic imaging, for instance Ultrasounds (US), Computer Tomography (CT) or Magnetic Resonance (MR). Tomography is a method of producing a three-dimensional image of the internal structures of a solid object (as the human body or the earth) by the

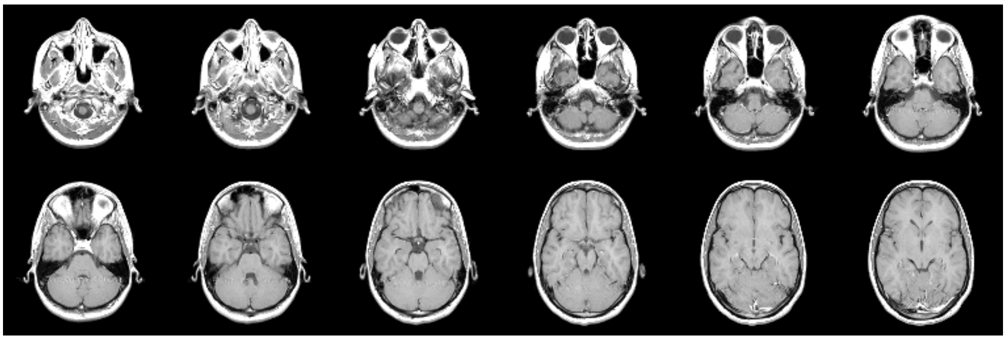


Figure 2.3. MRI Slices: examples of MRI slices. Starting from these images it is possible to extract high quality 3D surface models by means of Isosurface algorithms.

observation and recording of the differences in the effects on the passage of waves of energy impinging on those structures. The resulting 3D image is basically a sliced version of the original 3D object 2.3. This technique produce a volumetric structure which is not exactly what is needed in the context of ear 3D model extraction. What is really useful is the surface of the object. Fortunately using Isosurface based algorithms [42] it is possible to recover the surface of the object starting from the tomographic volumetric image. This method provides extremely high precision 3D point cloud, especially when MR is used to image the object. On the other hand the high cost of the imaging machines and the need of a medical operator could make this approach hard to adopt on a large scale. Nonetheless [10][8] acquired high resolution mesh of head,torso and ears using MRI; they also successfully used the obtained 3D models for HRTF computation.

In this chapter I described the basic binaural localization cues together with the main HRTF computation methods. To overcome the issues related to acoustical methods, I presented the state of the art alternatives. In particular the numerical prediction approaches take advantage of 3D models to obtain the highest degree of personalization. Indeed, the structural models rely on 2D images, therefore the antrophometric information is limited. Numerical prediction could also provide a new insight on the relation between human ear shape and correspondent HRTF. One can modify the 3D ear and recompute the HRTF so to relate changes in the signal to changes in the 3D shape. 3D models play a fundamental role in the overall process, this is true for the 3D acquisition system accordingly. Among the many existing systems, the best results are provided by the tomography imaging. Though, the associated costs are the highest (in the order of 1.5 milion euros). Good results are obtainable also by laser scanning the subject, the main drawback is the complexity of the procedure. The whole procedure is influenced by the movement of the subject. Ideally, he should stay still during all the scan session. Techniques based on multi view can overcome this problem since they do not need the subject to be still. The idea of reconstructing a 3D object by capturing images of it from different viewpoints is widely used in computer vision. The associated costs are orders of magnitude smaller than the ones associated to laser scanners (hundreds versus hundreds of thousands euros). These features make the multi view approach interesting in the context of binaural audio.

Chapter 3

Theoretical Background

This chapter introduces the theory and the terminology used throughout this thesis to extract the point clouds and the HRTFs filters. As a first step, I describe how to extract a 3D model of a subject using computer vision. Starting from the basics of projective geometry, I derive the fundamental tools needed to define the pin-hole model of a camera. The ideal model is then refined so to account for radial distortion. This is a typical degrading effect of real-world lenses, therefore has to be compensated. After having modeled a single camera, I use a couple of them to obtain the 3D point cloud exploiting Multiple View Theory. I first discuss the Epipolar Geometry and how a 3D point is related to its projections on the cameras imaging planes. After that, the actual 3D cloud extraction is presented. For the purpose, I use the *Direct Linear Transformation* (DLT) algorithm. A second fundamental step is the *Registration* of the 3D point clouds. To enrich its details, the same object can be imaged from many couples of viewpoints. Unfortunately, the trajectory of the cameras is unknown, therefore the 3D clouds are misaligned. The *Iterative Closest Point* algorithm is used to register the clouds to a common coordinate frame. The last part of this chapter is dedicated to *Boundary Element Method* theory. I will use this numerical technique to compute the personalized HRTF from the obtained 3D models. Starting from the Helmholtz equation I present three useful *Boundary Conditions* (BC) and a possible solution to the Helmholtz equation. This solution is a particular form of the full-space Green's function. I then introduce the *Boundary Integral Equation* (BIE) formulation for the Helmholtz equation. This formulation allows to obtain the value of the acoustical quantities (pressure or velocity) at any point in space, solving the Helmholtz equation only at the boundaries. The effective formulation by *Burton-Miller* is presented. To conclude, I show a discretized version of the problem. Different versions of the *Fast Multipole Method* (FMM) algorithm are used to solve the discretized problem (description of FMM is out of the scope of this work). This chapter is organized as follows: section 3.1 provides a background of some fundamental projective geometry concepts such as homogeneus coordinates, projective transformations, DLT algorithm and the pin-hole camera model. Also radial distortion, Epipolar geometry and 3D point cloud extraction are presented there. Section 3.2 descirbes the registration problem and the ICP algorithm. In section 3.3 the basics of BEM numerical techniques are presented together with a discretized version of the problem.

3.1 Basics of projective geometry

This Section reviews the concepts useful to extract a 3D representation of an object imaged from two views. The basic theory come from multiple view geometry in computer vision. For a comprehensive dissertation the reader is referred to [43]. Starting from the definition of the coordinate systems and projective transformations, I will describe how to extract the 3D coordinate of the point of the captured scene. I explain here how to solve reconstruction ambiguity so to obtain a point cloud from each pair of images.

3.1.1 Homogeneous coordinates

The representation of geometric entities in projective geometry is very common in computer vision. This is due to the fact that many non linear transformations, such as projections between images, become linear in the projective space. Homogeneous coordinates constitute an effective analytical tool for dealing with entities in the projective domain. Consider the implicit form of a line in the Euclidean space $\mathbb{R}^2$

$$ax + by + c = 0 \; ;$$

this equation can be interpreted as an inner product between the line parameter vector $\mathbf{l} = [a, b, c]^T$ and the vector $\mathbf{x} = [x, y, 1]^T$ representing a point on the line, thus obtaining $\mathbf{l}^T\mathbf{x} = 0$. Notice that the result still holds considering the vectors $k\mathbf{l}$ and $h\mathbf{x}$ with $k, h \neq 0$. This means that $k\mathbf{l}$ (or $h\mathbf{x}$ as well) defines a class of equivalence, whose representatives are called homogeneous vectors. The set of homogeneous vectors forms the projective space $\mathbb{P}^2 = \mathbb{R}^3 - [0, 0, 0]^T$, for which the following equivalence relation holds:

$$\mathbf{x} \cong \mathbf{y} \Longleftrightarrow \mathbf{x} = \lambda\mathbf{y}$$

for any $\lambda \neq 0$. The mapping from a point (x, y) in $\mathbb{R}^2$ to a point in $\mathbb{P}^2$ is always possible, leading to the vector $\mathbf{x} = [x, y, 1]^T$. The converse is not true, since the vector $[x, y, z]^T$ in $\mathbb{P}^2$ corresponds to the Euclidean point $(x/z, y/z)$ only if $z \neq 0$. The homogeneous points with $z = 0$ are called points at infinity.

Adding one dimension, we can introduce the homogeneous representation also for elements of $\mathbb{R}^3$, where planes play the same role as lines in $\mathbb{R}^2$. A plane is described by the equation

$$ax + by + cz + d = 0 \; ,$$

which can be rewritten as

$$\boldsymbol{\pi}^T\mathbf{X} = 0 \; .$$

The vectors $\boldsymbol{\pi} = [a, b, c, d]^T$ and $\mathbf{X} = [x, y, z, 1]^T$ are homogeneous in the projective space $\mathbb{P}^3 = \mathbb{R}^4 - [0, 0, 0, 0]^T$.

3.1.2 Projective transformations

A 2D projective transformation, also called *projectivity* or *homography*, is an invertible transformation $h : \mathbb{P}^2 \to \mathbb{P}^2$ for which a non singular 3×3 matrix $\mathbf{H}$ exists such that

$$h(\mathbf{x}) = \mathbf{H}\mathbf{x} \tag{3.1}$$

for every $\mathbf{x} \in \mathbb{P}^2$. A 3D projective transformation is defined analogously. Note that, since the elements of $\mathbb{P}^2$ are homogeneous quantities, eq.(3.1) continues to hold even if $\mathbf{H}$ is multiplied by a non-zero constant. Thus, also the transformation matrix $\mathbf{H}$ is to be regarded as a homogeneous quantity, that is, it is defined up to a scale factor. From this, it follows that a $2D$ projectivity has 8 degrees of freedom (dof), while a 3D projectivity has 15 dof.

A very important property of projective transformations is that inner products between points and lines (planes in $\mathbb{P}^3$) are preserved[1]. This property gives the transformation rule for lines (planes) under the point transformation $\mathbf{x}' \cong \mathbf{H}\mathbf{x}$:

$$\begin{aligned} & \mathbf{l}'^T\mathbf{x}' = \mathbf{l}^T\mathbf{x} \\ \Rightarrow \quad & \mathbf{l}'^T\mathbf{H}\mathbf{x} = \mathbf{l}^T\mathbf{x} \,. \end{aligned}$$

Since the last equation must hold for every $\mathbf{x} \in \mathbb{P}^2$, it follows that

$$\mathbf{l}'^T = \mathbf{l}^T\mathbf{H}^{-1} \;;$$

the same reasoning applies for planes in $\mathbb{P}^3$.

The structure of the transformation matrix is deeply related to the number of invariants under the transformation. In its most general form (8 dof in $\mathbb{P}^2$, 15 dof in $\mathbb{P}^3$), the transformation is simply said *projective.* Very severe distortions may arise from a generic projective transformation, since parallelism is lost. Example of invariants are given by the collinearity of three points, intersections and tangency of surfaces in contact. Passing to a more structured transformation (i.e., removing degrees of freedom), the invariants are conserved. Additionally, new invariants arise. An *affine* transformation (6 dof in $\mathbb{P}^2$, 12 dof in $\mathbb{P}^3$) is given by

$$\mathbf{H}_a = \begin{bmatrix} \mathbf{A} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \,,$$

where $\mathbf{A}$ is a non singular matrix, and $\mathbf{t}$ is a vector. Example of invariants under an affine transformation are parallelism of lines and planes. If $\mathbf{A}$ is further constrained to have the structure of a rotation matrix $\mathbf{R}$ multiplied by a scaling factor s, the transformation is said to be a *similarity* (4 dof in $\mathbb{P}^2$, 7 dof in $\mathbb{P}^3$):

$$\mathbf{H}_s = \begin{bmatrix} \mathbf{sR} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \,.$$

Measures of angles and ratios of lengths are invariant under a similarity. Finally, and *Euclidean* transformation is defined as

$$\mathbf{H}_e = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix} \,,$$

[1] Although some authors derive this property from the formal definition of projective transformations, it could as well be justified by imaging that the set of lines (or planes) forms the dual space of the vector space formed by points. With regards to this, points must be though as vectors, while lines (planes) must be though as covectors. Consequently, points transform covariantly while lines (planes) transform contravariantly, preserving the inner products.

which corresponds to a rigid motion induced by a rotation (the matrix $\mathbf{R}$) and a translation (the vector $\mathbf{t}$). Euclidean transformations have 3 dof in $\mathbb{P}^2$ and 6 dof in $\mathbb{P}^3$. The measures of length, area, volume, and angles are invariant under an Euclidean transformation.

Most important for the applications, projective transformations may be computed (up to a scale factor) from a set of point correspondences (at least 4 correspondences in $\mathbb{P}^2$ or 8 in $\mathbb{P}^3$).

3.1.3 Direct Linear Transformation (DLT) Algorithm

DLT is a simple linear algorithm for determining $\mathbf{H}$ given a set of four 2D to 2D point correspondences, $\mathbf{x}_i^n \leftrightarrow \mathbf{x}_j^n$, where $n = \{1, 2, 3, 4\}$ is the n-th correspondence. The transformation is given by the equation $\mathbf{x}_j^n = \mathbf{H}\mathbf{x}_i^n$. The equation may be expressed in terms of the vector cross product as $\mathbf{x}_i^n \times \mathbf{H}\mathbf{x}_j^n = \mathbf{0}$. Writing $\mathbf{x}_j^n = (x_j^n, y_j^n, z_j^n)^T$ and denoting $\mathbf{h}^{mT}$ the m-th row of $\mathbf{H}$; after some algebra we obtain that

$$\begin{bmatrix} \mathbf{0} & -z_j^n\mathbf{x}_i^T & -y_j^n\mathbf{x}_i^T \\ -z_j^n\mathbf{x}_i^T & \mathbf{0} & -x_j^n\mathbf{x}_i^T \end{bmatrix} \begin{pmatrix} \mathbf{h}^1 \\ \mathbf{h}^2 \\ \mathbf{h}^3 \end{pmatrix} = \mathbf{0} \tag{3.2}$$

that can be written

$$\mathbf{A}^n\mathbf{h} = \mathbf{0}$$

where $\mathbf{A}^n$ is now the 2×9 matrix of (3.2). Each point correspondence gives rise to two independent equations in the entries of $\mathbf{H}$. It's possible to build the set of equations $\mathbf{A}\mathbf{h} = \mathbf{0}$, where $\mathbf{A}$ is the matrix of equation coefficients built from the matrix rows $\mathbf{A}^n$ contributed from each correspondence. If $n = 4$, i.e if exactly 4 point correspondence are given, then $\mathbf{A}$ has dimension 8×9. Also $\mathbf{A}$ has rank 8 and thus has a 1-dimensional null-space which provides a solution for $\mathbf{h}$. Such a solution $\mathbf{h}$ can only be determined up to a non-zero scale factor. However, $\mathbf{H}$ is in general only determined up to scale, so the solution $\mathbf{h}$ gives the required $\mathbf{H}$. If more than four point correspondences $\mathbf{x}_i^n \leftrightarrow \mathbf{x}_j^n$ are given, then the set of equations $\mathbf{A}\mathbf{h} = \mathbf{0}$ is over-determined. Therefore an exact solution cannot be found (apart from the trivial and not useful zero solution). Given this, instead of demanding an exact solution, one could minimize the norm $\|\mathbf{A}\mathbf{h}\|$. The constraint $\|\mathbf{h}\| = 1$ is usually added to avoid the zero solution. This is identical to the problem of finding the minimum of the quotient $\|\mathbf{A}\mathbf{h}\|/\|\mathbf{h}\|$. The solution is known and corresponds to the unit singular vector corresponding to the smallest singular value of $\mathbf{A}$. DLT procedure can be expressed in algorithmic form:

Given $n \geq 4$ 2D to 2D point correspondences $\{\mathbf{x}_i^n \leftrightarrow \mathbf{x}_j^n\}$, determine the 2D homography matrix $\mathbf{H}$ such that $\mathbf{x}_j^n = \mathbf{H}\mathbf{x}_i^n$.

- For each correspondence $\mathbf{x}_i^n \leftrightarrow \mathbf{x}_j^n$ compute the matrix $\mathbf{A}^n$ from equation (3.2)
- Assemble the $n2 \times 9$ matrices $\mathbf{A}^n$ into a single $2n \times 9$ matrix $\mathbf{A}$.
- Obtain the SVD of $\mathbf{A}$. The unit singular vector corresponding to the smallest singular value is the solution $\mathbf{h}$. Specifically, if $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ with $\mathbf{D}$ diagonal with positive diagonal entries, arranged in descending order down the diagonal, then $\mathbf{h}$ is the last column of $\mathbf{V}$.

- The matrix **H** is determined from the structure of **h** in equation 3.2.

3.1.4 Camera models

In purely geometrical terms, a camera is a device for mapping a scene in the 3D world to a 2D representation on the image plane. Neglecting deformations due to real optics, the action of a camera may be effectively described by the so-called pinhole camera model. Within this model, a point is mapped from the 3D world to the image plane under central projection. With reference to Fig. 3.1, the centre of projection is called the *camera centre*. The line from the camera centre and perpendicular to the image plane is called the *principal axis*. The *principal point* is defined as the intersection between the principal axis and the image plane. The plane through the camera centre and parallel to the image plane constitutes the *principal plane*. According to Fig. 3.2, the *focal length* f is the distance from the camera centre to the principal point. The orthogonal distance of the point X from the principal plane is called the depth of X with respect to the camera. Let **X** and **x** be the homogeneous representations of the point in the 3D world and its projection onto the image plane; **X** and **x** are therefore homogeneous vectors of $\mathbb{P}^3$ and $\mathbb{P}^2$, respectively. An algebraic representation of the pinhole model is given as follows:

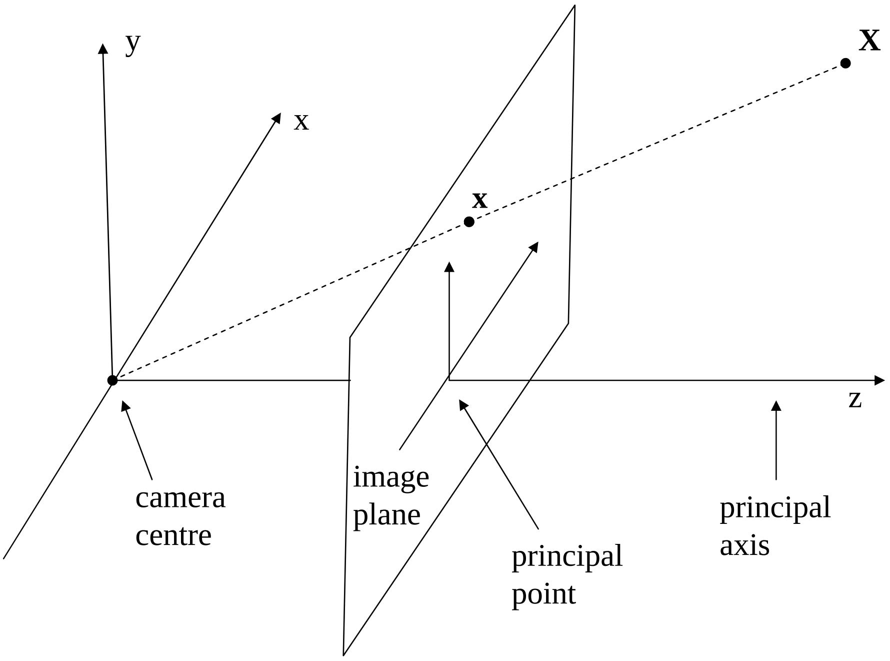


Figure 3.1. Standard pinhole model: **X** represents the point in the 3D world; **x** is the projection of **X** onto the image plane.

$$\mathbf{x} \cong \mathbf{P}\mathbf{X} \,, \tag{3.3}$$

where **P** is a 3×4 matrix which describes the mapping from $\mathbb{P}^3$ to $\mathbb{P}^2$ due to the action of the camera. Likewise the transformation matrix, it is easy to see that also

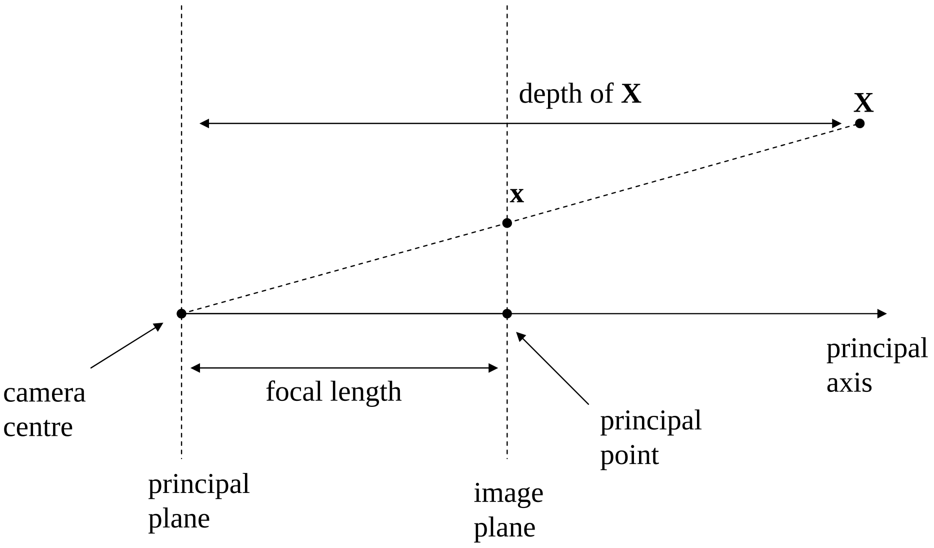


Figure 3.2. Standard pinhole model: focal length and depth of points.

the camera matrix **P** is defined up to a scale factor. As a consequence, a camera is characterized by 11 dof.

In the case of a finite camera (i.e.: the camera centre is a point at finite), the camera matrix may be decomposed as follows:

$$\mathbf{P} = \mathbf{K}[\mathbf{R}|\mathbf{t}] ,$$

where **R** is a 3D rotation matrix, **t** is a vector in $\mathbb{R}^3$, and **K** is the internal calibration matrix. The matrix $[\mathbf{R}|\mathbf{t}]$ accounts for the displacement and orientation of the camera with respect to the coordinate system of the 3D world. On the other hand, **K** accounts for the internal parameters of the camera. In its most general form, it is given by

$$\mathbf{K} = \begin{bmatrix} f_x & s & \Delta x \\ 0 & f_y & \Delta y \\ 0 & 0 & 1 \end{bmatrix} ,$$

where $(\Delta x, \Delta y)$ are the coordinates of the principal point in the image plane; (f_x, f_y) are the scale factors along the axes x and y; s is called *skew parameter* and accounts for a possible non-orthogonality between the axes x and y. In many applications, however, it may be reasonably supposed that the structure of **K** is diagonal, and furthermore that $f_x = f_y = f$ [**?**]. Therefore, the number of degrees of freedom of the camera matrix is likely to be reduced to 7 or 8.

Finally, the camera matrix enjoys many useful properties. In particular, two of these are fundamental for the comprehension of the analysis which is to follow in next Sections:

- the right null vector of the camera matrix is the homogeneous representation of the camera centre, **C**:

$$\mathbf{PC} = \mathbf{0} ;$$

- the third row of the camera matrix is the homogeneous representation of the principal plane.

Proofs are elementary, and they can be found in [43].

3.1.5 Radial Distortion

So far I assumed that the imaging process can be accurately described by means of a linear model. In general for real (non-pinhole) lenses this assumption does not hold. The most important deviation is generally a radial distortion. In practice this error becomes more significant as the focal length (and price) of the lens decreases [43]. Denote the image coordinates of a point under ideal (non-distorted) pinhole projection as $(\tilde{x}, \tilde{y})$, measured in units of focal-length. Thus, for a point $\mathbf{X}$ we have (see equation (3.3))

$$(\tilde{x}, \tilde{y}, 1)^T = [\mathbf{I}|\mathbf{O}]\mathbf{X}_{cam};$$

where $\mathbf{X}_{cam}$ is the 3D point in camera coordinates, related to world coordinates by $\mathbf{X}_{cam} = \mathbf{R}(\mathbf{X} - \mathbf{C})$, with $\mathbf{C}$ representing the coordinates of the camera center in the world coordinate frame.
The actual projected point is related to the ideal point by a radial displacement. Thus, radial (lens) distortion is modelled as

$$\begin{pmatrix} x_d \\ y_d \end{pmatrix} = L(\tilde{r}) \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}$$

where

- $(\tilde{x}, \tilde{y}, 1)$ is the ideal image position.
- (x_d, y_d) is the actual image position, after radial distortion.
- $\tilde{r}$ is the radial distance $\sqrt{(\tilde{x} - x_c)^2 + (\tilde{y} - y_c)^2}$ from the center (x_c, y_c) for radial distortion.
- $L(\tilde{r})$ is a distortion factor, which is a function of the radius $\tilde{r}$ only.

In order to use the pinhole model also with real cameras, the radial distortion has to be corrected. In pixel coordinates the correction is written:

$$\hat{x} = x_c + L(r)(x - x_c) \qquad \hat{y} = y_c + L(r)(y - y_c),$$

where (x, y) are the measured coordinates,$(\hat{x}, \hat{y})$ are the corrected coordinates, and (x_c, y_c) is the centre of radial distortion, with $r^2 = (x - x_c)^2 + (y - y_c)^2$. With this correction the coordinates $(\hat{x}, \hat{y})$ are related to the coordinates of the 3D world point by a linear projective camera. The radial distortion function $L(r)$ can be approximated with a Taylor expansion in the form $L(r) = 1 + k_1 r + k_2 r^2 + k_3 r^3 + ...$ The coefficients for radial correction $\{k_1, k_2, k_3, ..., x_c, y_c\}$ are considered part of the interior calibration of the camera, as I will better explain in .

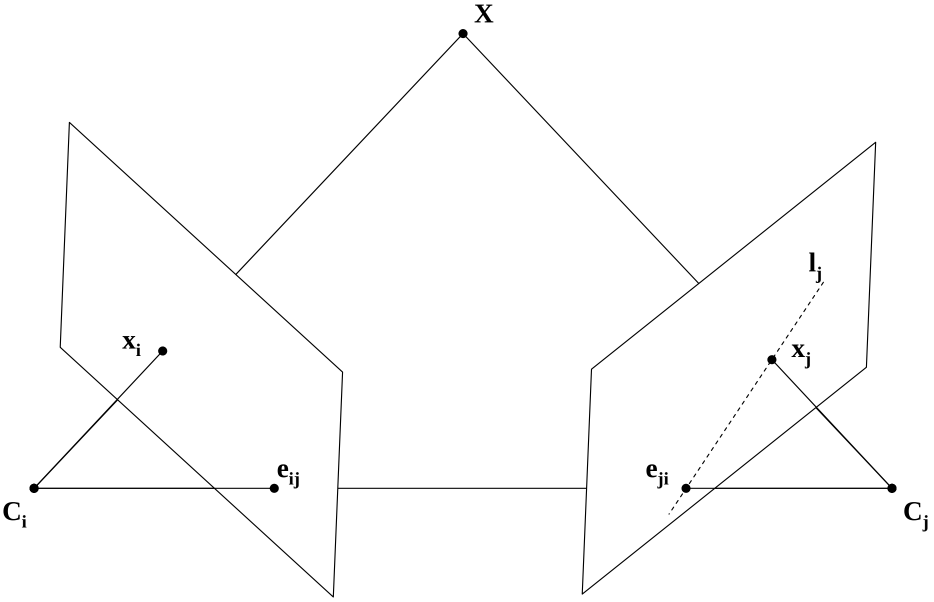


Figure 3.3. Epipolar geometry: the two camera centres are indicated by $\mathbf{C}_i$ and $\mathbf{C}_j$. A point $\mathbf{X}$ is projected onto $\mathbf{x}_i$ in the first view and $\mathbf{x}_j$ in the second view. The projections of the camera centres $\mathbf{e}_{ij}$ and $\mathbf{e}_{ji}$ are the epipoles. The epipoles and the point $\mathbf{X}$ identify the epipolar plane, whose projections on the images are called epipolar lines. For example, the epipolar line $\mathbf{l}_j$ on the second view joins $\mathbf{e}_{ji}$ with $\mathbf{x}_j$.

3.1.6 Epipolar geometry and reconstruction ambiguity

The epipolar geometry denotes the study of the constraints arising between two views of the same scene. In Fig. 3.3 the two-views geometry is depicted. Consider two cameras described by their projection matrices $\mathbf{P}_i$ and $\mathbf{P}_j$. A point $\mathbf{X}$ in the 3D world is then projected in the images points $\mathbf{x}_i = \mathbf{P}_i\mathbf{X}$ and $\mathbf{x}_j = \mathbf{P}_j\mathbf{X}$. It can be noticed that $\mathbf{X}$ and the camera centers $\mathbf{C}_i$ and $\mathbf{C}_j$ define the so-called *epipolar plane* of $\mathbf{X}$. The line joining the camera centres is called *baseline* and it lies on the plane, too. The projections $\mathbf{e}_{ij} = \mathbf{P}_i\mathbf{C}_j$ and $\mathbf{e}_{ji} = \mathbf{P}_j\mathbf{C}_i$ of the camera centres on the focal planes are called *epipoles*. The lines joining the epipoles and the projections of $\mathbf{X}$ are the *epipolar lines*.

The epipolar constraint depicted in Fig. 3.3 can be formalized in mathematical terms determining a relation between the correspondent pair of points in the two images. Consider the point $\mathbf{x}_i$ on the first image: it can be back-projected in the 3D world by means of a ray by inverting the system $\mathbf{x}_i \cong \mathbf{P}_i\mathbf{X}$. Since $\mathbf{P}_i\mathbf{C}_i = 0$, we get

$$\mathbf{X}(\lambda) \cong \mathbf{P}_i^{\dagger}\mathbf{x}_i + \lambda\mathbf{C}_i \ ,$$

where the superscript † denotes the pseudo-inverse operation and $\mathbf{C}_i$ is the centre of the first camera. In other words, each point on the ray originating from $\mathbf{C}_i$ and passing through $\mathbf{x}_i$ is projected in the same point on the first image. We can now consider the projection of $\mathbf{X}$ onto the image plane of the second camera, which is bound to lie on the epipolar line

$$\begin{aligned} l_j \quad : \quad & \mathbf{x}_j(\lambda) \cong \mathbf{P}_j \mathbf{X}(\lambda) \\ = \quad & \mathbf{P}_j \mathbf{P}_i^\dagger \mathbf{x}_i + \lambda \mathbf{P}_j \mathbf{C}_i \\ = \quad & \mathbf{P}_j \mathbf{P}_i^\dagger \mathbf{x}_i + \lambda \mathbf{e}_{ji} \; . \end{aligned}$$

The vector representation of l_j is given by $\mathbf{l}_j = [\mathbf{e}_{ji}]_\times \mathbf{P}_j \mathbf{P}_i^\dagger \mathbf{x}_i = \mathbf{F}\mathbf{x}_i$ where the matrix

$$\mathbf{F} = [\mathbf{e}_{ji}]_\times \mathbf{P}_j \mathbf{P}_i^\dagger \tag{3.4}$$

is the *fundamental matrix*, and $[\cdot]_\times$ denotes the cross product operator[2]. The fundamental matrix is a 3×3 homogeneous matrix with rank 2, and has 7 degrees of freedom. For any pair of corresponding points in the two images, it must be

$$\mathbf{x}_i^T \mathbf{F} \mathbf{x}_j = 0 \; .$$

As a consequence, it can be estimated when at least 7 point correspondences are given, without any kind of information about the scene geometry or the camera matrices. The epipoles can be computed up to a scale factor as the left and right null vectors of $\mathbf{F}$

$$\begin{aligned} \mathbf{F}\mathbf{e}_{ij} &= 0 \\ \mathbf{F}^T \mathbf{e}_{ji} &= 0 \; . \end{aligned}$$

We saw in eq.(3.4) how a pair of camera matrices $\mathbf{P}_i$ and $\mathbf{P}_j$ uniquely defines the fundamental matrix relative to the two views. Unfortunately, the converse is not true: the same fundamental matrix may be obtained by an infinite pairs of camera matrices. This fact can be easily demonstrated considering that the same projection $\mathbf{x}_i$ may be obtained as $\mathbf{x}_i \cong \mathbf{P}_i \mathbf{X}$ as well as $\mathbf{x}_i \cong \mathbf{P}_i \mathbf{H}\mathbf{H}^{-1} \mathbf{X}$, where $\mathbf{H}$ is some projective transformation of $\mathbb{P}^3$ (15 degrees of freedom). Therefore, the fundamental matrix determines the pair of camera matrices up to a projective ambiguity. It follows that the reconstruction of the 3D geometry is affected by the same kind of ambiguity; this means that the reconstructed scene will be given by an unknown projective transformation of the imaged scene: $\hat{\mathbf{X}} = \mathbf{H}^{-1}\mathbf{X}$. This ambiguity is known in the literature as the *projective reconstruction theorem.*

3.1.7 3D Point Cloud Extraction

Here I describe how to compute the position of a point in 3-space given its image in two views and the camera matrices of those views. The 3D point cloud can be obtained by repeating the procedure over as many couples of $\mathbf{x}_i$ and $\mathbf{x}_j$ as needed. It is assumed that there are errors only in the measured image coordinates, not in the

[2] The cross product operator is a short-hand notation to turn the vector $\mathbf{x} = [x_1, x_2, x_3]^T$ into the skew-symmetric matrix

$$[\mathbf{x}]_\times = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix} .$$

If $\mathbf{x}$ and $\mathbf{y}$ are points, the product $[\mathbf{x}]_\times \mathbf{y}$ yields the vector representation of the line passing through $\mathbf{x}$ and $\mathbf{y}$. Conversely, if $\mathbf{x}$ and $\mathbf{y}$ are lines, it yields their point of intersection.

projection matrices $\mathbf{P}_i$, $\mathbf{P}_j$. It is thus necessary to estimate a best solution for the point in 3-space. Since in affine and projective reconstruction there is no meaningful metric information about the object space, it is desirable to find a triangulation method that is invariant to projective transformations. Since there are errors in the measured points $\mathbf{x}_i$ and $\mathbf{x}_j$, the rays back-projected from the points are skew. This means that there will not be a point $\mathbf{X}$ which exactly satisfies $\mathbf{x}_i = \mathbf{P}_i\mathbf{X}$, $\mathbf{x}_j = \mathbf{P}_j\mathbf{X}$; and that the image points do not satisfy the epipolar constraint $\mathbf{x}_i\mathbf{F}\mathbf{x}_j = 0$. The key idea is to estimate a 3D point $\hat{\mathbf{X}}$ which exactly satisfies the supplied camera geometry, so it project as

$$\hat{\mathbf{x}_i} = \mathbf{P}_i\hat{\mathbf{X}} \quad \hat{\mathbf{x}_j} = \mathbf{P}_j\hat{\mathbf{X}} \tag{3.5}$$

and the aim is to estimate $\hat{\mathbf{X}}$ from the image measurement $\mathbf{x}_i$ and $\mathbf{x}_j$ In each image we have a measurement $\mathbf{x}_i = \mathbf{P}_i\mathbf{X}$, $\mathbf{x}_j = \mathbf{P}_j\mathbf{X}$ and these equations can be combined into a form $\mathbf{AX} = \mathbf{0}$, which is an equation linear in $\mathbf{X}$. First the homogeneous scale factor is eliminated by a cross product to give three equations for each image point, of which two are linearly independent. For example, considering the point $\mathbf{x}_i$, $\mathbf{x}_i \times (\mathbf{PX}) = \mathbf{0}$ and writing this out gives:

$$\begin{aligned} x_i(\mathbf{p}_i^{3T}\mathbf{X}) - (\mathbf{p}_i^{1T}\mathbf{X}) &= 0 \\ y_i(\mathbf{p}_i^{3T}\mathbf{X}) - (\mathbf{p}_i^{2T}\mathbf{X}) &= 0 \\ x_i(\mathbf{p}_i^{2T}\mathbf{X}) - y_i(\mathbf{p}_i^{1T}\mathbf{X}) &= 0 \end{aligned}$$

where $\mathbf{p}^{iT}$ are the rows of $\mathbf{P}$. These equations are linear in the componnets of $\mathbf{X}$. An equation of the form $\mathbf{AX} = \mathbf{0}$ can then be composed, with

$$\mathbf{A} = \begin{bmatrix} x_i\mathbf{p}_i^{3T} - \mathbf{p}_i^{1T} \\ y_i\mathbf{p}_i^{3T} - \mathbf{p}_i^{2T} \\ x_j\mathbf{p}_j^{3T} - \mathbf{p}_j^{1T} \\ y_j\mathbf{p}_J^{3T} - \mathbf{p}_j^{2T} \end{bmatrix} \tag{3.6}$$

where two equations have been included for each image, giving a total of four equations in four homogeneous unknowns.This is a redundant set of equations, since the solution is determined only up to scale. The DLT method can be used to solve this equation. It approximates the solution as the unit singular vector corresponding to the smallest singular value of $\mathbf{A}$, as shown in 3.1.3.

3.2 3D Point Clouds Registration

In section 3.1 I showed how to obtain the 3D point cloud of an object imaging it from two different views. Although, a single cloud could not be enough to represent all the details of the object. Hidden parts and self occlusions have to be compensated. Imaging the scene from many (couples of) viewpoints, one can capture all the information needed. In this way as many point clouds as (couple of) viewpoints are obtained. The problem is that the positions in the 3D space of the clouds are not coherent. This has nothing to deal with the movement of the object itself; it is due to the unknown movement of the cameras. These have to be moved in order to change the viewpoint, but the associated movement trajectory is not recorded. The idea is to find a geometric transformation that aligns the positions of two clouds. The operation of aligning two clouds is called *Registration*. I present here a classic registration algorithm known as *Iterative Closest Point* (ICP).

A 3D point cloud is a collection of points in a space of dimension 3. Therefore we define a cloud $\mathcal{P}$ of N_p points as:

$$\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{N_p}$$

With $\boldsymbol{p}_i = [x_i, y_i, z_i]$ where x_i, y_i, z_i are the Cartesian coordinates of the 3D points. When multiple point clouds represent the same object, one may be interested in merging the clouds to obtain a more complete representation of the scene. To effectively merge the clouds, they need to be aligned to a common coordinate frame. The basic idea is to keep a cloud still and to move the second to best-match it, according to some distance criteria. Let's call the fixed cloud *Model* $\mathcal{M} = \{\boldsymbol{m}_i\}_{i=1}^{N_m}$, and the moving one *Data* $\mathcal{D} = \{\boldsymbol{p}_j\}_{j=1}^{N_p}$. I will describe the most widely used registration algorithm called *Iterative Closest Point* (ICP)[44]. The transformation of the data shape to the model shape is assumed to be linear with a rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$. $\mathbf{R}$ and $\mathbf{t}$ will be applied to points in $\mathcal{M}$. The goal of the ICP algorithm is to find the transformation parameters, for which the error between the transformed data shape points and the closest points of the model shape gets minimal. The fundamental steps of ICP algorithm are:

- ***Closest Points Computation***: for every point $\boldsymbol{m}_i$ in $\mathcal{M}$ a correspondent point $\boldsymbol{p}_j$ in $\mathcal{D}$ is found. Once the correspondence is found $\boldsymbol{p}_i$ will denote the closest point in $\mathcal{D}$ to $\boldsymbol{m}_i$.
- ***Registration Computation***: the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ are computed.
- ***Registration Application***: $\mathbf{R}$ and $\mathbf{t}$ are applied to the model cloud $\mathcal{M}$

.

These steps are repeated until convergence of the error

$$E(\mathbf{R}, \mathbf{t}) = \frac{1}{N_p} \sum_{i=1}^{N_p} \left\| \boldsymbol{m}_i - \mathbf{R}\boldsymbol{p}_i + \mathbf{t} \right\|^2 \tag{3.7}$$

below a chosen threshold.
A point $\boldsymbol{p}_j$ in $\mathcal{D}$ is matched to one $\boldsymbol{m}_i$ in $\mathcal{M}$ by choosing the one that minimizes the euclidean distance defined as $d(\boldsymbol{p}_j, \boldsymbol{m}_i) = \sqrt{\boldsymbol{p}_j^2 - \boldsymbol{m}_i^2}$. Formally the distance between the point $\boldsymbol{p}_j$ and the model cloud $\mathcal{M}$ is

$$d(\boldsymbol{p}_j, \mathcal{M}) = \min_{i \in 1,2,\ldots,N_m} d(\boldsymbol{p}_j, \boldsymbol{m}_i) \tag{3.8}$$

The closest point $\boldsymbol{m}_i$ of $\mathcal{M}$ to $\boldsymbol{p}_i$ satisfies the equality $d(\boldsymbol{p}_i, \mathcal{M}) = d(\boldsymbol{p}_i, \boldsymbol{m}_i)$.
The registration computation solves the minimization problem

$$(\mathbf{R}, \mathbf{t}) = \arg\min_{\mathbf{R},\mathbf{t}} E(\mathbf{R}, \mathbf{t}) \tag{3.9}$$

with $E(\mathbf{R}, \mathbf{t})$ is the error defined in equation 3.7.
ICP is an iterative algorithm, therefore $\mathbf{R}$ and $\mathbf{t}$ have to be initialize Regarding $\mathbf{t}$

the center of mass equation (3.10) is often used. The quaternion method [A method for Registration of 3-D Shapes] or singular value decomposition (SVD) can be used to initialize $\mathbf{R}$.

$$\mathbf{t} = \frac{1}{N_{m'}} \sum_{i=1}^{N_{m'}} \mathbf{m}'_{\mathbf{i}} - \frac{1}{N_{p'}} \sum_{i=1}^{N_{p'}} \mathbf{p}'_{\mathbf{i}} \tag{3.10}$$

.

Once the registration for iteration step k has been computed it has to be applied to the data cloud $\mathcal{D}$ so that $\mathcal{D}_{k+1} = \mathbf{R}\mathcal{D}_k + \mathbf{t}$. The iteration can be stopped when the changes in the mean squared error $E(\mathbf{R}, \mathbf{t})$ fall below a chosen threshold $\tau > 0$:

$$E(\mathbf{R}, \mathbf{t})_k - E(\mathbf{R}, \mathbf{t})_{k+1} < \tau$$

.

3.3 Boundary Element Methods

To validate the 3D extraction procedure, I will compute the HRTF from the extracted 3D ear I will compare it to a groundtruth. To compute the HRTF I use numerical prediction methods that, as explained in section 2.1.3, require the 3D model as input. In particular, I use a software that implements *Boundary Element Methods* (BEM). In this section I present the basic theory behind BEM applied to acoustical simulations.

BEM is a numerical method for solving boundary-value or initial-value problems formulated by use of *boundary integral equations* (BIEs). In the BEM, only the boundaries - that is, surfaces for 3D problems or curves for 2D problems - of a problem domain need to be discretized. Solving acoustic wave problems is one of the most important applications of the BEM, which can be used to predict the HRTF resulting from the scattering processes of the sound off the ear, head and torso system.

3.3.1 Basic Equations In Acoustics

In 3D spaces, inhomogeneous *Helmholtz Equation* can be written as:

$$\nabla^2\phi + k^2\phi + Q\delta(\boldsymbol{x}, \boldsymbol{x}_Q) = 0, \quad \forall \boldsymbol{x} \in E \tag{3.11}$$

where $\phi = \phi(\boldsymbol{x}, \omega)$ is the complex acoustic pressure at point $\boldsymbol{x} \in \mathbb{R}^3$ and w circular frequency, $k = \omega/c$ is the wave number (c is the speed of sound), $Q\delta(\boldsymbol{x}, \boldsymbol{x}_Q)$ is a typical point source located at $\boldsymbol{x}_Q$ in E (figure 3.4) and ∇^2 is the Laplace operator. The acoustic domain E is considered to be isotropic and homogeneous and can be an infinite domain exterior to a body V or a finite domain interior to a closed surface. The *Boundary Conditions* (BCs) for the governing equation can be classified as follows:

- *Pressure is given*

$$\phi = \overline{\phi}, \quad \forall \boldsymbol{x} \in \mathcal{S} \tag{3.12}$$

- *Velocity is given*

$$q = \frac{\partial\phi}{\partial n} = \overline{q}, \quad \forall \boldsymbol{x} \in \mathcal{S} \tag{3.13}$$

- *Impedance is given*

$$\phi = Zv_n, \quad \forall \boldsymbol{x} \in \mathcal{S} \tag{3.14}$$

where v_n is the normal velocity, Z is the specific acoustic impedance, $\frac{\partial}{\partial n}$ is the derivative along the direction normal to the surface and the quantities with overbars indicate given values. For the boundary-value problem for acoustic waves, we need to solve governing equation 3.11, at a given frequency or wavenumber and under the BCs in equations 3.12,3.13,3.14.

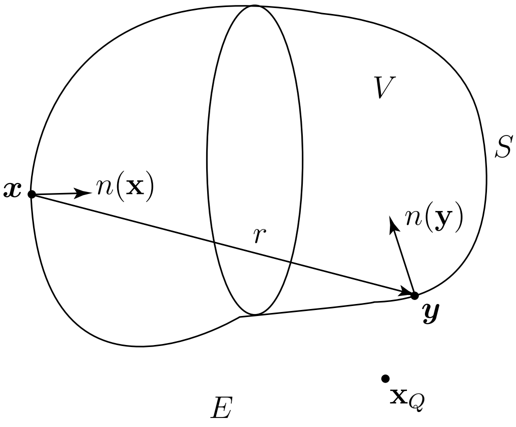


Figure 3.4. Solution Domain: the domain E of the exterior problem solution is sketched. S is the boundary surface.

There are two typical types of problems in acoustic wave analysis. One is called a radiation problem, in which a structure is in vibration and causes disturbances in the acoustic field outside or inside the structure. In this case, the velocity on the boundary S is specified in the acoustic analysis. Another type of acoustic wave problem is called a scattering problem, in which the structure stands still and an incoming disturbance (a plane incident wave or an incident wave from a point source) interacts with the structure and waves are scattered by the structure. For exterior (infinite domain) acoustic wave problems, in addition to the boundary conditions on S, the field at infinity must satisfy the *Sommerfeld* radiation condition. Basically, it says that any acoustic disturbances caused by the structure (either radiated or scattered) should die out at infinity based on energy considerations.

If we place a point source at a location $\boldsymbol{x}$ in an acoustic medium occupying the full space, then the mathematical representation for the response (acoustic disturbance pressure) at another point $\boldsymbol{y}$ is called the fundamental solution or the full-space Green's function for acoustic problems. This fundamental solution, denoted as $G(\boldsymbol{x}, \boldsymbol{y}, w)$, satisfies the following governing equation:

$$\nabla^2 G(\boldsymbol{x}, \boldsymbol{y}, w) + k^2 G(\boldsymbol{x}, \boldsymbol{y}, w) + \delta(\boldsymbol{x}, \boldsymbol{y}) = 0; \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^3 \tag{3.15}$$

in which the derivative is taken at field point $\boldsymbol{y}$ and the Dirac δ function represents the unit source at source point $\boldsymbol{x}$. $G(\boldsymbol{x}, \boldsymbol{y}, w)$ should represent an outgoing wave and have spherical (radial) symmetry. It has been found [45] that the fundamental solution for 3D acoustic wave problems $G(\boldsymbol{x}, \boldsymbol{y}, w)$ is:

$$G(\boldsymbol{x}, \boldsymbol{y}, w) = \frac{1}{4\pi r} e^{ikr} \tag{3.16}$$

where r is the distance between $\boldsymbol{x}$.

3.3.2 Boundary Integral Equation Formulations

To derive the BIE corresponding to Helmholtz equation 3.11, we apply the second Green's identity:

$$\int_E [u\nabla^2 v - v\nabla^2 u]\mathrm{d}E = \int_{S\cup S_R} [u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n}]\mathrm{d}S \tag{3.17}$$

in which E is a domain bounded by the boundary S of the structure and a large sphere S_R of radius R (with $R \to \infty$). Let $v(\boldsymbol{y}) = \phi(\boldsymbol{y})$ and $u(\boldsymbol{y}) = G(\boldsymbol{x}, \boldsymbol{y}, w)$, which satisfies 3.11. We have from 3.17:

$$\int_E [G\nabla^2 \phi - \phi\nabla^2 G]\mathrm{d}E = \int_{S\cup S_R} [G\frac{\partial \phi}{\partial n} - \phi\frac{\partial G}{\partial n}]\mathrm{d}S \tag{3.18}$$

After some math we can obtain the solution to the Helmholtz equation 3.11 for both radiation and scattering problems:

$$\begin{aligned}\phi(\boldsymbol{x}) = &\int_S [G(\boldsymbol{x}, \boldsymbol{y}, w)q(\boldsymbol{y}) - F(\boldsymbol{x}, \boldsymbol{y}, w)\phi(\boldsymbol{y})]\mathrm{d}S(\boldsymbol{y}) \\ &+ \phi^I(\boldsymbol{x}) + QG(\boldsymbol{x}, \boldsymbol{x}_Q, w), \quad \forall \boldsymbol{x} \in E\end{aligned} \tag{3.19}$$

with $F(\boldsymbol{x}, \boldsymbol{y}, w) \equiv \frac{\partial G(\boldsymbol{x},\boldsymbol{y},w)}{\partial n(\boldsymbol{y})}$ and ϕ^I is the incident wave, thus is not present for radiation problems.

Once the values of both ϕ and q are known on S, 3.19 can be applied to calculate ϕ everywhere in E, if needed. To get the values of ϕ and q on the surface S the following *Conventional Boundary Integral Equation* (CBIE) may be used:

$$\begin{aligned}c(\boldsymbol{x})\phi(\boldsymbol{x}) = &\int_S [G(\boldsymbol{x}, \boldsymbol{y}, w)q(\boldsymbol{y}) - F(\boldsymbol{x}, \boldsymbol{y}, w)\phi(\boldsymbol{y})]\mathrm{d}S(\boldsymbol{y}) \\ &+ \phi^I(\boldsymbol{x}) + QG(\boldsymbol{x}, \boldsymbol{x}_Q, w), \quad \forall \boldsymbol{x} \in E\end{aligned} \tag{3.20}$$

where the constant $c(\boldsymbol{x}) = 1/2$ if S is smooth around $\boldsymbol{x}$. It is well known that this CBIE has a major defect for exterior domain problems; that is, it has non-unique solutions at a set of fictitious eigenfrequencies associated with the resonating frequencies of the corresponding interior problems. This difficulty is referred to as the *fictitious eigenfrequency difficulty*. HRTF computation is indeed an exterior (scattering) problem, therefore another formulation has to be found.

A remedy to this problem is to use the normal derivative BIE in conjunction with this CBIE. Taking the derivative of integral representation of equation 3.19 with respect to the normal at the point $\boldsymbol{x}$ and letting $\boldsymbol{x}$ approach S, we obtain the following *Hypersingular Boundary Integral Equation* (HBIE) for acoustic wave problems:

$$\begin{aligned}\tilde{c}(\boldsymbol{x})q(\boldsymbol{x}) = &\int_S [K(\boldsymbol{x}, \boldsymbol{y}, w)q(\boldsymbol{y}) - H(\boldsymbol{x}, \boldsymbol{y}, w)\phi(\boldsymbol{y})]\mathrm{d}S(\boldsymbol{y}) \\ &+ q^I(\boldsymbol{x}) + QK(\boldsymbol{x}, \boldsymbol{x}_Q, w), \quad \forall \boldsymbol{x} \in E\end{aligned} \tag{3.21}$$

where $c\tilde{(\boldsymbol{x})} = 1/2$ if S is smooth around $\boldsymbol{x}$ and

$$K(\boldsymbol{x}, \boldsymbol{y}, w) \equiv \frac{\partial G(\boldsymbol{x}, \boldsymbol{y}, w)}{\partial n(\boldsymbol{y})} \tag{3.22}$$

$$H(\boldsymbol{x}, \boldsymbol{y}, w) \equiv \frac{\partial F(\boldsymbol{x}, \boldsymbol{y}, w)}{\partial n(\boldsymbol{y})} \tag{3.23}$$

$$\tag{3.24}$$

For exterior acoustic wave problems, one can use a linear combination of CBIE 3.20 and HBIE 3.21. The *Composite Hypersingular Boundary Integral Equation* (CHBIE) can be written as:

$$\text{CBIE} + \beta\ \text{HBIE} = 0 \tag{3.25}$$

where β is the coupling constant. This CHBIE formulation is called the Burton–Miller formulation for acoustic wave problems and was shown by Burton and Miller to yield unique solutions at all frequencies, if β is a complex number.

3.3.3 Discretization of the Boundary Integral Equations

To run software simulations a discretized version of the Burton-Miller formulation have to be used. The idea is to discretize the boundary S using finite elements. For 3D problems, as it is the HRTF prediction, surfaces of a domain will be discretized using surface elements of some kind. The most common elements are triangular or quadrilateral. In the context of personalized HRTF computation, these elements comes in the form of a 3D mesh model representing the subject's ear, head and torso. In each of the surface elements there are one or more *nodes*, these represent the discrete positions at which the acoustic quantities will be computed. BEM discrete numerical simulation works by computing the solution of the Helmoltz equation at the nodes, instead of computing them in the continuous space.

The discretized BIEs can be written as:

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{bmatrix} \begin{Bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{Bmatrix}, \quad \text{or} \quad \mathbf{A}\lambda = \mathbf{b} \tag{3.26}$$

where $\mathbf{A}$ is the system matrix; λ is the vector of unknown boundary variables at the nodes; $\mathbf{b}$ is the known vector containing contributions from the possible source term, the plane incident wave, or boundary conditions; and N is the number of nodes on the boundary. This system of equations is in complex numbers and A is, in general, a non symmetric and dense matrix. Direct solvers, such as Gauss elimination, require $\mathcal{O}(N^3)$ operations; iterative sovers speed the solution up to $\mathcal{O}(N^2)$. For these reasons, only relatively small models can be solved by use of the conventional BEM approach with either direct or iterative solvers.

A major improvement to the BEM theory is the *Fast Multipole Method* (FMM). FMM is an efficient algorithm that compute the solution of $\mathbf{A}\lambda = \mathbf{b}$ in $\mathcal{O}(N^1)$. The fundamental reason for the reduction in operations in the fast multipole BEM, is due to the fact that the Green's functions or the kernels in the BIEs can be expanded around an expansion point. This possibility greatly reduces the interactions between nodes, therefore decreasing the run time. A detailed description of FMM is out of the scope of this work, the interested reader is referred to [45].

3.3.4 HRTF Computation Using BEM

The BEM theory allows to compute the HRTFs starting from a 3D model of a subject. HRTFs computation is an exterior acoustic problem: the HRTFs are indeed generated by the scattering of the sound waves off the ear. Therefore, equation (3.25) can be used to record the acoustic pressure (generated by a source) at any point within a volume E enclosed between a sphere of radius $R \to \infty$ and the surface S of the 3D model. This formulation corresponds to the direct approach presented in 2. Following this approach one can position a number of point sources in the volume E and sense the pressure at the blocked ear canal. To obtain a meaningful set of HRTFs, the point sources are positioned on a spherical grid centered in the center of the head's 3D model. In this set-up, equation (3.25) should be computed once per each source location. To obtain all the HRTFs in one run, the reciprocal approach is exploited in this thesis. The reciprocity principle allows to turn the scattering problem into a radiation one: in this way the roles of sources and sensing points are swapped. More precisely the locations of the point sources in the direct approach are those at which the pressure will be sensed in the reciprocal. In the same way, the sensing point's location in the direct approach determines the position of the source in the reciprocal. Typically, in the reciprocal approach, the point sources are replaced with vibrating *meshes*. As explained in section 3.2, a 3D point cloud is a set of points in a 3D space. In a mesh model those points are considered to be vertexes of triangles. Practically a mesh model (simply mesh from now on) approximates the surface of an object using discrete surface elements, such as triangles. Mesh models can be used to simulate the radiation of a point source by making one specific triangle vibrate. This is achieved by correctly setting the normal velocity of the mesh which will vibrate generating the same pressure that a point source would generate. In the context of HRTFs computation this implies that, in the reciprocal approach, a mesh (or a set of meshes) close to the ear block canal will be put in vibration in order to simulate the presence of a point source. The pressure generated by the vibrating meshes will be recorded at the locations of the spherical grid. Once the pressure at the discrete spherical grid's locations is known, equation (2.1) can be used to compute the HRTFs.

Chapter 4

Methodology and System Implementation

In this chapter I describe the approach followed to extract the 3D model. The concept of an acquisition system based on multi view geometry and an implementation based on off-the shelf hardware are presented. The system is comprised of a *Capturing* and a *Projecting* device, respectively useful to acquire the scene and enrich its details. Leap Motion is used to image the 3D object from two different views: its two cameras are hardware synchronized so to shoot at the same instant. A Kinect is used to project an infrared pattern on the object surface, in this way it is textured and the illumination of the scene can be artificially controlled. Also, the calibration procedure needed to retrieve the cameras' intrinsic and extrinsic parameters is described in this chapter. Both the cameras are calibrated individually at first, a stereo calibration is then run to refine the estimated parameters. Moving the system around the subject's ear is possible to acquire many couples of images; those can be later processed and a 3D point cloud is extracted. The processing algorithm is not sensitive to the subject movements and can be run after the scan session. It works by matching couples of points in correspondent views and by triangulating those points. In this way the 3D position of each couples of points is found and the 3D point cloud is generated. To improve the quality of the 3D model, many clouds are extracted and, after an alignment procedure, fused together.

This chapter is divided as follows: section 4.1 describes the concept and features of an acquisition system based on multi view geometry. In section 4.2 the Leap Motion device and the calibration procedures are presented, the Kinect and its role in the system is defined in section 4.3. In section 4.4 the processing steps of the extraction algorithm are discussed in details.

4.1 Multi View System

To obtain the 3D model of the ear, torso and head of a subject is a critical task. Even though state-of-the-art systems can provide high quality results, the scanning process is complex and tedious. The scanners have to be handled by expert users or, in the case of tomography, by physicians and medical operators. Also, the high cost of the imaging machinery and associated software make the current solutions not suitable for a large scale consumer application. In this thesis I investigate the possibility of acquiring the 3D model using a easy-to-use and low-cost device. Even though this approach greatly simplifies the acquisition process, the resulting 3D models are less precise than the state-of-the-art. The impact of the 3D reconstruction imprecision on the numerically computed HRTF will also be studied in this thesis.

A first goal is to design a scanning procedure that is comfortable for the subject. A strong limitation of the current techniques is that the subject has to stay still for the duration of the scan session. This practically means that the subject has to focus so to avoid movements; this can be tricky considering the average duration of a scan session (about 5 minutes per ear). Also, in the case of laser scanners, the scanning device is often mounted on a supporting robotic arm, therefore the subject cannot freely choose his resting position because his ear has to be easily reachable by the arm-laser system. The solution I devise solves these problems by detaching the scanning and the processing stages. Exploiting Multi View theory, I am able to extract a 3D point cloud form a couple of acquired images. To get a richer 3D representation I need to acquire many couples of images form many point of view. The strength of this procedure is that I process the images with an algorithm that does not need the subject to remain still among shots. I will provide a detailed explanation of the algorithm in section 4.4. With the system I propose an average acquisition lasts less than one minute; the processing needed to obtain the 3D model from the acquisitions is independent from the scan phase and can be carried out in a later moment. Another fundamental goal is to keep the price low, preserving the quality of the results good enough for binaural audio applications. To achieve that, I exploit Multi View Theory of Computer Vision and I use cheap off-the-shelf hardware to build the acquisition device. The acquisition system I use consits of:

- *Capturing Device*: a system composed of two cameras for acquiring the subject's ear simultaneously from two different viewpoints. I exploit Epipolar Geometry to reconstruct the 3D object from these two views.

- *Projecting Device*: a device included able to artificially project light on the ear so to enrich its texture. This is needed because the processing algorithm works better with high textured objects.

The capturing and the projecting device are meant to be used together as part of a unique acquisition system. It is worth noting that light intensity is a crucial factor and unwanted light sources could degrade the quality of the captured images. Standard cameras are sensitive to visible light, therefore they have to be used in a light-controlled environment to avoid spurious illumination. Therefore, the capturing device has to cope with interference from environmental light. Being sensitive to infrared light only, infrared cameras can solve the problem. The capturing device takes advantage of these technology to effectively sense the scene. Accordingly, the projected pattern is infrared too.

In the next two sections I will describe the capturing and the projecting device more in detail.

4.2 Capturing Device

The capturing device plays the role of acquiring the scene from two different viewpoints. One fundamental assumption behind the Multi View theory is that the the 3D object has to be imaged (from two views) at the very same moment. Practically, this means that the two cameras have to shoot synchronously. This can be achieved using a couple of high-quality infrared cameras together with ad-hoc synchronization system. Unfortunately synchronization systems are complex and high-quality cameras are expensive, making this solution not appealing.

4.2.1 LeapMotion

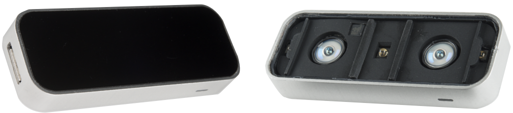


Figure 4.1. LeapMotion: the LeapMotion device. On the right the black cover plate is removed to show the three IR LEDs and the two lenses.

Leap Motion is the capturing device I use in my system. As shown in figure 4.1 it comprises two infrared cameras and three led. Being not sensitive to visible light, the Leap Motion can be used in any environment without having to take care about uncontrolled illumination. Another fundamental feature is that the two cameras are *hardware-synchronous*, meaning that they can capture the scene exactly at the same moment. The current market price of LeapMotion is 89.99 euros [*www.leapmotion.com*] making it an interesting solution also for large scale or consumer applications. This product hasn't been specifically designed for 3D extraction, therefore I had to build an ad-hoc software interface to achieve this goal.

The Leap Motion lenses have a wide filed of view of about 150° both in the x and y directions, unfortunately they suffer of high radial distortion. The Leap Motion Application Programming Interfaces (APIs) provide a couple *distortion maps* which I use to roughly compensate the distortion. I will describe the details of the rough undistortion procedure in section 4.4. This undistortion will be followed by a more precise radial distortion correction, carried out during the next calibration step. In order for the 3D extraction algorithm to work, I need to estimate the internal parameters matrix $\mathbf{P}$ of each camera, their radial distortion factor $L(r)$ and the fundamental matrix $\mathbf{F}$ f the stereo system.These quantities are defined in section 3.1.2. To get these information I have to *calibrate* the Leap Motion system. The next section is dedicated to the calibration procedure.

4.2.2 Calibration

The 3D reconstruction algorithm has to be provided with the knowledge of the matrices of the intrinsic parameters $\mathbf{P}_i$ and $\mathbf{P}_j$ of the left and right Leap Motion cameras. The fundamental matrix $\mathbf{F}$ of the stereo system is also needed. To compute them, I run a calibration procedure based on the algorithm in [46] and [47].The technique only requires the cameras to observe a planar pattern shown at a few different orientations.The position of the pattern in the 3D space is known (or can be estimated with high precision). The idea is to find a Homography $\mathbf{H}$ that relates the pattern imaged points $\mathbf{x}$ to the pattern 3D points $\mathbf{X}$. The practical procedure is described below:

- ***Calibration Object*** To correctly calibrate the system a *calibration object* is needed. This is a rectangular checkerboard of known square size. I printed a black and white, 12×7 squares checkerboard with square size $s = 25$ *mm* 4.2.

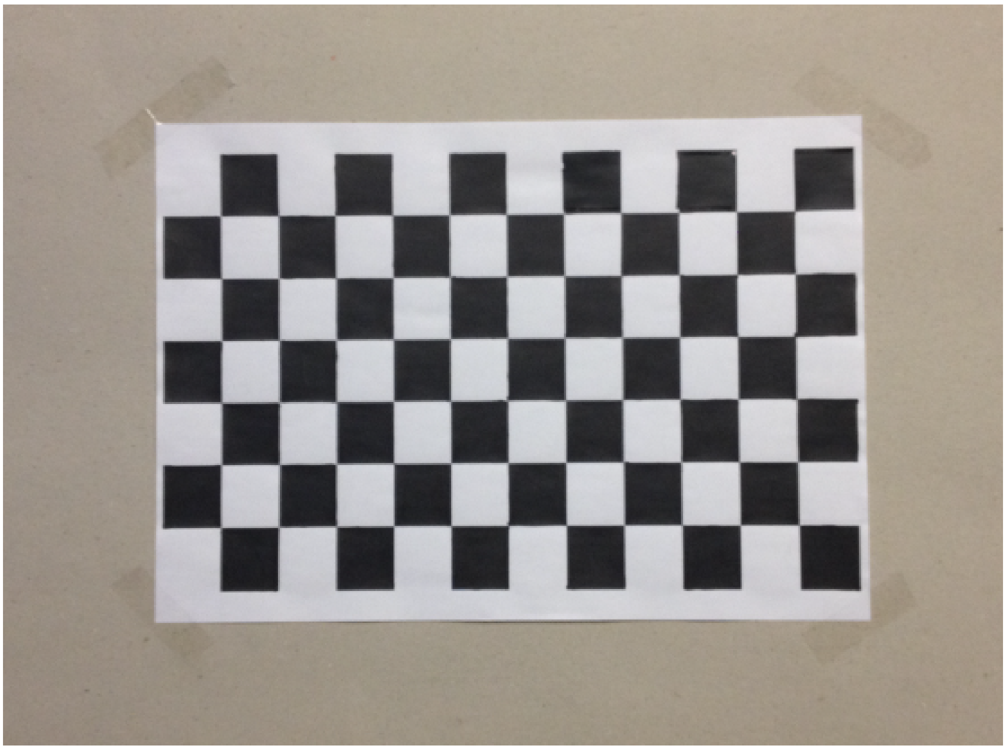


Figure 4.2. Calibration Object: The checkerboard used to calibrate the system is shown.

The knowledge of the square's real size allows to extract 3D model that not affected by scale ambiguity. To keep the pattern planar, I attached it to a hard surface which in my case was the hardcover of a book. High contrast among squares is fundamental to correctly estimate the position of the squares corners. Unfortunately, standard printing ink does not absorb infrared light, this makes the black squares look grayish when enlightened by infrared. To solve this issue I used an indelible ink to recolour the pattern. In this way the black squares actually look black both at visible and infrared light. The next calibration steps need as input images of the calibration object captured from different views. I acquired 25 images rotating and translating the checkerboard, the acquisition will be shown in chapter 5 .

- ***Single Camera Calibration*** In this step I calibrate the left and right cameras independently. The goal of this step is to compute the intrinsic parameters, in the form of matrices $\mathbf{P}_i$ and $\mathbf{P}_j$ of the cameras, and to correct the radial distortion. The algorithm proposed by [46] is based on iterative analysis of the relation between 3D pattern feature and the correspondent imaged feature. Basically, the idea is to recognize specific points in the calibration object, recognize correspondent points on the camera plane and iteratively estimate the intrinsic, exstrinsic and radial distortion parameters. The corners of the checkerboard's squares are easily recognizable using edge detection, straight line fitting and line intersection techniques [48]. Therefore I use those corners as feature points.

The calibration aims at minimizing the *Reprojection Error* defined as:

$$\sum_n d(\mathbf{x}_i^n, \hat{\mathbf{x}}_i^n)^2 \tag{4.1}$$

where $\mathbf{x}_i^n$ is the n-th measured point, $\mathbf{X}_i^n$ is a feature point of the calibration object and $\hat{\mathbf{x}}_i^n$ is the point $\mathbf{P}\mathbf{X}_i^n$, i.e. the point which is the exact image of $\mathbf{X}_i$ under $\mathbf{P}$. A good estimate of $\mathbf{P}$ can be obtained by solving the following minimization problem:

$$\min_{\mathbf{P}} \sum_n d(\mathbf{x}_i^n, \hat{\mathbf{x}}_i^n)^2 \tag{4.2}$$

Minimization is done in two steps: first *initialization*, and then *nonlinear iterative optimization*. The initialization step computes a closed-form solution for the calibration parameters not including any lens distortion. The DLT solution, or a minimal solution, may be used as a starting point for the iterative optimization. The non-linear optimization step minimizes the total reprojection error (in the least squares sense) over all the calibration parameters. The optimization is done by iterative Levenberg–Marquardt method [43]. The radial distortion coefficients $\{k_1, k_2, k_3, \ldots, x_c, y_c\}$ are also computed in this step. Since the lenses of LeapMotion device present a strong radial distortion I choose to approximate the distortion function $L(r)$ (defined in section 3.1.5) with a polynomial of degree 3.

The outputs of this step are the $\mathbf{P}$ matrix, the k_1, k_2, k_3 distortion coefficients and the (x_c, y_c) center for distortion. To compute these quantities for both the right and left cameras I run the calibration once per camera.

- ***Stereo System Calibration*** The goal of this step is to compute the Fundamental matrix $\mathbf{F}$ of the stereo system, meanwhile refining the estimates of $\mathbf{P}_i$ and $\mathbf{P}_j$. The result of the previous step are used as initial guess of a global optimization procedure. In particular the previously computed intrinsic parameters are the initial values for the stereo refined intrinsic parameters. The extrinsics are used to initialize the *pose* estimation parameters. *Pose* parameters determine the rigid motion (Rotation and translation) of the second camera with respect to the first one. The global stereo optimization procedure recompute all intrinsic and extrinsic parameters so as to minimize the reprojection errors on both camera for all calibration corners locations. Optimization is performed over a minimal set of unknown parameters. In particular, only one pose unknown (6 DOF) is considered for the location of the calibration grid for each stereo pair. This insures global rigidity of the structure going from left view to right view. Practically, the output of the stereo calibration are the $\mathbf{P}_i, \mathbf{P}_j$ and$\mathbf{F}$ matrices, meaning that the point to line mapping (explained in 3.1.6) from left to right cameras' planes is now known.

4.3 Projecting Device

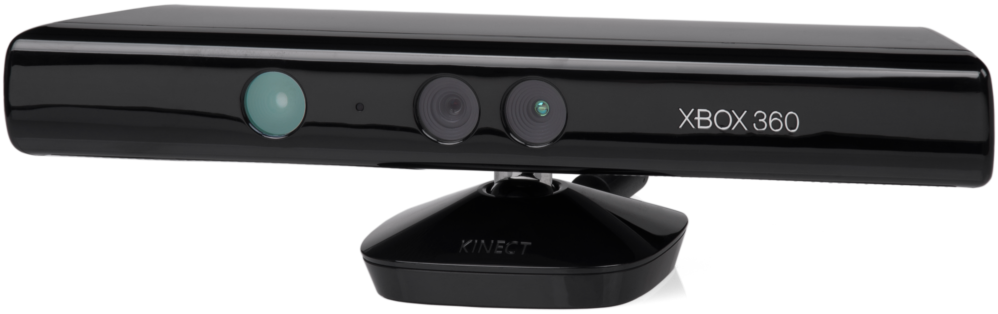


Figure 4.3. Kinect: the Kinect device. The leftmost lens covers the IR projector. The othe two lenses are an IR sensor and a standard camera but will not be used in the proposed acquisition system.

The projecting device plays the role of artificially controlling the illumination of the scene. More specifically, it aims at enriching the texture of the 3D object projecting a light pattern on it. As it will be explained in section 4.4 a fundamental operation in 3D reconstruction is to match point of the left images with point of the right ones. Flat or low detailed surfaces greatly degrades the matching results. The pattern to project has to have some specific features: for instance it has to be not repetitive so to avoid similar areas within the pattern area. It has to be detailed enough so to effectively texture the surface, it is also important that environment illumination does not influence the projection. Considering all these needs, the chosen projecting device is a Kinect by Microsoft (figure 4.3). Kinect comprises one camera sensitive to visible light, one sensitive to infrared light and an infrared projector. Although Kinect system already provides off-the-shelf technology and software for 3D extraction, I use the Kinect as a projecting device only. Indeed, 3D extraction using Kinect is not feasible in this case as the depth range of the IR camera is about $1m - 4m$ and its depth resolution is too low considering the dimensions of an average ear. For instance, capturing the ear at 1 m would result in depth resolution of 2 mm. [49] Considering that the ear depth is about 1 cm, that would mean quantize the ear depth on 5 levels, too few even for a low resolution application. The Kinect IR pattern is suitable for the purpose of enriching the details of the ear surface. It is a pseudo random speckle pattern generated by a single beam diffracted by a prisma. It also has the good feature of always remaining in focus. An example of Kinect pattern and projection on a real ear are shown in figure 4.4.

The distance between the 3D object and the Kinect greatly impacts on the performance of the algorithmic point matching. The closer they are the smaller are the projected speckles. If the Kinect is too close the projected pattern will look more like diffuse light than structured one, making the projection useless. On the other hand, if projected from too far, the speckels would be too big, adding too few detail to the object, (figure 4.4). Kinect is a consumer product, it is relatively small, easy to handle and not expensive. The model I use (Kinect 360) costs less than 150 Euros, making this solution palatable for a low-price application.

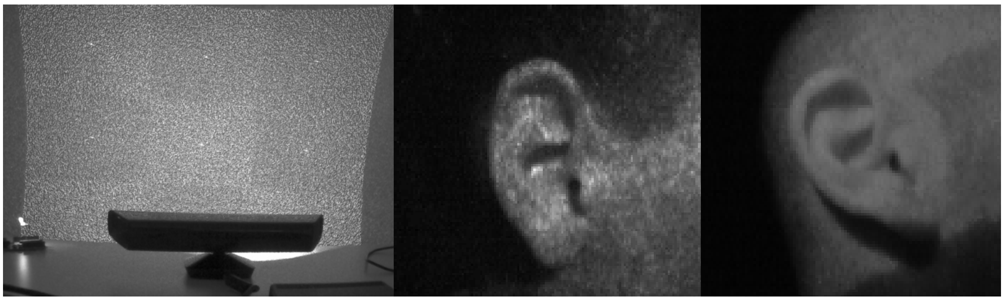


Figure 4.4. Projected IR Pattern: in the first square the Kinect IR is shown. The second square shows an example of correctly projected pattern: the ear is well textured and the speckles are clearly visible. In the third square an example of incorrect projection. The speckles are too big and create a diffused light effect, not texturing enough the ear surface.

4.4 3D Model Extraction

In this section I describe the algorithm I used to process the couples of 2D images. The goal is to extract a 3D point cloud representation of the object, more specifically of the ear of the subject. The algorithm implements the multiple view techniques explained in chapter 3. After the calibration of the acquisition device, accomplished as described in Section 4.2.2, the 3D reconstruction is obtained through the following steps:

- ***Images Acquisition*** The ear is captured from many couples of views using the LeapMotion and Kinect acquisition system.
- ***Points Matching*** i.e. finding a set of point correspondences between the left and right images.
- ***3D Triangulation*** Given the set of matching points, they are triangulated to extract the 3D point clouds.
- ***3D Point Cloud Fusion*** Looping over the previous three steps many point clouds are extracted. In this step they are registered to a common coordinate frame to obtain a single, more detailed, 3D cloud.

4.4.1 Images Acquisition

The first task is to capture the 3D object form many view. Even though conceptually this is not a difficult operation, there are some practical consideration to bare in mind.

As anticipated in section 4.2, I roughly compensate for the radial distortion using the *Distortoin Maps* provided by Leap Motion APIs. These maps come in the form of look-up tables in which each entry indicates where to find the corrected brightness value for the corresponding pixel in the raw image. More formally, let be $\boldsymbol{x} = (x, y)$ the pixel location in the undistorted image, $\boldsymbol{u} = (u, v)$ the pixel location in the distorted (raw) image and $\mathbf{L}_u, \mathbf{L}_v$ two matrices of size 64×64. Each entry

$\mathbf{L}_u(x, y)$ of the matrix $\mathbf{L}_u$ is a u pixel location, accordingly each entry $\mathbf{L}_v(x, y)$ of the matrix $\mathbf{L}_v$ is a v pixel location.

$$\mathbf{L}_u = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1N_v} \\ u_{21} & u_{22} & \cdots & u_{2N_v} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_u1} & u_{N_u2} & \cdots & u_{N_uN_v} \end{pmatrix} \qquad \mathbf{L}_v = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1N_v} \\ v_{21} & v_{22} & \cdots & v_{2N_v} \\ \vdots & \vdots & \ddots & \vdots \\ v_{N_u1} & v_{N_u2} & \cdots & v_{N_uN_v} \end{pmatrix} \tag{4.3}$$

Let's now call $\mathbf{I}_R$ the distorted image and $\mathbf{I}_U$ the distortion corrected image. To get the correct brightness value for the undistorted pixel $\boldsymbol{x}$ the following equation holds:

$$\mathbf{I}_U\ (x, y) = \mathbf{I}_R\ (\mathbf{L}_x(x, y), \mathbf{L}_y(x, y)) \tag{4.4}$$

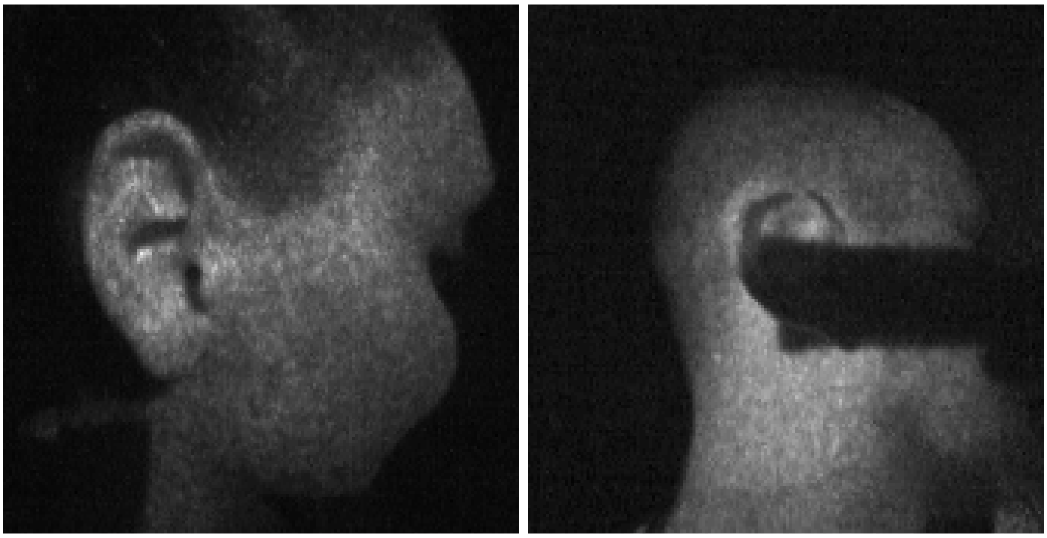
Notice that both $\mathbf{L}_u, \mathbf{L}_v$ have size 64×64, while the $\mathbf{I}_R$ and $\mathbf{I}_U$ images have size 640×240.This means that the 640 values of x and the 240 values of y have both to be mapped to 64 values. This mapping is easily obtained through:

$$x' = \frac{x}{640}64 \qquad y' = \frac{y}{240}64 \tag{4.5}$$

. Notice that the mapped values x' and y' can assume non-integer values, while the distortion maps $\mathbf{L}_u, \mathbf{L}_v$ are discrete grids. To obtain a continuous representation of the pixel locations I interpolate $\mathbf{L}_u, \mathbf{L}_v$ locations using *Bilinear interpolation*. Further distortion correction is performed on the basis of the distortion coefficients extracted during the preliminary calibration phase. After the first rough undistortion, I apply the radial distortion correction equation which can be found in 3.1.5.

Even though the distortion correction compensate well for pixels close to the center of distortion, the same is not true for pixels far from it. Practically, it means that the object to be captured should be kept close to center of distortion. In the Leap Motion images, it almost coincides to the actual center of the image. In imaging the ear I have to take care of keeping it close to the center, ideally the ear should be "centered" in the image center.

The orientation of LeapMotion with respect to the Kinect is noteworthy. In capturing the ear images I have to keep the Leap Motion out of the cone of light projected by the Kinect. Otherwise the LeapMotion would create a shadow on the ear's surface, degrading the enriching effect of the infrared pattern. In figure 4.5 example images resulting from a good and bad positioning of the acquisition system are shown. To acquire many views, I move both the Kinect and the LeapMotion using hands. This allows a great movement flexibility and enable the subject to choose a comfortable position. Practically the acquisition procedure starts with the subject sitting on a chair and choosing the most comfortable position. Obviously the ear has to be visible, so if the hair cover it, the hair has to be gathered-up or tied. Once the subject is ready the scan begins: the LeapMotion and the Kinect are moved around the ear avoiding the shadowing effect. The ear surface is complex and, for each view, there are some hidden parts that cannot be imaged. This is known as the *self occlusion* phenomenon. The movement of the acquisition system changes how the pattern illuminates the ear surface and the point of view of the cameras. This allows to capture a set of images in which the self occlusions of the ear surface are compensated. More precisely if a self occlusion is present in



(a) Correct Acquisition

(b) Incorrect Acquisition

Figure 4.5. IR Shadowing: on the left an example of correct acquisition. On the right, the black area is the shadow created by the LeapMotion being in the cone of IR light projected by the Kinect. Acquisition like this have to be discarded.

an image captured from one viewpoint, the same self occlusion will not be present in another image captured from a different view. If the set of captured images is large enough and the viewpoints are wisely chosen, the lack of information due to self occlusion can be greatly reduced. To guide the user in the acquisition a visual feedback is given. The scenes captured from the Leap Motion cameras are shown on a monitor as a continuous flows of frames (i.e a video sequence), this is fundamental to understand where to position the acquisition system. The scene is automatically captured at a frame rate of 1 second and stored in memory (i.e. in a laptop). As it will be explained in section 4.4.2 the subject does not need to remain still during the acquisition procedure. Moreover, an average scan session lasts less than a minute with this method. Before running the matching algorithm, as a preliminary stage, I manually select the best images among all the captured; this is needed in order to discard the shots which are redundant or not meaningful. In particular the images in which the ear is far from the center, or the projected pattern is not detailed enough have to be discarded. The same is true for those captured from too far or too close. To obtain good couples of images the pattern should be projected from 25-30 cm far from the ear and the Leap Motion should be kept at a distance of 10-15 cm. At these distances the IR LEDs of the Leap Motion illuminate the ear with a strong diffused light, covering completing the pattern projected by the Kinect. This has to be avoided since the details added by the projector are fundamental for the extraction. Unfortunately Leap Motion APIs do not provide a software procedure to turn the LEDs off; as a workaround I disassembled the Leap Motion and covered them with black insulating tape 4.6. In this way the LEDs diffused light is blocked and the cameras can sense the IR pattern.

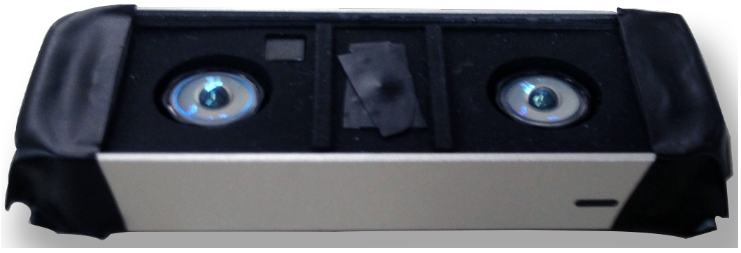


Figure 4.6. Leap Motion: the LeapMotion is disassembled and the LEDs are covered with black insulating tape.

4.4.2 Points Matching

The next step is to find points in the left and right images that represent the same 3D point. I call such couples of points *matching couples*. The idea is to exploit epipolar geometry as explained in section 3.1.6. Let $\mathbf{x}_i$ be a point on the left image and $\mathbf{l}_j = \mathbf{F}\mathbf{x}_i$ the correspondent epipolar line on the right image. The problem can be formulated as: *" Find the point* $\mathbf{x}_j$ *on the epipolar line* $\mathbf{l}_j$ *that best matches the point* $\mathbf{x}_i$ *".*

To mathematically formalize this problem the concept of *best match* has to be defined. Since $\mathbf{x}_i$ and $\mathbf{x}_j$ are projections of the same 3D point, areas of the left and right images close to $\mathbf{x}_i$ and $\mathbf{x}_j$ should look similar. In particular, the area centered in $\mathbf{x}_j$ should be closer to the one centered in $\mathbf{x}_i$ than all the areas centered in any other point lying on the epipolar line $\mathbf{l}_j$. This intuitive reasoning leads to express the *best match* as a comparison between areas of images. This problem is well known and it can be solved defining some *Similarity metric*. Let be $\boldsymbol{x}' = (x', y')$ a generic pixel lying on $\mathbf{l}_j$ and $\mathbf{N}_j(\boldsymbol{x}', s)$ the square area of image centered in $\boldsymbol{x}'$ having (odd) side s. In the same way, $\mathbf{N}_i(\boldsymbol{x}_i, s)$ is the area of left image centered in $\boldsymbol{x}_i$. Among the many existing I implement the *Normalized Inverse Sum of Squared Distances* (NISSD), which is a scalar defined as:

$$\text{NISSD}(\mathbf{N}_i, \mathbf{N}_j) \equiv \frac{1}{\frac{1}{s^2}\sum_{r=1}^{s}\sum_{c=1}^{s}(\mathbf{N}_i^{r,c}(\boldsymbol{x}_i, s) - \mathbf{N}_j^{r,c}(\boldsymbol{x}_j, s))^2} \tag{4.6}$$

where r, c are the row and column index respectively and $\mathbf{N}_i^{r,c}(\boldsymbol{x}_i, s)$ is the element of matrix $\mathbf{N}_i(\boldsymbol{x}_i, s)$ located at row r and column c.

As it can be seen NISSD is maximum (infinite) when $\mathbf{N}_i$ and $\mathbf{N}_j$ are equal. Now it is possible to mathematically express the *best match* problem as a maximization one in the following way:

$$\boldsymbol{x}_j = \arg\max_{\boldsymbol{x}'} \text{NISSD}(\mathbf{N}_i(\boldsymbol{x}_i, s), \mathbf{N}_j(\boldsymbol{x}', s)) \tag{4.7}$$

Even though this procedure works well for discrete pixel locations, it could be further improved looking for the maximum value of NISDD at subpixel resolution. The idea is to compute the differences at each locations of $\mathbf{N}_i$ and $\mathbf{N}_j$, and to perform interpolation. More precisely, I compute the *Normalized Inverse Squared Differences* (NISD), which is a matrix defined as:

$$\mathbf{N}^{r,c} = \text{NISD}(\mathbf{N}_i, \mathbf{N}_j) \equiv \frac{1}{\frac{1}{s^2}(\mathbf{N}_i^{r,c} - \mathbf{N}_j^{r,c})^2} \quad \forall\ r, c = \{1, 2, \ldots, s\} \tag{4.8}$$

Once the NISD matrix is known, NISD values between are interpolated using bicubic technique [48]. The highest similarity between $\mathbf{N}_i$ and $\mathbf{N}_j$ can now be found with subpixel precision. The overall maximization idea remains the same, the interpolation step only increases the resolution, leading to more precise points matching. Running this refined maximization technique for all the ear points $\boldsymbol{x}_i$ in the left image I find the $\boldsymbol{x}_j$ associated matching point in the right one. In figure 4.7 an example of matched points is shown.

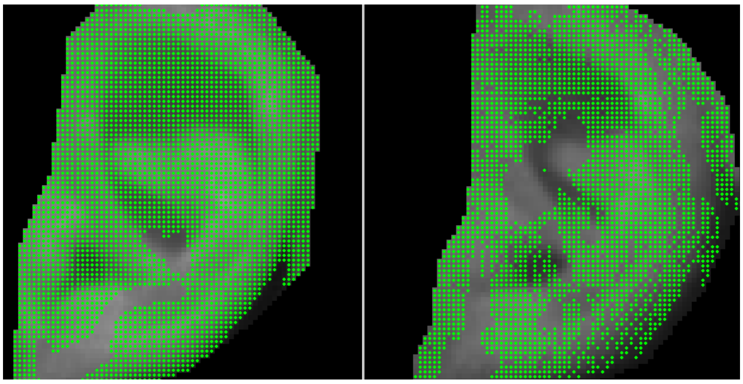


Figure 4.7. Matched Points: in the left and right images the green points have been matched by the matching algorithm. As it can be seen, not all the points on the left have been matched to a point on the right. The 3D cloud will be extracted by using the matched points only.

4.4.3 3D Triangulation

The goal of this step is to obtain the 3D coordinates of a point starting from the associated couple of matched points. To do so I implement the DLT algorithm to solve the system described in section 3.1.7. Running the triangulation for all the couples of matched points, I obtain the 3D point cloud associated to one acquisition. As discussed in Section 4.2.2, the 3D reconstruction is metric, i.e., is not affected by scaling ambiguity.

Even though the extracted cloud is already fairly clean, inaccuracies in the matching procedure or in the DLT algorithm can introduce noise. The outliers (i.e. 3D points not belonging to the 3D object but present in the extracted cloud) can be eliminated by means of median filtering. An effective way to remove the noise (outliers) is filtering the depth image associated to the 3D cloud [50]. To obtain the depth image, I set the brightness values of pixel $\mathbf{x}_i$ in the left image equal to the correspondent depth value in the 3D point cloud. I then compute the median of the brightness of a square area $\mathcal{N}_{depth}(\mathbf{x}_i, s)$ centered in $\mathbf{x}_i$ having side s. After that I set the brightness value of the pixel at location $\mathbf{x}_i$ equal to the median value. Moving $\mathcal{N}_{depth}$ over all the image I perform the same operation modifying the depth values accordingly. As a last step I substitute the new depth values (from the depth image) to the old ones in the 3D cloud. The result is a clean cloud in which the outliers are filtered out. Running triangulation and cleaning for the different couples of images, I obtain a 3D cloud for each couples of views.

4.4.4 3D Point Cloud Fusion

This step aims at fusing the 3D clouds extracted, so to obtain a complete representation of the 3D object. Even though all the clouds are metrically consiste3.2 the ICP algorithm can be used to effectively solve the registration problem. After having reconstructed a number C of clouds, I choose one of them as model $\mathcal{M}$ and I consider all the remaining $\mathcal{D}_c$ with $c = \{1, 2, \ldots, C-1\}$ as data. Running the ICP algorithm for all the $\mathcal{D}_c$, I register them to the common reference model $\mathcal{M}$. After the registration phase I can merge the clouds in a single data structure (a $V \times 3$ matrix where V is the total number of points).

4.5 BEM simulation set-up

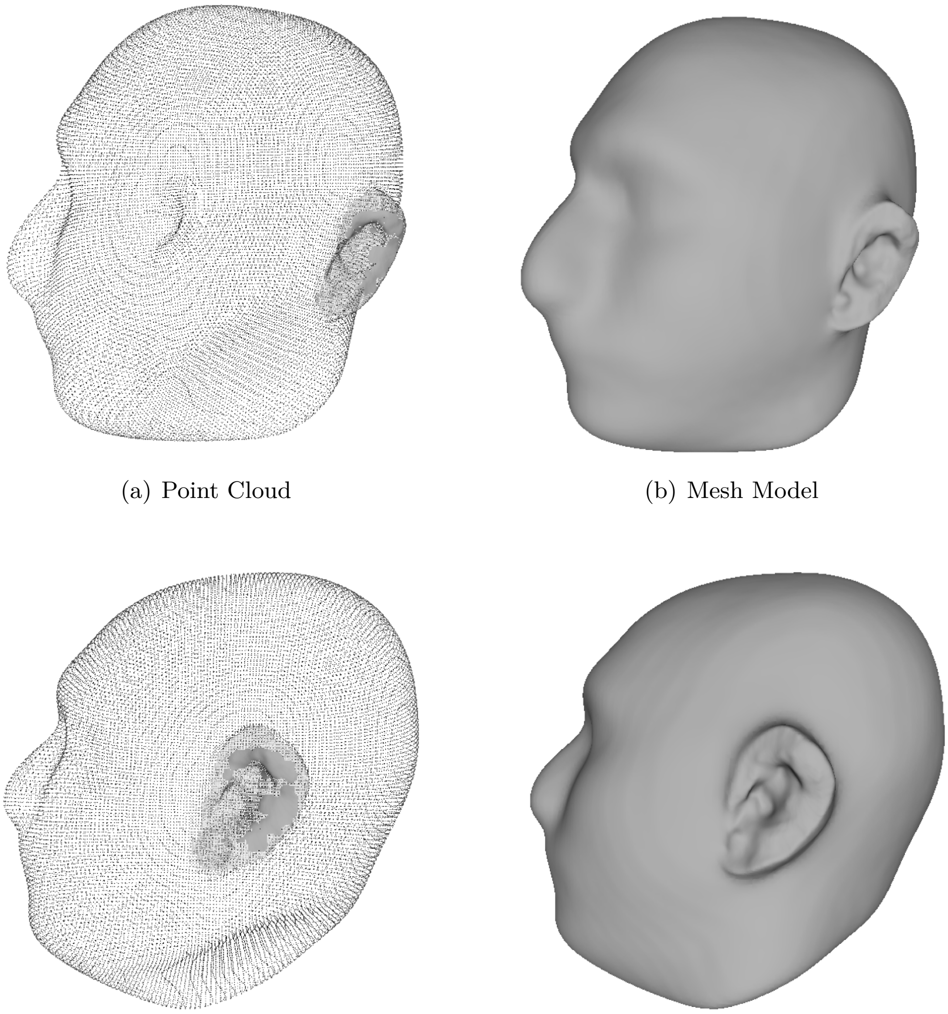
In this section I describe how to predict the HRTFs from the extracted 3D clouds. The meshing procedure is described together with the BEM implementation. In computing the HRTFs I follow the approach proposed by C.Jin in [10] who shared the SYMARE database with the Sound and Music Comptuing Laboratory of Politecnico di Milano. SYMARE database is comprised of 3D mesh models of 61 subjects' head, ear and torso (both ”glued” together and separately). Also the numerically computed HRIRs are present in the database. More detail on the SYMARE database can be found in [10]. The HRTFs are predicted using a software named Coustyx [http://ansol.us/Products/Coustyx/], by Ansol. Coustyx implements the BEM methods described in section 3.3, in particular the processing is speeded-up by using the FMM algorithm to solve the Burton-Miller formulation of the problem. To correctly run a simulation in Coustyx two preliminary steps have to be carried out:

- ***Point Cloud Meshing*** the point cloud is meshed and the ear is virtually ”glued” to a sample head model.
- ***Virtual Set-Up Generation*** the virtual positions of loudspeakers and microphones, the operative frequency band and some additional parameters are set.

Point Cloud Meshing The former task is carried out using *MeshLab meshlab.sourceforge.net*, a software for mesh editing. The goal is to create a model of head and ears by using the left ear extracted with the proposed method and a head and right ear model coming from the SYMARE database. The need of including a head which is not that of the analyzed subject is due to the lack of ground truth model. The only reference models are the two generated using the Laser Scanner, that do not include the head. The same holds for the right ear. To make the reference and the extracted models consistent, two random SYMARE subjects has been chosen. The head of those subject have been used to substitute the missing head and right ear in both the reference and the extracted models. In both cases, the point clouds of head and ears are then manually glued together to obtain a single cloud as shown in figure 4.8(a) and 4.8(c).

Coustyx does not support 3D model in the form of a point cloud. To correctly run a simulation the cloud has to be meshed. As explained in section 3.2, Practically a *mesh model* (simply mesh from now on) approximates the surface of an object using discrete surface elements, such as triangles. The problem of creating the set of triangles that best approximates the surface represented by a point cloud is known as *Surface Meshing problem*. The *Poisson Surface Reconstruction* algorithm is implemented in MeshLab and it's used to mesh the head and ear merged cloud. The meshes resulting form the application of the Poisson algorithm are shown in figures 4.8(b) and 4.8(d).

The meshes obtained in this way have a high number of vertexes; since the computational complexity of the BEM procedure is directly proportional to it, the mesh has to be *Decimated*. Decimation is an operation that reduces the number of triangles (therefore of vertexes) erasing the redundant vertexes and changing their connections [51]. After the decimation the model are "light" enough to be processed by Coustyx in a reasonable time (example in figure 4.9).



(a) Point Cloud

(b) Mesh Model

(c) Point Cloud

(d) Mesh Model

Figure 4.8. Point Cloud and Mesh Model: on the left the point cloud obtained merging the ear and head clouds. The ear cloud has higher point's density than the head. On the right the correspondent mesh model, used to run the numerical simulations.

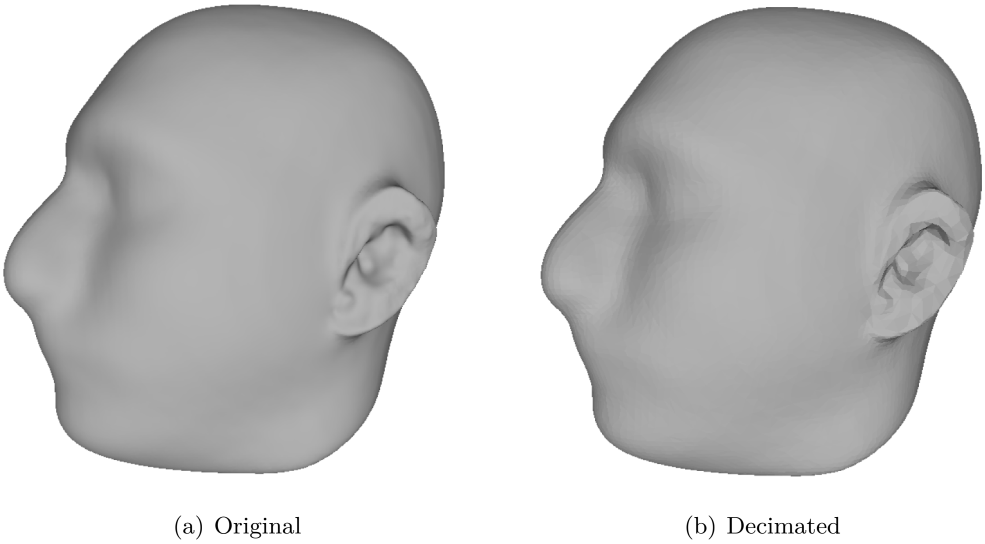

(a) Original (b) Decimated

Figure 4.9. Decimation: on the left the original mesh obtained merging the ear and head clouds. On the right its decimated version. Some fine details are lost in the decimation process but the number of points and triangles is greatly reduced. Original - Triangles=115604; Points=57804. Decimated - Triangles=23120; Points=11562.

Virtual Set-Up Generation In order to compute the HRTFs using Coustyx a number of parameters have to be input to the software. The fundamental are the ones related to the position of the sound sources and the microphones in the 3D space. Consider that Cosutyx solves the scattering problem defined in section 3.3.4, therefore some geometrical information regarding the 3D boundaries of the problem have to be provided too. The chosen virtual set-up is the one proposed in [10]. Using the acoustic reciprocity principle, it's possible to perform a simulation to determine all of the HRTFs in one go emulating the presence of a source on a surface mesh element that forms part of the blocked ear canal by setting a uniform normal velocity boundary condition on this surface element. Coustyx will consider that mesh as a vibrating element, simulating in this way the presence of a sound source thereby positioned. The HRTFs are then computed on a spherical grid of 1 m in radius and centered about the listener's head. The loudspeaker are placed at 393 positions on the sphere with an elevation greater than -45°. In figure 4.10 the spherical grid is shown. To create the needed set-up, the mesh model is imported in Matlab, the discrete sphere is created and the boundary conditions are set. From Matlab a Coustyx script is created and then run in the Cosutyx software environment. The resulting HRTFs are imported back in Matlab to be further analyzed.

In this chapter a 3D acquisition system and an extraction algorithm have been proposed. By using off-the-shelf hardware components, the cost of the system is kept low, making it suitable for large-scale consumer applications. The biggest strength of the whole procedure resides in its non-sensitivity to the subject's movements. As long as the ear is visible and close to the image center, the subject can freely move during the acquisition. This is an important feature, considering that to remain still for more than a few seconds is a difficult task. The fidelity of the extracted clouds

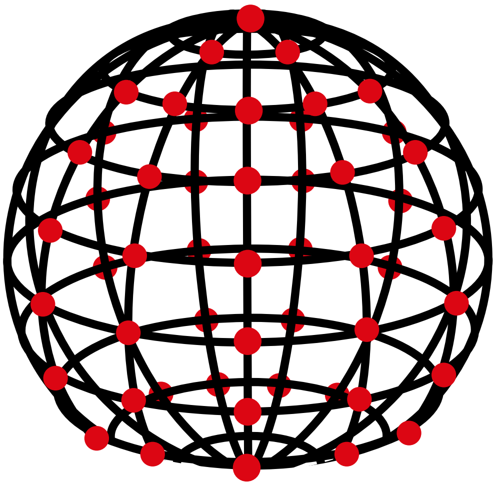


Figure 4.10. Spherical Grid: the discrete spherical grid is sketched. The red dots represent the positions of the virtual loudspeakers. The used grid comprises 393 virtual loudspeakers.

will be evaluated in chapter 5. As anticipated in chapters 2 a 3D anthropometric model can find different fields of application, in particular I investigate those related to binaural audio and anthropometric signal processing [52]. Therefore the precision of the 3D models has to be evaluated in this context. More precisely a 3D model is considered to be good when it can be effectively used to compute some acoustical properties (e.g. HRTFs).

Chapter 5

Results And Experimental Validation

In this chapter the proposed methodology is validated. The result of the processing steps are shown and critically analyzed, starting from the calibration procedure until the 3D clouds fusion. As a first step the *Mean Reprojection Error* resulting from the stereo calibration is investigated together with its impact on the 3D cloud. Practical considerations are given on how to position the Kinect and Leap Motion in order to avoid shadowing and maximize images quality. The results of the processing algorithm are also shown, together with the associated ICP alignment errors. In order to evaluate the quality of the geometrical 3D models obtained the *Hausdorff Distance* similarity metric is considered. The fidelity of the extraction method can be evaluated computing the geometrical similarity between an extracted model and a ground-truth (i.e. a model of the same subject obtained using a Laser Scanner). Another possible quality measure is the similarity between the HRTFs numerically computed starting from an extracted model and those extracted from the associated ground-truth model. Such similarity is computed using the *Spectral Distortion* metric. The acoustical relevance of the geometric similarity is further analyzed, a *Perceptually Weighted Hausdorff Distance* is defined and its relation with the Spectral Distortion is investigated.

This chapter is organized as follows: in section 5.1 the available dataset is described, section 5.2 presents the results of the stereo calibration procedure together with its 3D impact. The result of the ICP registration procedure are shown in 5.3. Section 5.4 defines the geometrical distance metric and the comparisons between extracted and ground-truth model are presented. The results related to the acoustical domain are presented in 5.5 and the *Spectral Distortion* similarity measure is defined. In the last section 5.6 the relation between geometrical and acoustical similarities is investigated.

5.1 Dataset Creation

In this section I describe how the dataset used in this thesis has been created and I give an overview of the analysis that have been carried out. To asses the quality of the point clouds extracted using the proposed method, some ground-truth models are needed. In this context a ground-truth model is an high-fidelity point cloud representing a subject ear. To obtain such high fidelity models, a Laser scanner has been used: two subjects $\mathcal{J}$ and $\mathcal{L}$ have been scanned both with the Laser and the Leap Motion. In this way high-resolution ear model $\mathcal{E}^{high}_{\mathcal{J}}$, $\mathcal{E}^{high}_{\mathcal{L}}$ and the lower resolution ear models $\mathcal{E}^{low}_{\mathcal{J}}$,$\mathcal{E}^{low}_{\mathcal{L}}$ are available for subjects $\mathcal{J}$ and $\mathcal{L}$ respectively. During the Laser Scansion procedure, the subject sits and lay the head back on a box so to keep it in rest position. Once the subject is ready the scan begins, during the session the subject does not have to move otherwise the scan has to be redone. Even though the Laser has been handled by expert user, some of the scans have been repeated due to the subject's involuntary movements. The low resolution clouds can be compared to the high resolution one to understand how similar they are; to do so the *Hausdorff Distance* (HD) will be used.

(a) $T_{\mathcal{L}}$ (b) $\mathcal{E}^{low}_{\mathcal{L}}$ (c) $\mathcal{ET}^{low}_{\mathcal{L}}$

Figure 5.1. 3D Model Composition: the steps followed to create the total low resolution 3D model of subject $\mathcal{L}$ are shown. For the sake of visualization only, in each step is shown a mesh surface instead of a point cloud.

The extracted models can be used to compute HRTFs by means of BEM simulations. As explained in 4.5 and with reference to figure 5.1 the ear is glued to a head model coming from the SYMARE database and then meshed. Since two subjects $\mathcal{J}$ and $\mathcal{L}$ have been laser scanned, two different head models $T_{\mathcal{J}}, T_{\mathcal{L}}$ have been randomly chosen. The head $T_{\mathcal{J}}$ is then used in the merging process with both the low and high resolution ear models $\mathcal{E}^{low}_{\mathcal{J}}$,$\mathcal{E}^{high}_{\mathcal{J}}$; the same is true for $T_{\mathcal{L}}$. The four head and ears model resulting from the merging are denoted as $\mathcal{ET}^{low}_{\mathcal{J}}$ $\mathcal{ET}^{high}_{\mathcal{J}}$ and $\mathcal{ET}^{low}_{\mathcal{L}}$ $\mathcal{ET}^{high}_{\mathcal{L}}$. Once the total mesh models are composed, the associated HRTFs are computed: this will result in a set of four HRTFs, two for each subject $\mathcal{J}$ and $\mathcal{L}$. Let's call $H^{low}_{\mathcal{J}}$ the HRTFs computed by using $\mathcal{ET}^{low}_{\mathcal{J}}$, accordingly $H^{high}_{\mathcal{J}}$ is the HRTFs computed by $\mathcal{ET}^{high}_{\mathcal{J}}$. With the same procedure $H^{low}_{\mathcal{L}}$ and $H^{high}_{\mathcal{L}}$ are obtained from $\mathcal{ET}^{high}_{\mathcal{L}}$ and $\mathcal{ET}^{high}_{\mathcal{L}}$. Once the high and low resolution HRTFs are available it is possible to compare them in order to understand how similar they are; to do so the *Spectral Distortion* (SD) will be used. Low values of spectral distortion imply good similarity between HRTFs.

As explained in chapter 1, psychoacoustical considerations suggest the existence of a relation between the ear morpohology and the features of the HRTFs. To investigate it, I introduce in section 5.6 the *Perceptually Weighted Hausdorff Distance* (PWHD), which computes the similarity of two 3D models weighting more its perceptually relevant parts. The idea is to compute the PWHD of two models, use these models to predict the associated HRTFs and compute the SD value. If some sort of correlation between PWHD and SD values exists, then the PWHD can be considered perceptually relevant geometrical measure. To carry out this analysis an ad-hoc dataset of head and ears 3D models has been built, the dataset is composed by

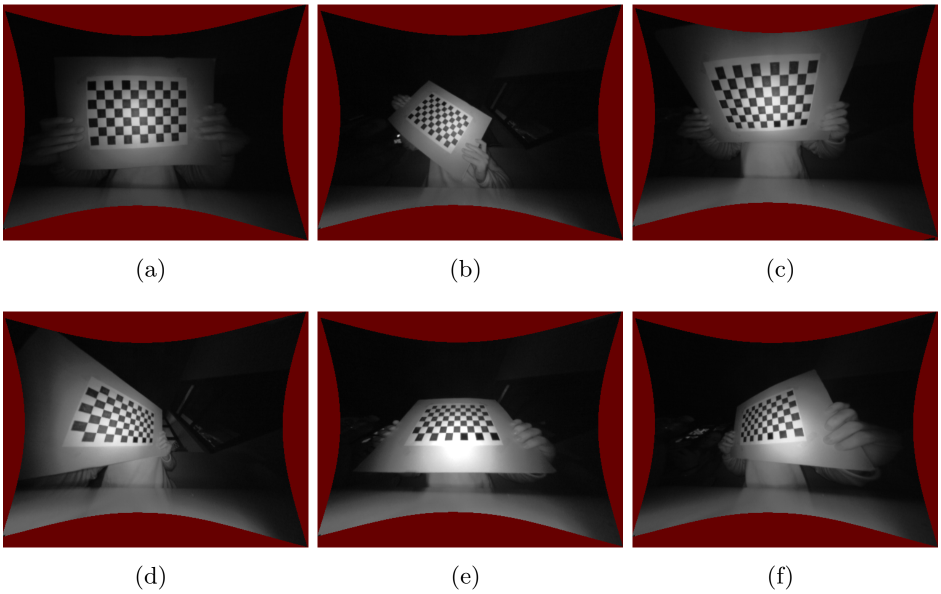
- A set $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{60}\}$ of 3D head and ears models taken from the SYMARE database
- Two head and ears models $\mathcal{ET}^{high}_{\mathcal{L}}, \mathcal{ET}^{high}_{\mathcal{L}}$. in which the ears have been laser scanned.
- Two head and ears models $\mathcal{ET}^{low}_{\mathcal{L}}, \mathcal{ET}^{low}_{\mathcal{L}}$ in which the ears have been extracted using the Leap Motion and the Kinect.

Together, all the $\mathcal{B}_i$ and the two $\mathcal{ET}^{high}_{\mathcal{L}}, \mathcal{ET}^{high}_{\mathcal{L}}$ models are denoted as the *target models*; $\mathcal{ET}^{low}_{\mathcal{L}}, \mathcal{ET}^{low}_{\mathcal{L}}$ will be called *test models*. The HRTFs of all these models are also available: in addition to $H^{high}_{\mathcal{J}}$,$H^{low}_{\mathcal{J}}$, $H^{high}_{\mathcal{L}}$ and $H^{low}_{\mathcal{L}}$, $H_{\mathcal{B}_i}$ are the HRTFs of the 60 models taken from SYMARE, The values of PWHD between both the test and each of the target models is computed together with the associated SDs. The correlation between PWHD and SD is then analyzed.

5.2 Calibration

In this section the method used to validate the approach is described. The results of the fundamental steps for 3D extraction are presented, starting from the calibration quality. To evaluate the effectiveness of the calibration I consider the *reprojection error* defined in Section 4.2.2. According to [43] a good calibration should minimize the reprojection error so to be below 1 px. As shown in figure 5.3 the reprojection error's mean resulting from the performed stereo calibration is equal to 0.31 px. The reprojection errors are averaged across the 31 shots so to account for the contributions of all the captured views. The calibration object used is the checkerboard presented in 4.2.2, and some of the taken shots are shown in figure 5.2.

The reprojection error is a distance defined on the cameras' planes which are 2-Dimensional. To understand how a 2D error impacts on the point cloud reconstruction, 3D reasoning has to be involved. With reference to figure 5.4, the idea is to consider the three triangulated points $\mathbf{x}_i$, $\mathbf{x}_j, \mathbf{X}$ and to displace the projected points $\mathbf{x}_i$, $\mathbf{x}_j$ of 0.31 px along some direction. The rays back projected from these points intersect in the point $\hat{\mathbf{X}}$; the euclidean distance between $\hat{\mathbf{X}}$ and $\mathbf{X}$ gives a measure of the reprojection error's impact on the 3D reconstruction. The calibration error is itself an euclidean distance (as defined in Section 4.2.2), therefore $\mathbf{x}_i$, $\mathbf{x}_j$ can be displaced in any position within a circle of radius 0.31 px centered in $\mathbf{x}_i$, $\mathbf{x}_j$ respectively (Figure 5.4). Let's call the displaced points $\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j$ such that $\left\|\mathbf{x}_i - \hat{\mathbf{x}}_i\right\| \leq 0.31$ px and $\left\|\mathbf{x}_j - \hat{\mathbf{x}}_j\right\| \leq 0.31$ px; the worst case scenario is when $\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j$ are displaced in opposite directions. This is verified when $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ are located at the extreme points of a diameter of a circle having radius 0.31 (Figure 5.4).
To obtain a consistent measurement of the error I match a couple of points $\mathbf{x}_i, \mathbf{x}_j$ and



(a) (b) (c)

(d) (e) (f)

Figure 5.2. Calibration Checkerboard: some of the 25 images used to calibrate the Leap Motion are shown here. The checkerboard orientation is changed among the shot, rotation and translation are allowed movements while bending (i.e. warping) is to avoid. The red areas are the result of the Leap Motion first undistortion and will not be considered by the calibration algorithm

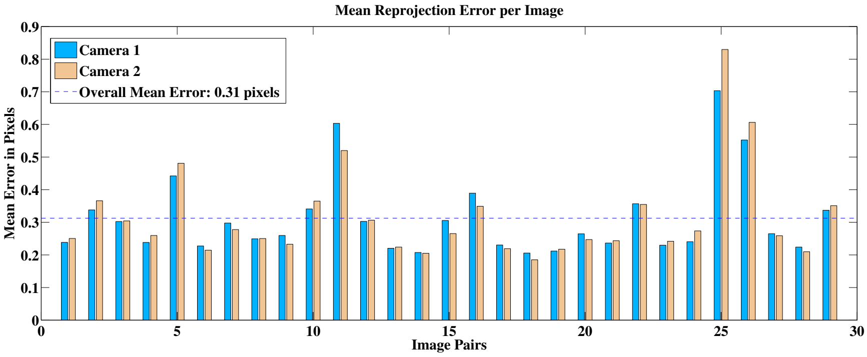


Figure 5.3. Reprojection error: each couple of bars represents the reprojection error for a couple of images. The average of all the reprojection errors is 0.31 px

extract the associated 3D point $\mathbf{X}$. The point $\mathbf{x}_i$ is then displaced of 0.31 px along the x direction so that $\hat{\mathbf{x}}_i = (x + 0.31, y)$. A set of points $\hat{\mathcal{X}}_j = \{\hat{\mathbf{x}}_j^n\}$ is obtained moving $\mathbf{x}_j$ along the circumference centered in $\mathbf{x}_j$ having radius r. Triangulating $\hat{\mathbf{x}}_i$ and each point in $\hat{\mathcal{X}}_j$ a set $\hat{\mathcal{X}}$ of 3D points is obtained. The euclidean distances between each point in $\hat{\mathcal{X}}_j$ and $\mathbf{X}$ indicate how the 3D error varies on the considered circumference. Figure 5.5 shows this distribution at three values of r. The result is that the maximum 3D error correspondent to the worst case scenario is 1.07 mm.

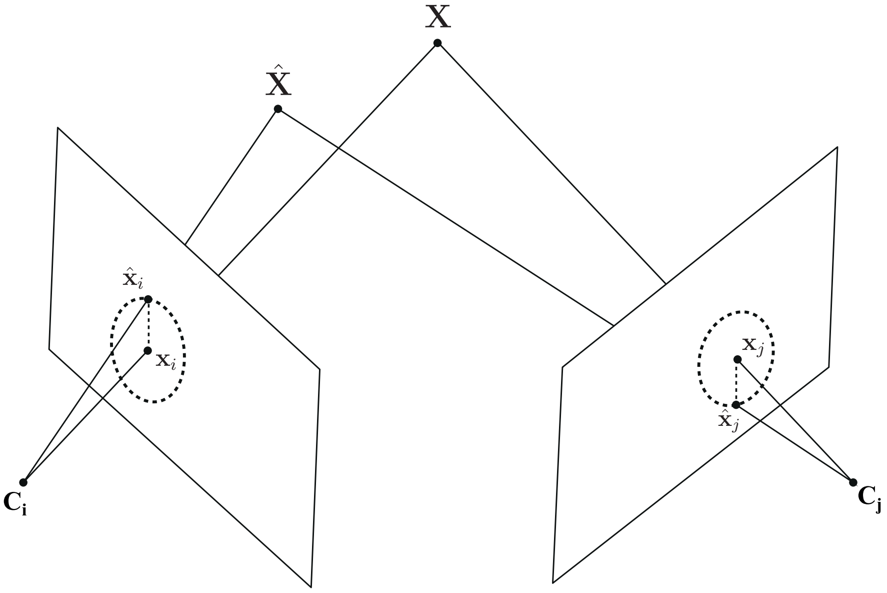


Figure 5.4. Calibration Error: the displaced points $\hat{\mathbf{x}}_j, \hat{\mathbf{x}}_i$ can fall in any location within the dashed circles of radius 0.31 px. The back projections of those points intercept in a 3D point $\hat{\mathbf{X}}$ which is different from $\mathbf{X}$. The euclidean distance between the two is the 3D counterpart of the calibration error

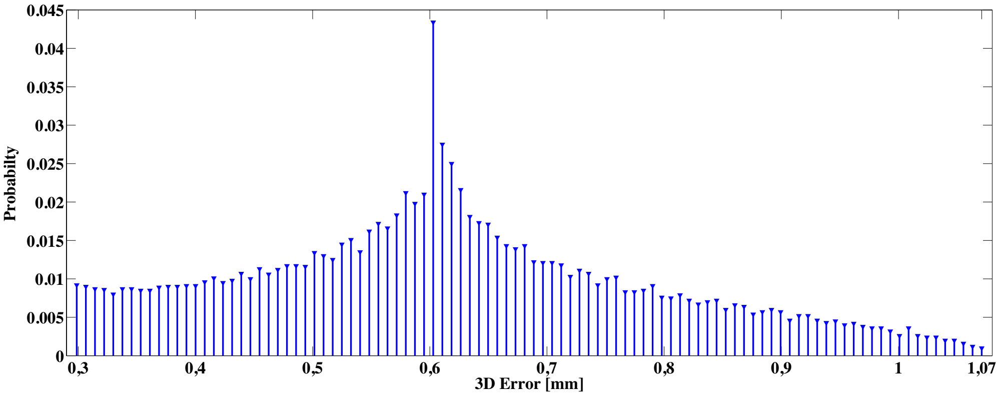


Figure 5.5. 3D Calibration Error Distribution: the probability distribution of the 3D calibration error is estimated. The point $\mathbf{x}_j$ is moved along the circumferences centered in $\mathbf{x}_j$ having radius r. r varies from 0 to 0.31 with step size of $\frac{0.31}{100}$. The maximum 3D error is equal to 1.071 [mm] and the mean value of the shown distribution is 0.6172 [mm]

5.3 Extracted 3D Models

In this section I show the 3D point clouds resulting from the extraction procedure. The processing algorithm is able to extract a point cloud from each pair of captured images. Running the algorithm for all the pairs of selected images provides many point clouds. The ICP algorithm is then used to align the clouds to a common coordinate frame. Figure 5.6 shows the importance of the registration procedure: in the top line there are the partial clouds extracted and in the bottom line there is the merged total cloud without ICP registration. As it is easy to see the total cloud does not resemble an ear even though the partial clouds are correctly extracted. Figure 5.7 shows the impact of ICP registration on the merge of the partial clouds. Being aligned to a common coordinate frame, the clouds are now coherent and the total merged cloud is representative of the scanned object (i.e the left ear).

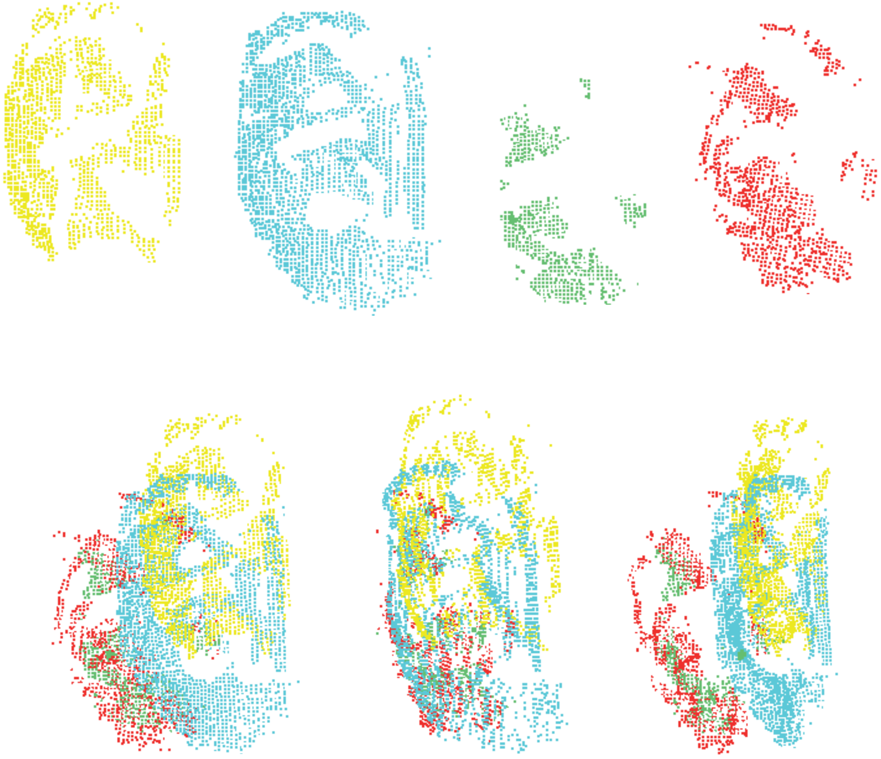


Figure 5.6. Point Cloud Fusion: the four clouds in the top row are extracted from four different views. In the bottom row the result of merging the clouds without ICP alignment.

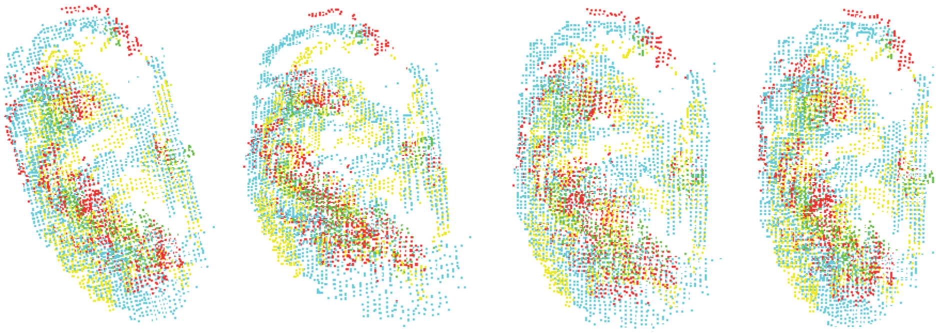


Figure 5.7. ICP alignment: running the ICP registration procedure the fusion of the clouds results in a coherent model. The lack of points due to self occlusions is greatly reduced here.

5.4 Geometrical Evaluation

In this section I present the methods used to evaluate the quality of the extracted 3D point clouds.

As explained in Section 5.1, to evaluate the quality of an extracted cloud I measure the similarity between it and its high-resolution counterpart. The laser scanned models are shown in figure 5.8, together with the associated clouds extracted with the proposed method.

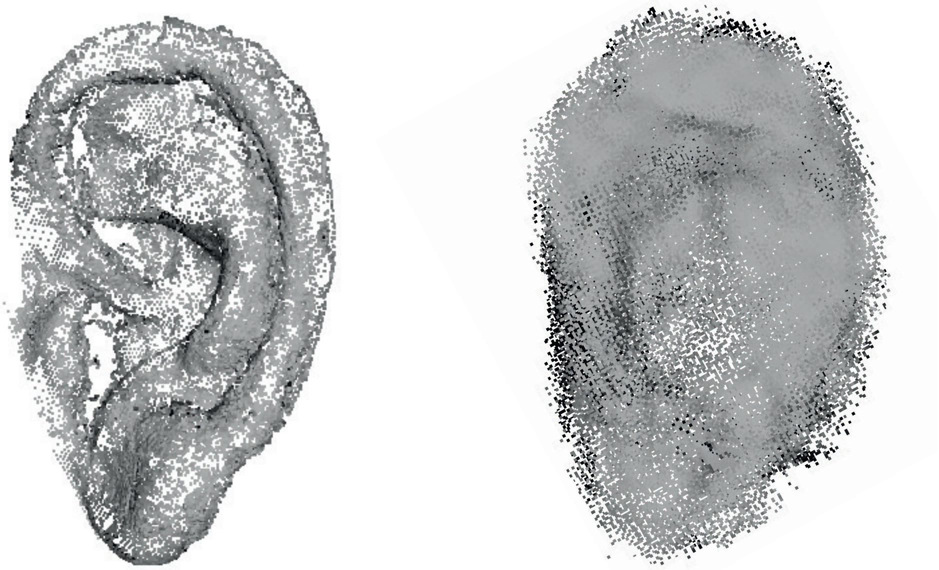


Figure 5.8. Laser Scan extraction and Proposed Method: the model on the left is extracted using a Laser Scanner, that on the right by using the proposed method. In the latter, the shape details are difficult to visualize due to the high point density.

To consistently compare the low and high resolution clouds a *Similarity Metric* is needed. The *Hausdorff Distance* [53] is widely used for this purpose. Let consider the distance $d(\boldsymbol{p}, \mathcal{X})$ of a point $\boldsymbol{p}$ from the cloud $\mathcal{X}$ as defined in (3.8). The *Hausdorff Distance* $h(\mathcal{A}, \mathcal{B})$ between the point clouds $\mathcal{A}$ and $\mathcal{B}$ is defined as:

$$h(\mathcal{A}, \mathcal{B}) = \max_{\boldsymbol{p} \in \mathcal{A}} d(\boldsymbol{p}, \mathcal{B}) \tag{5.1}$$

High values of h imply low similarity between the analyzed clouds. In this context the cloud $\mathcal{A}$ represent one of the low-resolution ear models, cloud $\mathcal{B}$ represents the correspondent high-resolution model. The results of the similarity evaluation for subjects $\mathcal{J}$ and $\mathcal{L}$ with the relative ground-truth models are shown in table 5.1. To understand which parts of the models $\mathcal{A}, \mathcal{B}$ are similar the most, a color is associated to the distance $d(\boldsymbol{p}, \mathcal{B})$ $\forall \boldsymbol{p} \in \mathcal{A}$. The color are then mapped to each point $\boldsymbol{p}$, as shown in figure 5.9.

Subject	Hausdorff Distance
$\mathcal{L}$	4.47 [mm]
$\mathcal{J}$	5,02 [mm]

Table 5.1. Hausdorff Distance: the values of $h(\mathcal{E}_{\mathcal{L}}^{low}, \mathcal{E}_{\mathcal{L}}^{high})$ and $h(\mathcal{E}_{\mathcal{J}}^{low}, \mathcal{E}_{\mathcal{J}}^{high})$ are shown.

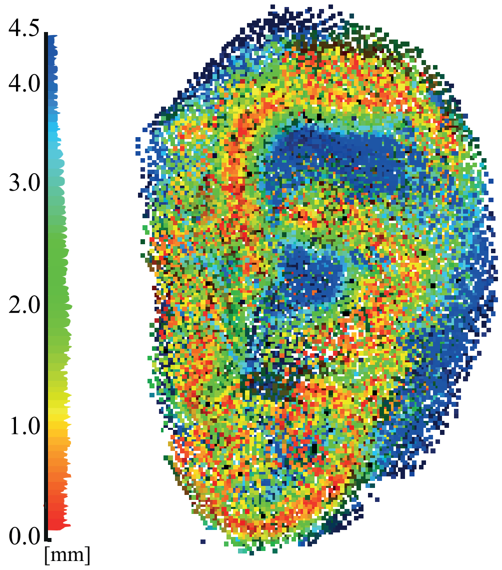


Figure 5.9. Hausdorff Distance Distribution: the distance between each point in $\mathcal{A}$ and it's closest point in $\mathcal{B}$ is mapped to a color. The resulting mapping and the associated histogram are shown here: blue means high distance, red means low distance.

5.5 Acoustical Evaluation

In this section the results of the HRTFs computation using the extracted models are presented. HRTFs are computed for both the high and low resolution models and then compared to evaluate the quality of extraction method from an acoustic perspective. The *Spatial Frequency Response Surface* (SFRS) visualization tool is introduced and the *Spectral distortion* metric is defined.

The HRTFs are computed using the method described in Section 4.5. At high frequencies the human sound localization ability is low, therefore the simulations run to compute the HRTFs have been frequency limited in a range of 20-5000 HZ. Also, the HRTFs magnitude already shows high variability among subjects within this range [54], therefore it can be considered a reasonable frequency interval for the HRTF computation. To the best of my knowledge, only the magnitude of the HRTFs is studied in the literature. Therefore, I focus the HRTFs analysis on the magnitude only without considering the phase. SFRS's present the same HRTF magnitude response information, except in a different coordinate system. Specifically, one SFRS is constructed for every frequency bin in the HRTF magnitude response, where magnitude is plotted against azimuth and elevation. In this manner, every HRTF is used in the computation of a single SFRS. More precisely

$$\mathbf{SFRS}_f(\theta,\phi) = \mid \mathbf{HRTF}(f,\theta,\phi) \mid \quad \theta \in \Theta \quad \phi \in \Phi \tag{5.2}$$

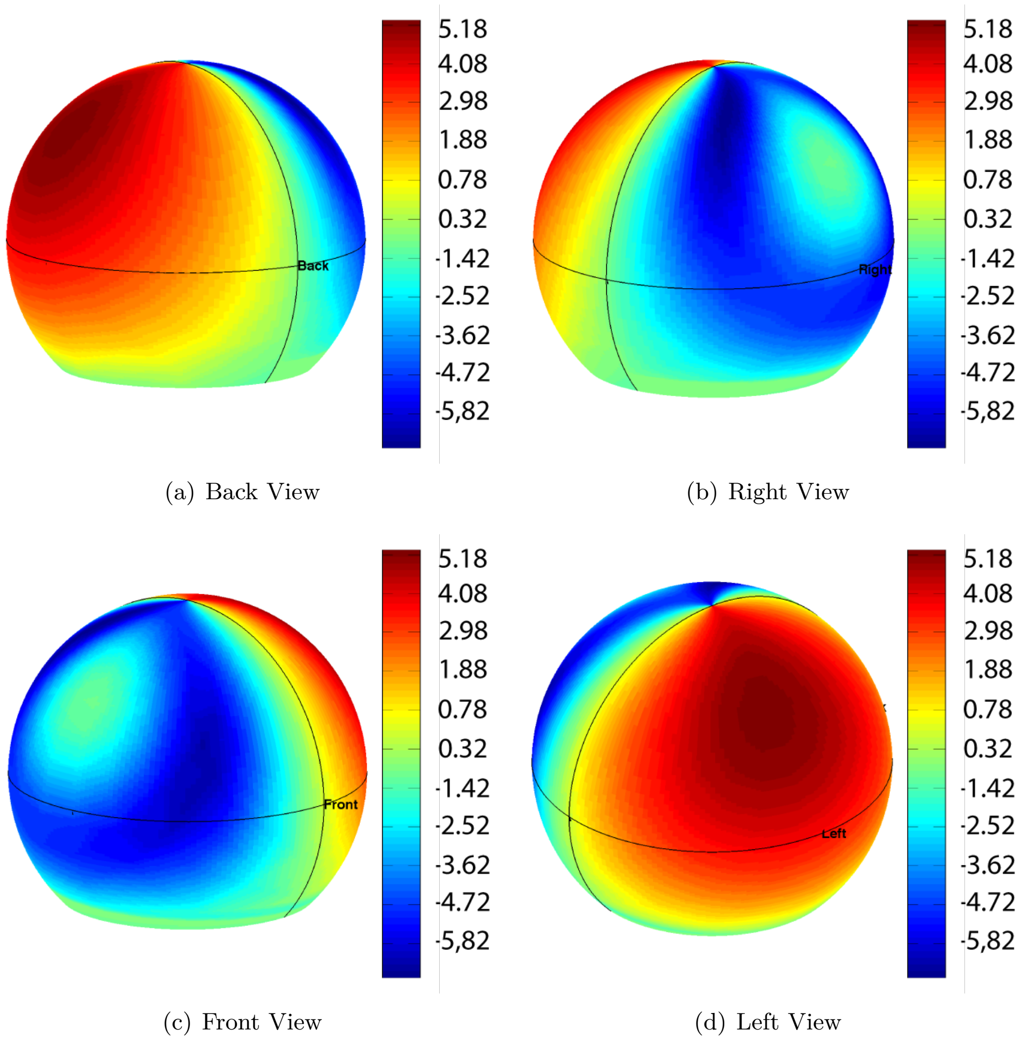
with Θ and Φ being the set of azimuthal and elevation positions respectively. The SFRS is usually plotted as a 2D image, in this thesis instead, I map it on the 3D sphere representing the microphones positions at which the HRTFs have been computed. In this way it is easier to inspect the global features of the SFRS. Examples of *Spherical SFRS* plots are shown in figure 5.10.
Having defined how to compute and visualize the HRTFs is now possible to asses the validity of the 3D extraction procedure from an acoustic analysis perspective.

The idea is to compare the HRTFs computed from the 3D models extracted using the Leap Motion with the ones predicted from the associated ground-truth Laser Scanned models. A first analysis is carried out by visually comparing the spherical SFRS plots at different frequencies. Visual inspection allows to easily recognize the main features, similarity and dissimilarities between the plotted SFRS; this is useful to rapidly understand whether the two computed HRTFs are comparable or not. Since there is one SFRSs per frequency, the comparison can be carried out at different frequency ranges. Doing so, it's possible to understand up to which frequency the low and the high resolution 3D models provide HRTFs that are similar enough. An example of spherical plotting comparison can be found in figure 5.11.

Visual inspection is certainly useful, although, in order to keep the analysis rigorous, a similarity metric has to be defined. I consider an error measure widely used in recent literature: *Spectral Distortion* SD

$$SD_{f_a}^{f_b}(H,\hat{H}) = \sqrt{\frac{1}{N_{a,b}} \sum_{i\in[f_a,f_b]} \left(20\log_{10}\frac{\mid H(f_i,\theta,\phi)\mid}{\mid \hat{H}(f_i,\theta,\phi)\mid}\right)^2} \quad [\text{dB}] \tag{5.3}$$

where $H(f_i,\theta,\phi)$ is the ground-truth reference HRTF at frequency bin f_i, $\hat{H}(f_i,\theta,\phi)$ is the reconstructed HRTF at frequency bin f_i, and $N_{a,b}$ is the number of available frequencies in the considered range $[\ f_a, f_b\ ]$. SD is a local measure, meaning that it evaluates the similarity between two HRTFs at a specific (θ,ϕ) location. To asses the overall similarity I compute the SD at all the (θ,ϕ) directions obtaining a SD



(a) Back View

(b) Right View

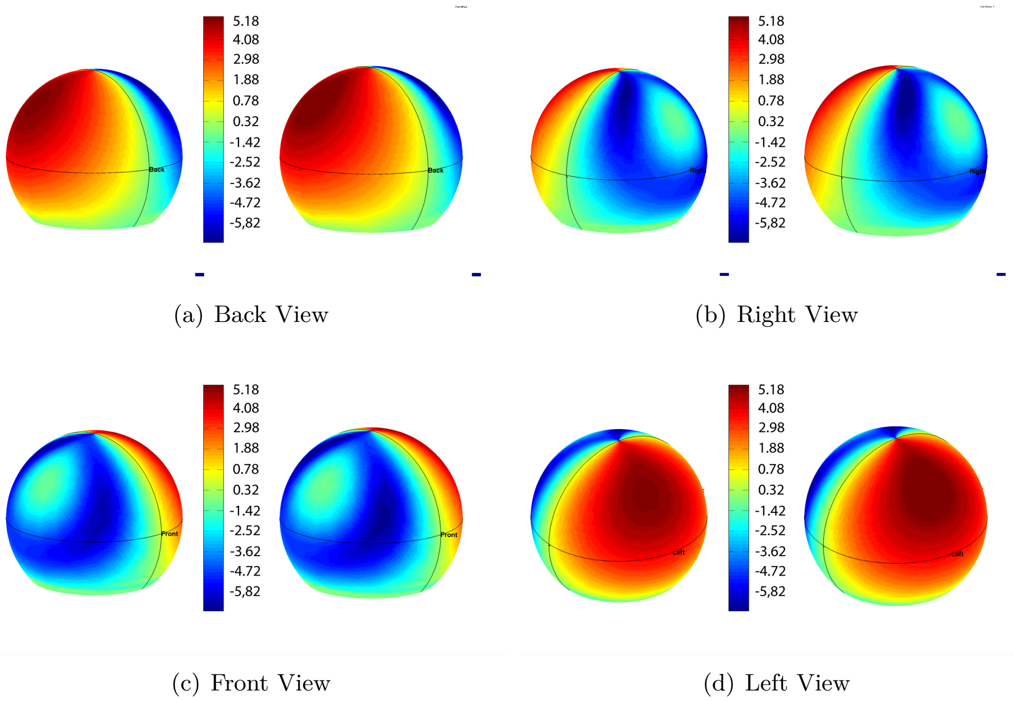
(c) Front View

(d) Left View

Figure 5.10. Spherical SFRS - subject $\mathcal{L}$: the interpolated SFRS at 1kHz is mapped to a sphere and shown from different viewpoints. The discrete sphere is the same of the SYMARE database. The back, right, Front and Left view are relative to the coordinates system presented in 2. All the values are in [dB]

surface. This is then mapped to the same sphere used for the spherical plots so to obtain a 3D distribution of the (SD). In figure 5.12 the SDs resulting from the comparison of $H_{\mathcal{J}}^{low}$ with $H_{\mathcal{J}}^{high}$ and $H_{\mathcal{L}}^{low}$ with $H_{\mathcal{L}}^{high}$ are shown. Spectral distortion provides meaningful information regarding the quality of an estimated HRTFs, as explained in [55] SD values below 5.7 dB results in low localization errors. The practical implication is that, as long as $SD(H, \hat{H}) \leq 5.7$ dB, a sound source could be positioned in space by filtering it with $\hat{H}$ instead of H, introducing a negligible error. In figure 5.13 another example of SD obtained comparing reference and predicted HRTFs is shown.

SD is a local measure, a global similarity metric between two HRTFs can be obtained combining the mean and the standard deviation of the SD. Let $E_{f_a}^{f_b}$ be the spatial mean of $SD_{f_a}^{f_b}(H, \hat{H})$, $\sigma_{f_a}^{f_b}$ its standard deviation and $[\ f_a, f_b\ ]$ the considered



(a) Back View

(b) Right View

(c) Front View

(d) Left View

Figure 5.11. Spherical SFRS Comparison - Subject $\mathcal{L}$: in each plot the sphere on the left is the SFRS computed using the Laser Scanned Model, that on the right is the one extracted by using Leap Motion and Kinect. SFRS is computed at 1kHz and shown at different viewpoints. All the values are in [dB]

frequency range such that:

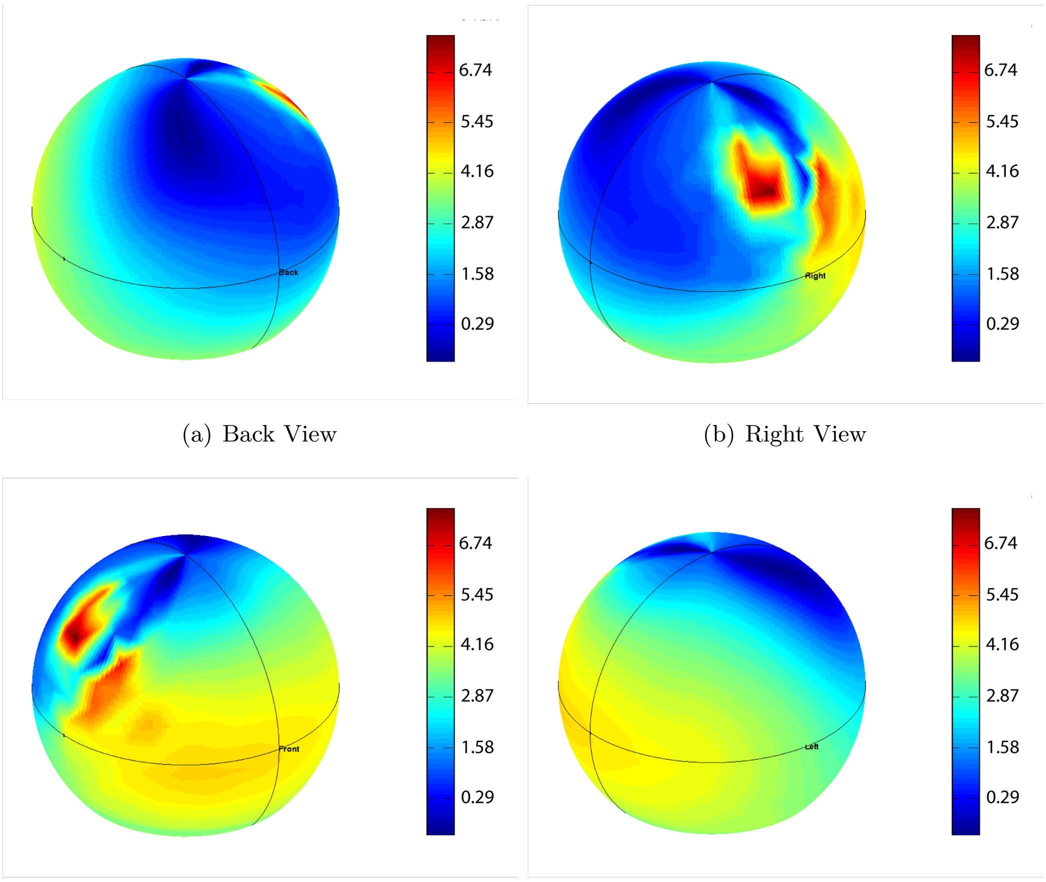
$$E_{f_a}^{f_b} = \frac{1}{N_\theta N_\phi} \sum_\theta \sum_\phi SD_{f_a}^{f_b}(H, \hat{H}) \qquad \theta \in \Theta \quad \phi \in \Phi \tag{5.4}$$

$$\sigma_{f_a}^{f_b} = \sqrt{\frac{1}{N_\theta N_\phi} \sum_\theta \sum_\phi (SD_{f_a}^{f_b}(H, \hat{H}) - E_{f_a}^{f_b})^2} \qquad \theta \in \Theta \quad \phi \in \Phi \tag{5.5}$$

with Θ and Φ being the set of azimuthal and elevation positions respectively and $N_\theta N_\phi$ their cardinality. Let's define the *Global Spectral Distortion* (GSD) as:

$$GSD_{f_a}^{f_b}(H, \hat{H})_\lambda = E_{f_a}^{f_b} + \lambda \sigma_{f_a}^{f_b} \tag{5.6}$$

where λ is a scalar value. In the evaluation I set λ is set equal to 1.5 so to weight more the dispersion of data than the average. A rigorous study on how to set the weight λ is out of the scope of this work and can represent a interesting point for further research. Notice that in both the analyzed cases the GSD is below the 5.7 dB threshold, considering that $GSD = SD + X$ where X is a positive term, this implies that also SD is less than 5.7 dB. The low SD and GSD values suggest that the fidelity of the low-resolution models is high enough to capture the main features of the HRTFs in the considered range. Since the selected frequency interval is large enough to present a high inter-subject HRTFs variability, the extracted 3D clouds can be effectively used to achieve a good degree of *HRTFs Personalization.*



(a) Back View

(b) Right View

(c) Front View

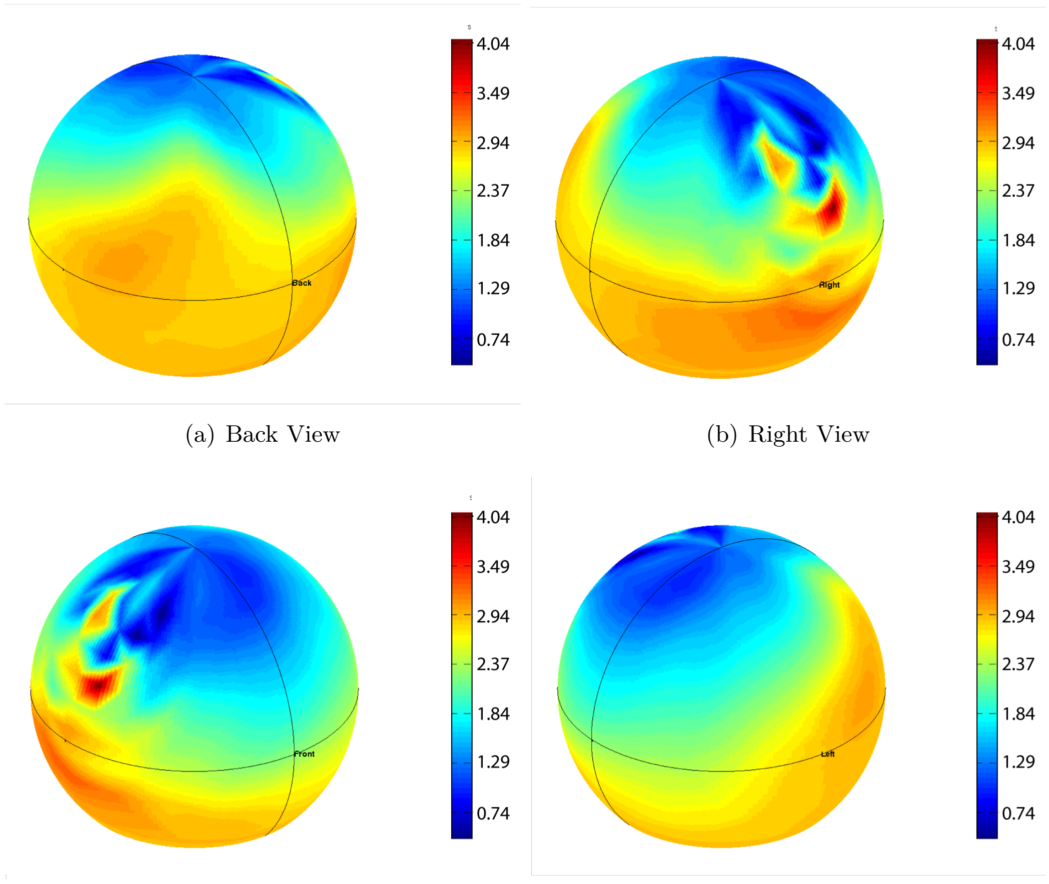
(d) Left View

Figure 5.12. Spherical SD - Subject $\mathcal{J}$: the SD of the left HRTFs is mapped to a sphere and shown from different views. Notice that the highest SD values are in the right part of the sphere; this is probably due to meshing imprecision in the right ear area. All the values are in [dB]. Frequency range: [20Hz – 5kHz] average SD value is 2.45 dB, the maximum is 6.74 dB

5.6 Psychocoustical Considerations

In this section the relation between similarity of 3D models and the associated HRTFs are investigated. The goal is to understand if a similarity in the geometrical domain implies a similarity in the sound features too. As explained in chapter 2, the HRTFs result from the scattering processes of the sound off the ear; this, in turn, depends on the ear anatomy. Therefore, it is realistic to expect a correlation between similarities in the two domains.

Even though the relation between human morphology and HRTFs features is not completely clear yet, many studies have shown that localization ability, especially in the vertical direction, is brought by the presence of the pinnae [56]. In particular pinnae resonances and diffraction inside the concha are seen to contribute to the HRTF spectral shape [57]. These psycho acoustical considerations have to be taken into account when evaluating the quality of the 3d model extraction. For instance,



(a) Back View

(b) Right View

(c) Front View

(d) Left View

Figure 5.13. Spherical SD - Subject $\mathcal{L}$: the SD of the left HRTFs is mapped to a sphere and shown from different views. Notice that the maximum value is below the threshold of 5.7 dB. All the values are in [dB]. Frequency range: [20Hz – 5kHz] average SD value is 2.11 dB, the maximum is 4.08 dB

consider the low resolution model $\mathcal{ET}_{\mathcal{L}}^{low}$ and the correspondent high-resolution model $\mathcal{ET}_{\mathcal{L}}^{high}$: by using the *Hausdorff Distance* it is possible to evaluate how similar the two models are, from a purely geometrical standpoint. These models can then be used to compute the HRTFs $H_{\mathcal{L}}^{low}, H_{\mathcal{L}}^{high}$ by means of BEM techniques; the SD value between the two is a measure of how similar the two models are from an acoustical standpoint. As explained in 5.1, the idea is to introduce the psycho acoustical intuition in the geometrical similarity computation step. Following the approach proposed by [33] the pinna contours are considered to be perceptually more important than other areas of the 3D model. In the Hausdorff Distance computation this information is lost because all the points of the cloud are equally weighted. For this reason, I refine that distance by introducing a weighting scheme. A scalar weight is assigned to each point of the 3D cloud: if the point belongs to a pinna contour its weight is equal to 1, otherwise it is equal to 0.5. The pinna contours are manually traced on the 3D model, an example of traced contours is shown in figure 5.14. By using this weighting schema the *Perceptually Weighted Hausdorff Distance* (PWHD) can be defined as:

$$Ph(\mathcal{A},\mathcal{B}) = \max_{\boldsymbol{p}\in\mathcal{A}} w(\boldsymbol{p})d(\boldsymbol{p},\mathcal{B}) \tag{5.7}$$

where $d(\boldsymbol{p},\mathcal{B})$ is defined in equation (3.8) and $w(\boldsymbol{p})$ is a scalar weight assigned to point $\boldsymbol{p}$.

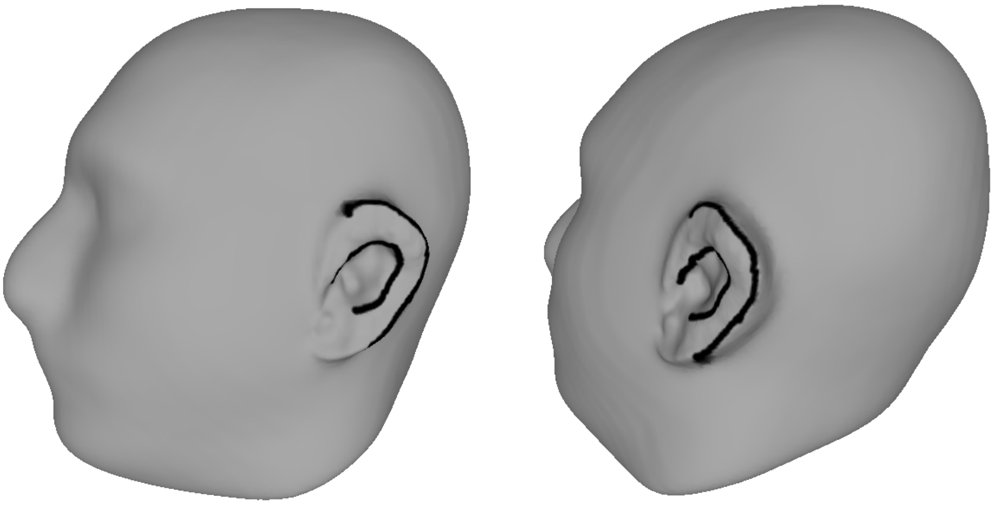
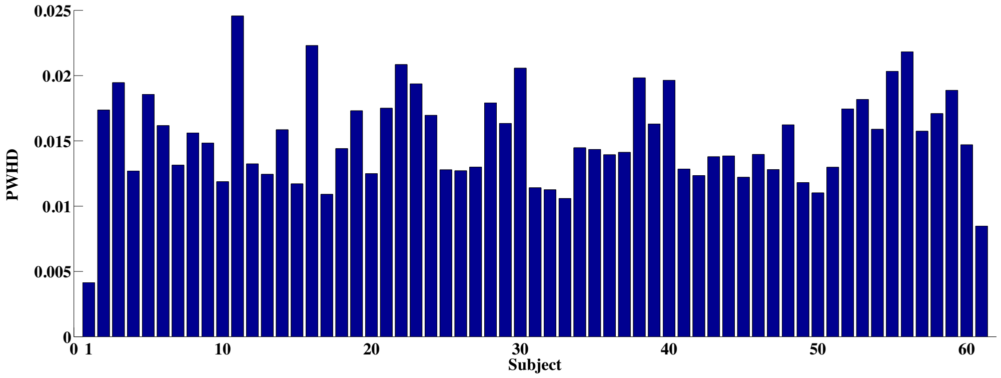

(a) Back View (b) Right View

Figure 5.14. Traced Contours - the black lines indicate the manually traced pinna contours.

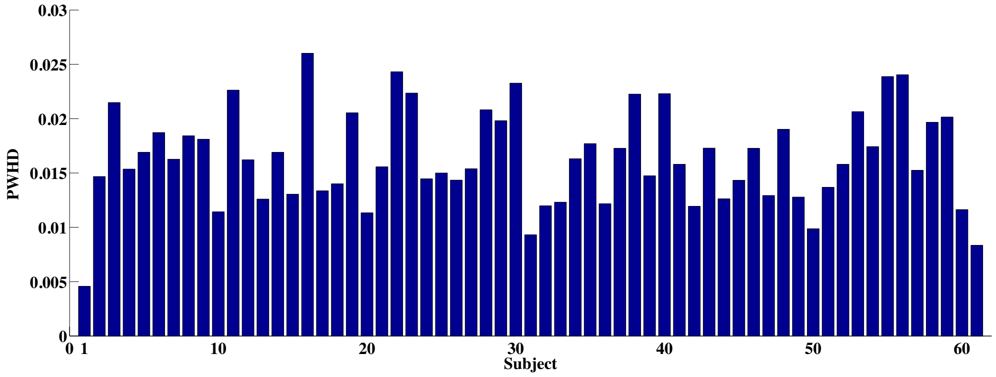
The PWHD is then used to compute the similarity between the test and the target models (as defined in 5.1). The PWHD values are then used to rank the target models on a similarity basis: the lower the PWHD between the target and the test, the higher will be the position of that target model in the rank. According to this reasoning, the first-ranked model is the most similar to the test in analysis. As it can be appreciated in figure 5.15(a), by computing the PWHD between $\mathcal{ET}^{low}_{\mathcal{L}}$ and all the target models the first ranked is $\mathcal{ET}^{high}_{\mathcal{L}}$, that is the high resolution version of the tested model. This means that, as expected, the low and high resolution models $\mathcal{ET}^{low}_{\mathcal{L}}$,$\mathcal{ET}^{high}_{\mathcal{L}}$ are the most similar models, since they provide the minimum PWHD value. The same is true with $\mathcal{ET}^{low}_{\mathcal{J}}$ and $\mathcal{ET}^{high}_{\mathcal{J}}$ as shown in figure 5.15(b).

The next logical step is to rank the target subjects considering the GSD metric. More precisely the HRTFs of the test and the target models are used to compute the GSD values and the target models are ranked accordingly. As it was for the PWHD comparison, the lower the GSD between the target and the test's HRTFs , the higher will be the position of that target model in the rank. As it can be seen in figure 5.16, subject number 1 (i.e. $\mathcal{ET}^{high}_{\mathcal{J}}$ when $\mathcal{ET}^{low}_{\mathcal{J}}$ is tested and $\mathcal{ET}^{high}_{\mathcal{L}}$ when $\mathcal{ET}^{low}_{\mathcal{L}}$ is tested and) is the one having minimum GSD.

One may ask if there exists a relation between the PWHD and the GSD values or between the associated rankings. Such relations would highlight a link between geometrical shapes and acoustic features; in particular it would mean that, by weighting more the psycho acoustically relevant shapes, it could be possible to link geometrical and acoustical similarity. To investigate the aforementioned relation the



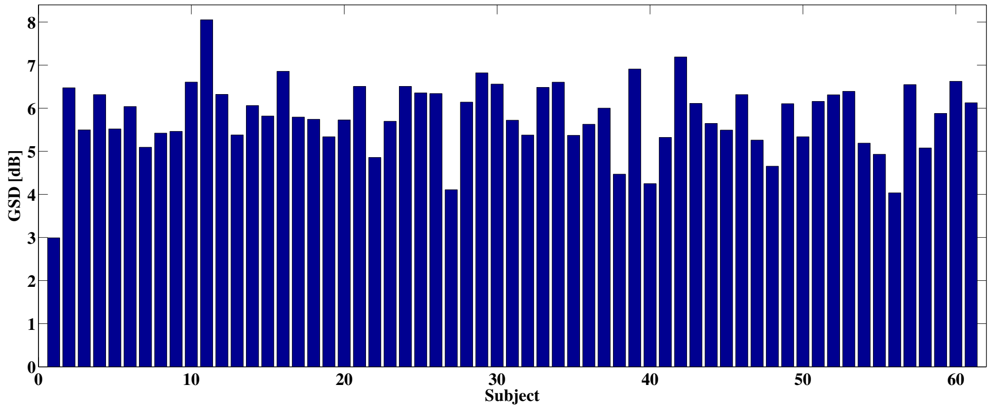
(a) Subject $\mathcal{L}$



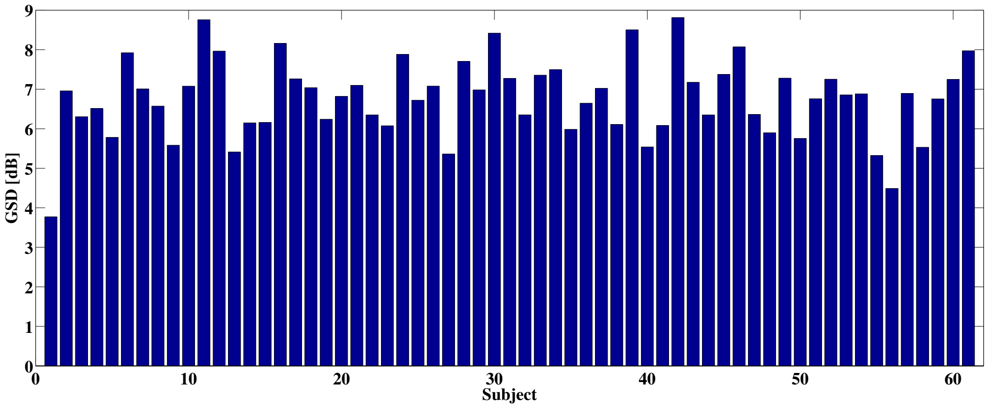
(b) Subject $\mathcal{J}$

Figure 5.15. PWHD Ranking: in the first chart the PWHD between $\mathcal{ET}_{\mathcal{L}}^{low}$ and 61 target models is computed. Subject 1 is the high-resolution model $\mathcal{ET}_{\mathcal{L}}^{high}$, and it is the one having minimum PWHD. In the second chart the same computation is carried out for the test model $\mathcal{ET}_{\mathcal{J}}^{low}$. Subject 1 is now model $\mathcal{ET}_{\mathcal{J}}^{high}$, and it is the one having minimum PWHD.

Pearson's correlation coefficient between the PWHD and GSD values is computed. Even though $\mathcal{ET}_{\mathcal{L}}^{high}$ is the first ranked in both the PWHD and GSD ranks, the correlation value of PWHD and GSD (computed from 20 Hz to 5000 Hz), by itself, is too low to indicate a strong link between the two metrics. In order to further investigate this relation and to make the analysis more rigorous, different frequency bins have been considered in computing the GSD. The targets are ranked on the basis of each sub-band GSD value: in particular, one rank is generated for each considered frequency interval. The position of a subject in the ranking is assigned with the same criteria used before. As it can be seen in the table related to subject $\mathcal{J}$ there are sub-bands in which the subject having the lowest value of GSD is the number 1, that is $\mathcal{ET}_{\mathcal{L}}^{high}$. This means that, for these sub-bands, $\mathcal{ET}_{\mathcal{L}}^{high}$ has both the minimum PWHD ans GSD values: for the sake of clarity, let's call this situation a *Match*. In certain sub-bands this happens also for subject $\mathcal{L}$. A match is an optimal situation because it implies that the two subjects that are similar the most in the geometrical domain are those that are similar the most in the acoustic domain too.



(a) Subject $\mathcal{L}$



(b) Subject $\mathcal{J}$

Figure 5.16. GSD Ranking: in the first chart the GSD between $H_{\mathcal{L}}^{low}$ and the HRTFs of 61 target models is computed. Subject 1 is the high-resolution HRTFs $H_{\mathcal{L}}^{high}$, and it is the one that provides minimum GSD. In the second chart the same computation is carried out for the test HRTFs $H_{\mathcal{J}}^{low}$. Subject 1 is now HRTFs $H_{\mathcal{J}}^{high}$, and it is the one that provides minimum GSD.

The sub-band division can be used also to study how the correlation of PWHD and GSD changes at the varying of the frequency. A high value of correlation in a certain band would imply that, for that frequency bin, the PWHD gives good information about the acoustic properties of the model. In other words, such correlation would highlight a relation between geometrical similarity, computed with PWHD, and the acoustical similarity, computed with the GSD. In figure 5.17 the results of this analysis are shown; below 2 kHz the high correlation values highlight the existence of a correlation between the PWHD and the GSD; in turns this implies a relation between similarity in the geometrical and the acoustical domains. These results also suggest that the resolution of the extracted models is enough to capture those parts of the ear that are perceptually relevant in the considered frequency range.

In this chapter the quality of the proposed methodology has been validated. The SD and GSD values between HRTFs computed from extracted and ground-truth

(a) Subject $\mathcal{L}$

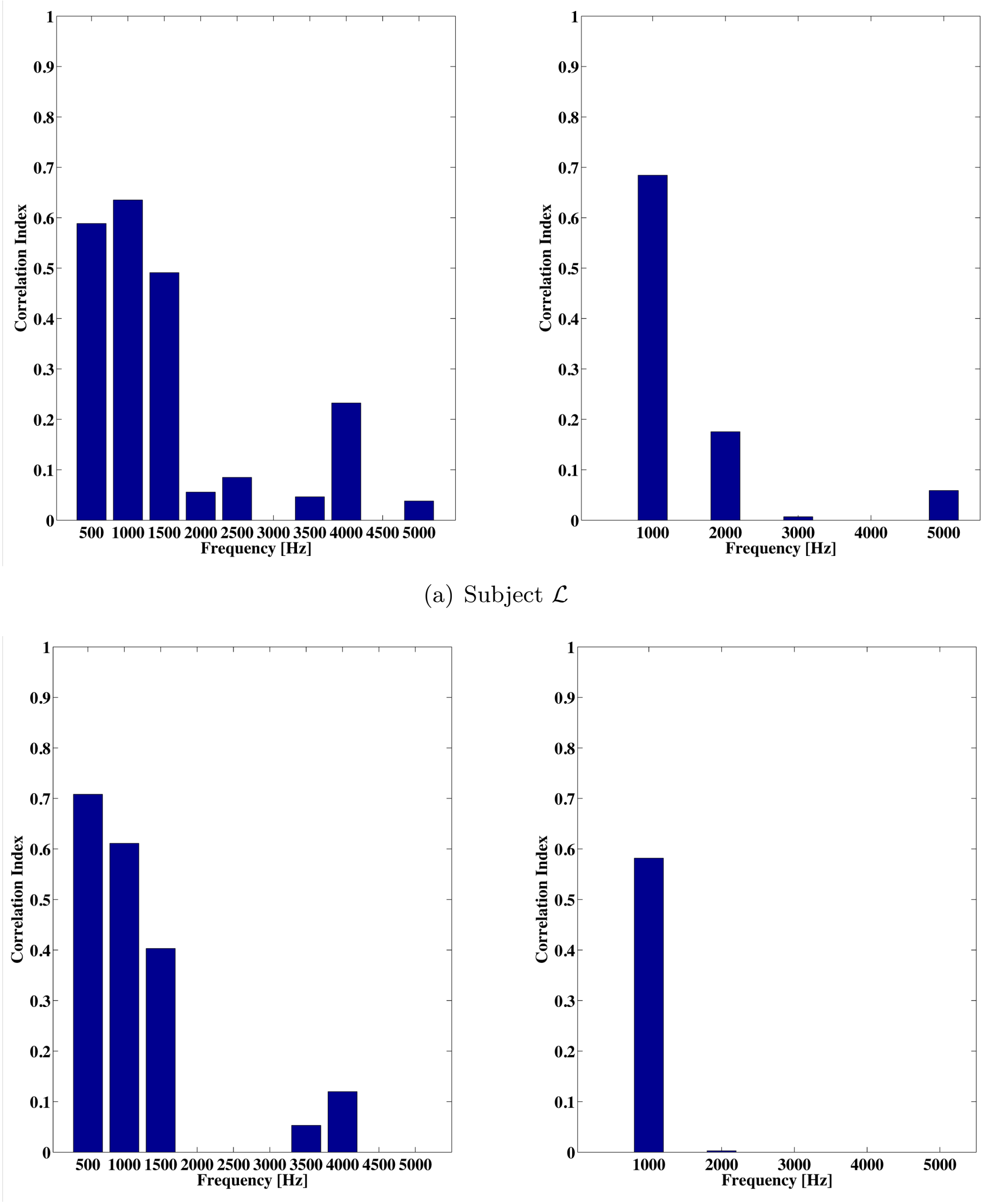
Frequency Bin	Subject Having Lowest GSD
20 Hz -500 Hz	1
500 Hz-1000 Hz	9
1000 Hz-1500 Hz	9
1500 Hz-2000 Hz	27
2000 Hz-2500 Hz	27
2500 Hz-3000 Hz	1
3000 Hz-3500 Hz	1
3500 Hz-4000 Hz	27
4000 Hz-4500 Hz	1
4500 Hz-5000 Hz	1
20 Hz-1000 Hz	9
1000 Hz-2000 Hz	27
2000 Hz-3000 Hz	1
3000 Hz-4000 Hz	50
4000 Hz-5000 Hz	1

(b) Subject $\mathcal{J}$

Frequency Bin	Subject Having Lowest GSD
20 Hz -500 Hz	1
500 Hz-1000 Hz	9
1000 Hz-1500 Hz	2
1500 Hz-2000 Hz	27
2000 Hz-2500 Hz	4
2500 Hz-3000 Hz	1
3000 Hz-3500 Hz	1
3500 Hz-4000 Hz	27
4000 Hz-4500 Hz	1
4500 Hz-5000 Hz	1
20 Hz-1000 Hz	9
1000 Hz-2000 Hz	27
2000 Hz-3000 Hz	1
3000 Hz-4000 Hz	50
4000 Hz-5000 Hz	1

Table 5.2. GSD Ranking: the target subjects having the lowest GSD are listed for each sub-band. The first 10 bons have band of 500 Hz starting from 20 Hz until 5000 Hz; in the last 5 the band is 1000 Hz.

models are low, implying an extraction precision that is suitable for the binaural processing context. The high geometrical similarity suggests that this technique can be adopted in cost-critical scenarios instead of high-end machinery. The overall cost of this system is indeed low, thanks to the exploitation of off-the-shelf hardware. The possibility to assign higher weights to the psychoacoustically most relevant anthropometric features allows to better evaluate the similarity between two models. The correlation between PWHD and GSD highlights the existence of a link between similarity in the geometrical and acoustical domains. This result is interesting and can be the starting point for further research; in conclusion the proposed 3D reconstruction method is able to provide models that can be effectively used to generate personalized HRTFs.



(a) Subject $\mathcal{L}$

(b) Subject $\mathcal{J}$

Figure 5.17. Sub-band Correlation Analysis: the PWHD and GSD values are correlated for each considered frequency bin. The frequency division is the same used in table 5.2.

Chapter 6

Conclusions And Future Work

In this thesis I have presented an alternative technique for acquiring a 3D model of the ear of a subject. Considering the increasing interest in binaural and Anthropometric signal processing, the importance of a fast and easy-to-use acquisition procedure has been stressed. The Multi View approach is the force that drives the design of both the algorithms and the acquisition system. This choice is promising since it decouples the processing and the scanning stages, greatly shortening the duration of a scan session. Indeed, the Multi View approach allows to extract a 3D cloud from a few couples of images that, by using the proposed acquisition system, can be captured in less than a minute. Moreover the proposed algorithm is not sensitive to the movements of the subject, issue that strongly degrades the results of the typical extraction procedures. The cost of the system is kept low by using off-the-shelf hardware: the Leap Moton and the Kinect devices together cost less than 200 euros and can be easily found on the market. The acquisition system is small, simple and easy to handle, making this implementation potentially interesting also for a large scale-consumer application.

To fairly evaluate the quality of the extraction, the ground-truth models have been acquired using a Laser Scanner; having those models it has been possible to compare them to the correspondent ones, extracted using the proposed method. Different evaluation metrics have been considered, both the geometrical and the acoustical domains have been investigated. The low values of *Spectral Distortion* and *Global Spectral Distortion* show that the 3D extraction method provides models that can be effectively used to synthesize the personalized HRTFs. The simulations have been frequency limited to 5kHz and in this range the spatial mean value of the SD is way below the threshold of 5.7 dB; this means that, potentially, neither a high-fidelity model nor acoustic measurements are needed to effectively predict the HRTFs in this frequency range. The relation between geometry and HRTFs has been studied by analyzing the *Perceptually Weighted Hausdorff Distance*, in this metric the perceptually relevant parts of the ear are weighted more than the rest of the model. Computing such distance between a test model and a set of target models, one is able to find the two subjects that are similar the most. The interesting result is that also the HRTFs of the geometrically ”matched” couple of subjects are the most similar; the SD of those HRTFs has indeed the smallest GSD. To further investigate the relation between geometrical ear shape and HRTFs features the GSD values have been computed at different frequency ranges. Correlating those values with the PWHD values a meaningful relation has been highlighted below 2 kHz. This result suggest that the existence of a link between similarity in the geometrical and acoustical domains.

To conclude, I mention a set of possible evolutions of the work presented in this thesis. The proposed approach could be used to rapidly create a large database of low-resolution 3D models of ears. Automatic selection of correctly acquired images could be included in the system together with automatic ear shape recognition. A large dataset could be statistically processed to gain a better insight on which parts of the ear surface are acoustically more relevant. Statistical analysis could also be beneficial to the creation of a parametric 3D model of the ear; this could be later fitted to the specific anthropometry of the subject. Also morphing between low and high resolution ears could be adopted as a technique to choose two most similar models. The *Large Deformation Diffeomorphic Metric Mapping* technique proposed by C.Jin already represents an interesting result in this sense.

Bibliography

[1] Richard O.Duda V. Ralph Algazi, Pierre L. Divenyi. Subject depedent transfer functions in spatial hearing. 1997.

[2] C.P. Brown and R.O. Duda. An efficient hrtf model for 3-d sound. In *Applications of Signal Processing to Audio and Acoustics, 1997. 1997 IEEE ASSP Workshop on*, pages 4 pp.–, Oct 1997.

[3] V. Ralph Algazi, Richard O. Duda, Ramani Duraiswami, Nail A. Gumerov, and Zhihui Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5), 2002.

[4] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4463–4467, May 2014.

[5] P. Bilinski, J. Ahrens, M.R.P. Thomas, I.J. Tashev, and J.C. Platt. Hrtf magnitude synthesis via sparse representation of anthropometric features. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4468–4472, May 2014.

[6] F. Grijalva, L. Martini, S. Goldenstein, and D. Florencio. Anthropometric-based customization of head-related transfer functions using isomap in the horizontal plane. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4473–4477, May 2014.

[7] C. Ahuja and R.M. Hegde. Fast modelling of pinna spectral notches from hrtfs using linear prediction residual cepstrum. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4458–4462, May 2014.

[8] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Computer simulation of hrtfs for personalization of 3d audio. In *Universal Communication, 2008. ISUC '08. Second International Symposium on*, pages 435–440, Dec 2008.

[9] Lawrence V. Hmurcik1 Manan Joshi, Navarun Gupta1. Modeling of pinna related transfer functions (prtf) using the finite element method (fem). 2013.

[10] P. Guillon, R. Zolfaghari, N. Epain, A. van Schaik, C.T. Jin, C. Hetherington, J. Thorpe, and A. Tew. Creating the sydney york morphological and acoustic

recordings of ears database. In *Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 461–466, July 2012.

[11] Dooyong Sung, Nara Hahn, and Kyogu Lee. Individualized hrtfs simulation using multiple source ray tracing method. In *Audio Engineering Society Conference: 49th International Conference: Audio for Games*, Feb 2013.

[12] Niklas Röber, Sven Andres, and Maic Masuch. *HRTF simulations through acoustic raytracing*. Univ., Fak. für Informatik, 2006.

[13] Reza Zolfaghari, Nicolas Epain, Craig T. Jin, Joan Alexis Glaunès, and Anthony I. Tew. Large deformation diffeomorphic metric mapping and fast-multipole boundary element method provide new insights for binaural acoustics. *CoRR*, abs/1401.7100, 2014.

[14] Laurent Guillon, Pierre ; Simon. *Individualisation des indices spectraux pour la synthèse binaurale recherche et exploitation des similarités inter-individuelles pour l'adaptation ou la reconstruction de HRTF*. PhD thesis, 2009. Reproduction de Thèse de doctorat Acoustique Le Mans 2009.

[15] D.N. Zotkin, J. Hwang, R. Duraiswaini, and L.S. Davis. Hrtf personalization using anthropometric measurements. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 157–160, Oct 2003.

[16] V.R. Algazi and R.O. Duda. Headphone-based spatial sound. *Signal Processing Magazine, IEEE*, 28(1):33–42, Jan 2011.

[17] J.C.R. Licklider. A duplex theory of pitch perception. *Experientia*, 7(4):128–134, 1951.

[18] S.A. Gelfand. *Essentials of Audiology*. Thieme Publishers Series. Thieme, 2009.

[19] DS Brungart. Auditory localization of nearby sources. iii. stimulus effects. *The Journal of the Acoustical Society of America*, 106(6):3589—3602, December 1999.

[20] S. Weinzierl, M. Vorländer, F. Zotter, H.J. Maempel, and A. Lindau. *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014:*. Universitätsbibliothek Technische Universität Berlin, 2014.

[21] V.R. Algazi, R.O. Duda, D.M. 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.

[22] B Masiero, P Dietrich, M Pollow, J Fels, and M Vorländer. Design of a fast individual hrtf measurement system. *Fortschritte der Akustik-DAGA 2012*, 2012.

[23] M. Pec, M. Bujacz, P. Strumillo, and A. Materka. Individual hrtf measurements for accurate obstacle sonification in an electronic travel aid for the blind. In *Signals and Electronic Systems, 2008. ICSES '08. International Conference on*, pages 235–238, Sept 2008.

[24] Peter Balazs, Bernhard Laback, and Piotr Majdak. Multiple exponential sweep method for fast measurement of head related transfer functions. In *Audio Engineering Society Convention 122*, May 2007.

[25] Sascha Spors, Hagen Wierstorf, and Jens Ahrens. Interpolation and range extrapolation of head-related transfer functions using virtual local wave field synthesis. In *Audio Engineering Society Convention 130*, May 2011.

[26] Khoa-Van Nguyen, Thibaut Carpentier, Markus Noisternig, and Olivier Warusfel. Calculation of head related transfer functions in the proximity region using spherical harmonics decomposition: comparison with measurements and evaluation. In *2nd Int. Symposium on Ambisonics and Spherical Acoustics*, pages –, Paris, France, May 2010. cote interne IRCAM: NGuyen10b.

[27] Dmitry N. Zotkin, Ramani Duraiswami, Elena Grassi, and Nail A. Gumerov. Fast head-related transfer function measurement via reciprocity. *The Journal of the Acoustical Society of America*, 120(4), 2006.

[28] Doris J. Kistler and Frederic L. Wightman. A model of headrelated transfer functions based on principal components analysis and minimumphase reconstruction. *The Journal of the Acoustical Society of America*, 91(3), 1992.

[29] R Jenison and K Fissell. A spherical basis function neural network for modeling auditory space. *Neural Computation*, 8(1):115–128, Jan 1996.

[30] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and zero modeling of head-related transfer functions. *Speech and Audio Processing, IEEE Transactions on*, 7(2):188–196, Mar 1999.

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

[32] C. Ahuja and R.M. Hegde. Fast modelling of pinna spectral notches from hrtfs using linear prediction residual cepstrum. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4458–4462, May 2014.

[33] 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.

[34] Yuvi Kahana, Philip A. Nelson, Maurice Petyt, and Sunghoon Choi. Numerical modelling of the transfer functions of a dummy-head and of the external ear. In *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Mar 1999.

[35] Brian F. G. Katz. Boundary element method calculation of individual head-related transfer function. i. rigid model calculation. *The Journal of the Acoustical Society of America*, 110(5), 2001.

[36] Makoto Otani and Shiro Ise. Fast calculation system specialized for head-related transfer function based on boundary element method. *The Journal of the Acoustical Society of America*, 119(5), 2006.

[37] Timothy Walsh, Leszek Demkowicz, and Richard Charles. Boundary element modeling of the external human auditory system. *The Journal of the Acoustical Society of America*, 115(3), 2004.

[38] Tian Xiao and Qing Huo Liu. Finite difference computation of head-related transfer function for human hearing. *The Journal of the Acoustical Society of America*, 113(5), 2003.

[39] Tamás Várady, Ralph R Martin, and Jordan Cox. Reverse engineering of geometric models—an introduction. *Computer-Aided Design*, 29(4):255 – 268, 1997. Reverse Engineering of Geometric Models.

[40] Fausto Bernardini and Holly Rushmeier. The 3d model acquisition pipeline. *Computer Graphics Forum*, 21(2):149–172, 2002.

[41] Heng Liu and Jingqi Yan. Multi-view ear shape feature extraction and reconstruction. In *Signal-Image Technologies and Internet-Based System, 2007. SITIS '07. Third International IEEE Conference on*, pages 652–658, Dec 2007.

[42] J.S. Suri, Kecheng Liu, S. Singh, S.N. Laxminarayan, Xiaolan Zeng, and L. Reden. Shape recovery algorithms using level sets in 2-d/3-d medical imagery: a state-of-the-art review. *Information Technology in Biomedicine, IEEE Transactions on*, 6(1):8–28, March 2002.

[43] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[44] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, Feb 1992.

[45] Yijun Liu. *Fast Multipole Boundary Element Method: Theory and Applications in Engineering*. Cambridge University Press, 2009.

[46] Zhengyou Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, Nov 2000.

[47] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 1106–1112, Jun 1997.

[48] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[49] Kourosh Khoshelham and Sander Oude Elberink. Accuracy and resolution of kinect depth data for indoor mapping applications. *Sensors*, 12(2):1437–1454, 2012.

[50] Qingxiong Yang, N. Ahuja, Ruigang Yang, Kar-Han Tan, J. Davis, B. Culbertson, J. Apostolopoulos, and Gang Wang. Fusion of median and bilateral filtering for range image upsampling. *Image Processing, IEEE Transactions on*, 22(12):4841–4852, Dec 2013.

[51] K.J. Renze and J.H. Oliver. Generalized unstructured decimation [computer graphics]. *Computer Graphics and Applications, IEEE*, 16(6):24–32, Nov 1996.

[52] R. Zolfaghari, N. Epain, C.T. Jin, J. Glaunes, and A. Tew. Large deformation diffeomorphic metric mapping and fast-multipole boundary element method provide new insights for binaural acoustics. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2863–2867, May 2014.

[53] Aspert N, D. Santa-Cruz, and T. Ebrahimi. Mesh: measuring errors between surfaces using the hausdorff distance. In *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, volume 1, pages 705–708 vol.1, 2002.

[54] Areti Andreopoulou, Agnieszka Rogińska, and Hariharan Mohanraj. Analysis of the spectral variations in repeated head-related transfer function measurements.

[55] Takanori Nishino, Shoji Kajita, Kazuya Takeda, and Fumitada Itakura. Interpolation of the head related transfer function on the horizontal plane. *J. Acoust. Soc. Jpn*, pages 91–99, 1999.

[56] Mark B. Gardner and Robert S. Gardner. Problem of localization in the median plane: effect of pinnae cavity occlusion. *The Journal of the Acoustical Society of America*, 53(2), 1973.

[57] Edgar AG Shaw. Acoustical features of the human external ear. *Binaural and spatial hearing in real and virtual environments*, 25:47, 1997.