UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Ingegneria dell'Informazione

Scuola di Dottorato in Ingegneria dell'Informazione
Indirizzo: Scienza e Tecnologia dell'Informazione

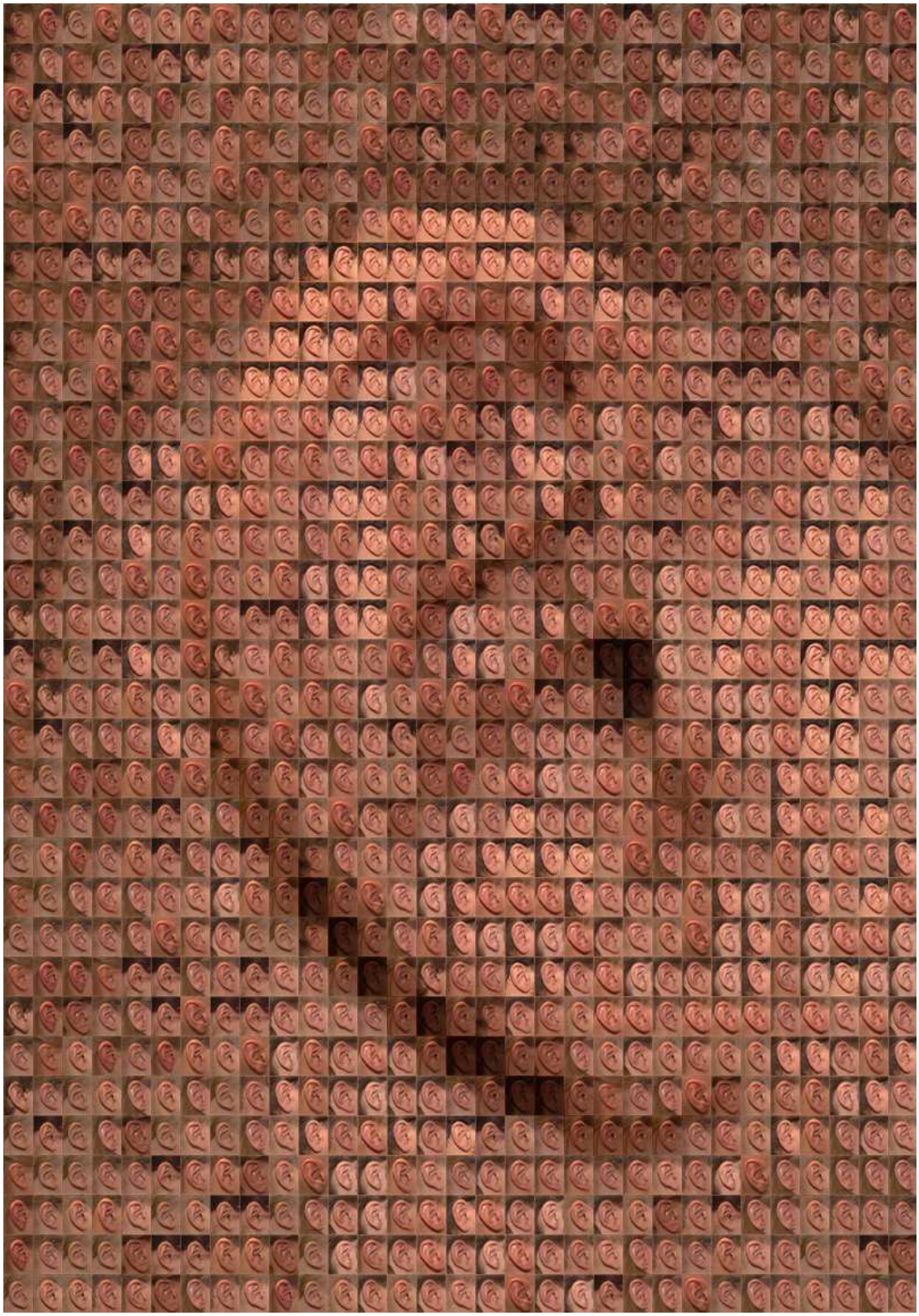
CICLO XXIV

TECHNIQUES FOR CUSTOMIZED BINAURAL AUDIO RENDERING WITH APPLICATIONS TO VIRTUAL REHABILITATION

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco

Supervisore: Ch.mo Prof. Giovanni De Poli

Dottorando: Simone Spagnol



Who speaks, sows. Who listens, reaps.

-Argentine proverb

Prefazione

Le interfacce multimodali rappresentano al giorno d'oggi un fattore chiave per l'abilitazione di un uso inclusivo delle nuove tecnologie. In questo contesto, sono di basilare importanza modelli realistici che descrivano il nostro ambiente, in particolare modelli che rappresentino accuratamente i fenomeni acustici e la comunicazione attraverso la modalità uditiva. Fra questi, i modelli per l'audio spaziale (o 3-D) sono capaci di offrire informazioni accurate sulla relazione tra la sorgente sonora e l'ambiente circostante, rappresentando un'informazione che non può essere sostituita da nessun'altra modalità. Tuttavia, essendo i sistemi multimediali attualmente focalizzati soprattutto sul *processing* grafico e integrati semplicemente con audio *stereo* o *surround*, l'odierna rappresentazione spaziale del suono tende ad essere semplicistica e ad aver poco potenziale interattivo. Inoltre, le tecnologie di auralizzazione si basano correntemente su dispositivi di riproduzione invasivi e/o costosi (ad es. *head-mounted display* e altoparlanti), responsabili di un'esperienza percettiva non integrata a causa di un vuoto mai colmato tra il mondo reale e quello virtuale.

Gli approcci di audio binaurale (ossia basati su riproduzione tramite cuffie) si collocano su un livello diverso. La maggior parte delle tecniche di *rendering* binaurale attualmente utilizzate in ricerca fanno affidamento sull'uso delle cosiddette *Head-Related Transfer Function (HRTF)*, ovvero particolari filtri che catturano le trasformazioni subite da un'onda sonora nel proprio percorso dalla sorgente al timpano, generalmente dovute a effetti di riflessione e diffrazione sul torso, sulla testa, sulle spalle e sui padiglioni auricolari dell'ascoltatore. Tale caratterizzazione permette di posizionare virtualmente una o più sorgenti sonore nello spazio circostante semplicemente filtrando i segnali desiderati attraverso un paio di HRTF, creando quindi una coppia di segnali da presentare ai canali sinistro e destro di un paio di cuffie. In questo modo, campi sonori tridimensionali con un alto senso di immersione possono essere simulati e integrati in strutture multimodali.

Purtroppo, importanti limitazioni si nascondono dietro tali tecniche. Innanzitutto, potrebbero richiedere grosse risorse computazionali nel caso in cui si vogliano simulare più sorgenti sonore nello spazio. In secondo luogo, i filtri HRTF vengono solitamente presentati sotto forma di segnali acustici registrati attraverso appositi manichini: ciò significa che le differenze antropometriche fra diversi soggetti non vengono prese in considerazione. Al contrario, alla pari dell'importanza della posizione relativa tra l'ascoltatore e la sorgente sonora, l'antropometria del soggetto

ha un ruolo chiave nella caratterizzazione della HRTF: sebbene le HRTF non individualizzate rappresentino un mezzo diretto ed economico per offrire una parvenza di percezione 3-D nella riproduzione via cuffie, l'ascolto del segnale risultante potrebbe frequentemente tradursi in evidenti errori di localizzazione quali percezione distorta dell'elevazione della sorgente, inversioni fronte-retro, e mancanza di esternalizzazione, specialmente in condizioni statiche. D'altro canto, misurare individualmente le HRTF di un numero significativo di soggetti comporterebbe un elevato dispendio di risorse e di tempo.

La modellazione strutturale delle HRTF rappresenta invece un'attraente soluzione a tutte le sopracitate limitazioni. Nello specifico, isolando i contributi alla HRTF di testa, padiglioni auricolari, canali uditivi, spalle e torso dell'ascoltatore in diverse componenti - ciascuna modellante un fenomeno acustico ben definito - la HRTF globale può essere ricostruita attraverso un'adeguata combinazione di tutti gli effetti considerati, grazie alla linearità della scomposizione.

Questa tesi presenta un modello strutturale utilizzabile per una riproduzione immersiva del suono, focalizzato in particolare sul contributo del padiglione auricolare (pinna) alla HRTF. La pinna gioca un ruolo fondamentale nella percezione dell'elevazione della sorgente grazie alle rilevanti modifiche spettrali che essa introduce nel suono che arriva al timpano. Tuttavia, la relazione tra i fenomeni acustici dovuti alla stessa - soprattutto risonanze e riflessioni - ed antropometria non ha ancora trovato una convincente rappresentazione nella letteratura. Una promettente corrispondenza tra i punti di riflessione teorici sulla superficie della pinna e le frequenze di una terna di *notch* spettrali presenti nella HRTF è invece discussa in questa tesi: tale risultato, sicuramente nuovo nel suo genere, apre le porte ad un'interessante forma di personalizzazione del modello strutturale, il quale include parametri relativi all'antropometria dell'utente oltre a parametri più strettamente correlati alla posizione della sorgente.

L'approccio proposto ha implicazioni anche in termini di trasmissione dei contenuti, poiché opera elaborando un segnale monofonico esclusivamente dalla parte del ricevitore (ad es. su un dispositivo terminale o mobile) per mezzo di filtri di basso ordine, permettendo così una riduzione dei costi computazionali. Grazie alla ridotta complessità, il modello può essere quindi utilizzato per rendere scene con molteplici oggetti audiovisivi in una varietà di contesti quali giochi per computer, cinema, *edutainment*, e qualsiasi altro scenario in cui spazializzazione realistica del suono e riproduzione personalizzata del suono siano requisiti importanti.

Tra questi, le specifiche aree di ricerca per le quali il suddetto modello è stato pensato sono quelle della riabilitazione virtuale (*virtual rehabilitation*) e della robotica riabilitativa (*rehabilitation robotics*), potenzialmente due dei più interessanti campi di applicazione per la ricerca nel design di interazione sonora (*sonic interaction design*). Lo scopo finale della ricerca in queste due aree è quello di facilitare la reintegrazione di pazienti con disordini neurologici (causati ad esempio da ictus) nella vita sociale e domestica aiutandoli a riottenere le abilità per compiere autonomamente le *activities of daily living* (ADLs, e.g. mangiare o camminare); nonostante ciò, una grossa mole di lavoro è tuttora richiesta per fronteggiare esigenze relative a hardware, software, design di sistemi di controllo, così come per la definizione di approcci efficaci per il

trattamento. Le ADL incorporano infatti task motori complessi per i quali i sistemi riabilitativi attuali mancano della raffinatezza richiesta nell'assistenza dei pazienti durante l'esecuzione degli stessi task. In particolare, è risaputo che un grosso numero di gradi di libertà deve essere usato nella riabilitazione assistita da robot, e che il feedback multimodale spesso gioca un ruolo centrale.

Nonostante l'esistenza di una varietà di sistemi per la riabilitazione che sfruttano ambienti virtuali multimodali con feedback visivo e aptico, l'uso consistente del feedback uditivo è tuttora raro. Un'analisi accurata della letteratura conferma tale ipotesi, dimostrando come il potenziale del feedback uditivo sia largamente sottostimato in tale contesto. Cinque diversi esperimenti, descritti in questa tesi, permettono lo studio del ruolo che nuovi tipi di feedback uditivo presentati durante la camminata o durante movimenti di tracciamento giocano nel miglioramento della *performance* in soggetti sani, costituendo una base per un futuro paragone con pazienti neurologicamente deficitari. In particolare, viene qui attestata l'utilità di un feedback sonoro relativo al task e della spazializzazione del suono nel coordinamento dei movimenti dell'utente durante semplici task di inseguimento. I risultati suggeriscono quindi come un feedback multimodale costruttivo e ben progettato possa essere usato sistematicamente per migliorare performance e *learning* in task motori complessi, grazie all'elevato livello di attenzione, coinvolgimento e presenza offerto all'utente. Tali studi rappresentano una novità nella letteratura sulla riabilitazione virtuale e/o assistita da robot, soprattutto per quanto riguarda l'utilizzo di tecniche di sonificazione per convogliare informazioni in uno scenario riabilitativo.

Ringraziamenti

Ho innanzitutto l'immenso piacere di ringraziare il mio supervisore, Prof. Giovanni De Poli, per il suo supporto durante questi tre anni di studi di Dottorato. Senza le sue interessantissime lezioni di Informatica Musicale non sarei probabilmente mai arrivato ad affrontare le complesse ma stimolanti tematiche sviluppate in questa tesi.

Questo progetto di ricerca non avrebbe prodotto i risultati sperati senza l'inestimabile assistenza, il supporto, e la guida del Dott. Federico Avanzini, a cui devo la mia più profonda gratitudine. Lo ringrazio di cuore per tutto il tempo che ha dedicato ai miei dubbi e alle nostre (sempre fruttuose) sessioni di *brainstorming*, e per aver deciso quali direzioni di ricerca valesse la pena esplorare. Ma soprattutto, grazie per il notevole numero di preziosi consigli, direi quasi da fratello maggiore, che hanno segnato il mio percorso accademico fin dagli inizi della mia Tesi Specialistica, contribuendo senza alcun dubbio ad aumentare la fiducia in me stesso dal punto di vista professionale.

Sono fortunato ad avere Michele Geronazzo come collega, stretto collaboratore e amico. Lo ringrazio per i continui scambi di idee che hanno caratterizzato la seconda metà del mio Dottorato e per aver condiviso con me (molti) successi, (poche) cadute, ed epiche "talismanate". Senza dimenticare le ispirate sessioni di lavoro nel bel mezzo delle maestose Dolomiti (aspettando con ansia la prossima occasione)!

Vorrei ringraziare tutte le persone che in questi tre anni hanno fatto parte del gruppo Sound and Music Computing al Dipartimento di Ingegneria di Informazione, per aver condiviso più che una semplice esperienza lavorativa con me e per tutti gli eventi conviviali passati insieme. Grazie in particolare al Dott. Enrico Marchetto per aver decisamente alleviato le tipiche preoccupazioni di un dottorando dall'alto della sua recente esperienza.

La mia gratitudine è dovuta al Dott. Ville Pulkki per la sua supervisione durante il mio periodo all'estero presso la Aalto University, così come per tutti gli scambi di idee e per la concessione di materiali e spazi che hanno reso possibile parte del lavoro incluso in questa tesi. Più in generale vorrei citare l'intero Spatial Sound Research Group, specialmente Marko Hiipakka e Javier Gomez Bolaños, per il loro gradito aiuto nella preparazione e nell'esecuzione dei miei strani esperimenti. Grazie anche ad Archontis Politis e Simeon Delikaris-Manias per le chiacchierate e i cicchetti condivisi nei migliori pub di Helsinki.

Ci tengo a ringraziare tutte le persone che hanno collaborato in remoto e coautorato gli articoli scritti durante il mio periodo di Dottorato. Una dovuta menzione vanno a: Prof. Giulio Rosati e Fabio Oscari al Dipartimento di Innovazione Meccanica e Gestionale, Università di Padova; Prof. Sunil K. Agrawal e Dott. Damiano Zanotto alla University of Delaware; e Prof. David J. Reinkensmeyer alla University of California Irvine. Cito anche il Prof. V. Ralph Algazi per il suo supporto con il CIPIC HRTF database. Inoltre, voglio rimarcare il contributo delle persone che hanno preso parte agli esperimenti riportati in questa tesi.

Un ringraziamento speciale a Barbara, Emanuele e Francesca (in rigoroso ordine alfabetico)

per essermi sempre stati vicini, sopportando ammirevolmente la mia incostanza cronica durante gli ultimi tre anni e oltre. Siete la prova vivente del fatto che la vera amicizia non abbia una data di scadenza.

Per ultimo, ma primo per importanza, vorrei esprimere il mio affetto e la mia gratitudine alla mia famiglia. Questa tesi non avrebbe mai visto la luce senza l'amore e l'indefesso supporto che mio padre Giorgio e mia madre Antonella mi hanno sempre dimostrato, e mai dimenticherò gli sforzi spesi senza esitazione alcuna per far sì che io diventassi quello che sono oggi. E come potrei non menzionare mia sorella Aurora: grazie per i sorrisi che mi doni ogni giorno. Sei la cosa più bella che la vita mi abbia mai riservato.

Abstract

Multimodal interfaces represent a key factor for enabling an inclusive use of new technologies by everyone. To achieve this, realistic models that describe our environment are of topical importance, in particular models that accurately describe the acoustics of the environment and communication through the auditory modality. Models for spatial (or 3-D) audio can provide accurate information about the relation between the sound source and the surrounding environment, and this information cannot be substituted by any other modality. However, being multimedia systems currently focused mostly on graphics processing and integrated with simple stereo or surround sound, today's spatial representation of audio tends to be simplistic and with poor interaction potential. Furthermore, current auralization technologies rely on invasive and/or expensive reproduction devices (e.g. head-mounted displays, loudspeakers), which cause the user to perceive a non-integrated experience due to an unbridged gap between the real and virtual worlds.

On a much different level lie binaural sound rendering approaches (i.e. based on headphone reproduction). Most of the binaural rendering techniques currently exploited in research rely on the use of the so-called Head-Related Transfer Functions (HRTFs), i.e. peculiar filters that capture the transformations undergone by a sound wave in its path from the source to the eardrum and typically due to reflection and diffraction effects on the torso, head, shoulders and pinnae of the listener. Such characterization allows virtual positioning of sound sources in the surrounding space by filtering the desired signals through a pair of HRTFs, thus creating left and right ear signals to be delivered by headphones. In this way, three-dimensional sound fields with a high immersion sense can be simulated and integrated within multimodal frameworks.

However, such techniques bear relevant limitations. First, they may request considerably large computational resources, especially in the case where one needs to simulate several sound sources in the surrounding space. Second, and most important, HRTF filters are usually presented under the form of acoustic signals recorded through dummy heads: this means that anthropometric differences among different subjects are not taken into account. Contrariwise, along with the critical relative position between listener and sound source, anthropometric features of the human body have a key role in HRTF characterization: while non-individualized HRTFs represent a cheap and straightforward mean of providing 3-D perception in headphone reproduction, listening to non-individualized spatialized sounds may likely result in evident sound localization

errors such as incorrect perception of source elevation, front-back reversals, and lack of externalization, especially in static conditions. On the other hand, individual HRTF measurements on a significant number of subjects is often both time- and resource-expensive.

Structural modeling of HRTFs ultimately represents an attractive solution to these shortcomings. As a matter of fact, if one isolates the contributions of the listener's head, pinnae, ear canals, shoulders, and torso to the HRTF in different subcomponents - each accounting for some well-defined physical phenomenon - then, thanks to linearity, he can reconstruct the global HRTF from a proper combination of all the considered effects.

This thesis presents one such model that can be employed for immersive sound reproduction, with a particular focus on the pinna contribution to the HRTF. The pinna plays a primary part in the perception of source elevation by introducing major spectral modifications, yet the relation between acoustic phenomena due to the pinna - mainly resonances and sound reflections - and anthropometry has not been understood up to date. Instead, a promising correspondence between reflection points on pinna surfaces and frequencies of notches occurring in the high-frequency range of the HRTF spectrum is formally found here. Such a relevant result allows for an interesting form of content adaptation and customization of the structural model, as it includes parameters related to the user's anthropometry in addition to the spatial ones.

The proposed approach has also implications in terms of delivery, since it operates by processing a monophonic signal exclusively at the receiver side (e.g., on a terminal or mobile device) by means of low-order filters, allowing for reduced computational costs. Thanks to its low complexity, the model can be used to render scenes with multiple audiovisual objects in a number of contexts such as computer games, cinema, edutainment, and any other scenario where realistic sound spatialization and personalized sound reproduction is a major requirement.

Remarkably, the specific areas for which the proposed model is thought for are those of virtual rehabilitation and rehabilitation robotics, two of the most potentially interesting application fields for research in sonic interaction design today. The final goal of research in these areas is to facilitate re-integration of patients with neurological disorders into social and domestic life by helping them regain the ability to autonomously perform activities of daily living (ADLs, e.g., eating, or walking); however, much work is still needed to address challenges related to hardware, software, control system design, as well as effective approaches for delivering treatment. As a matter of fact, ADLs embody complex motor tasks for which current rehabilitation systems lack the sophistication needed in order to assist patients during their performance. In particular, it is recognized that a large number of degrees of freedom ought to be used in robot-assisted rehabilitation, and that multimodal feedback often plays a key role in both forementioned application fields.

Although several rehabilitation systems which make use of multimodal virtual environments with visual and haptic feedback already exist, the consistent use of auditory feedback is less investigated. A thorough analysis of literature reported in this thesis confirms this impression, showing that the potential of auditory feedback is largely underestimated in such systems. Five

different proposed experiments allow investigation of the role that novel auditory feedbacks presented during gait training and tracking movements play in improving performance in healthy participants, providing a basis for a future comparison with neurologically injured patients. In particular, usefulness of task-related sound feedback and sound spatialization in coordinating the user's movements during simple target following tasks is attested. Results thus suggest that constructive and well-designed multimodal feedback can definitely be used to improve performance and learning in complex motor tasks, thanks to the high level of attention, engagement, and presence provided to the user. Such studies represent a novelty in the current literature on virtual rehabilitation and rehabilitation robotics, especially concerning the use of sonification techniques to convey information in a rehabilitation scenario.

Acknowledgments

First of all I am pleased to acknowledge my supervisor, Prof. Giovanni De Poli, for his support during these three years of doctoral studies. Without his extremely interesting Sound and Music Computing lectures I would probably never have approached the challenging topics developed in this Thesis.

This research project would not have been possible without the invaluable assistance, support, and guidance of Dr. Federico Avanzini, to whom I owe my deepest gratitude. I heartfully thank him for all the time he has dedicated to my doubts as well as to our (always fruitful) brainstorming sessions, and for ultimately deciding what directions of research might have been worth exploring. More important, an uncountable number of precious, elder-brotherlike advices from him have marked my academic path since the beginning of my Master Thesis, undoubtedly contributing to increase my own professional self-confidence.

I am fortunate to have Michele Geronazzo as a colleague, strict collaborator, and friend. I thank him for the continuous exchange of ideas that has taken place during the second half of my Doctorate and for having shared (many) successes, (few) failures, and epic fantasy board game matches with me. Not forgetting about the inspirational work sessions up in the awe-inspiring Dolomites (looking forward to the next one)!

I would like to acknowledge all the people who have been part of the Sound and Music Computing Group at the Department of Information Engineering over the past three years, for having shared more than a work experience with me and for the convivial lunches spent together. Special thanks to Dr. Enrico Marchetto for greatly alleviating the typical worries of a PhD student from the height of his recent past experience.

Deepest gratitude is due to Dr. Ville Pulkki for his supervision during my 6-month abroad period at Aalto University, and for all the exchanges of ideas as well as the provision of material and spaces that has made part of the work included in this Thesis possible. More in general I acknowledge the whole Spatial Sound Research Group, especially Marko Hiipakka and Javier Gomez Bolaños, for their help in setting up and performing the odd measurements. Also thanks to Archontis Politis and Simeon Delikaris-Manias for the drinks and chats we had together in the coolest Helsinki pubs.

Thank you to all the people who have remotely collaborated and co-authored the papers written during my doctoral period. A worthy mention goes to Prof. Giulio Rosati and Fabio Oscari at the Department of Mechanical Innovation and Management, University of Padova; Prof. Sunil K. Agrawal and Dr. Damiano Zanotto at the University of Delaware; and Prof. David J. Reinkensmeyer at the University of California Irvine. I also wish to acknowledge Prof. V. Ralph Algazi for his support with the CIPIC HRTF database and the accompanying pictures. I want to stress the contribution from the people that took part in the experiments reported in this thesis too.

Special thanks to Barbara, Emanuele and Francesca (in strict alphabetical order) for always

having been close to me, admirably tolerating my chronic fickleness during the past three years and beyond. You are living proof that true friendship does not have an expiration date.

Last but absolutely never least, I wish to express my love and gratitude to my family. This thesis would have not been possible without the unflagging love and support that my father Giorgio and my mother Antonella have constantly demonstrated to me throughout my life, and I will never forget about the effort they have unconditionally spent to let me become who I am today. And how could I not mention my sister Aurora: thank you for always making me smile. You are the most beautiful thing life has ever gifted me with.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The ability of the human auditory system to estimate the spatial location of sound sources in an acoustic environment has high survival value as well as a relevant role in several everyday tasks: detecting potential dangers in the environment, selectively focusing attention on one stream of information, and so on. Audition performs remarkably at this task, complementing visual information: as an example, it can provide localization information of targets that are out of sight.

Accordingly, in recent years spatial sound has become increasingly important in several application domains. Spatial rendering of sound is recognized to greatly enhance the effectiveness of auditory human-computer interfaces [14], particularly in cases where the visual interface is limited in extension and/or resolution, as in mobile devices [64]; it improves the sense of presence in augmented/virtual reality systems, and adds engagement to computer games.

The sound produced by an external source placed in the space around the listener is subject to diverse transformations along its path towards the listener's ears. Among these, we can count isofrequential energy loss due to the traveled distance (i.e. *path loss*) and frequency-dependent energy loss due to air absorption and/or to the screening effects of possible objects the wave encounters that cause reflections and diffraction of sound waves. Knowing just this information, the diversity of sound waves arriving at the listener's ears already comes out clear: the two paths are different. Still, another factor that prominently comes into play is the listener itself. As a matter of fact, sound waves are influenced by the active role of the listener's body in transforming them too, emphasizing the difference between the inputs to the two auditory channels.

It is indeed thanks to such difference that the listener can collect informations on the spatial location of the sound source, in such a way that he/she can discriminate its angular direction and distance with respect to his/her head within a certain error threshold. Auditory cues related to directional information include both binaural cues, such as interaural level and time differences, and monaural cues, such as the spectral coloration resulting from filtering effects of the human body, especially the external ear. All these features are summarized into the so-called *Head Related Transfer Functions (HRTFs)* [30], i.e. the frequency- and space-dependent acoustic transfer functions between the sound source and the eardrum. Binaural spatial sound can be

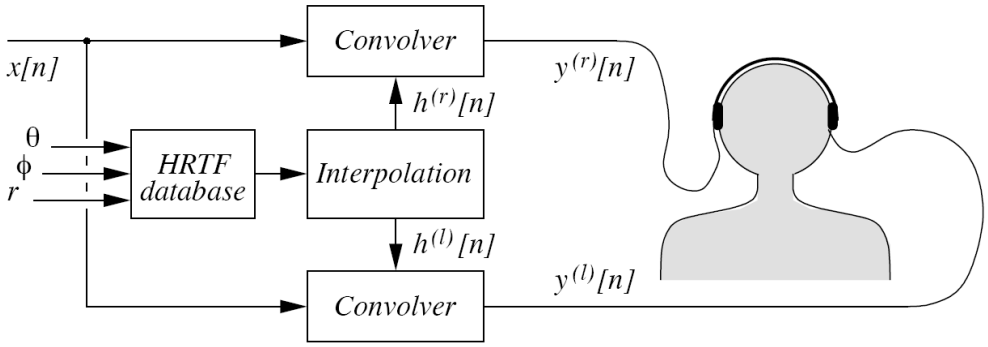


Figure 1.1: *Simplified* 3*-D audio reproduction system based on headphones and HRTFs.*

synthesized by convolving an anechoic sound signal with the corresponding left and right HRTFs.

Techniques for sound source spatialization follow different approaches [83]. Stereo is the simplest system involving spatial sound, but a correct spatial image can only be rendered along the central line separating the loudspeakers (the "sweet spot"). Surround systems based on multichannel reproduction, such as 5.1 or 10.2 systems, or Ambisonics [60], also suffer from similar "crosstalk" problems (i.e., the sound emitted by one loudspeaker is always heard by both ears). Crosstalk cancellation techniques commonly employed are effective only in a very limited listening region.

Wave-Field Synthesis is a currently active line of research. This method, initially proposed in [16], uses arrays of small and individually driven loudspeakers to reproduce a faithful replica of a desired spatial sound field. As a result, the spatial image is correct in the whole half-space at the receiver side of the array. Research in this direction is progressing rapidly, however wave-field methods require expensive and cumbersome reproduction hardware, which makes them suitable only for specific application scenarios (e.g., digital cinema [55]).

The most important distinction among spatialization techniques concerns the sound reproduction method, i.e. the use of loudspeakers opposed to headphone-based systems. Since binaural reproduction can be obtained either with loudspeakers or headphones [54], binaural techniques lie between the two groups and enable authentic auditory experiences if the eardrums are stimulated by sound signals bearing roughly the same pressure as in real life conditions [19]. Nevertheless, the use of headphone-based reproduction – onto which this thesis focuses on – in conjunction with head tracking devices grants a degree of interactivity, realism, and immersion that is not easily achievable with loudspeaker-based binaural systems, due to limitations in the user workspace and to acoustic effects of the real listening space.

The rendering scheme of Fig. 1.1 assumes the availability of a database of measured HRTFs. Acoustic measurement of individual HRTFs for a single subject is an expensive and cumbersome procedure, which has to be conducted in an anechoic chamber, using in-ear microphones, specialized hardware, and so on. Therefore individual HRTFs cannot be used in most real-word applications. Alternatively, generalized HRTFs are typically measured on so-called "dummy

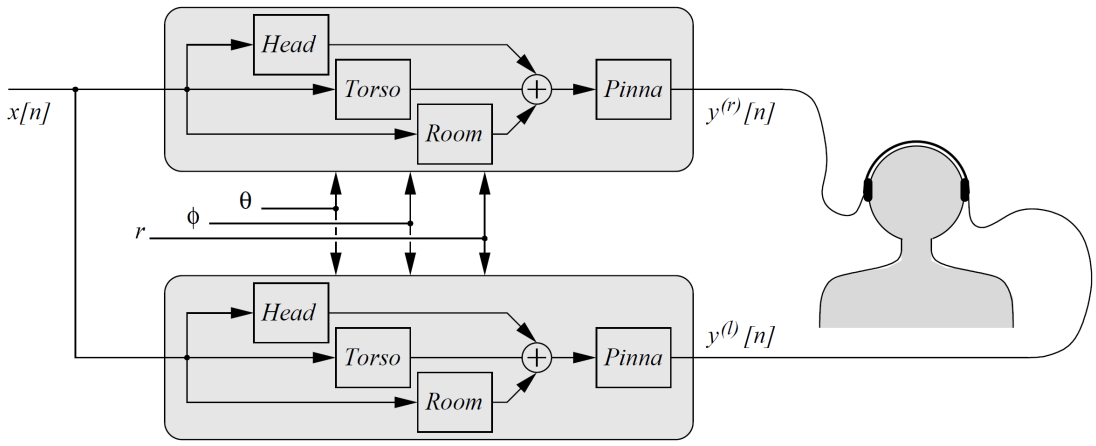


Figure 1.2: *Generalized 3-D audio reproduction system based on a structural HRTF model.*

heads", i.e. mannequins constructed from averaged anthropometric measures, representing standardized heads with average pinnae and torso. However, this limits to some extent the realism of the rendering: in fact one dummy head might sound more natural to a particular set of users than another, depending on anthropometric measures and also on technicalities in the measurement procedure.

A second problem is that HRTF measurements can only be made at a finite set of locations, and when a sound source at an intermediate location must be rendered, the HRTF must be interpolated. If interpolation is not applied (e.g., if a nearest neighbour approach is used) audible artifacts like clicks and noise are generated in the sound spectrum when the source position changes. Clearly this problem becomes even more severe in interactive settings, where both the listener and the sound sources are moving in the environment and the rendering must be dynamically updated.

In such context innovative techniques for binaural sound rendering are being developed. These are based on structural modeling of HRTFs [20], an extremely attractive and revolutionary approach based on the physical description of the most important effects involved in spatial sound perception (such as acoustic delays and shadowing due to head diffraction, reflections on pinna contours and shoulders, resonances inside pinna cavities and the ear canals, etc.) and, above all, a solution to the forementioned shortcomings. The main advantage of this approach is that the rendering algorithms' control parameters are definable as a function of the main anthropometric features; it is thus possible to adapt the designed spatial rendering algorithms to a specific subject, just by knowing some of his/her anthropometric quantities such as head radius, pinna shape, shoulder width, and so on. As Fig. 1.2 depicts, such an approach also bears advantages from the computational point of view because models are substructured in smaller blocks, each one simulating a single physical effect. This leads in particular to a greater efficiency in the case where one needs to simulate complex, multisource acoustic environments. A review of the

literature in the field of structural modeling of HRTFs, along with a roundup of basic concepts and findings on sound source localization, is provided in Chapter 4.

Following the structural modeling approach, this thesis primarily investigates the contribution of the head and of the external ear to the HRTF. The head will be assumed to be spherical and its contribution in the near field (i.e. the sound source lies within 2 m from the center of the head), summarized in the Near-Field Transfer Function (NFTF), studied by means of an analytical representation [42] and modeled through a low-order filter structure in Chapter 5, the results being objectively verified through spectral distortion measurements.

Concerning the external ear, its contribution to the HRTF, known as Pinna-Related Transfer Function (PRTF), will be extensively studied in the following chapters. While the pinna is known to play a primary role in the perception of source elevation, the relation between PRTF features – resonances associated to cavities and spectral notches resulting from reflections [12] – and anthropometry is not fully understood. Recent related works [85, 81, 115] adopt a physical modeling approach in which PRTFs are studied through computationally intensive simulation techniques, such as finite-difference time-domain (FDTD) methods, or boundary elements methods (BEM). Alternatively, the relationship between PRTF features and pinna geometry can be studied by directly analyzing real measured HRTFs, and by relating relevant extracted spectral features to known anthropometric data [137, 149].

In this thesis I will follow this latter approach. In Chapter 6 a number of experimental PRTFs will be estimated, both through direct measurement and through derivation from available HRTFs, and analyzed with the help of an algorithm that separates the resonant and reflective parts of the PRTF spectrum. In Chapter 7, focus will then be put on the relationship between PRTF notches and pinna contours: ray-tracing analysis performed on the notches' central frequencies extracted in the previous analysis step will be compared with a set of possible reflection surfaces directly recognizeable from the corresponding subject's pinna picture. Results of such an analysis will be discussed in terms of the reflection coefficient sign.

Based on these findings, a structural model for real-time HRTF synthesis which allows to control separately the evolution of different acoustic phenomena such as head diffraction, ear resonances, and reflections will be proposed in Chapter 8 through the design of distinct filter blocks. Parameters to be fed to the model will be derived either from analysis or from anthropometric features of the specific subject. Finally, objective evaluations of reconstructed HRTFs in the chosen spatial range will be performed through spectral distortion measurements. The results of this work are the first step towards the development of a fully parametric structural HRTF model that can be customized according to individual anthropometric data, which in turn can be automatically estimated through straightforward image analysis.

The application area of the presented work on spatial audio is that of technology-assisted motor rehabilitation. The integration of auditory feedback in rehabilitation devices is a topic that is rarely investigated in the literature of technology-assisted motor rehabilitation and that will be discussed in the first two chapters of this thesis in order to highlight how and why an effective

model for spatial sound rendering is required.

In Chapter 2 current trends, open issues and up-to-date uses of auditory feedback in such context will be reviewed, along with a discussion on a number of scenarios in which the use of auditory feedback can contribute to overcome some of the main current limitations of rehabilitation systems, in terms of user engagement, improved motor learning, acute phase rehabilitation, standardization of the rehabilitation process, and development of home rehabilitation devices. Then, in Chapter 3 the results of a set of novel preliminary experiments in which continuous auditory feedback is used to augment motor training exercises for upper limb and gait training in healthy subjects will be presented. In particular, the benefits that task-related spatial audio feedback can offer the user will be highlighted.

Chapter 2

Auditory Feedback in Virtual Rehabilitation and Rehabilitation Robotics

The final goal of the rehabilitation process is to facilitate re-integration of patients into social and domestic life, by helping them regain the ability to autonomously perform activities of daily living (ADLs, e.g., eating, or walking). However such activities embody complex motor tasks, for which current rehabilitation systems lack the sophistication needed in order to assist patients during their performance.

One of the domains being explored to enhance patient recovery during rehabilitation is to employ technological means for the rehabilitation treatment, mainly robotic systems (*rehabilitation robotics*) and virtual reality systems (*virtual rehabilitation*) [66]. The main difference between the two approaches is that robotic systems can assist the patient in completing the motor task, while a virtual reality system can provide the patient with augmented feedback only. From this point of view, virtual reality systems are more likely to be employed in the chronic phase (i.e. after approximately three months from the trauma), while robot-mediated rehabilitation can be delivered in the acute and sub-acute phases (typically less than three months from the trauma) to severely impaired subjects as well. However, in most cases the robot or haptic interface is not used in isolation and requires at least a computer interface and possibly also a virtual environment to deliver the rehabilitation therapy. This is why the two approaches are currently converging into one common direction, so that we can refer to both of them with the concept of *technology-assisted motor rehabilitation*.

Technology-assisted motor rehabilitation is today one of the most potentially interesting application areas for research in sonic interaction design (SID). The strong social implications, the novelty of such a rapidly advancing field, as well as its inherently interdisciplinary nature (contents combine topics in robotics, virtual reality, haptics, as well as neuroscience and rehabilitation) are some of the aspects that consolidate its challenging and captivating character. Such prospects justify the considerable amount of attention it has received in the last decade from researchers in the fields of both medicine and engineering, the purpose of their joint effort being

the development of innovative methods to treat motor disabilities occurring as a consequence of several possible traumatic (physical or neurological) injuries, e.g. stroke (discussed in Section 2.1).

Still, much work is needed to address challenges related to hardware, software, control system design, as well as effective approaches for delivering treatment [66]. In particular, although it is understood - as Section 2.2 will point out - that multimodal feedback can be used to improve the performance in complex motor tasks [37], a thorough analysis of the literature in this field, reported in Section 2.3, shows that the potential of auditory feedback is largely underestimated.

The work presented in this Chapter was published in [9] and has been accepted for publication in [11].

2.1 Post-stroke rehabilitation robotics

Hemiparesis/hemiplegia is the most common outcome of stroke, the third leading cause of death after cardiovascular diseases and cancer and the greatest cause of severe disability and impairment in the industrialized world. Every year in the U.S. and Europe there are 200 to 300 new stroke cases per 100.000 people, the 30% of whom survive with severe invalidity and marked limitations in daily activities, mainly deriving from impaired motor control and loss of dexterity in the use of the arm [95, 89]. The main characteristics observed in hemiparetic patients are weakness of specific muscles, abnormal muscle tone, abnormal postural adjustments, lack of mobility, incorrect timing of components within a pattern, abnormal movement synergies, loss of interjoint coordination, and loss of sensation [31].

The group most prone to cerebrovascular accidents or stroke (whose relative incidence doubles every decade) is the over 55 years old category [56]. According to the World Health Organization (WHO), by 2050 the proportion of people over 65 years old will have increased by more than 70% in industrialized countries and by more than 200% worldwide: as a consequence, stroke cases are going to increase further in the next decades [89]. Motor training after stroke is thus becoming a primary social goal, based on the increasing evidence that the motor system is plastic following stroke and can be influenced by motor training [125].

The fundamental rehabilitation goal in hemiplegic subjects is to promote recovery of lost functions so as to allow independence and early reintegration into social and domestic life. Traditional treatments are based on the use of physiotherapy which is heavily reliant on the therapist's training and past experience. Robotic therapy proposes itself to be a novel and realistic approach that can help therapists increase the intensity of treatments while operating safely within the human workspace and with the prospective of reducing costs during their work. In other words, robotic devices have the potential to help post-stroke automatic repetitive training in a controlled fashion. Mechanical devices for rehabilitation are, in fact, designed to interact with the patient, guiding his/her upper or lower limb through repetitive exercises based on a stereotyped pattern,

and providing force feedback for sensorimotor-type rehabilitative training [101].

The most commonly explored paradigm is to use a robotic device to physically assist the patient in completing desired motions of the arms, hands, or legs as he/she plays computer games presented on a screen. A variety of assistive control strategies have been designed (see review [99]), ranging from robots that rigidly move limbs along fixed paths, to robots that assist only if the patient's performance fails to remain within some spatial or temporal bound, to soft robots that form a model of the patient's weakness.

The rationale for physically assisted movement is that it provides novel sensory and soft tissue stimulation, demonstrates how better to perform a movement, and increases the motivation of the patient in therapy engagement [99]. However, an unintended and possibly negative effect of providing assistance is that patients may reduce their effort and participation during training, both for arm [187] and gait [76] rehabilitation. Such reduction has been hypothesized to explain the diminished benefits of robot-assisted gait training compared to conventional gait training, recently documented for chronic stroke patients who were ambulatory at the start of the robotic training [70]. In the extreme, if a patient is passive as a robot moves his/her limbs, the effectiveness of repetitive movement training is substantially reduced [73]. Yet, even a moderate reduction in patient effort may diminish training effectiveness, in case that the magnitude of the experienced efferent activity plays a role in provoking neural plasticity mechanisms.

For these reasons, one of the main goals of research in this area is to identify the mechanisms that determine engagement of the patient during robotic arm movement training after stroke, in order to optimize the design of rehabilitation robotic systems. A reasonable working hyphothesis is that patient engagement and effort are related to (and can be modulated by) sensory information delivered by the robotic system, and that more highly engaged patients are able to experience increased benefits from robot-assisted training [73].

In the past two decades there has been a rapid increase in the number of research groups and companies that develop robotic devices for assisting motor rehabilitation in people with disabilities, but recent reviews on the first Randomized Controlled Trials (RCTs) of upper-limb robot-assisted rehabilitation outlined that clinical results are still far from being fully satisfactory [94, 108, 131]. Indeed, even though motor recovery is usually greater in robot-assisted groups than in control groups, only few studies on acute and sub-acute phase rehabilitation showed some positive results at the functional level (i.e., in ADLs) [100], the summary effect size of all the studies being very close to zero. These results suggest that the therapy devices, exercises and protocols developed so far still need to be improved and optimized, one further issue being the development of home rehabilitation systems in order to help patients continue treatment after hospital discharge [142].

Still, the most fundamental problem that robotic movement therapy must address in order to progress is the lack of knowledge on how motor learning during neuro-rehabilitation works [139]. Specifically, many experimental results suggest that, after local damage to the motor cortex, rehabilitative training with active engagement of the participant can shape subsequent reorganization

in the adjacent intact cortex, and that the undamaged motor cortex may play an important role in motor recovery [126]. There is also evidence that kinematic error drives motor adaptation [175] and, moreover, that humans adapt to robot-generated dynamic environments in a way that appears to minimize a cost function in both error and effort terms [45].

Besides, it is still not sure how the central nervous system combines different kinds of simultaneous feedbacks such as proprioceptive and visual information or haptic feedback. It is known that visual and proprioceptive feedback may be combined in fundamentally different ways during trajectory control and final position regulation of reaching movements [153] and that when estimating the position of the arm, the brain selects different combinations of sensory input based on the computation in which the resulting estimate will be used [162]. Moreover, people tend to make straight and smooth hand movements when reaching an object [50], these trajectory features being resistant to perturbation, and proprioceptive as well as visual feedback may guide the adaptive updating of motor commands enforcing this regularity. Morris *et al.* [120] found that recall following visuohaptic training is significantly more accurate than recall following visual or haptic training alone, although haptic training is inferior to visual training. This result suggests that haptic training may be an effective tool for teaching sensorimotor skills that include a force-sensitive component in conjunction with visual feedback.

There is also evidence that the effect of auditory feedback in reaching tasks after chronic stroke depends on the hemisphere which was damaged by the stroke [140], and that a proper sound may help individuals in learning a motor task [173, 135, 155], but the precise ways that mental engagement, repetition, kinematic error and sensory information in general translate into a pattern of recovery is not well defined for rehabilitation [139]. Audio is used in many rehabilitation systems with the purpose of motivating patients in their performance; however, in all these systems the audio component plays mostly a marginal role, giving a positive (negative) feedback if the patient completes (does not complete) a task or reinforcing the realism of a virtual reality environment.

2.2 Audio and rehabilitation

2.2.1 Motivations and open issues

Strong motivations for integrating interactive sound into motor rehabilitation systems can be found by examining in some detail the most prominent current research challenges in the field of rehabilitation robotics, as described by Harwin *et al.* [66] in a recent study.

As already mentioned, the most important challenges are related to recovery of ADLs, in order to facilitate re-integration of patients into social and domestic life. The functional movements associated to ADLs typically involve very complex motor tasks and a large number of degrees of freedom of the implied limbs (e.g., ankle, arm, hand). On one hand, this requires the use of sophisticated sensors and actuators (in particular, multiple degrees-of-freedom robots

have to be used in the case of robot-assisted therapy). On the other hand, representing such complex motor tasks to the patient is a particularly challenging goal. The simple schematic exercises implemented in current rehabilitation systems help recovery of ADLs only to a limited extent.

ADLs rely on an essentially continuous and multimodal interaction with the world, which involve visual, kinesthetic, haptic, and auditory cues. Such cues integrate and complement each other in providing information about the environment and the interaction itself, both in complex tasks (e.g., walking, riding a bicycle) and in relatively simpler ones (e.g., a reach and grasp movement). To this regard, in order to effectively represent the environment and/or the user's movements, auditory feedback has to be used in conjunction with other modalities.

Engagement of the patient to the rehabilitation task is another fundamental aspect to consider. It is common sense that a bored patient may not be as motivated as an engaged patient. In the literature it is widely accepted that highly repetitive movement training in which the participant is actively engaged can result in a quicker motor recovery and in better re-organization [33]. Therefore an open research challenge is how to increase engagement and motivation in motor rehabilitation.

Several studies have shown that auditory feedback purposedly designed to be related with physical movement can result in attainment of optimal arousal during physical activity, reduction of the perceived physical effort, and improvement of mood during training [84]. Moreover, engagement is strictly related to the concept of presence, i.e. the perception of realism and immersion in a virtual environment, commonly used in virtual reality research. In this respect, it is known that faithful spatial sound rendering increases the realism of a virtual environment, even in a task-oriented context [69, 136]. Nonetheless, it must also be emphasized that auditory feedback can also be detrimental to patient engagement, if not properly designed. This is a general issue in sound design: users will typically turn audio off in their PC interface if the auditory feedback is monotonous or uninteresting/uninformative (e.g. sound objects unrelated to what happens on the virtual scene).

The use of interactive sound in rehabilitation systems is also motivated by technological challenges. The qualities of virtual reality and robotic systems in motor rehabilitation are counterbalanced by their disadvantages in terms of customizability and high costs, and designing low-cost devices and hardware-independent virtual environments for home rehabilitation systems is indeed one of the current challenges for technology-assisted rehabilitation.

In this context the auditory modality can be advantageous over the visual and haptic ones, in terms of hardware requirements and computational burdens. High quality sound rendering is comparatively less computationally demanding than 3D video rendering or haptic rendering, and can be conveyed to the patients through headphones or through a commercial home theater system, with no need for dedicated, expensive, and cumbersome equipment. In the context of home rehabilitation, auditory feedback may even be used as a sensory substitute for the visual and haptic modalities [10, 109].

Finally, auditory feedback may be in certain cases the only modality accessible to the patient,

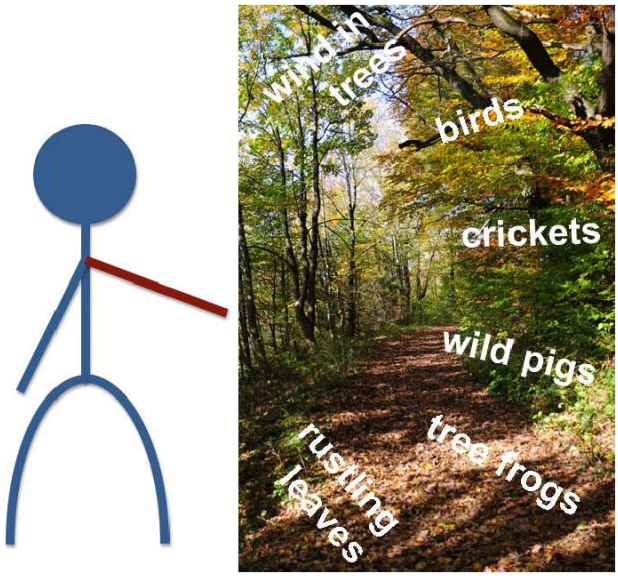


Figure 2.1: *PhysioSonic training scenario: forestal wildlife as a metaphor to absolute height of the arm (figure reproduced from [179]).*

whereas other modalities (especially the visual one) are not. A notable example is found in post-stroke neurorehabilitation treatment: in this case, it is demonstrated by many studies [143, 38, 39] that it is essential to start the rehabilitation process as soon as possible, since training during the acute and sub-acute phase has a greater impact in improving recovery of ADLs compared to robotic therapy in the chronic phase. However, since the acute phase patient has extremely limited motor and attentional capabilities, and in some cases a limited state of consciousness, this is not always possible. Auditory feedback may be successfully used in such situations, since it can still be perceived without requiring patients to keep their attention focused on a screen, and can be processed with relatively little cognitive effort.

2.2.2 Auditory feedback for continuous interaction

A few scenarios, methods and technologies from recent research on sound modeling and on sonic interaction design, which can be employed and applied in the context of motor rehabilitation tasks, are now examined.

In order to realize a fully interactive auditory feedback, suitable synthesis models which allow continuous control of audio rendering in relation to user gestures need to be used. One interesting scenario is provided by the PhysioSonic [179] system, that presents a generic model for movement sonification as auditory feedback, in which target movement patterns produce motivating sounds while negatively defined sounds are triggered by evasive movement patterns. Sonification is applied to intuitive attributes of bodily movements, and comes in the form of metaphorical sounds (e.g., the sound of a spinning wheel is associated to velocity). Furthermore,

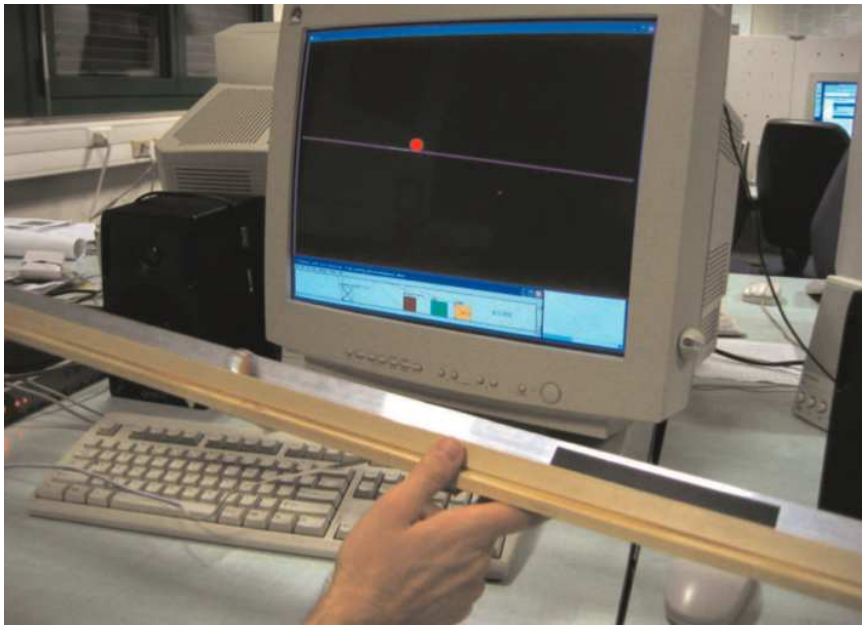


Figure 2.2: *The Ballancer: balancing a virtual glass marble on an aluminium track (figure reproduced from [135]).*

sounds can be chosen by each subject and change over time, thus reducing fatigue or annoyance. In an example the authors implement a system for the treatment of shoulder injuries, providing two different training scenarios for the abduction movement where the arm elevation and velocity are sonified into environmental sounds (see Fig. 2.1) and reproduction rate of a sound file, respectively. In both cases, all evasive movements add noise or creaking to the auditory feedback proportionally to their displacement.

One second relevant example of continuous and interactive auditory feedback related to user gesture is described in [135]: the Ballancer is a simple tangible interface composed by a track (approximately a 1-meter long) and an accelerometer that measures acceleration in the direction of the track's length thus allowing estimation of the track's tilt angle. The movement of a virtual ball, which rolls on the track and stops or bounces back when it reaches the extremities, is rendered in real-time both visually on a monitor (see Fig. 2.2) and sonically through a physically-based sound synthesis model which uses the state of the ball and the tilt angle as input controls. The task of the user of this interface is to balance and tilt the track in order to move the virtual ball to a target position on the track, and to stop it there. The experimental tests presented in [135] demonstrate that the presence of continuous auditory feedback (the rolling sound of the virtual ball) shortens the completion time for this task with respect to the case where no sound is provided. Therefore the auditory feedback effectively conveys information about such a complex gesture as tilting and balancing. Although the Ballancer is not conceived as an interface for motor rehabilitation, it highlights the potential of continuous auditory feedback in supporting

motor learning in complex tasks such as in ADLs.

Despite the abundance of literature on the use of Human-Computer Interaction (HCI) methods in the design and evaluation of input devices and interfaces [98], sound started to play a significant role in HCI research only in recent years, and yet few studies [146] were devoted to the application of HCI methods to the design and the evaluation of "new interfaces for musical expression". Orio *et al.* [128] started the investigation in this direction in 2001, focusing on the evaluation of controllers for interactive systems. The authors mention a target acquisition task that could be compared with the acquisition of a given pitch as well as a given loudness or timbre [178], proposing interesting analogies with HCI studies.

These works inspired a thread of research in the field of sound and music computing, focused on the analysis of simple HCI tasks (e.g. target acquisition) in the auditory domain [130]. Here the aim is not the comparison of different input devices but rather the evaluation of the influence of different kinds of feedback on the user's performance. As an example, de Götzen and Rocchesso [36] performed various tests to evaluate pointing/tuning tasks with multimodal feedback: their results suggest that when interaction involves any sort of kinesthetic feedback the performance is distinctly better with respect to free gesture interfaces, and that these improvements in performance are especially significant with high speeds of the target, i.e. when the target should be more difficult to hit. Furthermore, redundant feedback is needed when the task is difficult. These results support the idea of applying predictive HCI laws, along with multimodal feedback, in the field of technology-assisted motor rehabilitation, with the purpose of improving the patient's performance during rehabilitation tasks.

Recent research on novel musical interfaces provides a number of systems and approaches that could be directly applied to rehabilitative applications [138, 112]. Work in mobile sensor performance technology is particularly interesting in this respect. Small sensors (including microphones, accelerometers, and so on) are already being used to detect various kind of movements and gestures that can affect the produced auditory feedback, e.g. by changing the tempo of a musical accompaniment, or by controlling some expressive effects added according to input gesture [27, 26], thus transforming rehabilitation into a more engaging activity.

The application of all these results to the design of auditory feedback for motor rehabilitation systems must take into account the specificities of people involved, which can often be affected by various perceptual deficits. In particular, extensive experimental work is needed in order to assess the influence of auditory feedback on motor learning processes, to understand the effect of the combination of auditory feedback with other modalities, such as the visual and haptic ones, and to define criteria and guidelines for the design of the feedback, depending on the required motor task.

2.3 Current uses of auditory feedback in rehabilitation systems

In recent years auditory feedback has been exploited in various systems, both in the fields of rehabilitation robotics and virtual rehabilitation. The simplest possible use, which can be found in many systems discussed in the literature, consists in employing non-processed, pre-recorded samples of vocal or environmental sounds in order to improve the involvement of the patient in the task. As an example, Cameirao *et al.* [25] developed a neurorehabilitation system composed by a vision-based motion capture device augmented with gaming technologies: in this case audio has a rewarding function, in particular a "positive sound" is triggered whenever the patient accomplishes the goal of a specific game. In a similar way, speech and nonverbal sound is used by Louriero *et al.* [97] as a feedback modality, with the role of providing encouraging words and sounds during task execution, and congratulatory or consolatory words at the end of the exercise. Despite its simplicity, such use of sound has positive effects on patients' emotions and involvement.

A complementary approach to the use of auditory feedback is to actively guide the execution of a motor task, rather than simply triggering it as a response to the patient's performance. As an example, Masiero *et al.* [102] present a robotic device which includes simple auditory feedback: a sound signal is delivered to the patient and its intensity is increased at the start and the end phase of the rehabilitation exercise, in order to signal the patient the occurrence of these phases. According to the authors, this kind of feedback retains the power of maintaining a high level of attention in the patient. On the other hand, the feedback has no correlation with quality of performance. Colombo *et al.* [32] used a similar type of feedback to guide the user's movement in wrist and elbow-shoulder manipulators.

A more interactive use of sound can be found in the GenVirtual application [35]. This augmented reality musical game is designed as an aiding tool for patients with learning disabilities. Users of this system are instructed to imitate sound or color sequences in the GenVirtual environment, and auditory feedback is provided to help users memorizing such sequences. A similar approach can be found in [92], [91], and [61]. However, it has to be noted that no realistic interaction is provided between user and environment, even though sounds are more correlated to user movements with respect to the former examples. Moreover, auditory feedback is still realized in the form of triggered pre-recorded sounds.

In many systems, auditory feedback is intended as generation of soundscapes that can reinforce the verisimilitude and realism of a virtual environment, thus addressing aspects of sound design that are closer to SID research topics and in particular to aesthetic quality issues. To date, a plethora of environments have been developed, ranging from relatively simple driving scenarios (such as car, boat or airplane [79, 37]), to more complex ADLs [122, 71]. The latter work describes a system which allows patients to practice various everyday activities, such as

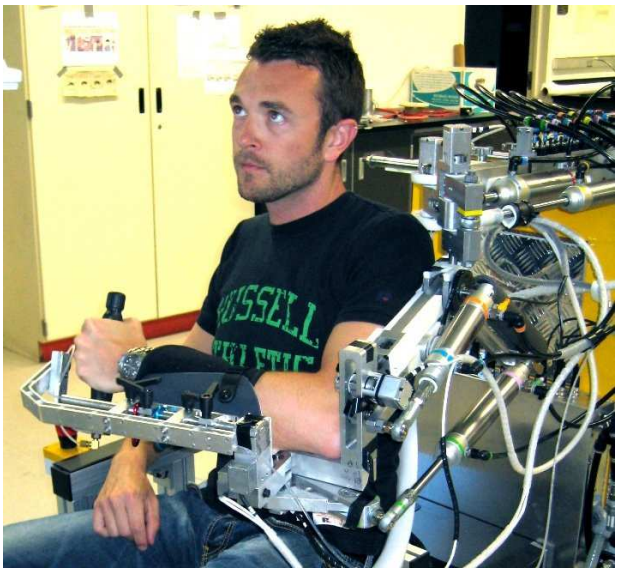


Figure 2.3: *Auditory feedback and engagement: experimental setup with the Pnew-WREX robotic system (figure reproduced from [155]).*

preparing a hot drink: here the role of auditory feedback is to render as realistically as possible the sounds of the virtual objects involved in the activity and manipulated by the patient (e.g. the kettle, the cup, and so on). However, a fully realistic sonic interaction is not achieved, because of the unidirectional and noncontinuous nature of the relation between user movements and sound generation.

Despite the great variety of uses assigned to auditory feedback, the studies discussed above do not generally provide a quantitative assessment of the effectiveness of sound, regarding patient's performance in the rehabilitation task. Schaufelberger *et al.* [152] are among the few authors who have provided such an assessment, although using healthy subjects. In their work, the use of short tonal sequences is experimentally evaluated in the context of an obstacle scenario. In particular, different distances from the obstacle and different obstacle heights are sonically rendered using different repetition rates and different pitches of a tonal sequence, respectively. Experimental results provide quantitative indications that, when auditory feedback is added to the visual one, subjects perform better both in terms of completion time and in terms of fewer obstacle hits.

Concerning upper limb movements in stroke patients, Maulucci *et al.* [104] used auditory feedback which informed subjects of the deviation of their hand from the ideal path in order to verify if auditory feedback could increase motor learning, and found that such training improved performance. In [155] it was studied how a relatively mild visual distractor affects performance errors during a common robot-assisted exercise (target tracking, see Fig. 2.3) with both healthy subjects and patients with chronic stroke. It was fond that in both cases performance was degraded by the distractor yet restored to near normal values when auditory feedback of the tracking error was provided.

Reference	Device	EC	AI	SO	SS	SP
Boian *et al.* [18]	Rutgers Ankle		X			
Colombo *et al.* [32]	Wrist Rehab. Device	X				
Colombo *et al.* [32]	Shoulder and Elbow Device	X				
Connor *et al.* [34]	AFF Joystick	X				
Johnson *et al.* [79]	Driver's SEAT		X			
Johnson *et al.* [80]	HapticMaster robot	X				
Kousidou *et al.* [90]	Salford Rehab. Exoskeleton			X		
Krebs and Hogan [91]	MIT-MANUS	X				
Loureiro *et al.* [97]	GENTLE/s				X	
McLaughlin *et al.* [107]	Phantom				X	
Nef *et al.* [122]	ARMin	X	X			
Reinkensmeyer *et al.* [139]	Pneu-WREX	X	X			
Reinkensmeyer *et al.* [139]	T-WREX	X	X			
Rosati *et al.* [143]	NeReBot	X				
Shing *et al.* [160]	Rutgers Master II	X				X
Wellner *et al.* [182]	Lokomat	X	X			X

Table 2.1: *List of the surveyed robotic/haptic devices that use some form of auditory feedback, and related literature. Columns from 3 to 6 list the typologies of auditory feedback that have been used, according to the classification given in Section 2.3.1, whereas column 7 indicates whether sound spatialization is used.*

It has to be noted that, despite the substantial amount of research, there are very few cases in which technology-assisted rehabilitation systems have made the step from research prototypes to a real-world application in a medical context. A relevant example is Vibroacoustic Sound Therapy (VAST), initially conceived for children with profound and multiple learning difficulties and recently developed with frail and mentally infirm elderly people in the context of an interactive multisensory environment (iMUSE) [44].

2.3.1 Comparative analysis of auditory displays in rehabilitation systems

To conclude this section, a quantitative analysis of current uses of auditory feedback in technology-assisted rehabilitation systems is provided. A detailed review of a large number of systems has been carried out. Specifically, the systems taken into account for this analysis have been collected based on the works referenced in two recent review articles [170, 72], on a related journal special issue [82], and on the 2006-2008 proceedings of two relevant international conferences: the ICORR (International Conference on Rehabilitation Robotics), and the ICVR (the International Conference on Virtual Rehabilitation). A total of 36 systems, described in 47 papers, have

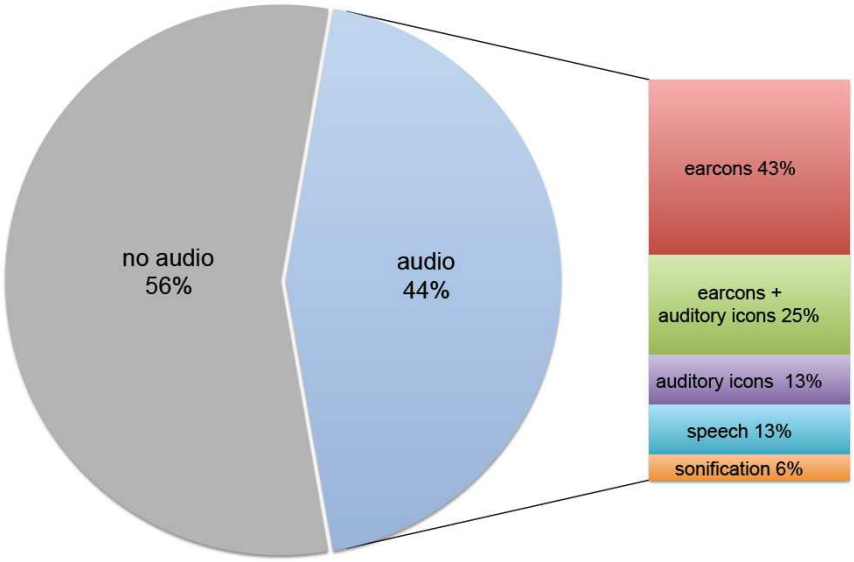


Figure 2.4: *Distribution of auditory feedback in the 36 reviewed rehabilitation systems.*

been selected. These systems have been grouped into four different clusters, representing four different macro-categories of auditory feedback:

- *auditory icons* (AI), pre-recorded everyday sounds (including environmental sounds often employed in virtual reality applications) mapped to computer events;
- *earcons* (EC), short pieces of music that characterize individual events;
- *sonification* (SO), the mapping of multidimensional datasets into an acoustic domain;
- *synthetic speech* (SS).

These categories correspond to those identified by McGookin and Brewster [106].

The quantitative results can be viewed in Fig. 2.4, whereas a more detailed report of the surveyed systems that use any kind of audio is given in Table 2.1. Such analysis pitilessly reveals that most of the systems do not make any use of auditory feedback. In addition, speech and sonification, despite being the two most attractive alternatives for SID, are used only in a small number of cases. Sound spatialization is also very little used, as the last column of Table 2.1 clearly shows. On the other hand, the vast majority of the systems that employ sound use it in the simplest possibly way, i.e. through pre-recorded samples triggered by a single event. As a result it emerges clearly that, although several systems exist which make use of multimodal virtual environments, the consistent use of auditory feedback is very little investigated, and its potential largely underestimated.

2.4 Conclusions

Very little attention to auditory feedback is paid in the robotic rehabilitation community. The majority of the reviewed systems do not utilize any auditory feedback, whereas the others exploit only a limited set of possibilities, such as earcons or auditory icons. Auditory feedback is mostly implemented in a virtual reality context, to reproduce realistic environmental sounds with the aim of increasing the user's sense of presence. Only in very few cases it is utilized to support the motor learning process, providing an augmented feedback to the user.

Although current technology-assisted rehabilitation systems exploit only a limited set of possibilities from SID research, several studies show that properly designed auditory feedback, able to provide temporal and spatial information, can improve engagement and performance of patients in the execution of motor tasks, can improve the motor learning process, and can possibly substitute other feedback modalities (as with visually impaired users). Moreover, the relatively limited computational requirements of audio rendering and the low costs of related hardware make it attractive to use auditory feedback in the context of home rehabilitation systems. In light of this, there is strong evidence that research in technology-assisted rehabilitation may only take advantage from a wary use of the know-how in sonic interaction design.

Chapter 3

Effects of Audio in Virtual Rehabilitation Tasks: Experimental Results

This chapter presents experimental results from a set of tests with healthy subjects that use auditory feedback to augment motor training exercises. The goal of the presented study is to investigate the role of sound in motor learning and motor control as a potential novel sensory information to be compared to both visual and proprioceptive modalities. The final aim of this work is to incorporate an optimized real-time auditory feedback related to one or more variables (e.g. position error or velocity) in augmented-feedback robotic or virtual rehabilitation systems, in order to improve clinical outcomes of therapy. In this context, the term auditory feedback denotes an audio signal automatically generated and played back to the user in response to an action or an internal state of the system. The design of auditory feedback requires a set of sensors to capture the system state, a feedback function to map sensor signals into acoustic parameters, and a rendering engine to generate audio accordingly [151].

An incentive to the following research is given by the observation that audio, just like video, is more direct and requires less attention than proprioception as input modality [156]. Thus, auditory feedback is not only potentially relevant as a stimulation in augmenting the patient's engagement and motivation but also as an additional or substitutive straightforward information with respect to video in improving performance and learning. The working hypothesis is then that properly designed auditory feedback could be used to:

1. aid patient motivation in performing task-oriented motor exercises;

2. represent temporal and spatial information that improves the motor learning process;

3. substitute other feedback modalities in case of their absence.

Five different experiments are presented, one for gait training and four for upper limb training, with the aim of investigating the effect of novel different auditory feedbacks in improving the

performance in non-disabled participants. Very preliminary results for one subject are presented for the gait training exercise in Section 3.1, where the effect of task-related and subject-related auditory feedbacks with or without the support of visual feedback and force field constraint is compared to the no-sound condition.

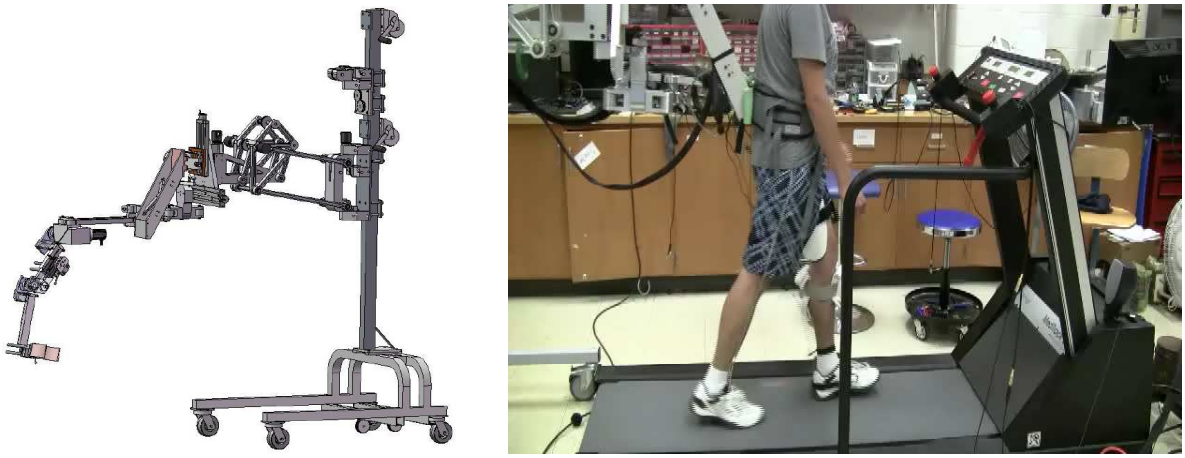
Concerning upper limb rehabilitation, the first experiment (Section 3.2) investigates whether continuous task-related auditory feedback can be more efficacious than error-related feedback in terms of performance during a common tracking task. In the second (Section 3.3), sensory substitution [103] is applied to compare different types of auditory feedback with their equivalent visual feedback, in order to find out whether mapping the same information on a different sensory channel (the visual channel) yields comparable effects to those gained in the first experiment. In the third experiment (Section 3.4) a continuously altered visuomotor transformation [116] between the controller and the screen is applied and kinematic information is mapped in order to investigate whether the task-related auditory feedback is more effective in the screen or in the controller's reference system. Finally, the fourth experiment (Section 3.5) compares performances of spatialized versus non-spatialized task-related auditory feedback.

The studies described in the following sections are the result of a joint work with the Department of Mechanical Innovation and Management, University of Padova, together with the University of Delaware (gait training experiment) and the University of California Irvine (upper limb training experiments). In these works my main task has been the development of the various auditory feedback modalities and their integration within the experimental setups.

The works presented in this Chapter were published in [144] (Section 3.2) and have been submitted for publication in papers [191] (Section 3.1) and [145] (Sections 3.2- 3.4). The experiment reported in Section 3.5 is still unpublished.

3.1 Auditory feedback for gait training

The aim of this experiment, which I refer to as experiment G1 (where G stands for gait), is to investigate whether the most commonly used combination of feedback in rehabilitation robotics (i.e., haptic and visual) can be either enhanced by adding auditory feedback or successfully substituted with a combination of kinetic guidance and auditory feedback, while helping a subject performing and learning an altered gait motor pattern. Adaptation of an altered motor pattern was in fact found to be necessary to achieve a semblance of functional walking following the neuromuscular impairments caused by stroke [129]. Auditory feedback is presented in conjunction with or in substitution to visual feedback and/or force-field constraint (FFC) provided by a robotic exoskeleton that the participant is wearing.



(a) A CAD model of the ALEX II device.

(b) Healthy subject during the experiment.

Figure 3.1: *Experiment G1: experimental setup (figure (a) reproduced from [186]).*

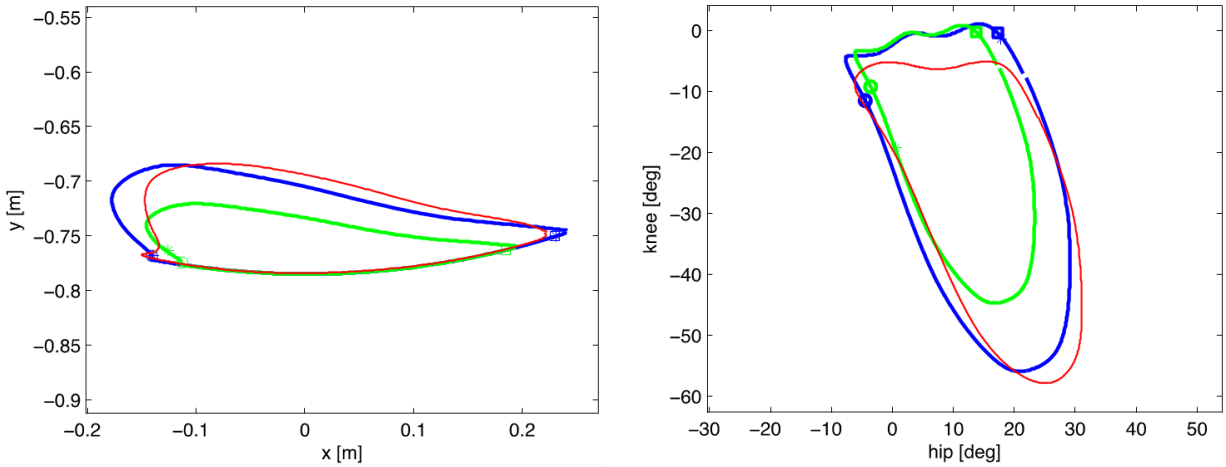
3.1.1 Subjects

The experiment is currently involving a significant number of healthy subjects (> 25), randomized into four groups based on the kind of feedback provided during gait training: group NF, group FS, group TR, and group SR (such division will be later explained). Conversely, early results reported in this thesis refer to a pilot study on a single healthy subject, aged 28 and caucasian, performing all of the single training sessions of the four groups.

3.1.2 Experimental setup

Fig. 3.1(a) reports a model of the robotic exoskeleton employed in this study, i.e. the ALEX II (Active Leg EXoskeleton) device developed at the University of Delaware. Without entering into detailed information about mechanics and control of the exoskeleton that can be found in [186], it is sufficient to report in this context that the device is a unilateral exoskeleton with two active DoF (hip and knee flexion and extension) that can accommodate either the left leg or the right leg of the subject. Additionally, the ALEX II provides 4 passive DoF at the trunk (vertical rotation and anterior/posterior, superior/inferior and side-to-side motions). The leg is mounted on a movable back support that has been designed to be used in conjunction with a traditional treadmill for robot-aided gait training.

The device can operate in zero-torque mode or with force field enabled. The inertia of the device is not actively compensated in the zero-torque mode, even though gravitational loads are. When the Cartesian-based force field is active, a target footpath is loaded into the controller, which represents the locus of points that the projection of the subject's malleolus onto the sagittal plane would pass through in an ideal gait cycle. The force-field behavior is modeled by a



(a) Cartesian space trajectories.

(b) Joint space trajectories.

Figure 3.2: *Experiment G1: Cartesian space and joint space trajectories. Blue curves: baseline trajectory. Green curves: template trajectory. Red curves: mean trajectory during a training time slice.*

non-linear virtual spring [186] that exerts a normal force towards the prescribed footpath if the deviation of the subject's foot from the target footpath exceeds an adjustable threshold. Conversely, no force is ideally exerted by the robot to the subject's leg if his/her ankle is sufficiently close to the prescribed footpath.

Fig. 3.1(b) pictures a more global view of the setup. The shoe worn by the subject's chosen leg (that in post-stroke patients will be the impaired lower limb) is instrumented with three pressure sensors mounted at the heel, ball, and toe of the foot that provide information about foot contact. These signals are used to pace the auditory feedback to subject's gait as well as for offline data processing. Each subject, after having fit into the exoskeleton, has to walk at regular speed on a treadmill. Subject motion in exoskeleton is recorded by the device itself, that continuously monitors the hip angle, knee angle, and ankle position in the subject's sagittal plane relative to the hip. The ankle position plotted against time gives rise to a peculiar trajectory in the sagittal plane, an example of which can be seen in Fig. 3.2(a).

The goal of the subject during training is to follow as much as possible one such trajectory (the *template* trajectory), guided by FFC plus auditory feedback and/or visual guidance (VG) presented on a screen placed in front of the treadmill, reporting the current position of the ankle on the Cartesian space template. The trajectory can also be visualized by the operator in the space of the two joints involved during gait training, hip and knee: Fig. 3.2(b) reports such an example.

Recorded data are sent from the robot real-time controller to a host PC, which runs the human-machine interface (HMI). A MATLAB script running on the same PC performs real-

time processing on the limited set of data (hip angle, knee angle, and ankle position) required for computing the auditory feedback, and sends them to a laptop via the OSC (Open Sound Control) protocol. A real-time graphical programming environment, Pure Data [132], is running on the laptop and is used for real-time audio synthesis. Sounds are presented to the subject by means of stereo speakers located in front of the treadmill.

Two different types of auditory feedback were designed:

- a *rhythmic feedback* triggering instrumental sounds during the subject's walk;
- *sonification* of the template trajectory based on formant synthesis of voice.[1]

The rhythmic feedback can be both subject-related and template-related. In the first case, a thumb piano key sample is triggered each time the subject lifts up his toe from the treadmill (toe-off, TO) and a metronome is played when the subject's foot hits again the treadmill (heel-strike, HS). Notes played by the thumb piano cycle among 4 different tones to make the auditory feedback more pleasant. In the second case, events TO and HS are not related to the subject's gait but to the template points where the expected TO and HS occur: rhythmic cues are triggered by the PC at a frequency that corresponds to the stride period of the target trajectory, yielding a regular rhythmic pattern. The use of a double-metronome allows participants to pace both footfalls to the rhythmic cues.

Concerning template sonification, each point of the joint space is mapped to a determined vocalized sound produced by a formant synthesis patch. In particular, the current hip angle controls the formants (i.e., the couple of frequencies that produce a vowel) of the sound, while the current knee angle is mapped to its fundamental frequency, which increases with knee flexion. By replicating the same sound at each gait cycle, the subject is able to easily reproduce the same footpath. However, informations about the mismatch between the current and the prescribed footpath are only provided by the force field.

3.1.3 Experimental protocol

In this experiment each participant in every group is asked to walk on the treadmill taking into consideration the information given by the perceived feedbacks. As already mentioned, what distinguishes the different groups is the kind of feedback provided to them during training:

- group NF: FFC plus VG, no auditory feedback;
- group FS: sonification plus FFC;
- group TR: template-related rhythmic feedback plus FFC;

[1] Sonification can be defined as a mapping of multidimensional datasets into an acoustic domain for the purposes of interpreting, understanding, or communicating relations in the domain under study [151]. As such, it can be thought of as the auditory equivalent of data visualization.

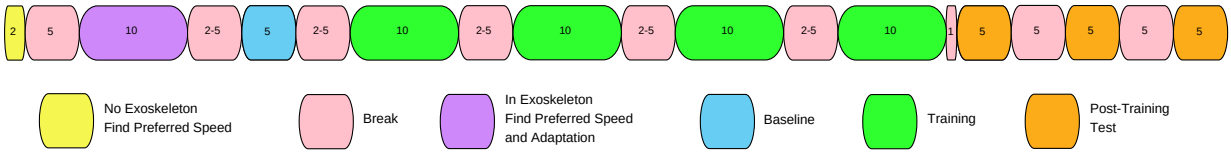


Figure 3.3: *Experiment G1: experimental protocol timeline for a single subject.*

- group SR: subject-related rhythmic feedback plus FFC and VG.

The full protocol for each subject is schematized in Fig. 3.3. Each subject walks for 2 minutes on the treadmill at his/her comfortable walking speed (CWS). Afterwards, his/her left leg is fitted to the device, and he/she walks for 10 minutes while the robot is controlled in zero-torque mode. During this warm-up session, the subject's CWS in the robot is determined, and the adjustments of the device can be slightly modified to improve his/her comfort. The treadmill speed is then maintained for the rest of the experiment. In this study, the robotic leg is attached to subject's non-dominant leg, i.e., the limb which is commonly recognized to be mainly involved in motion control. Indeed, several works on laterality and asymmetry in able-bodied gait corroborate the hypothesis that the non-dominant leg contributes to control tasks, support, and body weight transfer, while the contralateral limb is mainly responsible for propulsion [147].

In the following 5-minutes walk (baseline session, BSL), the hip and knee joint angles are recorded by motor encoders. By averaging data taken from the last 30s time-slice in this session, and mapping the resulting averaged footpath to the task-space (i.e., the Cartesian space), the subject's baseline footpath is derived. The target footpath is then computed by applying isotropic scaling to the set of points of the baseline footpath in the hip/knee joint space, with 0.8 as the scaling factor and the origin of the hip/knee axes as the external homothetic center. This method yields a stable yet challenging gait cycle, characterized by a shorter and shallower step. Similarly, the target stride period is computed from the average baseline stride period by comparing the relative positions of the heel-strike/toe-off points in the baseline footpath to the corresponding estimated positions in the target footpath. Notice that, being the treadmill speed equal to the baseline CWS, smaller steps in the prescribed trajectory result in a faster prescribed cadence (i.e., a shorter target stride period).

During training, subjects walk in the robotic exoskeleton, trying to match the target footpath. Training consists of four, 10-minutes long sessions, during which the force field (FFC) is always active (threshold 10mm, stiffness of non-linear spring $760000\mathrm{N}/\mathrm{m}^2$). Conversely, when provided, visual guidance and auditory feedbacks are turned on intermittently (i.e., during the first and third quarter of each training session) to prevent subjects from over-relying on extrinsic feedbacks.

Breaks are given to subjects between each pair of consecutive training sessions. Duration of the breaks is up to the subjects, ranging from 2 to 5 minutes. Minimal verbal cues are provided during early trainings, only if the subject finds it difficult to adapt to the force field. Participants

included in groups TR and SR are shown the current ankle position and the prescribed footpath during the first 40s of each training session. This approach is meant to provide subjects with minimal information about the goal movement. Therefore, these groups do not receive kinetic guidance and auditory feedback only, even though the amount of visual guidance is negligible if compared to groups NF and FS. Similarly, during the first 40s of each training session, people in group SR are provided with the prescribed cadence instead of the subject-triggered one. Thus, subjects in group SR do actually receive minimal informations about the cadence they are expected to walk at.

Post-tests consist of 3 sessions, 5 minutes each, the first of which starts 1 minute after the conclusion of the last training session. A 5-minute break is given between consecutive sessions, thereby the second and the third post-tests start 11 and 21 minutes after training, respectively. During these sessions, the robot is controlled in zero-torque mode and subjects are instructed to walk as normally as possible. These last sessions are regarded as measures of learning.

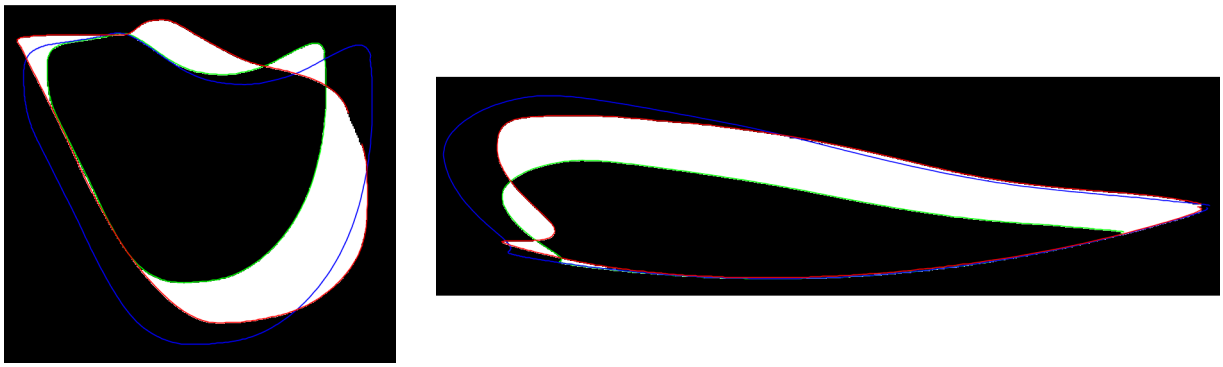
3.1.4 Data analysis

Data from encoders and pressure sensors are collected at 500 Hz and then filtered with a forward-backward 5-th order Butterworth filter ($f_c = 30$ Hz). Starting from these data, a set of variables which describes the participant's performance during training and post-test sessions (accuracy and precision measures) is computed. Data collected over a specific session (5 minutes for baseline and post-tests, 10 minutes for trainings) are first split into 30s time intervals. Then, metrics are computed within each time interval and subsequently averaged to yield a single value per session.

Concerning accuracy, three different error metrics are considered:

1. the normalized error area enclosed between the current trajectory and the template trajectory in the joint space,
$$JS_{\text{err}} = \frac{|A_i^{JS} - A_{\text{tmp}}^{JS}|}{|A_{\text{bsl}}^{JS} - A_{\text{tmp}}^{JS}|}, \tag{3.1}$$
where A_i^{JS} is the trajectory mean area in the i-th time slice, A_{tmp}^{JS} is the area enclosed by the template trajectory, and A_{bsl}^{JS} is the area enclosed by the baseline trajectory, all computed in the joint space reference system;

2. the normalized error area enclosed between the current trajectory and the template trajectory in the Cartesian space,
$$TS_{\text{err}} = \frac{|A_i^{TS} - A_{\text{tmp}}^{TS}|}{|A_{\text{bsl}}^{TS} - A_{\text{tmp}}^{TS}|}, \tag{3.2}$$
where A_i^{TS}, A_{tmp}^{TS}, and A_{bsl}^{TS} are the metrics respectively equivalent to A_i^{JS}, A_{tmp}^{JS}, and A_{bsl}^{JS}, but computed in the Cartesian reference system; and



(a) Calculation of JS_{err}.

(b) Calculation of TS_{err}.

Figure 3.4: *Experiment G1: two error metrics. The white area enclosed between the current trajectory (red curve) and the target trajectory (green curve) gives a measure of accuracy. The blue curve is again the baseline trajectory.*

3. the normalized stride period error,

$$T_{\text{err}} = \frac{|T_i - T_{\text{tmp}}|}{|T_{\text{bsl}} - T_{\text{tmp}}|}, \tag{3.3}$$

where T_i, T_{tmp}, and T_{bsl} are the average stride period in the i-th time slice, and the stride periods of the template and baseline, respectively. Single-stride periods are all calculated as the time intercurring between two subsequent TO events.

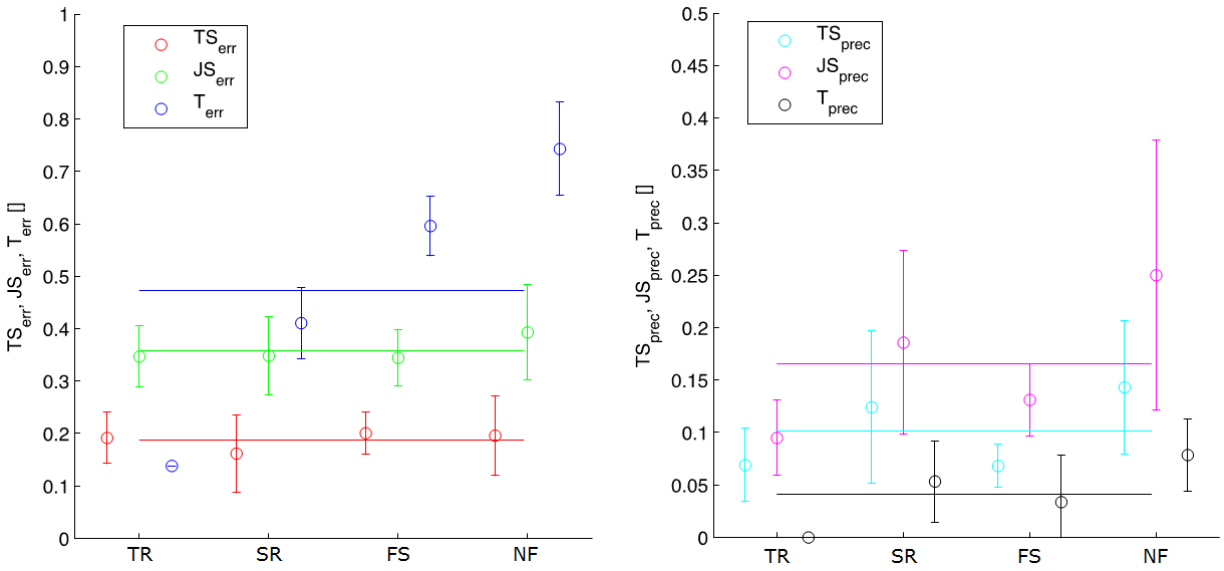
Fig. 3.4 reports a graphic representation of the above metrics. As an example, if the subject is walking on the baseline trajectory both JS_{err} and TS_{err} will be unitary, whereas if he/she follows perfectly the template trajectory these will be equal to zero. Similarly, a unitary value for T_{err} indicates that the subject is walking to the pace of the baseline trajectory.

Precision metrics JS_{prec}, TS_{prec}, and T_{prec} are defined just as JS_{err}, TS_{err}, and T_{err}, with the only difference that the average trajectory/stride period of the subject computed over the whole bout replaces the baseline trajectory/stride period. These can be seen as repeatability measures: a high value of such metrics indicates that gait in different time slices is highly variable around its mean, i.e. that repeatability is low.

3.1.5 Results and discussion

I shall now present and discuss the results of a pilot experiment involving one subject performing the training sessions for the four groups. Fig. 3.5 illustrates the mean and standard deviation among different time slices of the six defined metrics for each of the four feedback types.

It can be clearly seen that mode TR (template-based rhythm) gives the best results in all of the precision metrics and in T_{err}. These are easily explained by the “metronome” paradigm incorporated in such kind of regularly triggered feedback. Still, good results are also obtained in



(a) Accuracy metrics.

(b) Precision metrics.

Figure 3.5: *Experiment G1: results for one healthy subject.*

the other two metrics (JS_{err} and TS_{err}) even though visual feedback is absent and no extrinsic information about error is provided apart from FFC. Such result suggests that this kind of auditory feedback allows the subject to concentrate more on proprioception than on visual feedback, hence to base his/her own gait corrections on proprioceptive information.

Mode SR (subject-based rhythm plus visual guidance) yields very good results in TS_{err} probably thanks to visual information, whereas JS_{err} is comparable to mode TR. However, variance across time slices is greater, and this fact reflects itself onto the high values in the precision metrics: trajectories are not repeatable. T_{err} is also higher both in mean and variance, probably because of the prominent weight of the feedback component thanks to which the subject concentrates more on minimizing error rather than performing a correct gait. Still, T_{err} is lower than in mode NF (no auditory feedback), indicating that having a greater perception of his/her steps helps the subject avert his/her attention away from error-related feedback provided by visual guidance.

Similarly to mode TR, audio in mode FS (formant synthesis) cannot be regarded as an error-related feedback. Errors in JS_{err} and TS_{err} are comparable to both previous modes, while T_{err} is definitely worse. However, precision is almost as good as in mode TR, even without any rhythmic information: this result may be explained by the observation that thanks to trajectory sonification the subject has knowledge of results, and aims at replicating the correct trajectory by minimizing differences between what he/she had learned to be the correct vocalization and what he/she hears during gait, regardless of step cadence. Although not being the best in the viewpoints of both accuracy and precision, such sonification-based feedback modality is the most innovative, having

never been investigated in previous literature.

Along with correctness of movement, repeatability is a fundamental goal for gait training: as a matter of fact, it can mean that the subject has created his/her own motor pattern and follows it with a high degree of precision. A proper task-based feedback may help the subject in performing repetitive exercises, even in presence of a desired altered template, and could lead to improved learning. Furthermore, task-related feedback can easily be integrated in everyday life, where external information on error is unavailable. High precision does not occur for modes SR and NF, where the subject executes several different attempts in order to follow a template visualized on the screen; this result highlights how motor control is in these two modes more influenced by feedback than feedforward. As a consequence, all of the robotic rehabilitation systems that make use of video as the principal feedback modality are liable to induce the subject gaining excellent performances (i.e. low errors) without really learning motor control, that needs to be mainly based on feedforward information in order to reach a higher level of effectiveness.

The reported observations are currently being verified at the University of Delaware on a significant number of subjects. Clearly, the comprehensive results will have to be interpreted by considering retention effects (i.e. performance in post-training phases) too: a previous study [86] has revealed that, even in mode NF, short-term gait modifications lasting up to 2 hours could be induced in healthy individuals. Thus, it has to be investigated whether the same effect can be provided or even improved by the use of the auditory channel; in other words, the conclusion advanced in [86] that *multiple feedback modalities allow better adaptation to a new gait pattern* needs to be extended to audio. Still, the preliminary results of the presented pilot experiment have indicated that the auditory system can already be seen as a useful information channel for the subject during gait.

3.2 Auditory feedback for arm training: task-related feedback

We now move to experiments for upper limb training through simple target tracking exercises augmented with multimodal feedback. The aim of this first experiment, which is referred to as experiment T1 (where T stands for tracking), was to find out whether a task-related auditory feedback is able to provide informations that help the subject improving performance more than position-error-related auditory feedback or visual feedback alone.

3.2.1 Subjects

A total of 20 healthy subjects participated to the experiment. They were aged between 21 and 29 (mean age 22.95 ± 1.99), 55% male and 45% female, caucasian and right-handed. All the

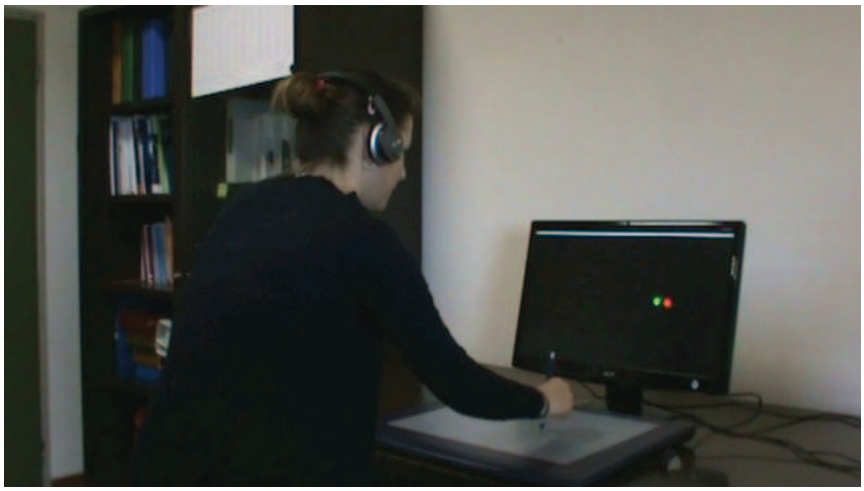


Figure 3.6: *Experiments T1 and T4: experimental setup.*

participants had normal vision with no color blindness, and no hearing problems based on their self-report.

3.2.2 Experimental setup

As pictured in Fig. 3.6, each participant was provided with a pair of common headphones that presented auditory feedback and a Wacom pen as his/her controller. During the whole experiment, each subject was sitting in front of a desk with a Full HD screen in the middle of it, and a Wacom pen tablet suitably calibrated in order to match the screen size right in front of him/her. The screen was backed by a blank wall.

The main application was implemented in MATLAB. Two color-filled, 25-pixel-radius dots were displayed on the screen: one representing the controller's position (green dot) and one for target position (red dot). Each participant was asked to perform a tracking exercise of the target's movement on a horizontal line while at the same time controlling the green dot as accurately as possible. Obviously, control of the pen required movement of the right upper limb.

Two different types of target motion were envisaged:

- a *fixed-length* profile, where the length of each left-right-left movement cycle was set to 60% of screen size for all iterations within the same session, corresponding to a range of motion for the subject's hand of nearly 300 mm;

- a *random-length* profile, where at each iteration the length of the segment pseudo-randomly varied from 20% to 90% of the screen size. At the end of the session, the total distance traveled by the target was the same as in the first case.

In both cases, the trajectory of the target had a minimum-jerk profile, i.e. considering a fifth-degree polynomial function

$$q(t) = a_0 + a_1(t-t_i) + a_2(t-t_i)^2 + a_3(t-t_i)^3 + a_4(t-t_i)^4 + a_5(t-t_i)^5, \tag{3.4}$$

where t_i is the start time of the trajectory, onto which the following contour conditions are imposed:

- end time $t_f = \frac{|q_f - q_i|}{v}$;
- initial position $q_i = q(t_i)$ and final position $q_f = q(t_f)$;
- initial velocity $\dot{q}_i = \dot{q}(t_i) = 0$ and final velocity $\dot{q}_f = \dot{q}(t_f) = 0$;
- initial acceleration $\ddot{q}_i = \ddot{q}(t_i) = 0$ and final acceleration $\ddot{q}_f = \ddot{q}(t_f) = 0$.

The coefficient values that uniquely determine the desired trajectory are then

$$\begin{aligned}
a_0 &= q_i \\
a_1 &= \dot{q}_i \\
a_2 &= \frac{1}{2}\ddot{q}_i \\
a_3 &= \frac{20(q_f - q_i) - (8\dot{q}_f + 12\dot{q}_i)T + (3\ddot{q}_f - 2\ddot{q}_i)T^2}{2T^3} \\
a_4 &= \frac{30(q_i - q_f) + (14\dot{q}_f + 16\dot{q}_i)T + (3\ddot{q}_f - 2\ddot{q}_i)T^2}{2T^4} \\
a_5 &= \frac{12(q_f - q_i) - 6(\dot{q}_f + \dot{q}_i)T - (\ddot{q}_f - \ddot{q}_i)T^2}{2T^5},
\end{aligned} \tag{3.5}$$

where v is the mean velocity in each segment, kept constant during the exercise, and $T = t_f - t_i$.

Just as in the gait training experiment, auditory feedback was developed in Pure Data. Target's and subject's data (positions and velocities in the X and Y directions) were sent in real-time to Pure Data through the OSC protocol. Two different types of auditory feedback were designed:

- a *task-related* auditory feedback simulating the sound of a rolling ball;
- an *error-related* auditory feedback performing formant synthesis of voice.

For task-related feedback, the velocity of the target was applied as a simple gain factor onto the output of a pink noise generator filtered through a bandpass filter with 200-Hz center frequency and Q factor equal to 7. Concerning error-related feedback, the position errors between indicator and target in both axes were used to control the parameters of the same formant synthesis patch used in the previous experiment. Specifically, the X-axis position error was mapped onto the

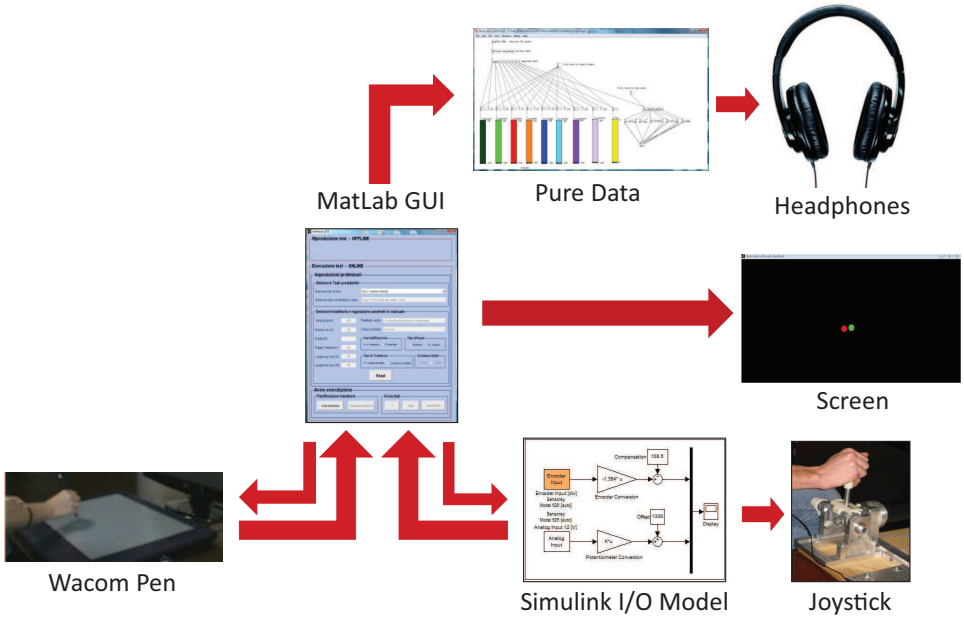


Figure 3.7: *Functioning scheme of the target tracking system. The Wacom pen is used in experiments T1 and T4; the joystick is used in experiments T2 and T3.*

amplitude and fundamental frequency of a synthetic vocalized sound, while the Y-axis position error controlled the formants of the sound.

Spatial sound information was added to both feedbacks using the `earplug~` Pure Data external, which offers 3-D sound rendering based on non-personalized head-related transfer functions [30], by fitting the target X-axis position to the azimuth angle parameter. The described auditory feedbacks were provided in turn to the user through headphones. A simple scheme of the system's architecture is reported in Fig. 3.7, where the Simulink model and the joystick were not used in this experiment.

3.2.3 Experimental protocol

All the participants were asked to complete six different tasks. For each task, the subject had to draw a trajectory onto the tablet with the pen in order to follow the target on the screen. The six tasks were:

- task A: fixed-length trajectory, no auditory feedback;
- task Br: random-length trajectory, no auditory feedback;
- task C: fixed-length trajectory, task-related auditory feedback;
- task Dr: random-length trajectory, task-related auditory feedback;

- task E: fixed-length trajectory, error-related auditory feedback;
- task Fr: random-length trajectory, error-related auditory feedback.

Each task lasted 80 seconds and consisted of 13 repetitions of the left-right-left movement cycle. The mean velocity of the target was set to 400 pixels per second. During each task, target and subject's indicator position and velocity were sampled at a frequency of 300 Hz. Each subject executed all tasks in a randomly-generated sequence, after a first warm-up task without target where the participant could get acquainted with the tablet. During the three seconds preceding each task, a countdown was simulated through a sequence of three tonal beeps.

3.2.4 Data analysis

For each participant, the integral of relative velocity (i.e., the difference between subject's and target's velocities), the weighted position error along the horizontal direction (X-axis), and the mean distance between subject indicator and target were measured. Each measure was calculated for every left-right and right-left segment, then it was averaged over the whole task for each subject.

The *integral of relative velocity* for the $k-th$ segment is defined as

$$R_{\text{vel}}(k) = \frac{1}{L_k}\int_{t_k}^{t_{k+1}} |\vec{v}_r| dt, \tag{3.6}$$

where $|\vec{v}_r| = |\vec{v}_s - \vec{v}_t|$ is the norm of the relative velocity vector, L_k is the length of segment k, whereas t_k and t_{k+1} are the beginning and end times of the segment. R_{vel} was calculated using the rectangle method:

$$\sum_{h=1}^{N} \frac{\sqrt{\left(v_{x,s}\left(h\right) - v_{x,t}\left(h\right)\right)^2 + \left(v_{y,s}\left(h\right) - v_{y,t}\left(h\right)\right)^2} \cdot dt}{L_k} \tag{3.7}$$

where N is the number of samples in the segment. The R_{vel} parameter measures the extra distance traveled by the subject while following the target, accounting for the movements made by the subject to correct tracking errors. A null value of this metric indicates that the velocity profile of the target has been exactly reproduced by the subject, even though the average position error (in terms of a constant offset measured by the second metric) may be not null.

The position error along the X-axis was weighed with the sign of target velocity and normalized to target radius R. The *average weighed position error* for segment k is defined as:

$$e_x(k) = \frac{1}{N}\sum_{h=1}^{N} \frac{\left(x_s\left(h\right) - x_t\left(h\right)\right) \cdot sign\left(v_{x,t}\left(h\right)\right)}{R}. \tag{3.8}$$

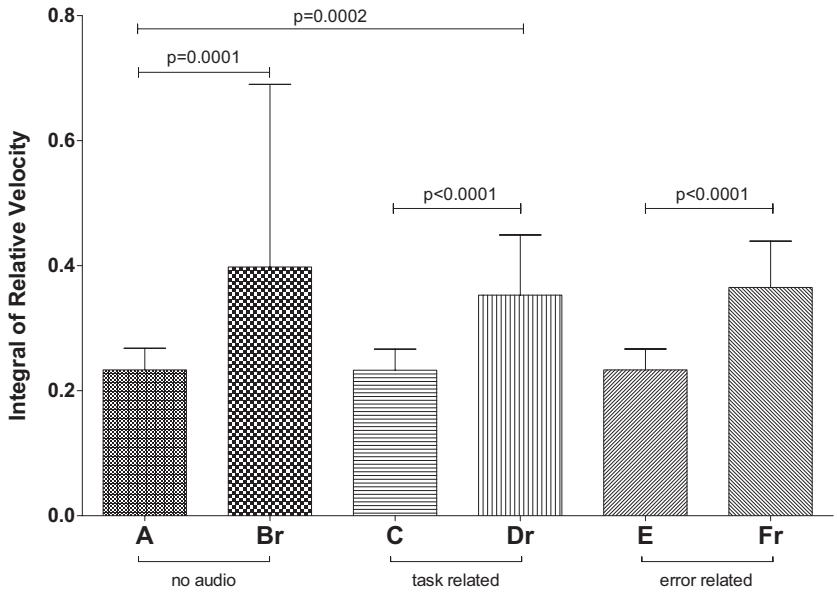


Figure 3.8: *Experiment T1: statistical analysis on integral of relative velocity.*

This formula takes into account the direction of motion of the target, thus showing whether the subject leads (positive error) or lags (negative error) the target during the exercise. Lead error is defined as the tracking error when the subject indicator anticipates the target (i.e. leads the target motion), while lag error is the tracking error when the subject indicator follows the target. Formally, positive terms in the summation in Eq. (3.8) contribute to lead error (e_x^{lead}) calculation, while negative terms contribute to lag error (e_x^{lag}) calculation. A null value in this metric indicates that the subject had an average null delay with respect to target motion, even though the distance traveled around the target (which is measured by the first metric) may be not null.

Finally, the average distance normalized to the dot radius, defined as

$$d_m(k) = \frac{1}{N}\sum_{h=1}^{N}\frac{\sqrt{\left(x_s\left(h\right)-x_t\left(h\right)\right)^2+\left(y_s\left(h\right)-y_t\left(h\right)\right)^2}}{R} \tag{3.9}$$

for segment k, was also calculated. This measure roughly indicates the accuracy of movement.

A comparison between paired data was performed (D'Agostino and Pearson omnibus normality test), resulting in a Gaussian distribution for tasks A-C-E-Fr (integral of relative velocity), A-Br-C-Dr-E-Fr (weighed position error and lead error), A-C-E (lag error), and A-Dr-E (mean distance). Consequently, either parametric or non-parametric (Wilcoxon) paired t-tests were performed in order to compare the measures among different tasks.

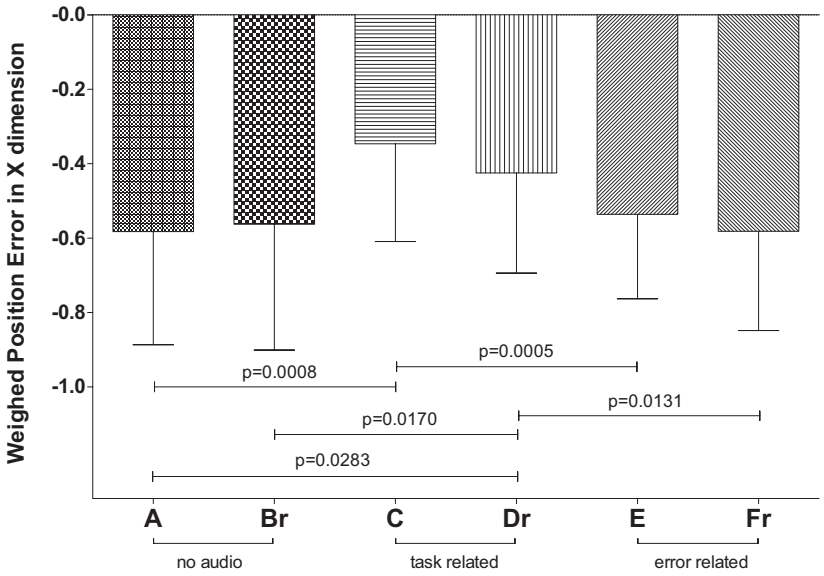


Figure 3.9: *Experiment T1: statistical analysis on weighed position error.*

3.2.5 Results and discussion

The main result of the statistical analysis on the integral of relative velocity was that, as one may expect, the fixed-length task is always significantly better executed than the corresponding random-length task, regardless the audio modality employed (see Fig. 3.8): the subjects made significantly greater corrections in the random-length tasks with respect to the corresponding fixed-length task for every audio modality. On the other hand, no statistically significant difference was found, in terms of extra distance traveled around the target, when the audio modality was changed while keeping the same trajectory type, indicating that the audio modality did not affect the number of corrections made by the subject while tracking the target.

Conversely, statistical analysis on average weighed position error revealed that, in terms of tracking delay, there is no significant difference between fixed and random length tasks within the same auditory feedback modality (see Fig. 3.9), indicating that the trajectory type did not affect the average tracking delay. However, task C presented a significantly smaller negative error with respect to tasks A and E, while task Dr did the same with respect to Br and Fr. In other words, task-related auditory feedback (C and Dr) helped the subjects to significantly reduce average tracking delay with respect to error-related auditory feedback (E and Fr) and to no audio (A and Br), both in the easier (fixed length) and in the more complex (random length) tasks.

By decomposing the previous measure into lead and lag error, it was found how subjects averagely tend to lag during the task, especially for the random-length tasks (see Fig. 3.10). This result can be easily justified by the unpredictability of these tasks and the subject's physiological response delay to multisensory feedback. In this context, it is important to underline that in the

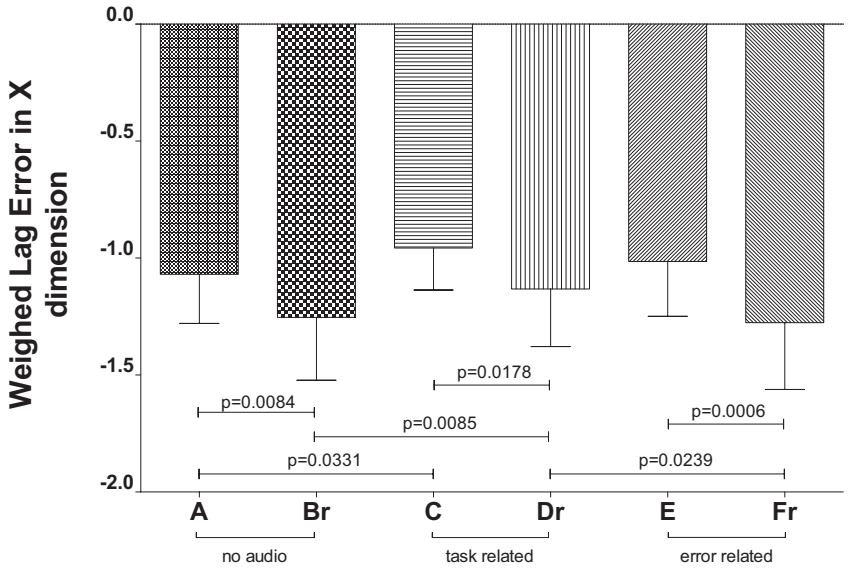


Figure 3.10: *Experiment T1: statistical analysis on lag error.*

statistical analysis lag error in tasks C and Dr is statistically lower compared to the corresponding tasks A and Br, whereas this does not happen for lead error. Task-related auditory feedback thus involves actions that aim at reducing lag error.

Conversely, error-related feedback seems to act more as a deterrent to performing tracking errors rather than an augmentation of the visual information available to the user. This hypothesis may explain the deterioration of the performance with respect to the task-related case in all of the analyzed statistics except the lead error: indeed, this kind of feedback seems effective only in indicating the target's deceleration phase by warning the user not to overshoot the end point of the segment.

The effects of both error-related and task-related feedback on performance and motor learning are often discussed in the literature, sometimes with contrasting results. As an example, according to Timmermans *et al.* [177] rehabilitation technology should provide both knowledge of results as well as knowledge of performance. A combination of error-based augmented feedback and feedback on correct characteristics of the performed movement is thus advisable to enhance both motor learning and motivation. The results of experiment T1 suggest instead that even task-related feedback alone can be beneficial to users' performance.

3.3 Auditory feedback for arm training: sensory substitution of audio

The purpose of the following experiment, which is referred to as experiment T2, was to prove whether equal information to auditory feedback onto a different sensory channel gives the same positive effects in users' performance. The experimental setup of this experiment was very similar to that of the previous one, the only substantial difference being the controller.

3.3.1 Subjects

A total of 22 healthy subjects participated to the experiment (mean age 23 ± 1.66, 81.8% male and 18.2% female). They were caucasian and right-handed, except for one subject who was left-handed. Again, all the participants had normal vision with no color blindness, and no hearing problems.

3.3.2 Experimental setup

As shown in Fig. 3.11, the subjects sat on a chair with an haptic 2-DoF semiactive joystick (see [28]) fixed on the right side of it. The experimental setup was also composed by a PC host and a pair of common headphones that presented auditory feedback. The 2-DoF joystick was used in passive mode (i.e. free motion). According to [28], manipulation of the joystick along the first DoF, which involves movement along the X-axis, was measured through a Baumer BHK incremental encoder with $8000 imp./rev.$, while the second DoF, which was disconnected from the passive actuator in order to reduce friction and inertia, was measured using a $10k\Omega$ single-turn conductive plastic potentiometer with a $10V$ power supply. A Sensoray 626 I/O module was managed in an external computer through a real-time software (Simulink R2010b) running at $100Hz$. A Graphical User Interface (GUI) in MATLAB allowed to send UDP packets in real-time to an audio synthesis patch through the OSC protocol. The scheme of the setup can be again related to Fig. 3.7, the joystick replacing the tablet.

As in the previous experiment, two color-filled, 25-pixel radius dots were represented on the screen, one for controller position (green dot) and one for target position (red dot), and each participant was asked to perform a tracking exercise of the target's movement as a horizontal left-to-right movement. In this experiment, all tasks shared a *fixed-length* trajectory with a minimum-jerk velocity profile as in Eqs. (3.4)$-$ (3.6), corresponding to a range of motion for the subject's hand of 150 mm.

auditory feedback was again developed in Pure Data. In addition to the previous two types of auditory feedback, adapted to and improved for this experiment, a third type of feedback related to velocity error was designed. Ultimately, the three auditory feedback modalities were:

- a *task-related* auditory feedback simulating the sound of a rolling ball;

Figure 3.11: *Experiments T2 and T3: experimental setup.*

- a *position-error-related* auditory feedback performing formant synthesis of voice;
- a *velocity-error-related* auditory feedback simulating DJ scratching.

Task-related auditory feedback was improved through the use of a bandpass filter with Q factor equal to 9 and 500-Hz center frequency, slightly variable with the velocity input in the latter version. For position-error-related auditory feedback, the position error between the indicator and the target in X-axis was used alone to control the parameters of the formant synthesis patch. Specifically, the X-axis position error was mapped onto the amplitude, the fundamental frequency and the couple of formants in order to generate a sound which changes in frequency for small errors and vowel for medium/large errors, resulting in a more straightforward feedback.

Velocity-error-related auditory feedback was designed as a cubic polynomial profile of the X-axis velocity error applied onto the output of a pink noise generator filtered through a bandpass filter, set up as in the task-related audio signal. In addition, a dead zone and a sign control were added to activate feedback only in presence of medium-to-large errors and when the controller was moving away from the target. All of the auditory feedbacks were again binaurally spatialized through the `earplug~` Pure Data external.

For this experiment visual alterations equivalent to the described auditory feedbacks were also designed in MATLAB. Basically, a progressive alteration of the screen's background color, fading from black to light blue proportionally to the current mapped quantity, i.e. X-axis position error, X-axis target velocity, or X-axis velocity error, was introduced.

3.3.3 Experimental protocol

Each participant was asked to complete seven different tasks. For each task, the participant had to grasp the joystick handle performing a horizontal movement with the aim of following the

target on the screen. The seven tasks were:

- task A: no audio/color feedbacks;
- task B: position-error-related color feedback;
- task C: velocity-error-related color feedback;
- task D: task-related color feedback;
- task E: position-error-related auditory feedback;
- task F: velocity-error-related auditory feedback;
- task G: task-related auditory feedback.

Each task lasted about 90 seconds and consisted of 15 repetitions of the left-right-left movement cycle. The mean velocity of the target was set to 500 pixels per second. Each subject executed all tasks in a randomly-generated sequence, after a first warm-up task without target, where the participant could get acquainted with the device. During the three seconds preceding the beginning of each task, a countdown was simulated through a sequence of three tonal beeps.

3.3.4 Data analysis

For this experiment the weighed position error e_x, the lead and lag position errors e_x^{lead} and e_x^{lag}, the integral of relative velocity R_{vel}, and the mean distance d_m were calculated as in the previous experiment, see Eqs. (3.6)$-$ (3.9). Four participants, who misinterpreted the execution of one or more color feedback tasks, were excluded from the analysis.

A comparison between paired data (D'Agostino and Pearson omnibus normality test) was performed, resulting in a Gaussian distribution for tasks A-B-C-D-F-G (integral of relative velocity, weighed position error and lead error), A-B-D-F-G (mean distance) and A-D-F-G (lag error). Thus, either parametric or non-parametric (Wilcoxon) paired t-tests were performed in order to compare performance parameters among different tasks.

3.3.5 Results and discussion

The comparison of the integral of relative velocity between tasks A and B in Fig. 3.12 shows that the addition of the error-related color feedback increases the extra total distance traveled by the subject. Moreover, each color feedback modality (B, C and D) induces significantly greater trajectory corrections with respect to the corresponding substituted audio modality (E, F and G). Concerning audio tasks, results confirm those found in experiment T1 (no significance on this metric if compared to the first task), including the new sound modality (velocity-error related

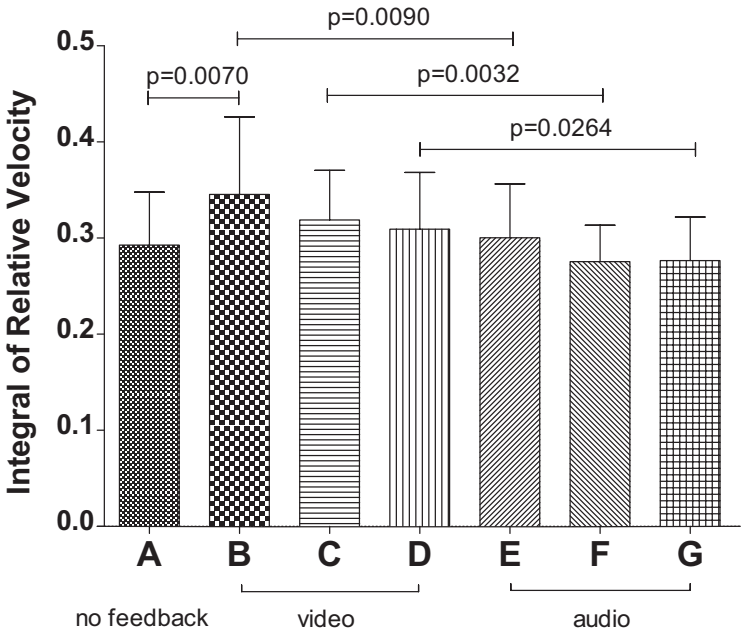


Figure 3.12: *Experiment T2: statistical analysis on integral of relative velocity.*

audio): the auditory feedbacks do not significantly alter the extra total distance traveled by the subject while tracking the target.

A similar trend for color tasks is also found in the statistical analysis on the weighed position error data, shown in Fig. 3.13. Indeed, background-color alteration degrades tracking performance in terms of lag from the target, resulting in an increased weighed error for each task both with respect to task A and to audio tasks. As a consequence, providing the same information on task or error through vision does not bring upgrades in performance, suggesting that visual information cannot be augmented through the same channel in the experienced motion tracking tasks. Replacing auditory feedback with a background color transformation on the screen leads to results that are even worse than having the original visual feedback alone. This finding indicates that, in these tests, the visual channel is already saturated by the target following task, so that the background color variation turns out to be a distraction rather than a useful additional information for the user. Instead, two separated information channels (visual and auditory), if properly coordinated, work in a parallel fashion and can contribute to performance enhancement, as already pointed out.

In addition, the absolute better performance of task-related auditory feedback, as found in the experiment T1, is confirmed. This last result shows how task-related auditory feedback is effective independently of the controller used in the experiments: as a matter of fact, the use of two very different input devices (pen tablet in T1 and joystick in T2) yielded totally analogous results.

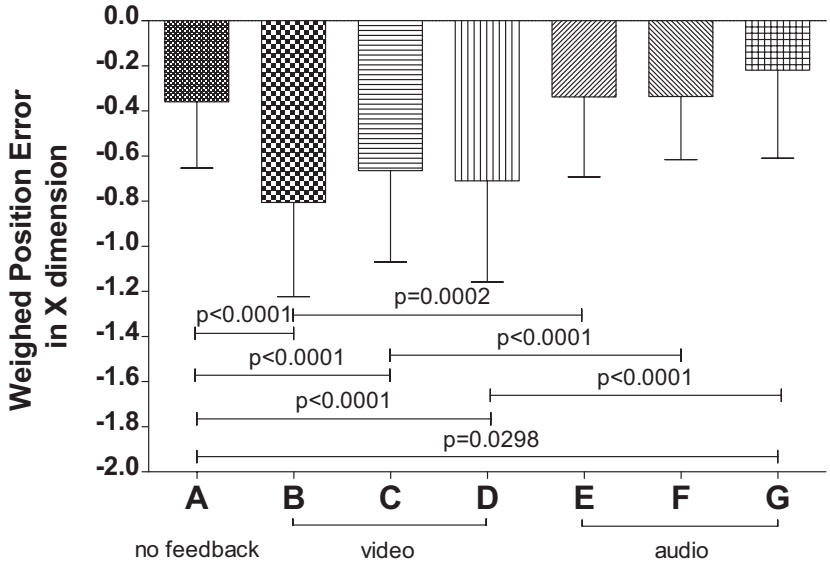


Figure 3.13: *Experiment T2: statistical analysis on weighed position error.*

3.4 Auditory feedback for arm training: visuomotor transformations in task-related feedback

The aim of this third experiment, which is referred to as experiment T3, was to find out in which reference system (video or controller) for target tracking as an input to a task-related auditory feedback better helps the subject improving performance in the context of a continuously variable visuomotor transformation. The experimental setup was almost identical to that of experiment T2.

3.4.1 Subjects

A total of 47 healthy subjects participated to the experiment (mean age 24.04 ± 2.77, 78.7% male and 21.3% female). They were caucasian and right-handed, except for two subjects who were left-handed. This time too, all the participants had normal vision and no hearing problems. Subjects were randomized into four groups based on the kind of feedback provided during the experiment: 11 subjects belonged to group NF, 12 to group ER, 12 to group TR-V, and 12 to group TR-J. Such division will be made clearer in the following.

3.4.2 Experimental setup

In this experiment the same hardware and software equipment of experiment T2 was exploited, i.e. the joystick fixed on the right side of the subject, the PC host running the Simulink model as

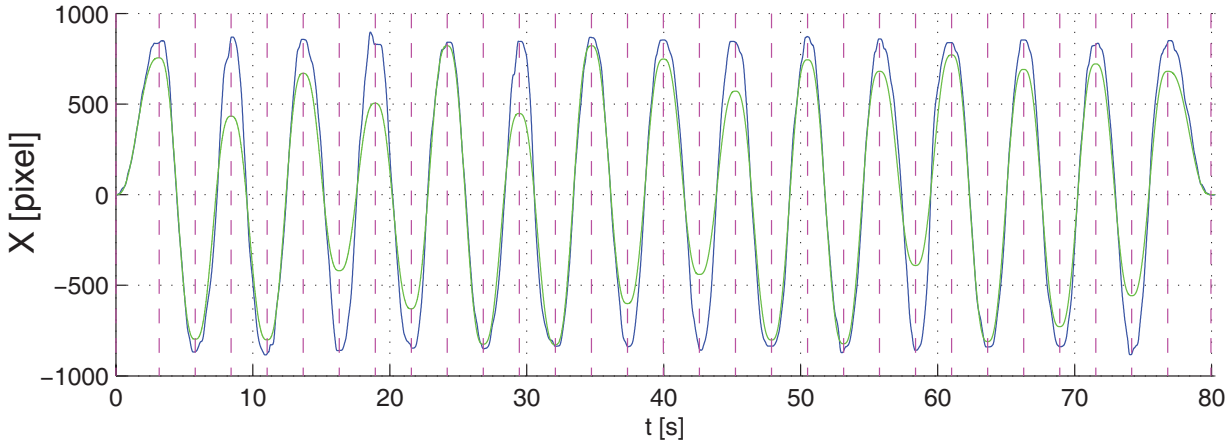


Figure 3.14: *X position versus time of target (green line) and subject (blue line) in one representative trial of the visuomotor transformation task of experiment T3. Subject position has been converted into pixels for the purpose of this chart. Dashed violet lines indicate the beginning of each trajectory segment. Despite the variable amplitude of target motion, the subject tends to make a fixed amplitude motion, due to the presence of the visuomotor transformation.*

well as the MATLAB GUI representing the two dots (target and current position) on the screen, and headphones as pictured in Fig. 3.11. Each participant was asked to perform a tracking exercise exactly as required in experiment T2.

The target movement displayed on the screen had a minimum-jerk velocity profile, in which the length of each segment:

- in the first phase (*warm-up task*), was kept constant as in experiment T2;
- in the second phase (*visuomotor transformation task*), pseudo-randomly varied from 20% to 90% of screen size; in addition, in this phase the scale between the video and the joystick was changed at each iteration, in such a way that the required motion of the joystick remained fixed along all segments (as in the warm-up task): owing to the alteration of the introduced video-joystick scale, the random-length motion of the target visualized in this phase corresponded to the same fixed-length target motion of the subject's hand used in the warm-up.

Fig. 3.14 depicts the X position versus time of the target (green line) and of the subject (blue line) in one representative run of the visuomotor transformation task. It is clearly shown in the figure that, despite the variable amplitude of target motion, the subject tends to make a fixed amplitude motion, due to the presence of the visuomotor transformation. We can summarize by saying that, in this modality, the target motion of the arm had a fixed length, while the motion of the target displayed on the screen had a randomly-variable length.

Three different auditory feedbacks were used:

- an *error-related* auditory feedback performing formant synthesis of voice; in this modality, the mapped quantity was the position error on the X-axis, measured on the screen;
- a *video-task-related* auditory feedback, simulating the sound of a rolling ball by mapping the target velocity in screen scale;
- a *joystick-task-related* auditory feedback, simulating the sound of a rolling ball by mapping the target velocity in joystick scale.

By using the last two modalities, it is tested whether the efficacy of task-related audio in reducing the average tracking error, as observed in experiments T1 and T2, was induced by an augmented description of the visualized task or by audio rendering of the target motion of the arm.

3.4.3 Experimental protocol

For each task, the participant had to grasp the joystick handle performing a horizontal movement with the aim of following the target on the screen. In this experiment each participant in every group was asked to complete the same single task. No information on the visuomotor transformation were provided to the subjects. As mentioned before, what distinguished the various groups was the kind of auditory feedback provided to them during the exercise:

- group NF: no auditory feedback;
- group ER: error-related auditory feedback (in the video reference system);
- group TR-V: video-task-related auditory feedback;
- group TR-J: joystick-task-related auditory feedback.

The warm-up task was made of 20 repetitions of the fixed-length, fixed scale target trajectory. After a 5 minutes rest, a sequence of three tonal beeps signaled the beginning of the visuomotor transformation task. This task consisted in 15 repetitions of the left-right-left movement cycle with random-length, visually altered trajectory. The mean velocity of the target on the screen was set to 500 pixels per second.

3.4.4 Data analysis

The weighed position error e_x, the lead and lag position errors e_x^{lead} and e_x^{lag}, the integral of relative velocity R_{vel}, and the average distance d_m were calculated as described in the previous experiments using the visual scale.

A comparison between unpaired groups (D'Agostino and Pearson omnibus normality test) was performed, revealing a Gaussian distribution for all the examined cases except for the lead

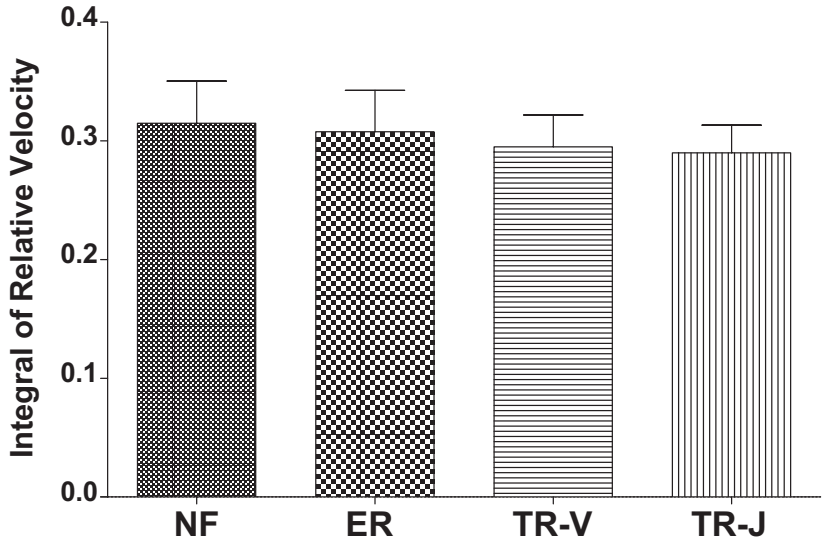


Figure 3.15: *Experiment T3: statistical analysis on integral of relative velocity.*

error of group ER. Thus, either parametric or non-parametric (Mann-Whitney) unpaired t-tests were performed in order to compare the participants' performance measures among different groups (i.e., among different feedback modalities).

3.4.5 Results and discussion

The histogram of the integral of relative velocity, shown in Fig. 3.15, reports no statistically significant difference in task execution between different groups, i.e. the extra total distance traveled by the subject's hand during the task is not influenced by the auditory feedback provided, and is comparable to that of group NF (no auditory feedback).

The results of the average normalized distance (Fig. 3.16) show instead a significant difference between group ER and TR-V and TR-J respectively, resulting in a better average accuracy of the movement when task-related auditory feedback (both in the joystick and in the video reference system) is provided rather than error-related auditory feedback. However, group NF presents an average distance statistically equivalent to the groups with auditory feedback, even if group TR-V actually lightly improves accuracy compared to group NF.

Regarding the average weighed position error reported in Fig. 3.17, one can observe that in presence of the visuomotor transformation, the error-related auditory feedback yields significantly greater average tracking delays with respect to all other modalities. In other words, providing position-error related information through sound, despite being substantially equivalent to the absence of auditory feedback in experiments T1 and T2, may be detrimental during learning of a novel visuomotor transformation.

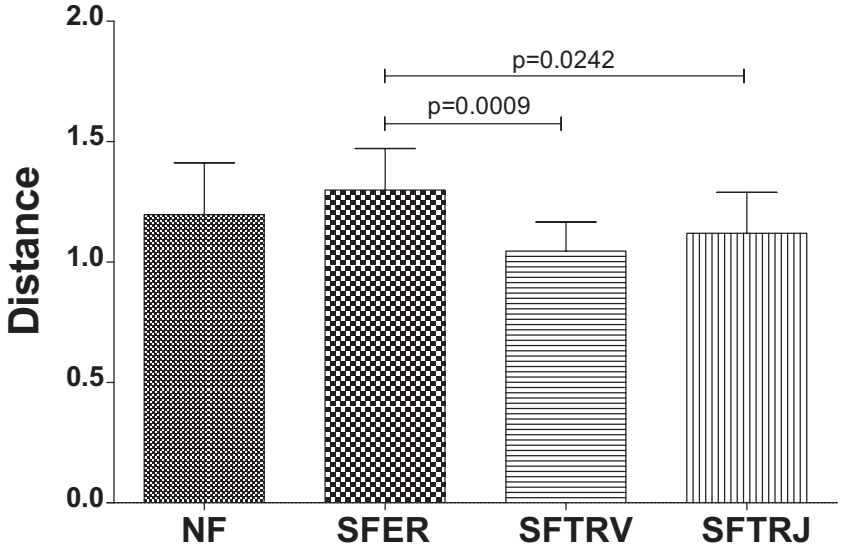


Figure 3.16: *Experiment T3: statistical analysis on average tracking distance.*

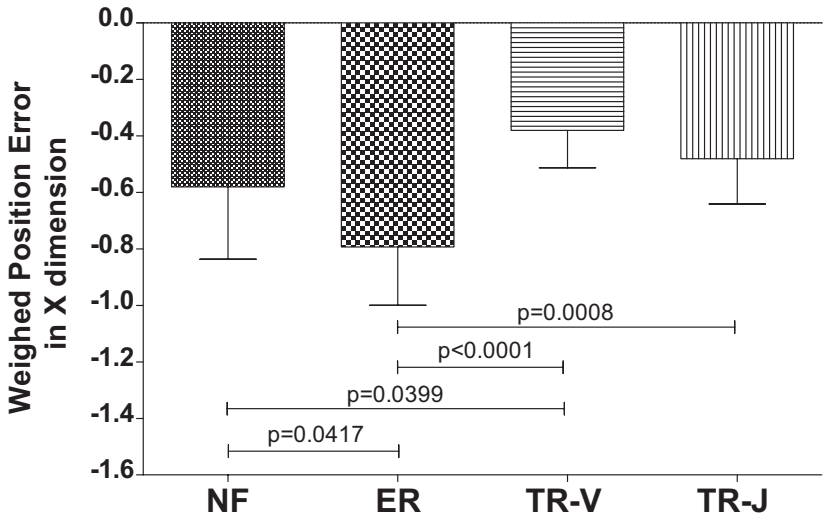


Figure 3.17: *Experiment T3: statistical analysis on weighed position error.*

On the other hand, providing task-related information through sound during learning of a novel visuomotor transformation can be beneficial if the auditory information is consistent with that provided by visual feedback, yielding reduced average tracking delay with respect to no auditory feedback. This trend is due to a consistent different influence on lag error (histogram and significances are similar to those reported in Fig. 3.17) as opposed to a comparable level of lead error (figure not included), which is however not sufficient in counteracting the lag. On the contrary, providing auditory information related to expected arm movement is not likely to bring benefits in presence of a novel visuomotor transformation with respect to no auditory feedback.

To sum up, it comes out from this experiment that task-related feedback is effective in the context of a visuomotor transformation explicitly designed to confuse the user, although in minor measure with respect to the proportional case. The video-related feedback, being consistent with what the user actually sees, is more effective in reducing tracking delay with respect to the joystick-scale audio, which provides information on the effective target motion of the arm. This result is particularly interesting, as more correct information on desired arm motion were provided in the joystick-related modality, which in turn yielded worse results. On the contrary, performance was improved by enhancing task information that were inconsistent with the desired arm motion. This finding suggests that the subject tends to expect information on task rather than on motor command from extrinsic feedback. Secondly, video-related audio provides additional information in accordance with the sensory channel onto which the user's attention is already focused, following a visual dominance principle.

In other words, the user manages to compensate the mismatch between the two movement ranges by relying mostly on the visual feedback, yet the sensory augmentation given by visual-scale auditory feedback contributes to increase performance with respect to the condition where the auditory channel is not used. Conversely, creating a conflict between the audio and video modalities leads the user to maintain attention focused onto the visual input [13], obtainig results comparable to those gained in absence of the audio signal. However, joystick-related feedback leads to a slightly better performance than the video-only condition. This result is in agreement with [141], where it is stated that misleading or noisy feedback increases coordination variability although saturating toward the level without feedback at most.

An improved performance achieved from video-related auditory feedback during a continuous visuomotor perturbation may indicate a more effective continuous learning of the scale variation. Thus, a properly designed task-related auditory feedback continuously provided to the user may lead to enhanced learning in rehabilitation exercises. This last hypothesis requires further investigation that could be addressed in future research.

3.5 Auditory feedback for arm training: effect of sound spatialization

The aim of this last experiment, which is referred to as experiment T4, was to find out whether the information given to the user by spatialized task-related auditory feedback helps the subject improving his/her performance more than non-spatialized task-related auditory feedback.

3.5.1 Subjects

A total of 16 healthy subjects participated to the experiment. They were aged between 19 and 42 (mean age 26.31 ± 6.46), 50% male and 50% female, caucasian and right-handed. All the participants had again normal vision with no color blindness, and no hearing problems.

3.5.2 Experimental setup and protocol

The setup for this experiment was much analogous to that of experiment T1, i.e. the participants used the Wacom pen tablet and wore headphones, and can thus be again related to Fig. 3.6. Both fixed-length and random-length profiles for the movement of the target were envisaged. The sole relevant difference with respect to experiment T1 lied in the choice of the auditory feedbacks. In particular, a single *task-related* feedback simulating the sound of a rolling ball was realized in Pure Data. The only differences with respect to the task-related feedback used in experiment T1 were in the bandpass filter's parameters, having 300-Hz center frequency and Q factor equal to 10: such a choice yielded a much lighter timbre to the rolling sound. The task-related feedback was then either spatialized through the usual binaural rendering patch or not, resulting in two different auditory feedbacks to be compared against each other as well as to the no-sound condition.

All the participants were asked to complete the following six different tasks, presented in a random order:

- task A: fixed-length trajectory, no auditory feedback;
- task Br: random-length trajectory, no auditory feedback;
- task C: fixed-length trajectory, non-spatialized auditory feedback;
- task Dr: random-length trajectory, non-spatialized auditory feedback;
- task E: fixed-length trajectory, spatialized auditory feedback;
- task Fr: random-length trajectory, spatialized auditory feedback.

Just as in experiment T1, each task lasted 80 seconds and consisted of 13 repetitions of the left-right-left movement cycle.

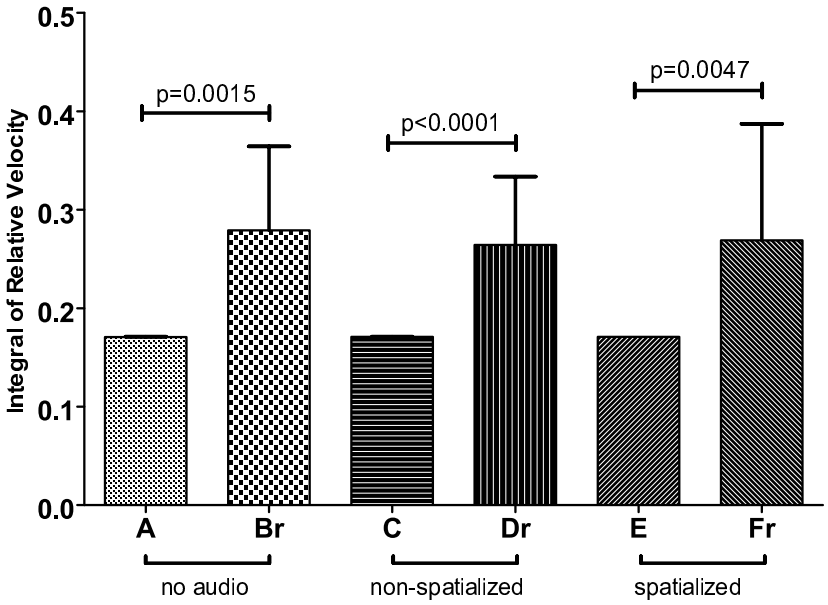


Figure 3.18: *Experiment T4: statistical analysis on integral of relative velocity.*

3.5.3 Data analysis

Identically to all of the three previous experiments, the weighed position error e_x, the lead and lag position errors e_x^{lead} and e_x^{lag}, the integral of relative velocity R_{vel}, and the mean distance d_m were calculated.

A comparison between paired data (D'Agostino and Pearson omnibus normality test) was performed, resulting in a Gaussian distribution for tasks Br-C-Dr-E-Fr (integral of relative velocity), A-Br-Dr-E-Fr (weighed position error and lead error), A-Dr-E-Fr (mean distance) and A-Br-Dr-E (lag error). Consequently, either parametric or non-parametric (Wilcoxon) paired t-tests were performed in order to compare performance parameters among different tasks.

3.5.4 Results and discussion

Similarly to previous findings, the only relevant result of the statistical analysis on the integral of relative velocity, reported in Fig. 3.18, was that the fixed-length task is always significantly better executed than the corresponding random-length task, independently of the audio modality. The same result was also found by analyzing the mean distance measure.

Conversely, no significant difference between fixed and random length tasks within the same auditory feedback modality was evidenced by the statistical analysis on the average weighed position error, see Fig. 3.19. In this case, it was the auditory feedback modality that made the

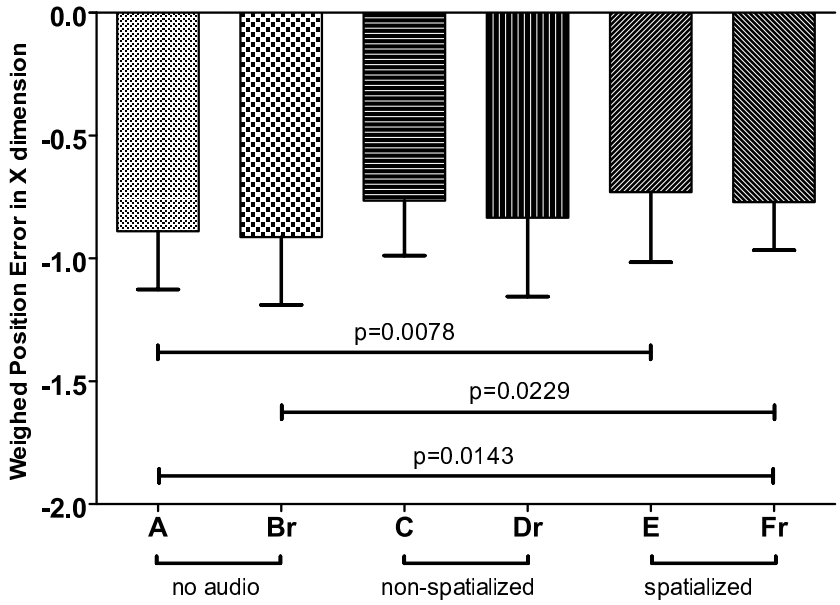


Figure 3.19: *Experiment T4: statistical analysis on weighed position error.*

difference. Both the fixed-length audio tasks C and E presented a smaller negative error with respect to task A, and the same applied to random-length audio tasks Dr and Fr with respect to task Br. However, only the spatialized audio tasks reported significant difference with respect to the no-audio tasks, while non-spatialized ones did not. In other words, only spatialized task-related auditory feedback (tasks E and Fr) helped the subjects to significantly reduce average tracking delay with respect to having no auditory feedback, both in the fixed length and in the random length tasks. Non-spatialized auditory feedback lied between the other two modalities in such terms, even though not reporting significant difference with respect to spatialized auditory feedback.

The careful observer will note that Fig. 3.19 exhibits smaller differences in average tracking error values between tasks A-E and Br-Fr if compared to the results for experiment T1 reported back in Fig. 3.9 for the equivalent feedback couples A-C and Br-Dr. This may be partly due to the slightly different settings of the rolling sound. However, the statistically significant upgrade given by the spatialized task-related auditory feedback is preserved.

While analysis of lag error did not add much with respect to the previous measure, lead error (reported in Fig. 3.20) was found to be statistically different both between fixed-length and equivalent random-length tasks and among fixed-length tasks themselves. In particular, lead error in task A was significantly lower than in tasks C and E. This result is harder to interpret than the previous ones; still, it could be advanced that the lead error component was greater in

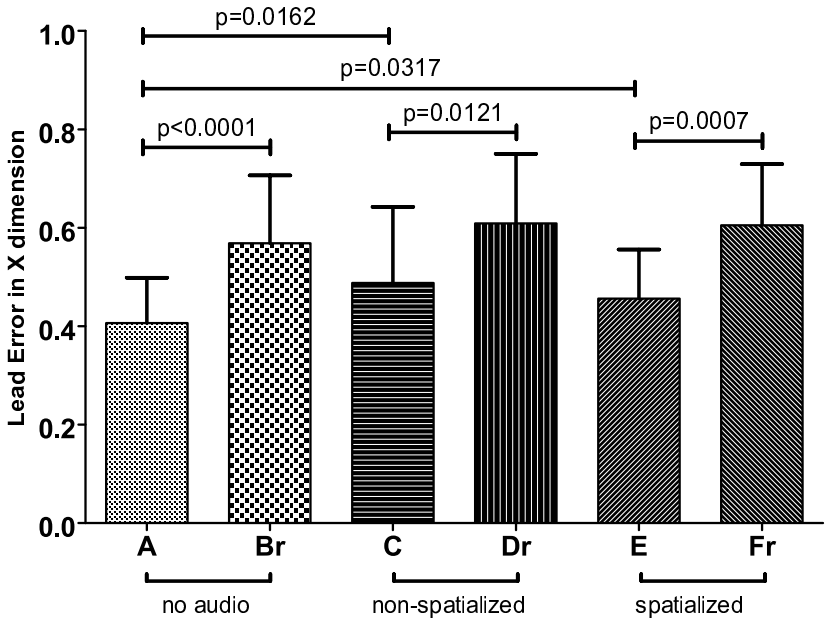


Figure 3.20: *Experiment T4: statistical analysis on lead error.*

random-length tasks because of the sudden, unpredictable deceleration phase for short segments, whereas in fixed-length tasks lead error was lower but tended to increase in presence of consistent auditory feedback because subjects felt more confident in executing the task, tending sometimes to lead the target's movement. Complementarily to experiment T1, it was thus found how task-related auditory feedback involves actions that aim at increasing lead error.

To sum up, whereas differences between spatialized and non-spatialized task-related auditory feedback were not seen to be particularly marked, spatialized feedback led to a statistically better performance than with no audio that could not be attested for non-spatialized feedback. The use of generalized HRTFs could of course have limited the realism of the spatialization in some subjects, psychoacoustically resulting in a trivially panned version of the non-spatialized feedback. It has indeed to be mentioned that half of the subjects (8 out of 16) informally reported no significant difference between the two audio modalities, and that 4 among them explicitly felt that the rolling auditory feedback was confusing, preferring the condition with no audio. However, the other half felt that spatialization added useful information to the task, by helping in particular during the most sudden acceleration and deceleration phases and by letting the subject better concentrate on the task. Along with improvements in the monophonic signal, a required step towards a better rendering of the used feedback is thus the exploitation of customized HRTFs.

3.6 Conclusions

The results of experiments T1 to T4 confirm that auditory augmentation of visual feedback is overall beneficial to the user's performance in upper limb movement tasks, even in presence of a novel visuomotor transformation. In other words, the addition of a secondary sensory channel that faithfully represents the information already provided by the visual channel helps the user having a stronger perception of the task, allowing for improved sensory-motor coordination. Such result lies in accordance with [141], which states that coordination variability with more than one sensory modality is smaller than with one modality only. This suggests that the performer can easily manage to integrate visual and auditory information online during task execution by tending to optimize the signal statistics.

The rolling ball paradigm for the proposed task-related feedback is obviously included in this class of continuous auditory cues, being a straightforward and intuitive mean of providing velocity profiles through the auditory channel, and indeed remarkably enhances performance. This may also be likely due to the fact that provision of feedback through the auditory system allows better parallel processing, even in cases where the information seems redundant. As a matter of fact, rather than acting as a confounding influence, auditory feedback enhances visuomotor control because it provides similar information [155].

Task-related auditory feedback proved to be effective in reducing the average tracking error, even though it did not affect the number of trajectory corrections made by the subject while attempting to follow the target (integral of relative velocity). Such result is consistent with the observation that this audio modality can be considered as a feedforward input for the subject's motor control. Conversely, providing error-related information through sound in presence of visual feedback (through both formant synthesis reflecting position error and scratching effects reflecting velocity error, as of experiment T2) did not affect tracking performance. This result may be explained by considering that error-related audio presents redundant information with respect to the visual modality, rather than providing an augmentation of the visual information available to the user. In addition, one may argue that the subject may expect to receive or elaborate error related information from video rather than from the auditory sensory channel, and this may lead the subject to disregard the information received through sound.

The effect of spatialization was also found to be overall beneficial to the user, although the information provided by generalized HRTFs could not be unanimously appreciated. In light of this, the next chapters of this thesis will focus on how to improve the currently exploited spatial sound rendering through a customized HRTF model not involving any cumbersome measurement. The use of customized HRTFs will be expected to ultimately positively augment the gap between performances in the no-audio and spatialized audio conditions.

The influence of auditory feedback was studied on healthy subjects first to characterize the normative response of the human motor system to auditory feedback, yet these experiments should be adapted to a post-stroke scenario in order to attest the absolute effectiveness of auditory

feedback in rehabilitation contexts. However, these results definitely provide a basis for a future comparison with post-stroke patients.

An important implication of these findings is that more and more attention should be paid to incorporating effective forms of auditory feedback during robot-assisted movement training. To date, although there are attempts to use sound in a more sophisticated way, auditory feedback is underutilized in most robotic therapy systems, playing a role as background music or signifying only task completion in most cases (as already discussed in Section 2.3). Understanding the real potential of audio in rehabilitation contexts requires further investigation that will be addressed in future research, that should thus examine how auditory feedback can best be crafted to improve engagement, performance and learning in rehabilitation exercises, the ultimate goal being enhancement and acceleration of motor adaptation and motor recovery.

Chapter 4

Binaural Perception and Rendering: Previous Work

At the beginning of the last century, Lord Rayleigh's studies on the scattering of sound waves by obstacles gave birth to the extensive and still partially misunderstood field of 3-D sound. Within the context of his notable Duplex Theory of Localization [169], a commonly known formula that approximates the behaviour of sound waves diffracting around the listener's head provided indeed a first glance of the today-called head-related transfer function (HRTF). Alas, despite the importance and applicative potential of such a centenary theory, most of the efforts towards efficient modeling of HRTFs were spent in the last few decades only.

Formally, the HRTF at one ear is defined as the frequency-dependent ratio between the sound pressure level (SPL) $\Phi(\theta, \phi, \omega)$ at the eardrum and the free-field SPL at the center of the head $\Phi_f(\omega)$ as if the listener were absent:

$$H(\theta, \phi, \omega) = \frac{\Phi(\theta, \phi, \omega)}{\Phi_f(\omega)}, \tag{4.1}$$

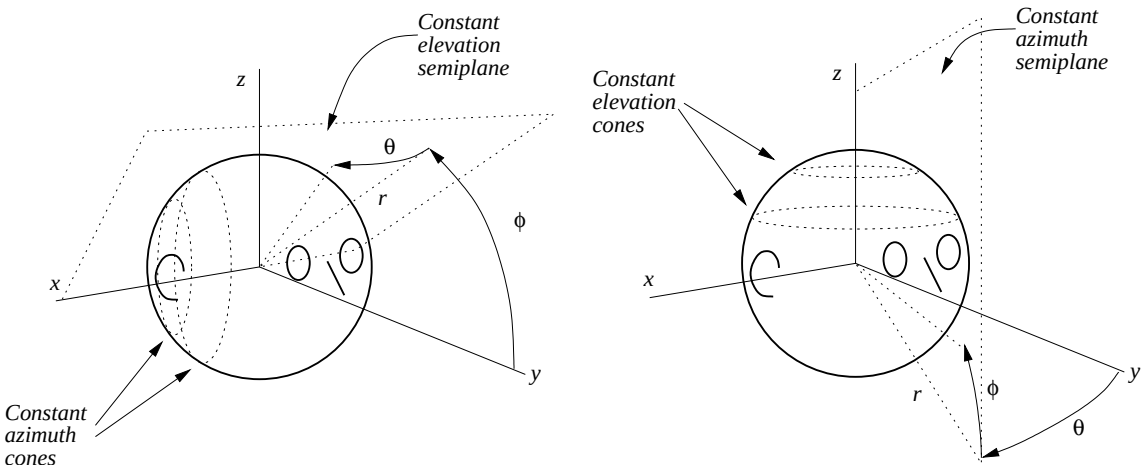
where (θ, ϕ) indicates the angular position of the source relative to the listener, and ω is angular frequency. The HRTF can alternatively be seen as the Laplace transform of the free-field compensated impulse response relative to the path of the sound wave from the source to the eardrum, the head-related impulse response (HRIR). This means that the HRTF contains all of the information relative to sound transformations caused by the human body, in particular by the head, external ears, torso and shoulders. Clearly, a left and a right HRTF exist, one per ear: apart from perfect symmetries, these two HRTFs are different. Such characterization allows virtual positioning of sound sources in the surrounding space: consistently with its relative position to the listener's head, the emitted signal can be filtered through the corresponding pair of HRTFs creating left and right ear signals to be delivered by headphones [30]. In this way, three-dimensional sound fields with a high immersion sense can be simulated and integrated into a great variety of contexts.

Unfortunately, recording individual HRTFs of a specific listener requires specific facilities, expensive equipment, and delicate audio treatment processes. For these reasons non-individualized (or generalized) HRTFs, e.g. measured on *dummy heads* (mannequins constructed from average anthropometric measures), are used in most applications. A series of experiments were conducted by Wenzel *et al.* [183] in order to evaluate the effectiveness of non-individualized HRTFs for virtual acoustic displaying. A very similar perceived horizontal angular accuracy in both real conditions and with 3-D sound rendering was obtained by employing generalized HRTFs; however the experiments showed that the use of generalized functions increases the rate of front-back reversals (i.e. a sound in the front is perceived in the back, or *vice versa*). Also, Begault *et al.* [15] compared the effect of generalized and individualized HRTFs applied onto a speech sound in static and dynamic conditions. Their results showed that source localization with generalized HRTFs in static conditions is marginally deteriorated with respect to the individualized case in the horizontal dimension, while head motion is crucial to reduce angular errors in the vertical dimension and to avoid reversals.

To sum up, while non-individualized HRTFs represent a cheap and straightforward mean of providing 3D perception in headphone reproduction, listening to non-individualized spatialized sounds is likely to result in evident sound localization errors such as incorrect perception of source elevation, front-back reversals, and lack of externalization [117] that cannot be fully counterbalanced by additional spectral cues, especially in static conditions [176]. In particular, elevation cues cannot be characterized through generalized spectral features. Hence, alongside critical dependence on the relative position between listener and sound source, anthropometric features of the human body have a key role in HRTF characterization.

Throughout the last decades, low-order rational functions and series expansions were proposed as tools for HRTF modeling. Albeit the straightforward nature and intrinsic simplicity of both techniques, real-time HRTF rendering requires fast computations which cannot undergo the complexity of filter coefficients and weights, respectively. Oppositely, structural modeling [20] ultimately represents an attractive solution to these shortcomings. If one isolates the contributions of the user's head, pinnae and torso to the HRTF in different subcomponents, each accounting for some well-defined physical phenomenon, then thanks to linearity he/she can reconstruct the global HRTF on-the-fly from a proper combination of all the considered effects. Relating each subcomponent's temporal and/or spectral features in the form of digital filter parameters to the corresponding anthropometric quantities would then yield a HRTF model which is both economical and individualizeable. As a further advantage, the intuitive nature of physical parameters enforces the chance to relate the model to simple anthropometrical measurements.

In this Chapter some of the most relevant findings and issues in the contexts of spatial sound localization (Section 4.1) and HRTF modeling (Sections 4.2, 4.3, 4.4) are outlined. In the following I will refer to an unique spatial coordinate system, the *interaural polar* system reported in Fig. 4.1(a). It is one of the two spherical coordinate systems found in literature, the other one being the *vertical polar* system reported in Fig. 4.1(b). In the interaural polar coordinate system



(a) Interaural polar coordinate system.

(b) Vertical polar coordinate system.

Figure 4.1: *The two spherical coordinate systems considered in literature.*

the origin coincides with the interaural midpoint, the elevation angle ϕ goes from $-180°$ to $180°$ with negative values below the horizontal plane and positive values above, while the azimuth angle θ ranges from $-90°$ at the left ear to $90°$ at the right ear. The third dimension, distance r, is the Euclidean distance between the observation point and the origin. Also with respect to Fig. 4.1, in the following I will refer to plane xy as the *horizontal* plane, plane yz as the *median* (or *sagittal*) plane, and plane xz as the *frontal* plane.

4.1 Spatial source localization

Spatial cues for sound localization can be categorized according to the involved polar coordinate. As a matter of fact, each coordinate is thought to have one or more dominant cues in a certain frequency range associated to a specific body part, in particular:

- azimuth and distance cues at all frequencies are associated to the head;
- elevation cues at high frequencies are associated to the pinnae;
- elevation cues at low frequencies are associated to torso and shoulders.

Based on known concepts and results, the most relevant cues for sound localization are now discussed.

4.1.1 Azimuth cues

Back in 1907, Lord Rayleigh studied the means through which a listener is able to discriminate at a first level the horizontal direction of an incoming sound wave. Following his famous Duplex

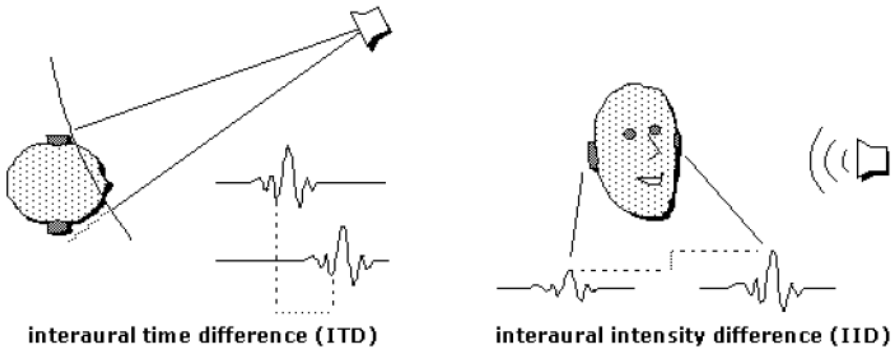


Figure 4.2: *Interaural time difference and interaural level difference.*

Theory of Localization [169], azimuth cues can be reduced to two basic quantities thanks to the active role of the head in the differentiation of incoming sound waves, i.e.

- *Interaural Time Difference* (ITD), defined as the temporal delay between sound waves at the two ears;
- *Interaural Level Difference* (ILD), defined as the ratio between the instantaneous amplitudes of the same two sounds, also known as *Interaural Intensity Difference* (IID).

Fig. 4.2 schematically sketches both concepts. ITD is known to be frequency-independent below 500 Hz and above 3 kHz, with a theoretical ratio of low-frequency ITD versus high-frequency ITD of $3/2$, and slightly variable at middle range frequencies [93]. Conversely, frequency-dependent shadowing and diffraction effects introduced by the human head cause ILD to greatly depend on frequency. These two points will be further discussed in the next section.

Consider a low-frequency sinusoidal signal (say up to 1.5 kHz approximately). Since its wavelength is greater than the head dimensions, ITD is reduced to a phase lag $\Delta\varphi < 2\pi$ between the signals arriving at the ears [17]. For this reason ITD is seen as a robust cue for horizontal perception in the low-frequency range. Conversely, ILD is not thought to be a robust cue because low frequency components trespass the head without causing significant attenuation on the opposite side with respect to the source. Specularly, a high-frequency sinusoidal signal (above 1.5 kHz) yields an ITD that is greater than a period. Being the human ear phase-sensitive only, ITD turns out to be useless in the high-frequency range, apart from detection of sound onsets. Nevertheless, the considerable shielding effect of the human head on high-frequency waves makes ILD the most relevant cue in such spectral range.

Still, the information provided by ITD and ILD can be ambiguous. If one assumes a spherical geometry of the human head as in Fig. 4.1, a sound source located in front of the listener at azimuth θ and a second one located at the rear, at azimuth $180 - \theta$, provide in theory identical ITD and ILD values. In practice, ITD and ILD will not be identical at these two azimuth angles because

1. the human head is clearly not spherical;

2. all subjects exhibit slight asymmetries with respect to the median plane;

3. ear canals are not located in the horizontal plane as in Fig. 4.1 but lie below and behind the x axis [2].

Nonetheless their values will be very similar, and *front-back confusion* is in fact often observed experimentally [23, 21]: listeners operate reversals in azimuth judgements, erroneously locating sources at the rear instead of at the front (or *vice versa*, less frequently). It can be argued that this asymmetry may originate from a sort of ancestral survival mechanism, according to which if something can be heard but not seen then it must be at the rear. However, this kind of reversal is highly listener-dependent.

4.1.2 Elevation cues

Directional hearing in the median vertical plane has long been known to bear little resolution compared with the horizontal plane [185]. For the sake of record, the threshold for detecting changes in the direction of a sound source (known as "localization blur") along the median plane was found to be never less than 4°, reaching a much larger threshold ($\approx 17^\circ$) for unfamiliar speech sounds, as opposed to a localization blur of approximately $1^\circ - 2^\circ$ in the horizontal plane for a vast class of sounds [17]. Such a poor resolution is motivated by two basic observations:

- the need of high-frequency content (above $4-5$ kHz) for accurate vertical localization [180, 68, 8];

- the theoretically nonexistent interaural differences between the signals arriving at the left and right ear in the sagittal plane.

Indeed, if a source is located outside the horizontal plane, ITD- and ILD-based localization becomes problematic. As Fig. 4.3 sketches, sound sources located in the far field at all possible points of a conic surface pointing towards the ear of a spherical head produce the same ITD and ILD values. These surfaces, that generalize the forementioned concept of front-back confusion for elevation angles, are known as *confusion cones* and represent a potential hump for accurate perception of sound direction.

Nonetheless, it is undisputed that vertical localization ability is brought by the presence of the pinnae [53]. Even though localization in any plane involves pinna cavities of both ears [119], determination of the perceived vertical angle of a sound source in the median plane is essentially a monaural process [67]. The external ear plays an important role by introducing peaks and notches in the high-frequency spectrum of the HRTF, whose center frequency, amplitude, and bandwidth greatly depend on the elevation angle of the sound source [159], to a remarkably minor extent

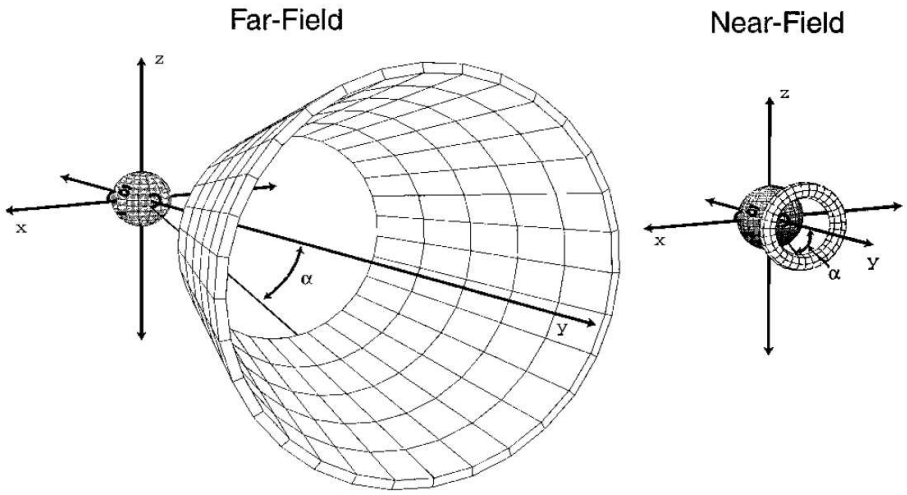


Figure 4.3: *Cone of confusion and torus of confusion (figure reproduced from [22]).*

on azimuth [96], and are almost independent on distance between source and listener beyond a few centimeters from the ear [24]. Following two historical theories of localization, the pinna can be seen both as a filter in the frequency domain [17] and a delay-and-add reflection system in the time domain [12] as long as typical pinna reflection delays for elevation angles, clearly detectable by the human hearing apparatus [189], were seen to produce spectral notches in the high-frequency range.

Additionally to reflections, pinna resonances and diffraction inside the concha were also seen to contribute to HRTF spectral shaping. Shaw [158] identified six resonant modes of the pinna (see Fig. 4.4) excited at different directions which clearly produce the most prominent HRTF spectral peaks: an omnidirectional resonance at 4.2 kHz (mode 1), two vertical resonances at 7.1 and 9.6 kHz (modes 2 and 3), and three horizontal resonances at 12.2, 14.4, and 16.7 kHz (modes 4, 5, and 6).[1] These results find accordance in a more recent study by Kahana *et al.* [81] on BEM-based numerical simulation of baffled pinna responses.

Concerning diffraction effects, Lopez-Poveda and Meddis [96] motivated the slight dependence of spectral notches on azimuth through a diffraction process that scatters the sound within the concha cavity, allowing reflections on the posterior wall of the concha to occur for any direction of the sound. Presence of diffraction around the tragus area has also been recently hypothesized by Mokhtari *et al.* [114, 115].

Nevertheless, the relative importance of major peaks and notches in elevation perception has been disputed over the past years.[2] A recent study [75] showed how a parametric HRTF

[1]The reported center frequencies were averaged among 10 different pinnae. Vertical modes are excited by sources above the head; horizontal modes by sources in the vicinity of the horizontal plane.

[2]In this context, it is important to point out that both peaks and notches in the high-frequency range are perceptually detectable as long as their amplitude and bandwidth are sufficiently marked [118], which is the case for most

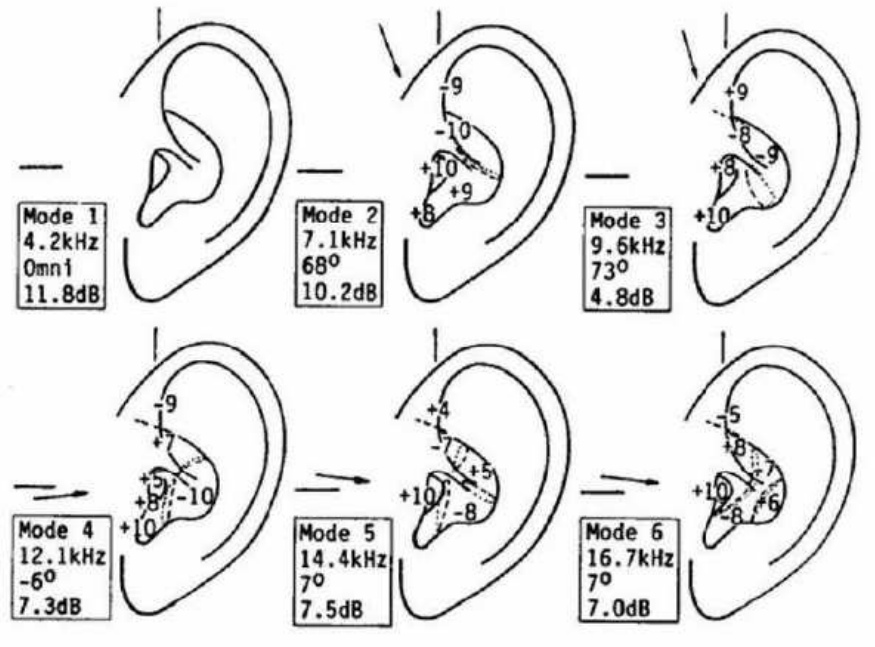


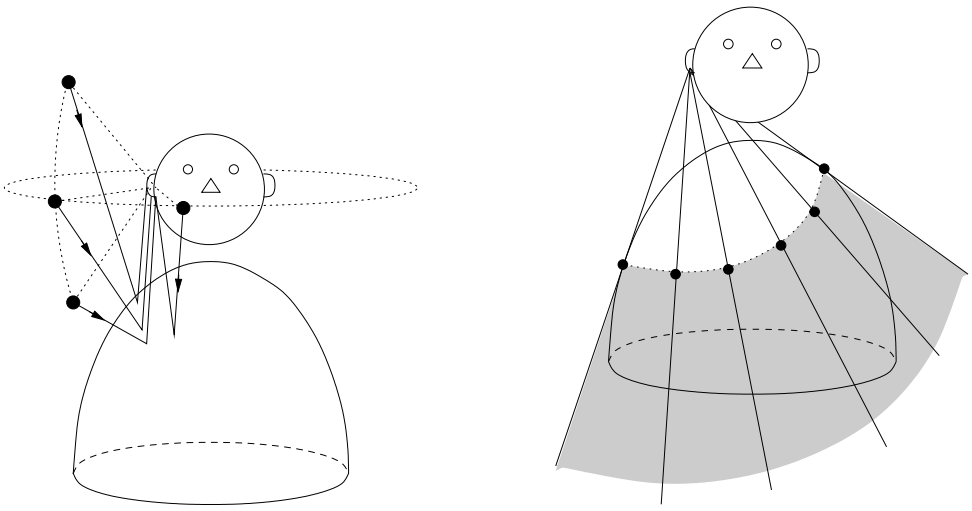
Figure 4.4: *The six pinna resonance modes identified by Shaw (figure reproduced from [158]).*

recomposed using only the first, omnidirectional peak in the HRTF spectrum (corresponding to Shaw's mode 1) coupled with the first two notches yields almost the same localization accuracy as the corresponding measured HRTF. Additional evidence in support of the lowest-frequency notches' relevance is given in [118], which states that the threshold for perceiving a shift in the central frequency of a spectral notch is consistent with the localization blur on the median plane. Also, in [68] the authors judge increasing frontal elevation apparently cued by the increasing central frequency of a notch, and determine two different peak/notch patterns for representing the above and behind directions.

In general, hence, both pinna peaks and notches seem to play an important function in vertical localization of a sound source, but it is difficult without extensive psychoacoustic evaluations to ascertain how importantly these features work as spatial cues. It is also generally considered that a sound source has to contain substantial energy in the high-frequency range for accurate judgement of elevation, because wavelengths longer than the size of the pinna are not affected. One could roughly state that the pinnae have a relatively little effect below 3 kHz.

While the role of the pinna in vertical localization has been extensively studied, the role of torso and shoulders is less well understood. Their effects are relatively weak if compared to those due to the head and pinnae, and experiments to establish the perceptual importance of the relative cues have produced mixed results in general [20, 8, 2]. As can be seen from Fig. 4.5(a), shoulders

measured HRTFs.



(a) Shoulder reflections. (b) Torso shadowing.

Figure 4.5: *Effects of torso and shoulders.*

disturb incident sound waves from all directions other than below at frequencies lower than those affected by the pinna by providing a major additional reflection, whose delay is proportional to the distance from the ear to the shoulder when the sound source is directly above the listener [87], that sums up with the direct sound. Such reflection translates into a series of comb-filter notches in the frequency domain [6]. Complementarily, the torso introduces a shadowing effect for sound rays coming from below (see Fig. 4.5(b)). Torso and shoulders are also commonly seen to perturb low-frequency ITD, even if it is questionable whether they may help in resolving localization ambiguities on a cone of confusion [87].

However, as Algazi *et al.* remarked in [2], when a signal is low-passed below 3 kHz elevation judgement is very poor in the sagittal plane if compared to a broadband source, but proportionally improves as the source is progressively moved away from the median plane, where performance is more accurate in the back than in the front. This result suggests the existence of low-frequency cues for elevation that, although being overall weak, are significant away from the median plane.

4.1.3 Distance cues

Distance estimation of a sound source (see [190] for a comprehensive review on the topic) is even more troublesome than elevation perception. At a first level, when no other cue is available, sound intensity is the first variable that is taken into account: the weaker the intensity, the farther the source should be percepted. Under anechoic conditions, sound intensity reduction with increasing distance can be predicted through the inverse square law: intensity of an omnidirectional sound source will decay of approximately 6 dB for each doubling distance [14]. Still, a distant blast and a whisper at few centimeters from the ear could produce the same sound pres-

sure level at the eardrum. Having a certain familiarity with the involved sound is thus a second fundamental requirement [52].

However, the apparent distance of a sound source is systematically underestimated in an anechoic environment [110]. On the other hand, if the environment is reverberant, additional information can be given by the proportion of reflected to direct energy, the so-called *R/D ratio*, which functions as a stronger cue for distance than intensity: a sensation of changing distance occurs if the overall intensity is constant but the R/D ratio is altered [14]. Furthermore, distance-dependent spectral effects also have a role in everyday environments: higher frequencies are increasingly attenuated with distance due to air absorption effects.

Literature on source direction perception generally lies its foundations on a fundamental assumption, i.e. the sound source is sufficiently far from the listener. In particular, previously discussed azimuth and elevation cues are distance-independent when the source is in the so-called *far field* (approximately more than 1.5 m from the center of the head) where sound waves reaching the listener can be assumed to be plane. On the other hand, when the source is in the *near field* some of the previously discussed cues and HRTF features exhibit a clear dependence on distance. By gradually approaching the sound source to the listener's head in the near field, it was observed that low-frequency gain is emphasized; ITD slightly increases; and ILD dramatically increases across the whole spectrum for lateral sources [24, 23, 21]. The following conclusions were drawn:

- elevation-dependent features are not correlated to distance-dependent features;
- ITD is roughly independent of distance even when the source is close;
- low-frequency ILDs are the dominant auditory distance cues in the near field.

It should be then clear that ILD-related information needs to be considered in the near field, where dependence on distance cannot be approximated by a simple inverse square law.

For small distances the concept of cone of confusion becomes inapplicable. Assuming an acoustically transparent head, i.e. that its effects on the sound are frequency-independent, one can analyze near-field iso-ITD and iso-ILD curves on the horizontal plane. Fig. 4.6 shows such contours spaced every 50 μs for ITD and every 1 dB for ILD. By superposing a given iso-ITD region with the corresponding iso-ILD region and rotating their intersection around the interaural axis we obtain a toroidal solid where ITD and ILD have approximately the same value, known as *torus of confusion* (see Fig. 4.3) [161]. As expected, as the source moves away laterally from the listener tori degenerate into cones of confusion. The same happens when the source is in the vicinity of the median plane.

Removing the previous assumption on the acoustical transparency of the head, strong dependence of ILD on frequency would lead to think of the torus of confusion as a purely abstract concept in presence of a broadband source. However, as [161] argues, the head can be assumed to be acoustically transparent below 500 Hz at least, while for medium and high frequencies

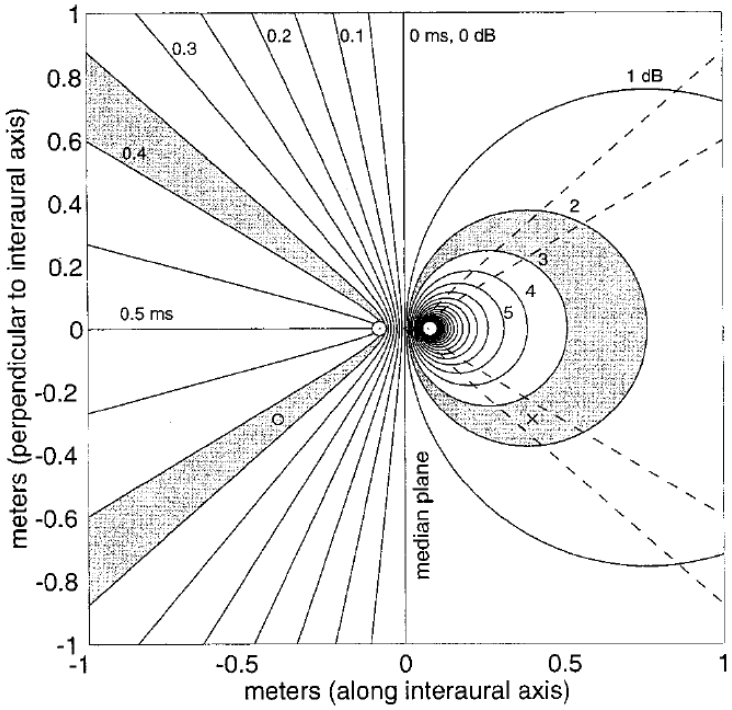


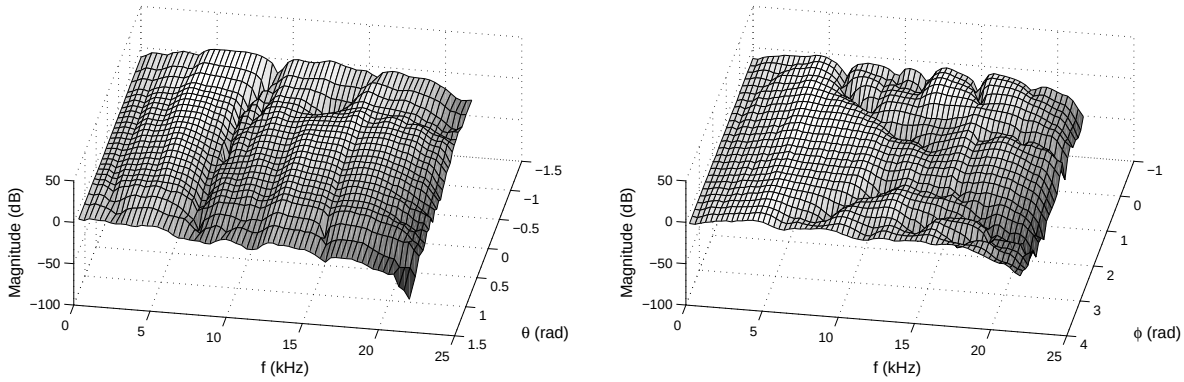
Figure 4.6: *Iso-ITD (left) and iso-ILD (right) contours as a function of spatial location for an acoustically transparent head (figure reproduced from [161]).*

part of the information conveyed in the ILD covaries with the information conveyed in the ITD. Hence, if the source is broadband, combining spatial information in the ILDs in different frequency bands will restrict the source location to the same torus of confusion, since mid- and high-frequency ILDs contain spatial information similar to the information conveyed by ITD.

Finally, it has to be remarked that switching from a static to a dynamic environment where the source moves with respect to the listener and/or *vice versa*, both source direction and distance perception become much eased. The tendency to point towards the sound source in order to minimize interaural differences, even without visual aid, is commonly seen and openly disambiguates any front/back confusion [184]. Active motion helps especially in azimuth estimation and to a lesser extent in elevation estimation [176]. Furthermore, thanks to the *motion parallax* effect, slight translations of the listener's head on the horizontal plane can help discriminating source distance: if the source is near, its angular direction will drastically change after the translation (reflecting itself onto interaural differences), while for a distant source this will not happen.

4.2 HRTF modeling techniques

As should be already clear, the HRTF is a function of four variables: three spatial coordinates and frequency. As Fig. 4.7 depicts, it is a quite complicated function, and it may significantly vary



(a) Azimuth dependence in the horizontal plane. (b) Elevation dependence in the median plane.

Figure 4.7: *Example HRTF magnitude plots.*

from person to person. Actually, the most effective systems for 3-D sound synthesis have large tables of FIR filter coefficients derived from HRIR measurements stored for individual subjects. The desirability of replacing such tables by functional approximations is thus well recognized; in light of this, different approaches to HRTF modeling have been pursued in the literature.

Being unable to factor the HRTF into an azimuth-dependent part and an elevation-dependent part, researchers have applied various filter design, system identification, and neural network techniques in attempts to fit multiparameter models to experimental data (see e.g. [43]). Unfortunately, many of the resulting filter coefficients are themselves rather complicated functions of both azimuth and elevation, and models that have enough coefficients to be effective in capturing individualized directional cues do not provide significant computational advantages.

However, one can argue on a physical basis that a relatively small number of physical parameters could suffice in completely determining the HRTF. This suggests that the intrinsic dimensionality of HRTFs might be small, and that their complexity primarily reflects the fact that we are not viewing them correctly. In the search for simpler representations, several researchers have applied series expansions such as principal component analysis (PCA) to the log magnitude of the HRTF [88], or to the complex HRTF itself [29], or again to the HRIR [74]; or such as surface spherical harmonics (SSH) to the magnitude and unwrapped phase of the HRTF, and to the HRIR [47]. These analyses produce each a directionally-independent set of basis functions and a directionally-dependent set of weights for combining the basis functions. In all of these cases, it has been found that a relatively small number of basis functions are sufficient to represent the HRTF/HRIR. Thus these techniques have proved to be a valuable tool for studying the characteristics of the data. Furthermore, it may be possible to relate them to anthropometric measurements and to scale them to account for individual differences. Unfortunately, series expansions still require significant computation for real-time synthesis when head or source motion is involved because weights are relatively complex functions of azimuth and elevation that must

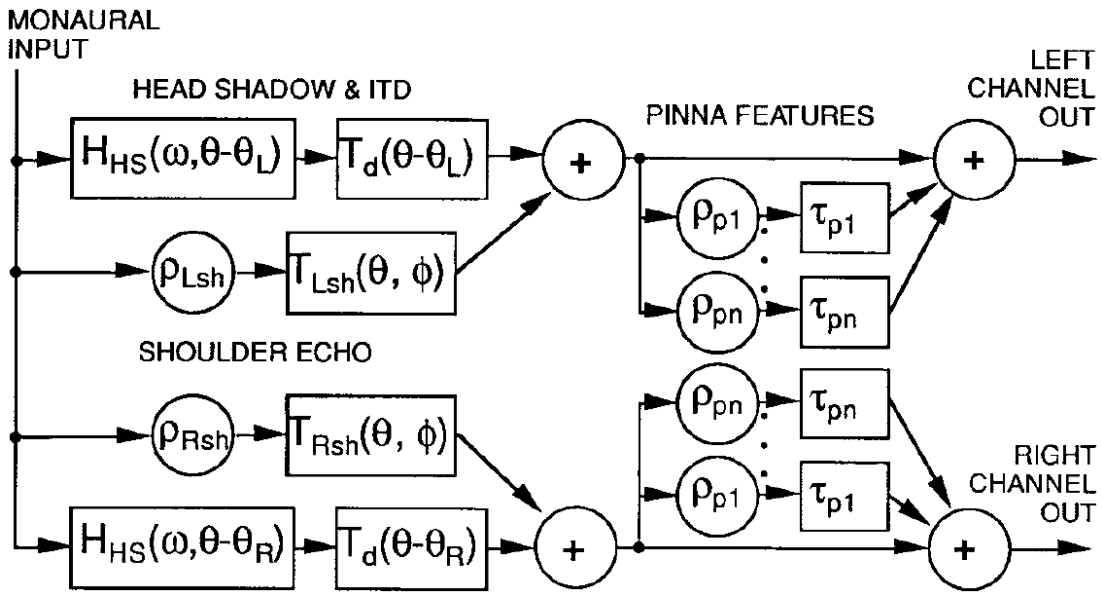
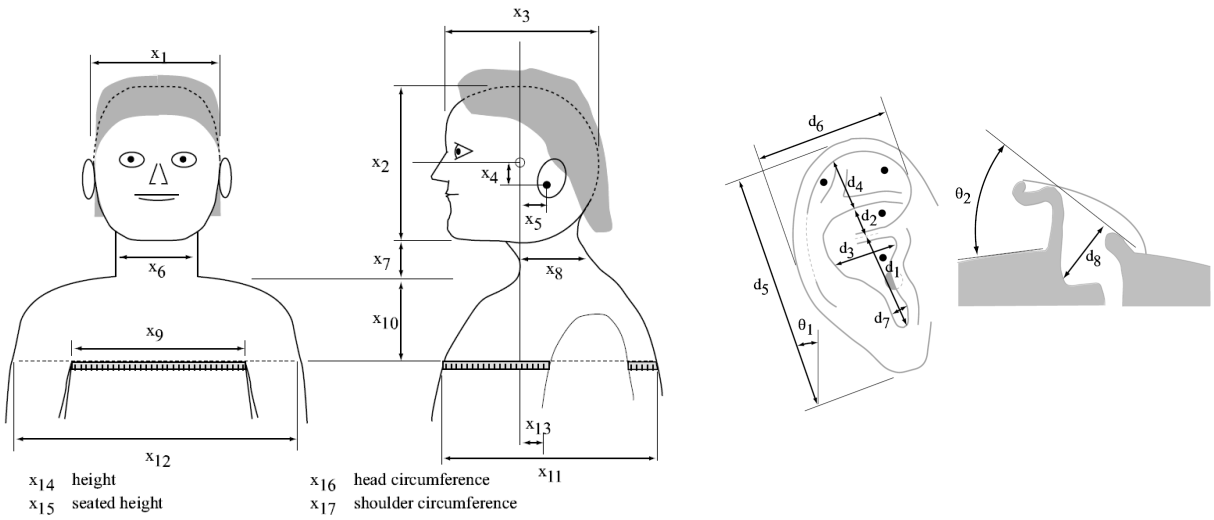


Figure 4.8: *Brown and Duda's complete structural HRTF model (figure reproduced from [20]).*

be tabulated, and because resynthesized HRTFs must be inverse-Fourier transformed to obtain the corresponding HRIRs needed to process the signals.

As one possible alternative to rendering approaches based on directly measured HRTFs or on forementioned models, the use of structural models represents an attractive solution to synthesize individual HRTFs or build an enhanced generalized HRTF model. In structural models the contributions of the listener's head, pinnae, shoulders and torso to the HRTF are isolated and arranged in different subcomponents each accounting for some well-defined physical phenomenon. The linearity of these contributions allows reconstruction of the global HRTF from a proper combination of all the considered effects [5]. Furthermore, room effects can also be incorporated into the rendering scheme: in particular, early reflections from the environment can be convolved with the pinna model, depending on their incoming direction. The choice of the room model is flexible to the specific application and not only directed at reproducing a realistic room behaviour, but also at introducing sound externalization [15]. However, the room model is not strictly correlated to the HRTF model and will not be treated in this thesis. A synthetic block scheme of a generic binaural audio system based on a structural model was depicted back in Fig. 1.2. Similarly, Fig. 4.8 reports a well-known complete and detailed structural model [20], some of whose blocks will be discussed in a while.

Above all, structural modeling open the doors for an interesting form of content adaptation to users' anthropometry. In fact, parameters of the rendering blocks sketched in Fig. 1.2 and Fig. 4.8 can be estimated from real data, fitted, and finally related to anthropometric measurements. Still, given the great variety of head and pinna shapes amongst the entire world population, fixing a subset of anthropometric parameters that fully characterize a specific listener is a challenging



(a) Head and torso parameters.

(b) Pinna parameters.

Figure 4.9: *A set of 27 anthropometric parameters for the head, torso, and pinna (figures reproduced from [7]).*

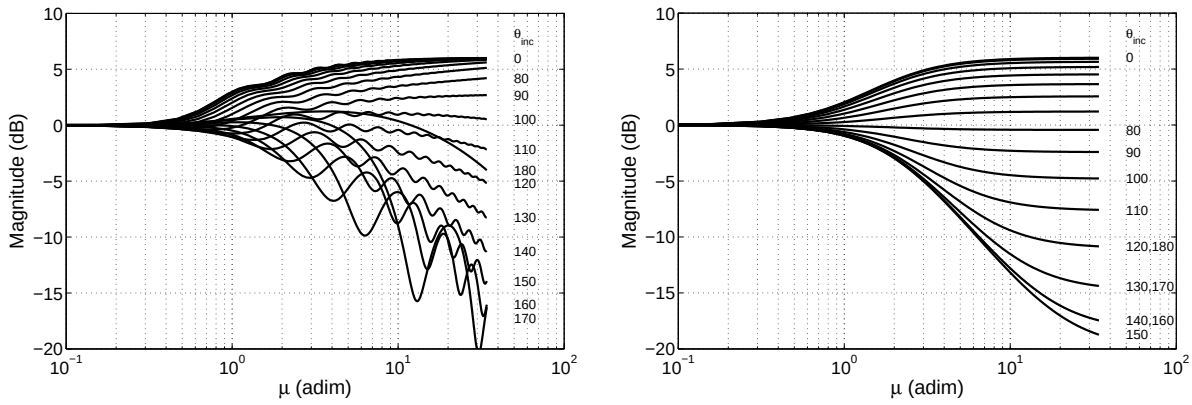
task. Following the studies by Genuit back in the early eighties on features of the human body that contribute to HRTF characterization, Algazi *et al.* [7] proposed 27 different parameters (17 for head and torso and 10 for the pinna, see Fig. 4.9) that can be used for HRTF fitting using regression methods or other techniques. In this way, a generic structural HRTF model can be adapted to a specific listener, allowing further increase of the quality of audio experience thanks to an enhanced realism of the sound scene.

4.3 Head and torso models

The following two Sections build up a short review of known head, torso, and pinna models that can be found in the literature along with some results and comments.

4.3.1 The spherical head model

As already mentioned in the previous Section, presence of the head implies diffraction of the sound wave around it, and a screening effect on high-frequency components. The simplest model of the head that can be found in the literature is that of a rigid sphere [169]. Within the assumption of an infinitely distant source from the center of the head, the response related to a fixed observation point on the sphere's surface can be described by means of the following transfer



(a) Exact solution.

(b) First-order approximation (after [20]).

Figure 4.10: *Magnitude response of a sphere for an infinitely distant source.*

function, based on Lord Rayleigh's diffraction formula:[3]

$$H(\mu, \theta_{\text{inc}}) = \frac{1}{\mu^2} \sum_{m=0}^{\infty} \frac{(-i)^{m-1}(2m+1)P_m(\cos\theta_{\text{inc}})}{h'_m(\mu)}, \tag{4.2}$$

where θ_{inc} is the incidence angle, i.e. the angle between rays connecting the center of the sphere to the source and the observation point, and μ is the normalized frequency, defined as

$$\mu = f\frac{2\pi a}{c}, \tag{4.3}$$

where *c* is the speed of sound[4] and a is the sphere radius, possibly the only subject-dependent parameter of the model.

Fig. 4.10(a) shows the magnitude of the transfer function on a dB scale against normalized frequency for 19 different values of θ_{inc}. By analyzing such plot the following considerations can be drawn:

- independently of the incidence angle, the magnitude response is unitary up to $\mu = 1$, which for a standard 8.75 cm-radius spherical head [65] corresponds to about 625 Hz;

- in case of normal incidence ($\theta_{\text{inc}} = 0°$) a high-frequency gain of 6 dB, equal to a double SPL with respect to a free-field response, is observed;

[3]Here P_m and h_m represent, respectively, the *Legendre polynomial* of degree *m* and the *m*th-order *spherical Hankel function*. h'_m is the derivative of h_m with respect to its argument.

[4]Speed of sound varies according to the medium and atmospheric conditions in which the wave travels; in dry air at $20°$C it is equal to 343.2 m/s.

- gain tends to decrease with increasing incidence angle: the response for $\theta_{\text{inc}} = 100^\circ$ is almost flat, and as the source is further moved towards the contralateral[5] side of the head the SPL is more and more attenuated while wider and wider oscillations due to wave propagation around the sphere along different directions are introduced in the high-frequency spectrum;

- however, the minimum response does not correspond to $\theta = 180^\circ$: if the source is antipodal to the observation point, waves that travel in different directions around the sphere constructively combine at the observation point producing the so-called *bright spot*.

A first-order approximation of the transfer function produced by Eq. (4.2) was proposed by Brown and Duda [20]. It is a single-pole, single-zero minimum-phase analog filter of the form

$$H(s, \theta_{\text{inc}}) = \frac{1 + \alpha\tau s}{1 + \tau s}, \quad 0 \leq \alpha(\theta_{\text{inc}}) \leq 2 \tag{4.4}$$

where

$$\tau = \frac{2a}{c}, \tag{4.5}$$

and

$$\alpha(\theta_{\text{inc}}) = 1 + \frac{\alpha_{\text{min}}}{2} + \left(1 - \frac{\alpha_{\text{min}}}{2}\right) \cos\left(\frac{\theta_{\text{inc}}}{\theta_{\text{min}}}\pi\right) \tag{4.6}$$

is a coefficient that controls the asymptotic high-frequency gain: if $\alpha = 2$, a 6-dB boost at high frequencies is introduced, while if $\alpha < 1$ high frequencies are cut down. Brown and Duda claimed that parameters $\alpha_{\text{min}} = 0.1$ and $\theta_{\text{min}} = 150^\circ$ provide a good overall match to the ideal solution shown in Fig. 4.10(a) and an attenuated bright spot. The magnitude curves resulting from this parameter choice are reported in Fig. 4.10(b).

Typically, in spherical models the two observations points (i.e. the ear canals) are assumed to be diametrically opposed, such that a direct correspondence between incidence angles ($\theta_{\text{inc}}^{(l)}$ and $\theta_{\text{inc}}^{(r)}$ for the right and left ears, respectively) and the azimuth angle θ exists in the horizontal plane. As an alternative model, the spherical-head-with-offset-ears model described in [2] was obtained by displacing the ears backwards and downwards by a certain offset, introducing a nonlinear mapping between ($\theta_{\text{inc}}^{(l)}$, $\theta_{\text{inc}}^{(r)}$) and θ in the horizontal plane and elevation dependency on a cone of confusion. Such model was found to provide a good approximation to elevation-dependent patterns both in the frequency and time domains, particularly replicating a peculiar X-shaped pattern along elevation (due to the superposition of two different propagation paths around the head) commonly seen in measured contralateral HRIRs.

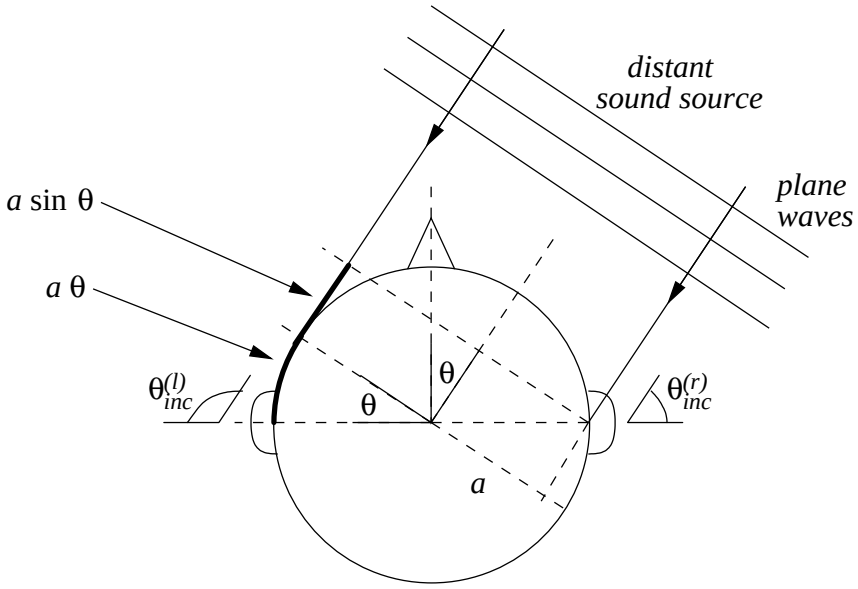


Figure 4.11: *ITD computation for a spherical head model on the horizontal plane.*

4.3.2 ITD and anthropometry

The filter structure in Eq. (4.4) introduces a group delay

$$\tau_g = \frac{a}{2c}(1-\alpha) \tag{4.7}$$

at low frequencies which is not sufficient in accounting for the correct ITD alone. However, ITD information can be hived off from all the remaining information contained in the HRTF through different possible approximations. The most popular amongst them is based on the simplified spherical geometry reported in Fig. 4.11 [17]:

$$\text{ITD} = \frac{a(\sin\theta + \theta)}{c}, \tag{4.8}$$

for a far-field sound source placed in the horizontal plane, taking the right ear canal as reference. Indeed, to a first degree of approximation, in the interaural-polar coordinate system the ITD is frequency-independent and depends on azimuth θ alone [2]. Eq. (4.8) can be split into two delay components, $\tau_{\text{ITD}}^{(l)}$ for the left ear and $\tau_{\text{ITD}}^{(r)}$ for the right ear, by applying Woodworth and Schlosberg's frequency-independent formula [188],

$$\tau_{\text{ITD}} = \begin{cases} -\frac{a}{c}\cos\theta_{\text{inc}} & \text{if } 0 \le |\theta_{\text{inc}}| < \frac{\pi}{2} \\ \frac{a}{c}\left(|\theta_{\text{inc}}| - \frac{\pi}{2}\right) & \text{if } \frac{\pi}{2} \le |\theta_{\text{inc}}| < \pi \end{cases}; \tag{4.9}$$

[5]The source is positioned on the *ipsilateral* side of the sphere if the ray-traced sound wave normally meets its surface on a point belonging to the hemisphere that has its pole in the observation point; *contralateral* if the wave meets the sphere on a point belonging to the opposite hemisphere.

that is, ITD $= \tau_{\text{ITD}}^{(l)} - \tau_{\text{ITD}}^{(r)}$.

Comparison of predicted ITD against measured ITD reveals a good match in the high-frequency range. However, ignoring the already mentioned 50% increase of ITD at low frequencies with regard to high frequencies could be detrimental to correct ITD estimation which highly relies on the lowest frequency range information. Still, this is not a big deal as long as $\tau_{\text{ITD}}^{(l)}$ and $\tau_{\text{ITD}}^{(r)}$ are modeled each as a frequency-independent delay line and coupled with the group delay τ_g induced by the head filter: as a matter of fact, the sum of the two delays at $\theta_{\text{inc}} = 0°$ provides exactly the required 50% additional low-frequency delay [20].

Note that Eq. (4.4) is a function of the head radius a. This is a critical parameter: for instance, a sphere having the same volume of the head approximates its behaviour much better than a sphere with diameter equal to the interaural distance [85]. Hence, in order to fit the spherical head filter model to a specific listener, parametrization of a on the subject's anthropometry should be performed. In [3] Eq. (4.8) is compared to a number of real ITD measurements for a specific subject, and the best head radius for that subject is defined as the value that corresponds to the minimum mean least squares distance between the two estimates for different azimuth angles on the horizontal plane. Then, a linear model for estimating the head radius given the three most relevant anthropometric parameters for the head, i.e. width, height, and depth (parameters x_1, x_2, and x_3 in Fig. 4.9(a), respectively),

$$a_{\text{opt}} = w_1 x_1 + w_2 x_2 + w_3 x_3 + b, \tag{4.10}$$

is fitted to ITD-optimized radii of 45 different subjects through linear regression, yielding optimal weights

$$w_1 = 0.26, \quad w_2 = 0.01, \quad w_3 = 0.09, \quad b = 3.2 \text{ cm}. \tag{4.11}$$

This result highlights how head height is a relatively weak parameter in ITD definition with respect to head width and depth.

To sum up, the spherical model of the head provides an excellent approximation to the magnitude of a measured HRTF. Although facial features contribute to HRTF coloring in a different way across subjects, a recent study [113] highlighted how there is roughly no difference between FDTD-simulated magnitude responses on an unmodified KEMAR head and on a head shape morphed towards a sphere in the median plane. However, the spherical model is far less accurate in predicting ITD, being the latter actually not constant around a cone of confusion, but variable by as much as 18% of the maximum interaural delay [40]. In other words, ITD is a function of elevation as well as azimuth. Elevation dependence can be integrated in Eq. (4.8) by introducing a further term which takes into account the decrease in ITD as the source moves away from the horizontal plane:

$$\text{ITD} = \frac{a(\sin\theta + \theta)}{c} \cos\phi. \tag{4.12}$$

Indeed, a simple cosine dependence of the elevation angle was found to be accurate enough for simulation purposes [150].

Following an alternative approach, Duda *et al.* [40] managed to improve ITD estimation accuracy by considering an ellipsoidal head model that can account for the ITD variation and be adapted to individual listeners. Despite the good result, the analytical solution for the ITD is far complicated, and no explicit model for the ellipsoid-related transfer function was proposed. Conversely, models for the head as a prolate spheroid were studied in [124, 77] as the sole alternative analytical model to a sphere. Although adding nothing new in the ITD's point of view, comparison of spheroidal HRTFs against spherical HRTFs revealed a different behaviour in head-induced low-frequency ripples in the magnitude response at the contralateral ear, which is closer to responses of a KEMAR head for the spheroidal case [78]. Still, this model has been very little studied, and consistent advantages over the spherical model have not been made clear.

4.3.3 Inclusion of distance dependence

When the assumption of an infinitely distant source does not hold, dependence on distance can no longer be ignored. Having defined the normalized distance to the source ρ as the ratio between the absolute distance from the center of the sphere and the sphere radius

$$\rho = \frac{r}{a}, \tag{4.13}$$

the pressure on the spherical surface caused by a sinusoidal point source at an arbitrary distance greater than the sphere radius can be evaluated by means of the following function [134]:

$$H(\rho, \mu, \theta_{\text{inc}}) = -\frac{\rho}{\mu} e^{-i\mu\rho} \sum_{m=0}^{\infty} (2m+1) P_m(cos\theta_{\text{inc}}) \frac{h_m(\mu\rho)}{h'_m(\mu)}, \tag{4.14}$$

for each $\rho > 1$.

Different considerations on HRTF behaviour for changing distances can be drawn by analyzing Fig. 4.12, which reports the magnitude of the new transfer function for $\theta_{\text{inc}} = 0^\circ$ and $\theta_{\text{inc}} = 150^\circ$, where the mean frequency response has maximum and minimum gain, respectively, and six different normalized distances:

- as the source approaches the sphere (ρ tends to 1) the response on the ipsilateral side increases, while the response on the contralateral side decreases almost exponentially on a dB scale;
- as frequency rises, the difference between magnitude responses for different distances slightly decreases on the ipsilateral side and slightly increases on the contralateral side;
- combination of the previous two points motivates the dramatical ILD boost at small distances across the whole frequency range;
- however, if absolute gain is overlooked, the responses for a same incidence angle maintain a common behaviour along the frequency axis.

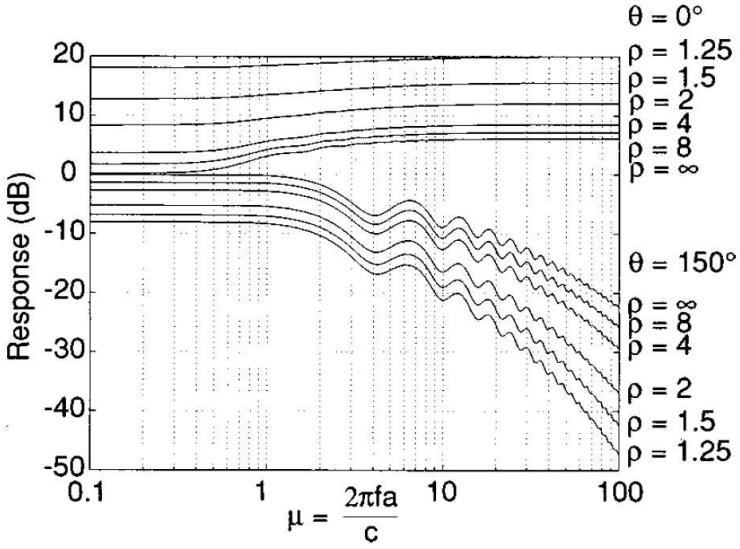


Figure 4.12: *Effect of distance on the magnitude response of a spherical head (figure reproduced from [42]).*

No low-order filter model has actually been proposed to approximate Eq. (4.14). Furthermore, the infinite sum does not allow construction of a finite algorithm to evaluate the function, while computation of spherical Hankel functions and Legendre polynomials requires high computational costs. The solution to all of these shortcomings is provided in [42] by means of a recursive algorithm where the latter functions are developed iteratively, allowing a relatively fast evaluation. The resulting equation becomes

$$H(\rho,\mu,\theta_{\text{inc}}) = \frac{\rho}{i\mu}e^{-i\mu}\sum_{m=0}^{\infty}(2m+1)P_m(cos\theta_{\text{inc}})\frac{Q_m(\frac{1}{i\mu\rho})}{\frac{m+1}{i\mu}Q_m(\frac{1}{i\mu})-Q_{m-1}(\frac{1}{i\mu})}, \tag{4.15}$$

where complex polynomials P_m and Q_m are recursively computed through the following equations:

$$Q_m(z) = -(2m-1)zQ_{m-1}(z) + Q_{m-2}(z), \tag{4.16}$$

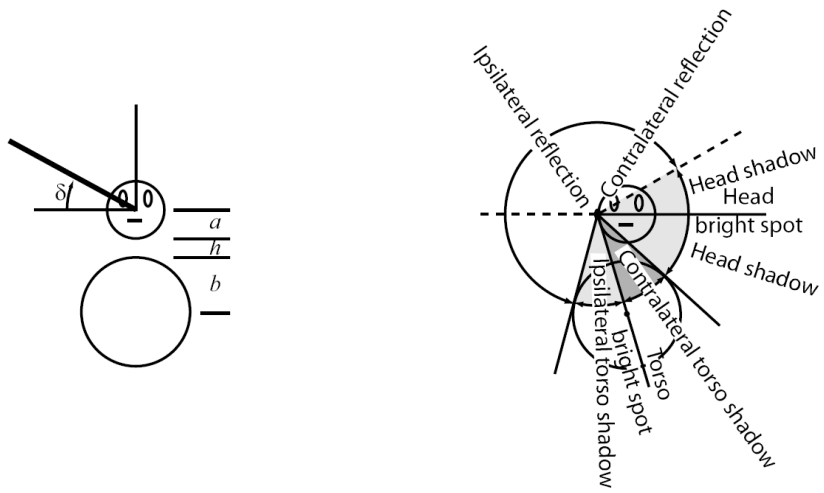
$$P_m(x) = \frac{2m-1}{m}xP_{m-1}(x) - \frac{m-1}{m}P_{m-2}(x), \tag{4.17}$$

having fixed initial conditions

$$Q_0(z) = z, \quad Q_1(z) = z - z^2, \quad P_0(x) = 1, \quad P_1(x) = x. \tag{4.18}$$

Iteration on m stops when the fractional change falls below a user-supplied threshold for two successive terms, evading the infinite sum in Eq. (4.15). The code of the recursive algorithm can be found in [42].

As already mentioned, while the magnitude of the ILD increases dramatically at the closest distances, ITD generally increases by no more than $10\%-12\%$ [24]. A similar geometry to that



(a) Frontal view of the model. (b) Zones for the right-ear response.

Figure 4.13: *The snowman head-and-torso model (figures reproduced from [6]).*

of Fig. 4.11 that takes into account near-field effects is reported in [42], yielding a closed form for ITD at whatever distance on the horizontal plane. Still, it was conjectured that such small changes in the ITD probably do not provide significant information about distance.

4.3.4 Inclusion of the torso

Similary to the head, in previous works the torso has been approximated by a sphere too. Coaxial superposition of the two spheres of radius a and b, respectively, separated by a distance h that accounts for the neck, gives birth to the known *snowman model* [6] represented in Fig. 4.13(a). The far-field behaviour of the snowman model has been studied in the frontal plane both by direct measurements on two rigid spheres and by computation through multipole reexpansion, a method that extends the classical solution for a single sphere in Eq. (4.2) to scattering by multiple spheres [4]. Taking Fig. 4.13(b) as reference, such studies revealed that:

- the snowman model exhibits two bright spots, one associated to the head and one due to the torso;
- in the reflection zone the response is dominated by the comb-filter patterns produced by torso reflections;
- in the head shadow zone no relevant torso effects are seen;
- in the contralateral torso shadow zone the combined result of head shadow and torso shadow produces complicated notch patterns and significant high-frequency loss;

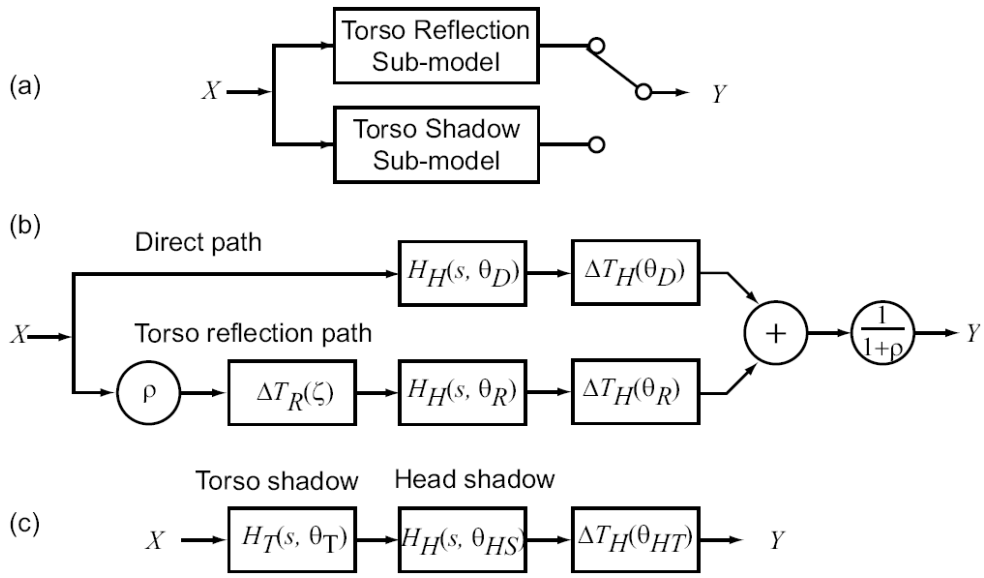


Figure 4.14: *The snowman head-and-torso filter model (figure reproduced from [6]). (a) Major components; (b) the torso reflection sub-model; (c) the torso shadow sub-model.*

- in the ipsilateral torso shadow zone responses are more or less flat.

What's more relevant in this context, a structural head-and-torso model has been derived from the snowman model [6]. Its structure, reported in Fig. 4.14, distinguishes the two cases where the torso acts as a reflector or as a shadower, switching between the two filter sub-structures (b) and (c) as soon as the source leaves or enters the torso shadow zone, respectively. The torso reflection sub-model includes:

- a direct component that arrives from the direction of the source (incidence angle θ_D), gets diffracted by a spherical head filter, and time delayed; and

- a reflected component that arrives at the head from a different direction (incidence angle θ_R) after being reflected from the torso with reflection coefficient ρ, additionally delayed by ΔT_R because of the longer reflection path.

At the end of the filter chain, a scale factor allows continuity when switching between the two sub-models.

Conversely, the torso shadow sub-model has an unique path that includes two spherical filters, one for the head and one for the torso, and the usual time delay, all appropriately tuned to the relative incoming sound direction. All of the spherical filters H_H and H_T in the model are of the form described in Eq. (4.4), with filter H_T parameterized on radius b instead of a, whereas the time delays ΔT_H are of the form described in Eq. (4.9). Reflection coefficient ρ is assumed constant for simplicity, and ΔT_R is analytically derived by ray-tracing arguments.

The frequency response of the snowman filter model has a definitely similar behaviour to the analytical solution, i.e. strong ripples on the ipsilateral side and significant shadowing on the contralateral side. The only significant difference lies in the torso bright spot, which is absent in the filter model. However, since a human torso is not spherical, such a behaviour is actually not expected in measured HRTFs; furthemore, torso effects at very low elevations greatly depend on the subject's posture.

Additionally to the spherical model, an ellipsoidal model for the torso was also studied in combination with the usual spherical head. This was done either by ray-tracing analysis [2] or through the BEM [4]. Such model is able to account for different torso reflection patterns and further breaks up the symmetry that leads to the torso bright spot. Listening tests confirmed that this HAT approximation and the corresponding measured HRTF gave similar results, showing larger correlations away from the median plane. Also, the ellipsoidal torso can be easily customized for a specific subject by defining control points for its three axes directly on the subject's torso [4], whereas a spherical torso is hardly personalizable.

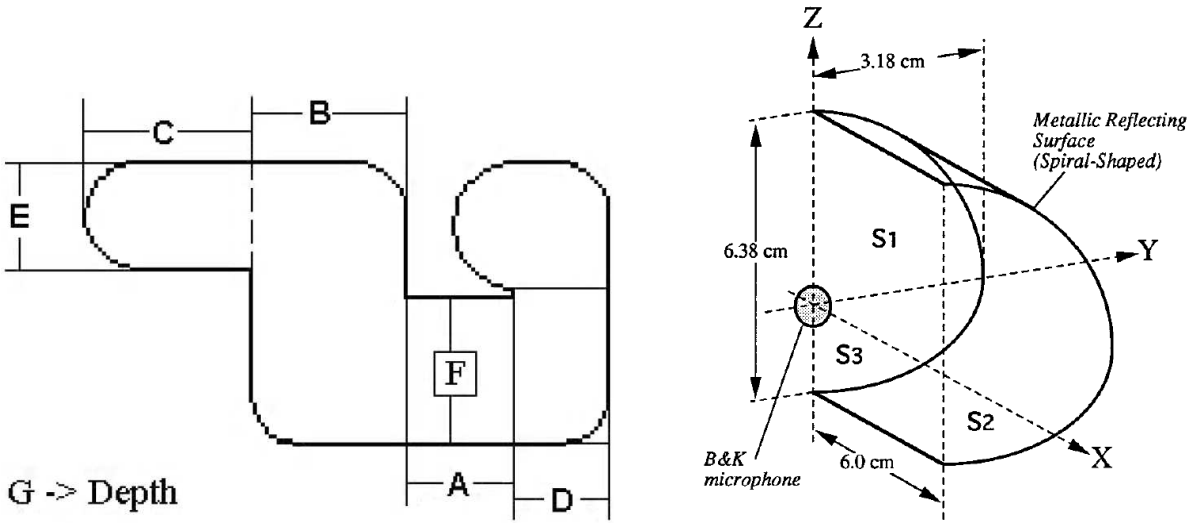
In conclusion, the addition of either a spherical or an ellipsoidal torso to the spherical head brings the overall behaviour of the model closer to that of a real HRTF.

4.4 Pinna models

Different physical and structural models of the pinna have been proposed in the past. The former class aims at recreating the physics lying behind the production of the forementioned spectral patterns either by approximating the pinna as a cavity configuration or as a reflecting surface. Examples of the first approach are the simple geometric (cylindrical or rectangular) concha/pinna models by Teranishi and Shaw [174], which progressively led to Shaw's notable flange-and-cavity model in Fig. 4.15(a) [157], and the recent "three-step" model by Takemoto *et al.* [172], simulated through the Finite-Difference Time Domain (FDTD) method, which qualitatively recreates typical peak/notch patterns along the median plane. The second approach is best exemplified by the rigorous diffraction/reflection model by Lopez-Poveda and Meddis [96] based on diffraction theory applied to both a half-cylinder shape (see Fig. 4.15(b)) and a realistic concha shape. Despite the objectively good approximations that physical models can provide, their main drawback is the difficulty in introducing effective customization to the physical structure.

4.4.1 Time-domain structural pinna models

The history of pinna structural models, a new one of which will be presented in this thesis, begins with Batteau's reflection theory [12]. According to such theory, high-frequency components which arrive at the listener's ear are typically reflected by the concha wall and rim, provided that their wavelength is small compared to the pinna dimensions. Due to interference between the



(a) Shaw's flange-and-cavity pinna model.

(b) Lopez-Poveda's concha model.

Figure 4.15: *Two notable physical models of the pinna and concha (figures reproduced from [158, 96]).*

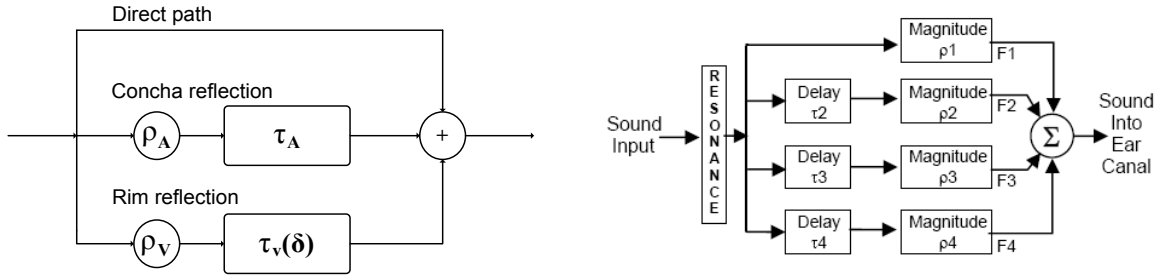
direct and reflected waves, sharp notches can be observed in the incoming sound's spectrum at high frequencies with a periodicity of $1/\tau_i$, where τ_i is the time delay of the i-th reflection.

Following Batteau's observations, Watkins [181] designed a very simple double-delay-and-add time-domain model of the pinna (see Fig. 4.16(a)) where the considered reflection paths are characterized by fixed reflection coefficients ρ_A and ρ_V, a fixed time delay $\tau_A = 15\mu$s and an elevation-dependent time delay τ_V calculated from empirical data. The fit with experimental data was found to be reasonably good; still, beside considering a very limited amount of reflections,

- no method for extracting parametric time delays and gain factors was proposed;
- fixed reflection coefficients overestimate the effective number of notches in the spectrum;
- simple delay-and-add approximations were proven to be inadequate to predict both the absolute position of the spectral minima and the relative position between them [96];
- the model lacks the description of pinna resonant modes: since pinna cavities act as resonators the frequency content of both the direct and the reflected sound waves is significantly altered.

Nonetheless, the pioneering novelty of such model is undisputed.

Watkins's model has accordingly been improved by Faller *et al.*, whose model [49], reported in Fig. 4.16(b), consists in a reflection structure represented by four parallel paths, each modeled by a time delay τ_i and a magnitude factor ρ_i, cascaded to a low-order resonator block. The model



(a) Watkins's model. (b) Faller's model.

Figure 4.16: *Time-domain structural models of the pinna (figure (b) reproduced from [49]).*

parameters are fitted by decomposing each specific measured HRIR into four scaled and delayed damped sinusoidal (DDS) components using a procedure based on the second-order Steiglitz-McBride (STMCB) algorithm, and associating the delay and scaling factor of each component to the corresponding parameters of its associated path in the model. A more recent version of the model [48] exploits an adaptation of the Hankel Total Least Squares (HTLS) decomposition method instead of the STMCB algorithm to extract a heuristic number of DDSs from measured HRIRs. Multiple regression analysis was used in order to link the former model parameters to eight measured anthropometric features [62]. Unfortunately, as well as providing no cloudless evidence of the physics behind the scattering phenomenon (no clear relation between model parameters and human anthropometry was explicitly found), the considered measures can only be acquired through the use of a 3-D laser scanner. Regardless of such particular concerns, this work surely endorses the pinna model as a "resonance-plus-delay" architecture.

4.4.2 Frequency-domain structural pinna models

Despite the intuitive nature of multipath HRIR structures, poor temporal resolution of the human auditory system has led to a progressive abandon of these models [149]. A different approach for reflection modeling, acting both in the time and frequency domains, was pursued by Raykar *et al.* [137]. Robust digital signal processing techniques are used here to extract the frequencies of the spectral notches due to the pinna alone: first the autocorrelation function of the HRIR's windowed LP residual is computed; then, frequencies of the spectral notches are found as the local minima of the group-delay function of the windowed autocorrelation. This procedure is reported in Fig. 4.17.

Spectral peaks are extracted in parallel by means of a linear prediction analysis, yielding results which match quite well the pinna resonant modes reported by Shaw, further justifying the "resonance-plus-delay" approach. What's more, the authors advanced a ray-tracing argument (borrowed from [68]) to show that the estimated spectral notches, each assumed to be caused by its own reflection path, are related to the shape of the concha and crus helias, at least on the

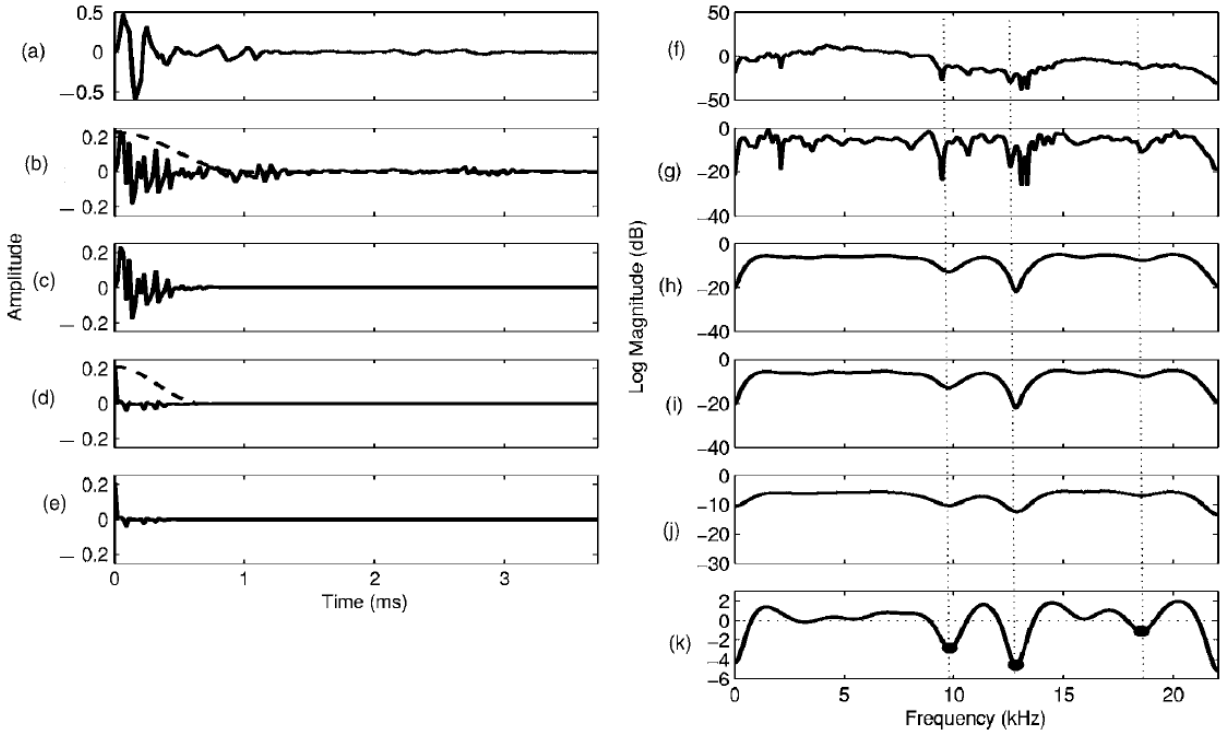


Figure 4.17: *Signal processing steps for extracting the pinna spectral notch frequencies. (a) original HRIR signal; (b) 12th order LP residual; (c) windowed LP residual; (d) autocorrelation function of the windowed LP residual; (e) windowed autocorrelation function; (f), (g), (h), (i), and (j) log magnitude spectrum corresponding to signals in (a), (b), (c), (d), and (e) respectively; (k) group-delay function of the windowed autocorrelation function. The local minima in the group-delay function, zero thresholded, are shown (figure reproduced from [137]).*

frontal side of the median plane. However, there is no clear one-to-one correspondence between pinna contours and notch frequencies in the available plots.

Finally, the approach followed by Satarzadeh *et al.* [149] approximates the pinna behaviour at elevations close to zero degrees through a structural model composed of two low-order bandpass filters and one comb filter, which respectively approximate the two strongest resonances (Shaw's resonance modes 1 and 4) and one main reflection. The two second-order bandpass filters and the comb filter are interconnected as in Fig. 4.18, the latter taking the form

$$H_{comb} = 1 + \rho_r \exp(-st_d), \tag{4.19}$$

where ρ_r is a frequency-dependent reflection coefficient which strongly attenuates low-frequency notches, coming over one of Watkins's model forementioned limitations, and t_d is the time delay of the considered reflection. Here t_d is strictly correlated to the frequency of the comb filter's first tooth, f_0, estimated from the spacing of consecutive notches in the measured spectrum. The model was proved to have sufficient adaptability to fit both rich and poor real notch patterns.

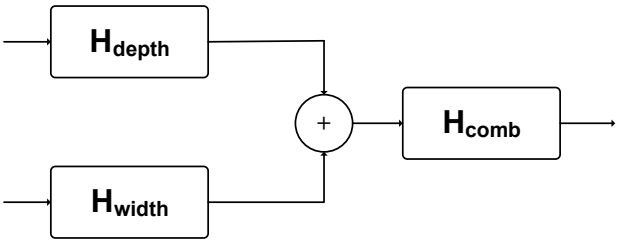


Figure 4.18: *Satarzadeh's pinna filter model.*

What's more relevant, encouraging correspondences with anthropometry were found in this work: depending on whether the reflection coefficient is positive or negative, the distances inferred from ray tracing put the point of reflection either at the back of the concha or at the edge of the rim. In addition, a cylindrical approximation to the concha is exploited in this work for fitting the model parameters to anthropometric quantities. Specifically, depth and width of the cylinder uniquely define the first resonance, while the second resonance is thought to be correlated to the main reflection's time delay, depending on whether the concha or the rim is the significant reflector. Though the anthropometric significance of resonance parameters is not robust, Satarzadeh claimed that if the pinna has an approximately cylindrical shaped concha and a structure with a dominant reflection area (concha or rim), such an anthropometry-based filter provides a good fit to experimental measurements.

Still, the most severe limitation of Satarzadeh's model is that no directions of the sound wave other than the frontal one are considered. Moreover, the presence of an unique reflection (and thus a single delay-and-add approximation) limits the generality of the representation. Nonetheless it represents, in my humble opinion, the only valuable anthropometry-based pinna model available up to date.

Chapter 5

Spherical Transfer Functions and Distance Modeling

Let's consider a dynamic scenario where the listener is free to move his/her head with respect to the virtual source to be rendered, and *vice versa*. It is clear that real-time computation of HRTFs is needed in order to track these movements with enough reactivity, possibly avoiding any discontinuity in the resulting rendered sound. Furthermore, the possibility of having to simulate a complex acoustic environment that includes several independent sound sources, and/or reflections coming from the environment, has to be taken into account.

Relatively simple HRTF-like filter structures for sources in the far field have been proposed to date (e.g., Brown and Duda's first-order filter, see Eq. (4.4)). These turn out to be impracticable in the near field, having no parametrization on source distance. Moreover, point-to-point real-time evaluation of Eq. (4.14) using the algorithm found in [42] is computationally still too expensive. As a consequence, a proper approximation to distance rendering on the spherical head model has to be introduced in order to grant a faster computation.

In this Chapter such an approximation is used to represent a collection of sample analytical responses. The earless head of the listener is conceptually isolated and treated as a rigid sphere; therefore its transfer function will be referred to as *spherical transfer function*, or STF, throughout the chapter. Furthermore, focus is put on sources located in the near field, for which real-time computation of HRTFs turns out to be troublesome. In Section 5.1 I make use of a well known powerful analytic tool, namely Principal Component Analysis (PCA), in order to look for common trends and possible systematic variability in a set of STFs. Then, in Section 5.2 the indications given by PCA open the door to a deeper analysis of distance rendering, which yields the novel low-order model presented in Section 5.2.3.

The work presented in Section 5.1 of this Chapter was published in [163]. The remaining sections refer to a still unpublished work.

5.1 Spherical transfer functions and PCA

Principal Component Analysis is a widely used procedure which makes use of linear combinations to reduce the dimensionality of an input data set. Its main goal is to provide an efficient representation of a set of correlated measures – in this instance, a set of vectors. PCA has already been used in previous works concerning HRTF modeling [29, 88], with the initial data set consisting in magnitude responses collected from a set of measured HRTFs. However, instead of applying the technique to experimental data, in this work it will be exploited to investigate possible trends in a collection of STF magnitudes sampled from Eq. (4.14) on a discrete set of frequencies. It will be shown that, thanks to the little correlation between source distance and frequency arising from the PCA analysis, distance dependence in STFs can be decoupled from far-field diffraction effects during the rendering process.

5.1.1 Principal Component Analysis

Without delving into deep technicalities (which can be found in [41]), in this context suffice it to say that given a set of n real-valued vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each of dimension δ, and defining its *covariance matrix* $\mathbf{S}$ as

$$\mathbf{S} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^t, \tag{5.1}$$

it can be seen that the best p-dimensional representation (with $p \leq \delta$) of the data set is obtained by taking as basis vectors the p eigenvectors of $\mathbf{S}$ that correspond to the p largest eigenvalues.[1] Each vector $\mathbf{x}_k$ is then projected onto the space defined by the basis vectors as follows:

$$\mathbf{a}_k = \mathbf{C}^t \mathbf{x}_k, \tag{5.2}$$

where $\mathbf{C}$ is a matrix, the columns of which are the basis vectors.

We call principal component the set of weights $\{a_{ki}\}$, $k = 1, \ldots, n$, associated to basis vector i. Obviously the number of principal components is equal to the number of basis vectors; these are conventionally labeled in such a way that the first component, PC_1, is the one that captures the direction along which the original data retains the maximum variability, while the following components $PC_2, \ldots, PC_p$ reflect increasingly smaller variations. An example of PCA applied to a 3-D data set can be seen in Fig. 5.1.

Now, given the set of p-dimensional vectors $\mathbf{a}_k$, $k = 1, \ldots, n$, an estimate of each original data vector can be reconstructed by the inverse equation:

$$\tilde{\mathbf{x}}_k = \mathbf{C}\mathbf{a}_k. \tag{5.3}$$

[1] An alternative formulation of PCA requires the mean of all vectors in the data set to be subtracted from each one of them before constructing the covariance matrix. However, as the data set that will be taken into consideration is already well-centered, inclusion of the mean turns out to be unnecessary.

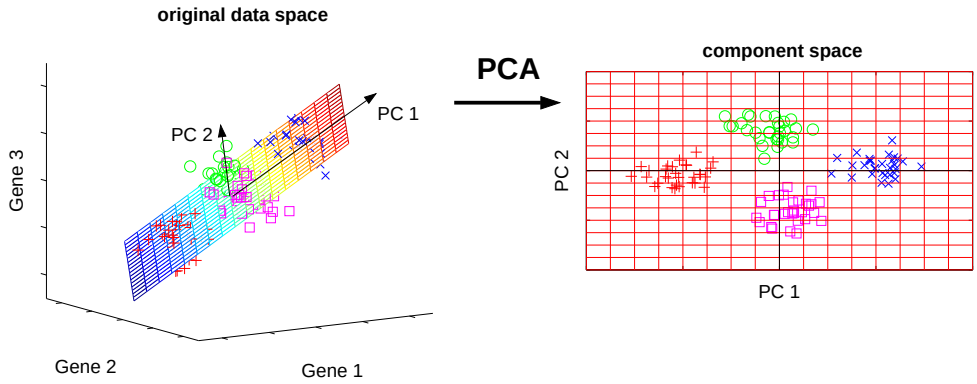


Figure 5.1: *Principal Component Analysis applied to a* 3*-D data set (figure reproduced from [154].*

Clearly, by increasing the dimension p of the representation the approximation improves. Thus, when dealing with PCA, the main design goal is to extrapolate the value p for which the trade-off between accuracy and data dimensionality is maximized.

5.1.2 PCA analysis of STFs

In this case, the initial collection is chosen to be a set of "representative" STFs for sound sorces located at different distances and incidence angles with respect to the observation point. Being Eq. (4.14) dependent on two spatial parameters only, the elevation angle is not considered and attention is restricted to points lying on the horizontal plane. Therefore for sake of simplicity θ_{inc} is assumed to be the incidence angle at the right ear canal, with $\theta_{\text{inc}} = 0°$, $\theta_{\text{inc}} = 90°$, and $\theta_{\text{inc}} = 180°$ corresponding to a sound source facing the right ear, in front of the head, and facing the left ear, respectively.

The set of STFs is sampled by fixing the head radius to the standard value $a = 8.75$ cm and varying the following parameters:

- 19 linearly spaced θ_{inc} values, from $0°$ to $180°$, at $10°$-angle increments;
- 7 exponentially spaced distance values, $\rho = 1.25, 1.5, 2, 4, 8, 16, 32$ (the last one approximating the far field response), where ρ is the normalized distance defined in Eq. (4.13));
- 100 linearly spaced frequency points from 100 Hz to 10 kHz, at 100-Hz increments.

A set of $19 \times 7 = 133$ STFs is obtained, of which only the dB magnitude responses – all reported in Fig. 5.2 – are considered next. Indeed, the transfer function of an ideal sphere appears to be minimum phase for all ranges and incidence angles [42]. In addition, when considering interaural differences for binaural hearing, approximated ITD models (e.g. Woodworth's formula,

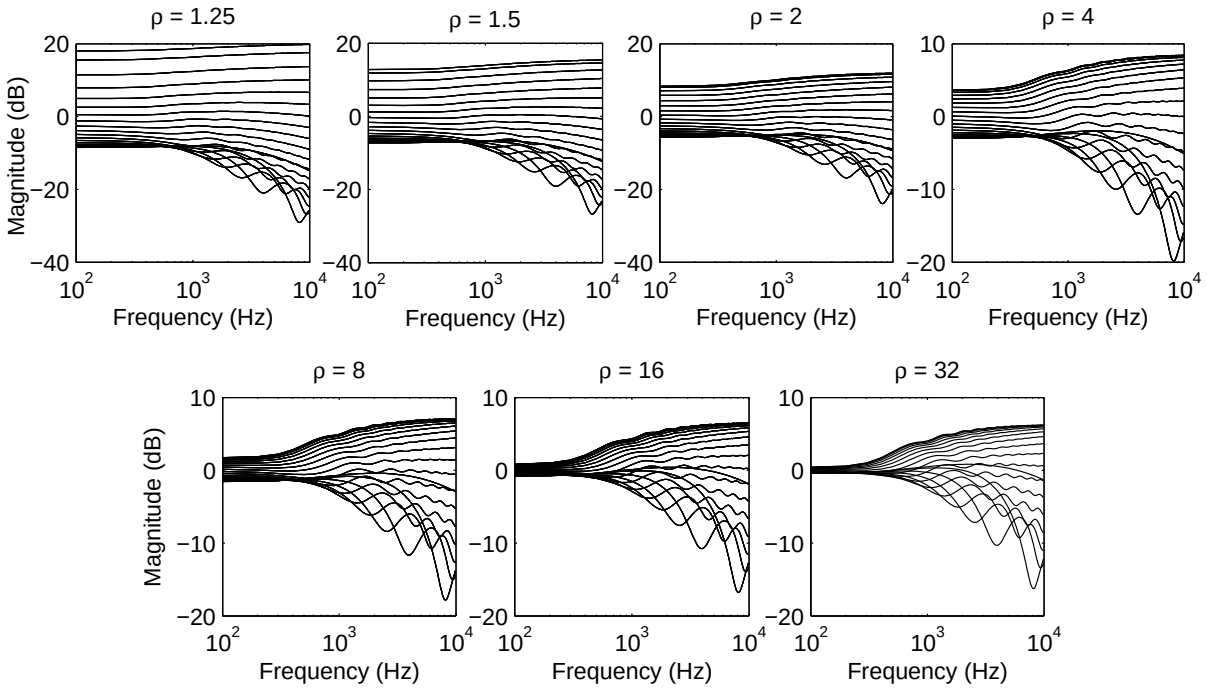


Figure 5.2: *The* 133 *STF vectors considered for PCA.*

Eq. (4.9)) can be used to simulate phase lag between right and left ear canal as a simple delay line. ITD effects can therefore be cascaded to the STF synthesis process.

At this point PCA is applied to the set of $n = 133$ real-valued vectors $\mathbf{x}_1$,...,$\mathbf{x}_n$, each of dimension $\delta = 100$. The first six basis vectors $BV_1 \dots BV_6$ of the analysis are sketched in Fig. 5.3, where μ is the normalized frequency defined in Eq. (4.3). After the first one which accounts for the general slope of the majority of STFs (with a positive weight for ipsilateral sources and a negative weight for contralateral ones – see Fig. 5.4), each successive basis vector introduces more and more ripples in the frequency response, starting from the most prominent in BV_2. The keen observer will note that BV_2's slope heavily resembles that of STFs for $\theta_{\text{inc}} = 170°$, BV_3 has a very similar frequency behaviour to STFs for $\theta_{\text{inc}} = 160°$, and so on. This means that the greatest variance in STFs appears for fixed distances along the angular range at contralateral source positions.

By investigating the trend of principal components PC_2 to PC_6 with the varying of distance and incidence angle we obtain a deeper insight of the analysis. As expected from the observations reported in Section 4.3.3, weights' moduli are amplified by decreasing distance; furthermore, Fig. 5.4 shows that each component emphasizes its corresponding basis vector only for a limited range of incidence angles, regardless of distance. This last observation further confirms that after the first basis vector which retains the average behaviour of the STF, those from the second onwards provide each a particularized description of the rippled high-frequency behaviour of

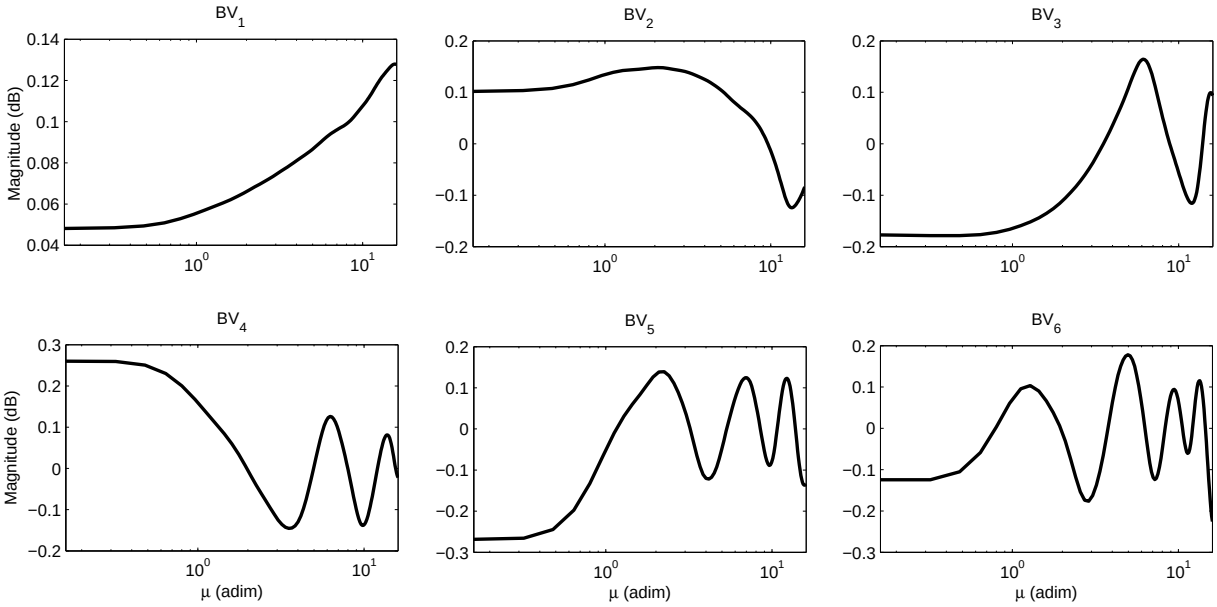


Figure 5.3: *The first six basis vectors from PCA applied to the STF collection.*

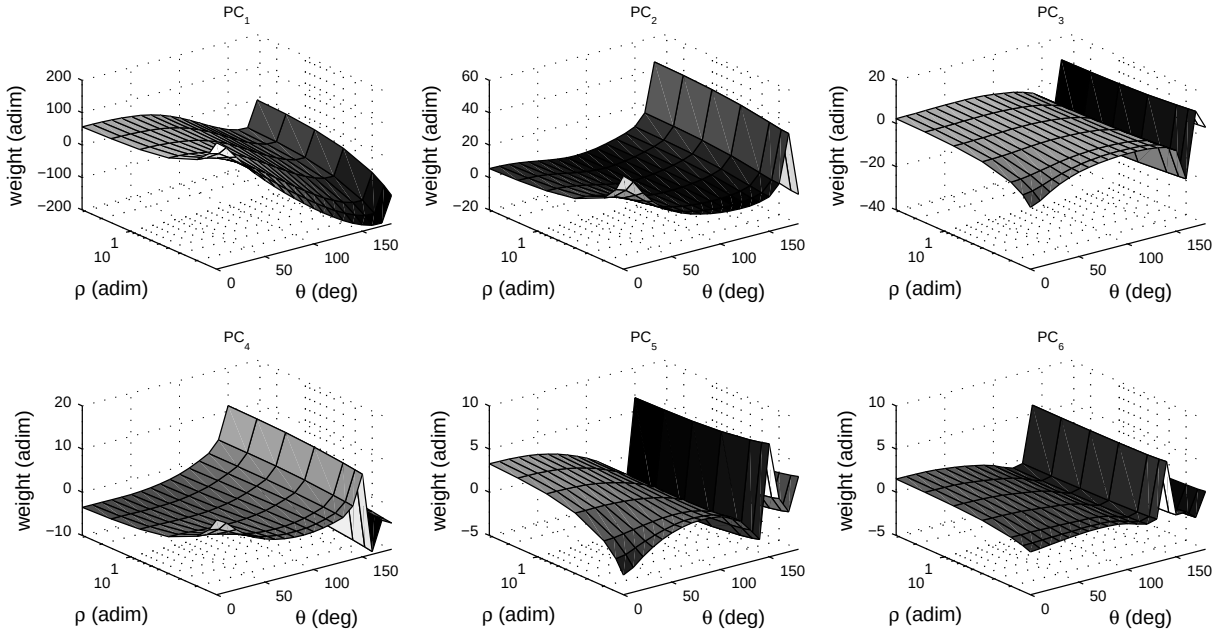
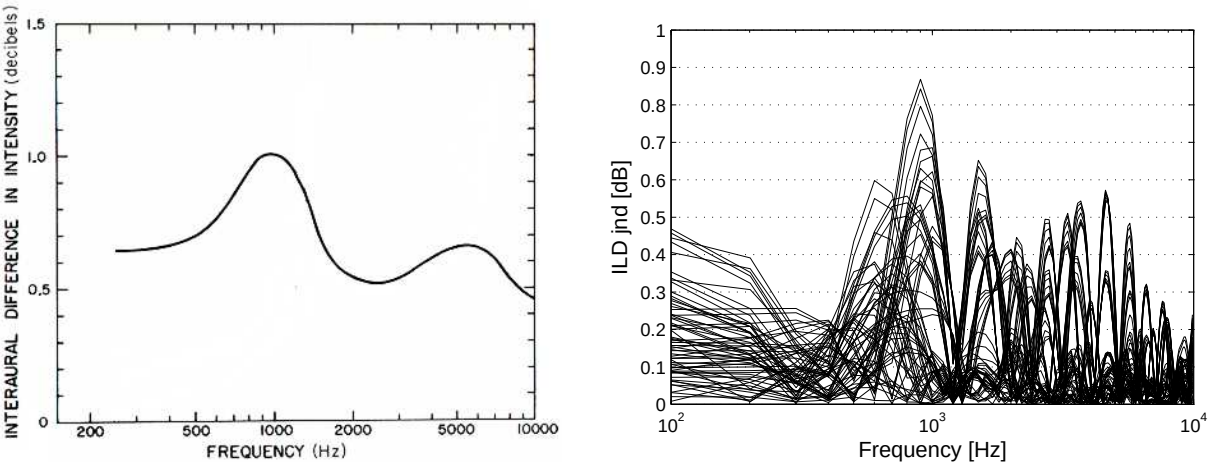


Figure 5.4: *The first six principal components from PCA applied to the STF collection.*

contralateral STFs, which varies according to the incidence angle. Also, note that principal components PC_2 to PC_6 present increasingly smaller weights; this point motivates the greater importance of component PC_{i-1} relative to PC_i.



(a) ILD jnd as a function of frequency.

(b) ILD error functions with $p = 7$.

Figure 5.5: *ILD jnd and PCA reconstruction optimality (figure (a) reproduced from [111]).*

5.1.3 STF reconstruction optimality

An interesting point is the number of principal components (parameter p) that can grant a hypothetically flawless reconstruction of the spherical transfer function by means of Eq. (5.3). To this end, a proper psychoacoustic principle is needed in order to theoretically quantify the maximum tolerable error, so as to extract the minimum p that meets its constraints.

Mills [111] presents a psychoacoustical result which can be used in this context. In particular the ILD jnd (just noticeable distance) curve as a function of frequency in Fig. 5.5(a) represents a safe upper bound on the approximation error, owing to no sensitivity of the human hearing apparatus to small changes in ILD. After having checked that the absolute error between all ILDs derived from a complementary pair of original HRTFs (same distance parameter and sum of incidence angles equal to 180 degrees, assuming diametrically opposite ear canals) and those reconstructed after PCA approximation turns out to lie under the jnd function, it can be stated that there is no significant information loss in the PCA approximation. Note that the jnd function has not been defined for very low frequencies; nevertheless, the dominant localization feature in this frequency range being ITD, ILD information appears to be relevant just for detecting very close distances.

As can be seen from Fig. 5.5(b) the minimum value p for which the total error introduced by the PCA approximation remains below the jnd curve is $p = 7$. As a consequence, it can be said that the intrinsic dimensionality of the STF representation is small, thus a good approximation of a spherical head model can, on a theoretical basis at least, be reached without expensive modeling nor an excessive waste of computational resources.

5.2 Near-Field Transfer Functions

For modeling purposes, one could think of the basis vectors arising from the PCA analysis as the magnitude responses of six filters to be designed, each weighed with a coefficient (i.e. the corresponding principal component) dependent on distance and incidence angle. However, such an analysis cannot be directly used for designing a filter model of the spherical head because, since the dB magnitude responses of the STFs were considered, weights $\mathbf{a}_k$ in Eq. (5.3) refer to a logarithmic scale instead of a linear scale. PCA was then applied to both linear magnitude responses and complex responses of STFs, yet the same conclusions as in the previous case were drawn for the following reasons:

- if linear magnitude responses are considered, PCA extracts basis vectors with negative values that cannot be seen as filters;
- if complex responses are considered, the new representation needs a greater number of basis vectors to yield a good reconstruction and weights are complex themselves.

It is therefore necessary to follow an alternative approach, having however in mind that the PCA analysis has clearly indicated that angular dependence of STFs is much greater than distance dependence in the transfer function frequency behaviour. Decoupling distance information from frequency is thus the primary goal towards the design of a cheap and effective model for the head.

In order to study the impact of distance, a given STF can be normalized to the corresponding far field spherical response yielding a new transfer function, which I refer to as Near-Field Transfer Function (NFTF):

$$H_{NF}(\rho, \mu, \theta_{\text{inc}}) = \frac{H(\rho, \mu, \theta_{\text{inc}})}{H(\infty, \mu, \theta_{\text{inc}})}. \tag{5.4}$$

Fig. 5.6 reports the 133 NFTFs corresponding to the STFs used for the previous section's PCA analysis. In all of the seven plots, each referring to a different normalized distance, the response for incidence angle $\theta_{\text{inc}} = 0^\circ$ is the uppermost in level; as the angle grows up to $\theta_{\text{inc}} = 180^\circ$, magnitude decreases.

From these plots it becomes clear that the rippled behaviour of contralateral STFs is not correlated to source distance at all. NFTFs are very regular functions that slightly decay with frequency, in an approximately monotonic fashion. Furthermore, the magnitude boost for small distances is evident in ipsilateral NFTFs while it is less prominent in contralateral NFTFs. Note also that for $\rho = 1.25$ the response completely crosses the 0-dB threshold at an angle that is far smaller than for the remaining distances, i.e. $\theta_{\text{inc}} \approx 65^\circ$: this behaviour is explained by the intuition that, as source distance decreases, the angular range for which a direct ray can reach an observation point on the sphere becomes narrower and narrower.

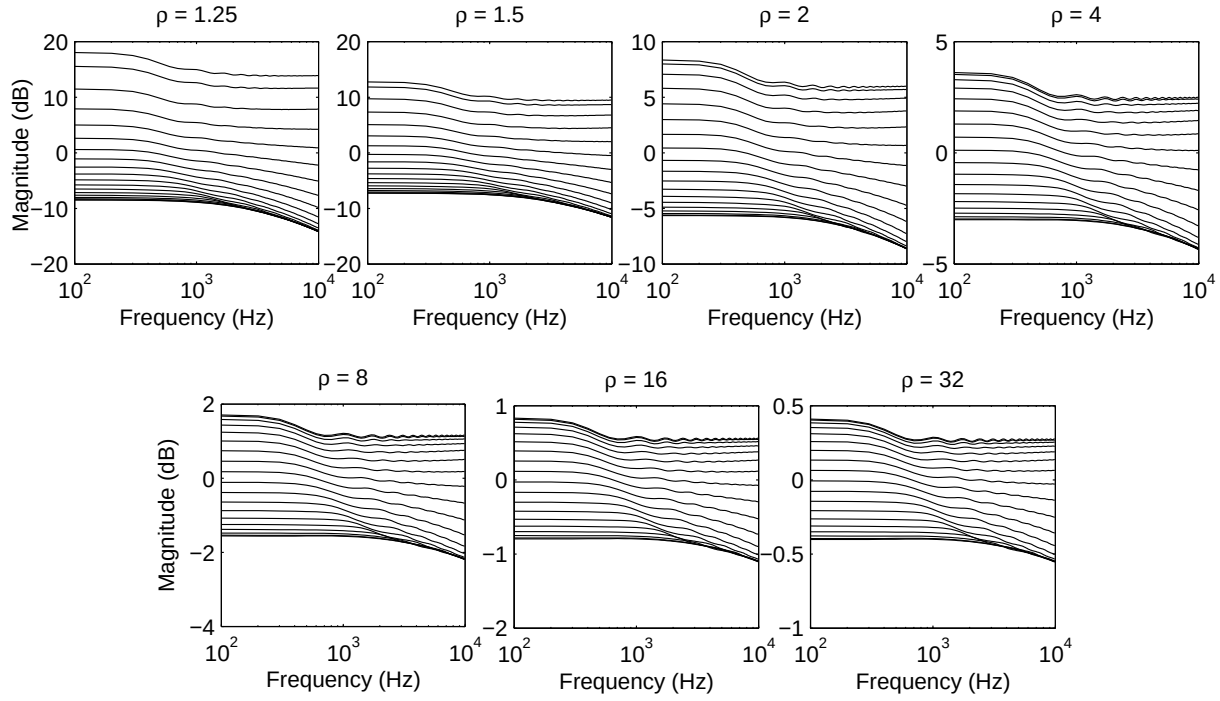


Figure 5.6: *Analytical NFTFs as functions of distance and incidence angle.*

5.2.1 DC gain of NFTFs

Now, let's look more closely at how the DC gain G_0 varies in NFTFs as the source moves away along a given angular direction. For each of the 19 incidence angles, $\theta_{\text{inc}} = 0^\circ - 180^\circ$ at 10-degree steps, Eq. (4.14) is sampled at DC ($\mu = 0$) for a great number of different, exponentially increasing distances, specifically

$$\rho = 1.15^{1+\frac{k-1}{10}}, \quad k = 1, \ldots, 250, \tag{5.5}$$

and its absolute value calculated, yielding DC gain

$$G_0(\theta_{\text{inc}}, \rho) = H_{NF}(\rho, 0, \theta_{\text{inc}}). \tag{5.6}$$

Fig. 5.7 plots DC gains as functions of distance and incidence angle.

Note that, if attention is focused on a single incidence angle, gain looks like an exponential function of distance, either for small incidence angles (where gain decreases with distance) and for contralateral positions of the source (where gain increases with distance). The only angles for which an exponential trend is not perfectly seen are those included in the range $30^\circ \leq \theta_{\text{inc}} \leq 60^\circ$, for which DC gain first slightly increases and then suddenly decreases as distance grows. Nevertheless, in order to model distance dependence of NFTFs at DC we can think of approximating it as a sum of two exponentials for all the 19 different incidence angles. The need of two functions

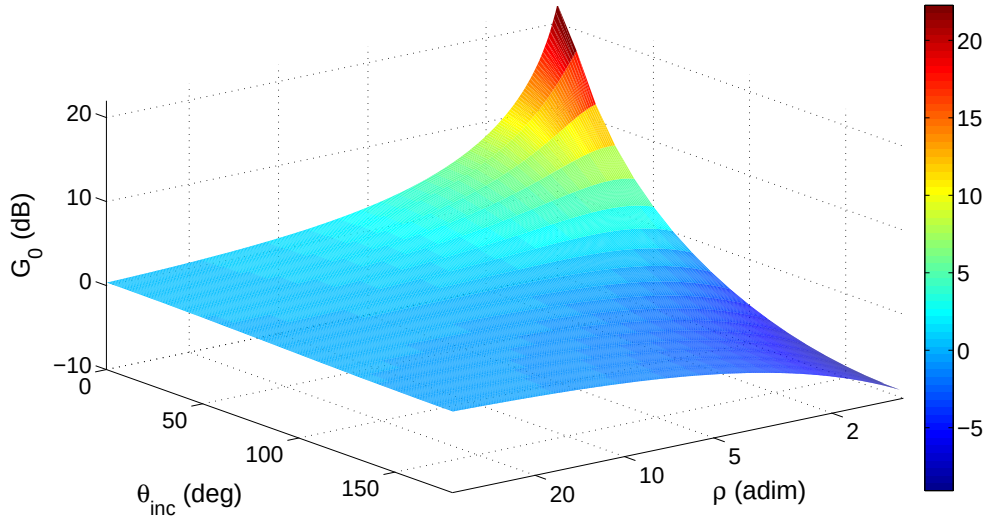


Figure 5.7: *NFTF gain at DC.*

is justified by both the asymptotic behaviour of the gain function, that tends to 1 for $\rho \to \infty$, and by the higher number of DoF that two exponentials functions can grant. The exponential fit, that will thus have the form

$$G_0^{\text{app}}(\theta_{\text{inc}}, \rho) = a_{\theta_{\text{inc}}} \exp(b_{\theta_{\text{inc}}}\rho) + c_{\theta_{\text{inc}}} \exp(d_{\theta_{\text{inc}}}\rho), \quad \theta_{\text{inc}} = 0°, 10°, \ldots, 180°, \tag{5.7}$$

is found with the help of the MATLAB Curve Fitting Toolbox (`cftool`).

Coefficients $a_{\theta_{\text{inc}}}$, $b_{\theta_{\text{inc}}}$, $c_{\theta_{\text{inc}}}$, and $d_{\theta_{\text{inc}}}$ for each of the 19 incidence angles are reported in Table 5.1, as well as the RMS (root mean square) error measure between real and approximated DC gains for each incidence angle at the 250 evaluated distances. The latter values confirm the overall excellent fitness of the exponential functions, especially from $\theta_{\text{inc}} = 70°$ onwards where $RMS(G_0, G_0^{\text{app}}) < 0.01$. The first two angles, $\theta_{\text{inc}} = 0°$ and $\theta_{\text{inc}} = 10°$, are less well approximated because DC gain seems to have an over-exponential inverse dependence on distance for small angles instead.

By investigating the trend of the four coefficients in Table 5.1, one could notice that an exponential function could also be fitted to at least three of them in order to fully parameterize G_0^{app} on incidence angle additionally to distance. However, the discontinuity appearing between $\theta_{\text{inc}} = 50°$ and $\theta_{\text{inc}} = 70°$, due to the passage from a sum of exponentials theoretically tending to $+\infty$ as $\rho \to 1$ to a sum of exponentials tending to $-\infty$, suggests that a simple linear interpolation between adjacent functions such as the one that follows could suffice to effectively model DC gain for intermediate incidence angles:

$$G_0^{\text{app}}(\theta_{\text{inc}}, \rho) = \left(\left\lceil\frac{\theta_{\text{inc}}}{10}\right\rceil - \frac{\theta_{\text{inc}}}{10}\right) G_0^{\text{app}}\left(\left\lfloor\frac{\theta_{\text{inc}}}{10}\right\rfloor 10, \rho\right) + \left(\frac{\theta_{\text{inc}}}{10} - \left\lfloor\frac{\theta_{\text{inc}}}{10}\right\rfloor\right) G_0^{\text{app}}\left(\left\lceil\frac{\theta_{\text{inc}}}{10}\right\rceil 10, \rho\right). \tag{5.8}$$

The effective fitness of such an approximation on a dB scale will be objectively evaluated at

θ_{inc}	$a_{\theta_{\text{inc}}}$	$b_{\theta_{\text{inc}}}$	$c_{\theta_{\text{inc}}}$	$d_{\theta_{\text{inc}}}$	$RMS(G_0, G_0^{\text{app}})$
0°	598.5	-3.583	1.748	-0.0237	0.2692
10°	61	-2.087	1.513	-0.01542	0.1342
20°	9.544	-1.115	1.318	-0.00876	0.0504
30°	2.971	-0.6695	1.203	-0.00504	0.0233
40°	1.173	-0.4082	1.121	-0.0025	0.0211
50°	0.4993	-0.216	1.044	-0.0003	0.0261
60°	0.6159	-0.0329	0.5805	0.01028	0.0337
70°	-5.083	-2.587	1.083	-0.0024	0.0099
80°	-1.587	-1.403	1.022	-0.0005	0.0017
90°	-1.073	-0.9476	0.9798	0.00074	0.0052
100°	-0.9161	-0.7297	0.9483	0.00157	0.0073
110°	-0.8496	-0.6075	0.9233	0.00218	0.0083
120°	-0.815	-0.5312	0.9031	0.00265	0.0087
130°	-0.7944	-0.4808	0.8868	0.00301	0.0089
140°	-0.7811	-0.4464	0.8738	0.0033	0.0089
150°	-0.7722	-0.423	0.864	0.0035	0.0089
160°	-0.7664	-0.4077	0.857	0.00365	0.0089
170°	-0.7631	-0.3991	0.8529	0.00373	0.0088
180°	-0.7567	-0.3924	0.8525	0.00372	0.0086

Table 5.1: *Coefficients for Eq. (5.7) and approximation fitness.*

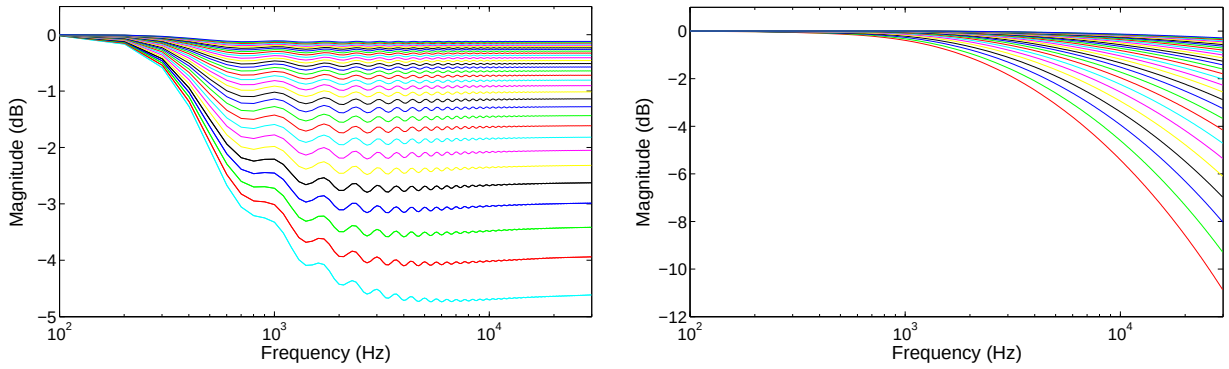
the end of the analysis process, in Section 5.2.3, even for incidence angles different from those considered up to now.

5.2.2 Frequency dependence in NFTFs

The behaviour of NFTFs at DC having been checked, it remains to be studied how much NFTFs depend on frequency and how such dependence can be cheaply modeled. In order to do this the DC gain G_0 can act as a further normalization factor, thus the following operation is performed for a set of NFTFs computed at the already considered 250 distances and in the frequency range up to 30 kHz, sampled at 100-Hz steps (assuming again head radius $a = 8.75$ cm):

$$\hat{H}_{NF}(\rho, \mu, \theta_{\text{inc}}) = \frac{H_{NF}(\rho, \mu, \theta_{\text{inc}})}{G_0(\theta_{\text{inc}}, \rho)}. \tag{5.9}$$

Fig. 5.8 shows the frequency behaviour of normalized NFTFs for the two extreme incidence angles, $\theta_{\text{inc}} = 0^\circ$ and $\theta_{\text{inc}} = 180^\circ$, and a downsampled number of distances (1 each 8 in Eq. (5.5)), the smallest always corresponding to the lowest NFTF in magnitude.



(a) $\theta_{\text{inc}} = 0°$. (b) $\theta_{\text{inc}} = 180°$.

Figure 5.8: *Frequency behaviour of normalized NFTFs for two different incidence angles. In both cases, magnitude increases with distance.*

In both plots, magnitude monotonically increases with distance at all frequencies although tending to the 0-dB threshold at most. This means that DC is always the frequency point of the NFTF where the gain is maximum. However, note the different high-frequency trend for the two considered angles: at $\theta_{\text{inc}} = 0°$ the magnitude plot looks like that of a high-frequency shelving filter, whereas at $\theta_{\text{inc}} = 180°$ a lowpass behaviour is observed. For intermediate incidence angles, the response for a specific distance ρ gradually morphs from that of a shelving filter to that of a lowpass filter as the angle increases, the faster rate being observed for small distances.

In light of such result, one could think of approximating the magnitude plot of the normalized NFTF through a shelving or lowpass filter, depending on incidence angle and distance. Unfortunately, two lawful observations complicate such design process:

- the switch from a shelving to a lowpass filter at a given incidence angle needs to be smooth in order to avoid listening artifacts;
- a first-order lowpass filter excessively cuts high frequencies with respect to the maximum 10-dB decay observed in the normalized NFTF plots.

These shortcomings can be solved, although at the cost of precision loss, by always approximating a normalized NFTF through a first-order high-frequency shelving filter. The implementation chosen for the filter is that found in [192],

$$H_{\text{sh}}(z) = 1 + \frac{H_0}{2}\left(1 - \frac{z^{-1} + a_c}{1 + a_c z^{-1}}\right), \tag{5.10}$$

where

$$a_c = \frac{V_0 \tan\left(\pi \frac{f_c}{f_s}\right) - 1}{V_0 \tan\left(\pi \frac{f_c}{f_s}\right) + 1}, \tag{5.11}$$

$$V_0 = 10^{\frac{G_\infty}{20}}, \tag{5.12}$$

and f_s is sampling frequency.

Now it has to be highlighted how the two key parameters of the shelving filter, cutoff frequency f_c and asymptotic high-frequency gain G_∞, can be extracted from the normalized NFTFs in order to yield a satisfactory approximation. First, the asymptotic gain is calculated as

$$G_\infty(\theta_{\text{inc}}, \rho) = 20 \log_{10} \left| \hat{H}_{NF} \left(\rho, 30000 \frac{2\pi a}{c}, \theta_{\text{inc}} \right) \right| \quad \text{[dB]}, \tag{5.13}$$

that is, the (negative) dB gain of the NFTF at 30 kHz. The choice of such a high frequency point is needed to best model the slope of near contralateral NFTFs in the range of interest for the HRTF, i.e. up to 15 kHz.

Second, taking as reference the previously computed asymptotic gain, the cutoff frequency is calculated as

$$f_c(\theta_{\text{inc}}, \rho) = \min_f \left| \left(20 \log_{10} \left| \hat{H}_{NF} \left(\rho, \frac{2\pi a}{c} f, \theta_{\text{inc}} \right) \right| - \frac{2}{3} G_\infty(\theta_{\text{inc}}, \rho) \right) \right| \quad \text{[Hz]}, \tag{5.14}$$

that is, the frequency point where the normalized NFTF presents a negative dB gain of approximately two thirds of the asymptotic gain. This point is heuristically preferred, after a number of trials with different values, to the point where the gain is $\frac{G_\infty}{2}$ in order to minimize differences in magnitude between a shelving filter and a lowpass filter for contralateral NFTFs.

The quality of the shelving filter approximation is attested through a measure widely used in recent literature [123, 133, 48]: spectral distortion

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(20 \log_{10} \frac{|H(f_i)|}{|\tilde{H}(f_i)|} \right)^2} \quad \text{[dB]}, \tag{5.15}$$

where H is the original response (here $\hat{H}_{NF}$), $\tilde{H}$ is the reconstructed response (here H_{sh}), and N is the number of available frequencies in the considered range, that in this case is limited between 100 Hz and 15 kHz. Mean spectral distortion between original normalized NFTFs and designed shelving filters, averaged among the responses for the 50 smallest distances (where the approximation is most challenging) was found to never exceed 1 dB at each of the 19 incidence angles.

The variation of parameters G_∞ and f_c along distance and incidence angle was also studied. In this context too, it was surprisingly noticed how both parameters bear an exponential growth (in the case of cutoff frequency) or decrease (in the case of high-frequency gain) as the source approaches. Thus, similarly to what was done for DC gains, a sum of two exponential functions was fitted as follows to the evolution of G_∞ and f_c along distance at given incidence angles:

$$G_\infty^{\text{app}}(\theta_{\text{inc}}, \rho) = k_{\theta_{\text{inc}}} \exp(l_{\theta_{\text{inc}}} \rho) + m_{\theta_{\text{inc}}} \exp(n_{\theta_{\text{inc}}} \rho), \quad \theta_{\text{inc}} = 0^\circ, 10^\circ, \ldots, 180^\circ, \tag{5.16}$$

θ_{inc}	$k_{\theta_{\text{inc}}}$	$l_{\theta_{\text{inc}}}$	$m_{\theta_{\text{inc}}}$	$n_{\theta_{\text{inc}}}$	$RMS(G_\infty, G_\infty^{\text{app}})$ [dB]
$0°$	-11.26	-1.118	-1.436	-0.1061	0.0612
$10°$	-8.544	-0.9573	-1.25	-0.09473	0.0476
$20°$	-7.586	-0.9905	-1.239	-0.09872	0.0594
$30°$	-38.62	-2.339	-1.839	-0.1573	0.0912
$40°$	-112.7	-2.858	-1.789	-0.1693	0.0844
$50°$	-95.55	-2.386	-1.807	-0.1753	0.082
$60°$	-70.2	-1.895	-2.258	-0.1826	0.0909
$70°$	-53.15	-1.51	-3.001	-0.1664	0.1133
$80°$	-42.73	-1.238	-3.768	-0.1376	0.146
$90°$	-36.84	-1.08	-4.468	-0.1162	0.1789
$100°$	-33.85	-1.018	-4.958	-0.1043	0.2024
$110°$	-32.84	-1.023	-5.105	-0.09919	0.2116
$120°$	-33.06	-1.062	-4.953	-0.09865	0.2085
$130°$	-33.85	-1.111	-4.676	-0.1012	0.1998
$140°$	-34.81	-1.157	-4.415	-0.1053	0.1902
$150°$	-35.7	-1.195	-4.213	-0.1099	0.182
$160°$	-36.43	-1.224	-4.081	-0.114	0.1758
$170°$	-36.89	-1.243	-4.006	-0.1168	0.1721
$180°$	-34.43	-1.19	-3.811	-0.1131	0.1608

Table 5.2: *Coefficients for Eq. (5.16) and approximation fitness.*

$$f_c^{\text{app}}(\theta_{\text{inc}}, \rho) = p_{\theta_{\text{inc}}} \exp(q_{\theta_{\text{inc}}}\rho) + r_{\theta_{\text{inc}}} \exp(s_{\theta_{\text{inc}}}\rho), \quad \theta_{\text{inc}} = 0°, 10°, \ldots, 180°. \tag{5.17}$$

Table 5.2 and Table 5.3 summarize fitness scores and function parameters' values for each of the two quantities.

The functional approximation of G_∞ is overall excellent, never exceeding a mean RMS error of 0.25 dB in the considered angular directions. Similarly, the approximation provided by f_c^{app} yields a mean RMS error that is below half the actual frequency resolution of 100 Hz for more than 70% of the incidence angles, the most problematic being the intermediate ones where the frequency behaviour of the normalized NFTF is halfway between those of a shelving and a lowpass filter. As a matter of fact, for these angular directions the normalized NFTF cutoff frequency for small distances suddenly explodes from a relatively low value (< 1 kHz) to a high value (≈ 10 kHz), yielding a very acute slope in the distance-dependent curve that a sum of two exponentials cannot approximate without losing precision in the following points.

Contrarily to the approximation of G_0, in these two cases no consistent trend is seen across incidence angles for any of the exponential functions' coefficients. An interpolation of adjacent polynomials analogous to that in Eq. (5.8) is thus definitely required to correctly model parameters G_∞^{app} and f_c^{app} for intermediate angular values.

θ_{inc}	$p_{\theta_{\text{inc}}}$	$q_{\theta_{\text{inc}}}$	$r_{\theta_{\text{inc}}}$	$s_{\theta_{\text{inc}}}$	$RMS(f_c, f_c^{\text{app}})$ [Hz]
$0°$	483.8	-0.7549	494.3	0.00046	21.02
$10°$	410.1	-0.6986	492.7	0.00059	20.15
$20°$	443.5	-0.7844	493.6	0.00053	21.22
$30°$	$2.339e^7$	-8.839	535.1	-0.0032	35.31
$40°$	$4.923e^7$	-8.23	553.3	-0.00476	46.31
$50°$	$1.087e^6$	-4.284	612.1	-0.00828	52.38
$60°$	$8.433e^4$	-1.804	656.7	-0.00339	115.5
$70°$	$2.46e^4$	-0.6892	1452	-0.01873	180.6
$80°$	$1.084e^4$	-0.3194	5639	-0.00938	126.4
$90°$	3677	-0.2464	$1.067e^4$	-0.00284	57.53
$100°$	846.6	-0.1363	$1.272e^4$	-0.00053	46.57
$110°$	$1.333e^4$	0.00024	$3.718e^5$	-5.851	36.26
$120°$	6913	-2.113	$1.32e^4$	$2.044e^{-7}$	18.53
$130°$	1629	-0.6621	$1.322e^4$	-0.00033	38.34
$140°$	815.8	-0.3119	$1.377e^4$	-0.00018	34.79
$150°$	$1.478e^4$	-0.0007	46.94	0.05192	21.95
$160°$	$1.558e^4$	0.00027	-1291	-0.6957	32.13
$170°$	$1.622e^4$	0.00038	-1787	-0.5135	34.00
$180°$	$1.641e^4$	0.0005	-2109	-0.5297	35.18

Table 5.3: *Coefficients for Eq. (5.17) and approximation fitness.*

5.2.3 A model for distance rendering

The analysis performed in the previous section allows straightforward construction of a filter model for the rendering of distance, that can be easily integrated with an infinite-distance spherical model of the head following one of the implementations available in the literature. In fact, if the latter is modeled through a filter $H^{\infty}_{\text{sphere}}$ that takes the incidence angle θ_{inc} as input (as for instance in Brown and Duda's model, see Eq. (4.4)), the information given by the NFTF can be provided by a cascade of a multiplicative gain G_0 and a shelving filter H_{sh} as made clearer by the following equations, that are neither more nor less than an approximated combination of Eq. (5.4) and Eq. (5.9):

$$H_{head}(\rho, \mu, \theta_{\text{inc}}) = H_{dist}(\rho, \mu, \theta_{\text{inc}}) H^{\infty}_{\text{sphere}}(\infty, \mu, \theta_{\text{inc}}), \tag{5.18}$$

$$H_{dist}(\rho, \mu, \theta_{\text{inc}}) = G_0^{\text{app}}(\theta_{\text{inc}}, \rho) H_{\text{sh}}(\mu, G_{\infty}^{\text{app}}(\theta_{\text{inc}}, \rho), f_c^{\text{app}}(\theta_{\text{inc}}, \rho)). \tag{5.19}$$

The general filter structure is sketched in Fig. 5.9. Here the head radius a can be freely chosen previous to the rendering process in order to correctly tune parameters ρ and μ, allowing to stretch or extend the frequency and distance axes of the STF/NFTF with respect to the

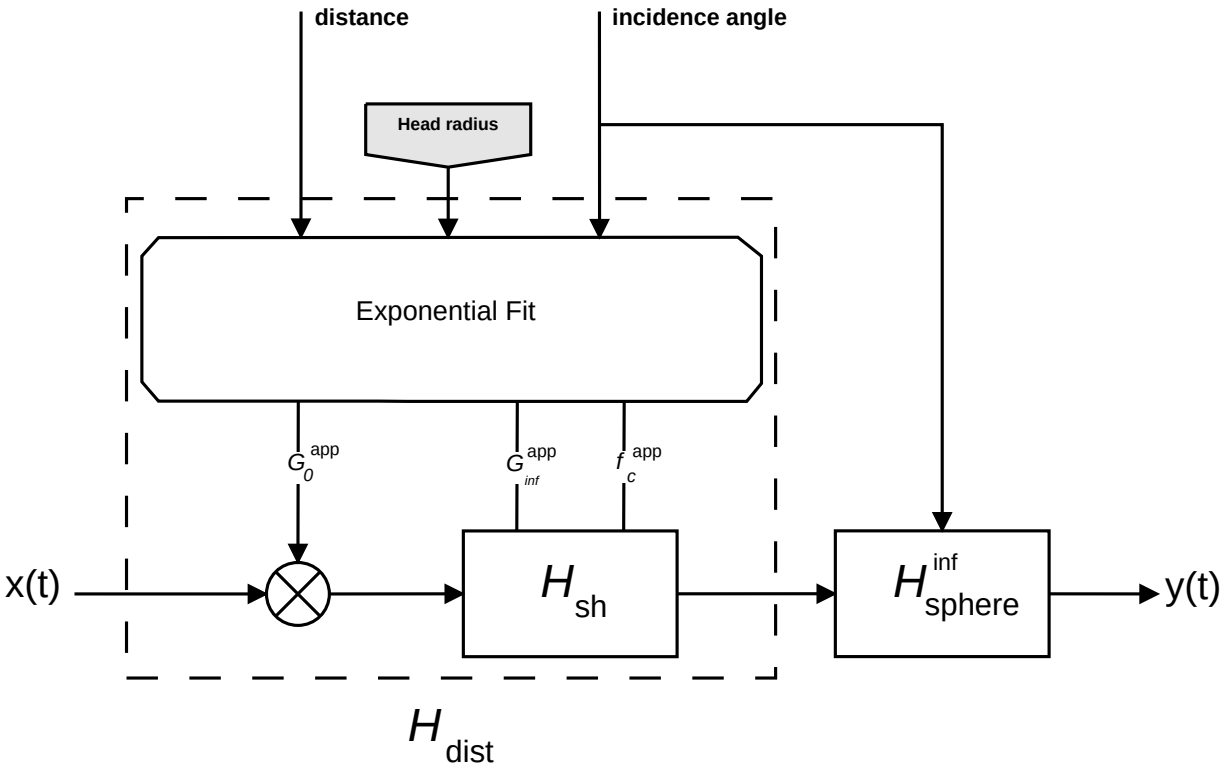


Figure 5.9: *A model for a spherical head including distance dependence.*

case of standard 8.75-cm radius. The head radius parameter thus represents a first raw way of customizing a HRTF model on the subject's anthropometry. Based on distance and incidence angle information, the "Exponential Fit" computation block linearly interpolates functions G_0^{app}, G_∞^{app} and f_c^{app} using Eq. (5.7), Eq. (5.16), and Eq. (5.17) respectively; afterwards, $G_0^{\text{app}}(\theta_{\text{inc}}, \rho)$ is used as multiplicative factor whereas $G_\infty^{\text{app}}(\theta_{\text{inc}}, \rho)$ and $f_c^{\text{app}}(\theta_{\text{inc}}, \rho)$ are feeded as parameters to the shelving filter.

A legitimate question is the overall goodness of model H_{dist}, that is, whether all the introduced approximations objectively unsettle the magnitude response of original NFTFs as computed through Eq. (4.14) and Eq. (5.4). NFTFs resulting from the above model and corresponding to the normalized distances and incidence angles of the analytical NFTFs in Fig. 5.6 are reported in Fig. 5.10. From direct comparison of the two figures it can be seen how the general shape of NFTFs is well reproduced, even though evident errors in the DC gain of a couple of responses for $\rho = 32$ are clearly recognizeable. In addition, some of the responses for the greatest distances do not exhibit a smooth transition between consecutive incidence angles as in the analytical responses around $\theta_{\text{inc}} = 60°$.

In order to have a quantitative indication of the model's accuracy and to better explain the above dissimilarities, the usual spectral distortion measure was calculated either for spatial locations that were used during the analysis process and new spatial locations, thanks to the functional

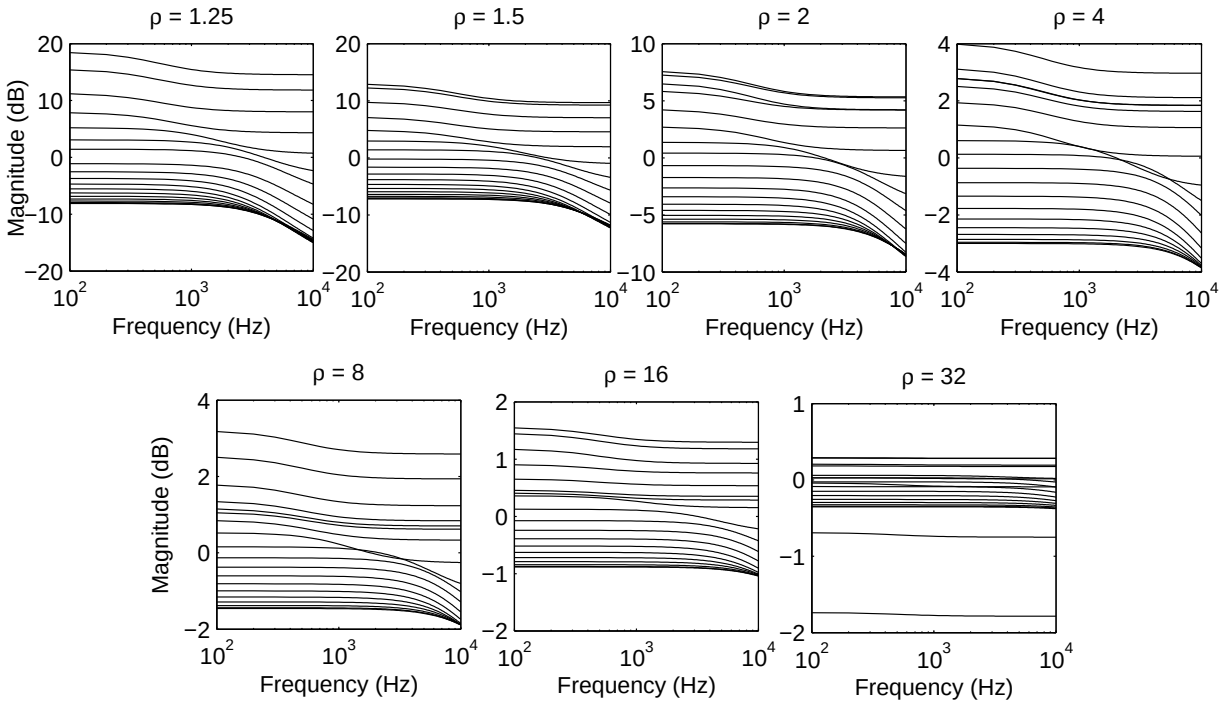
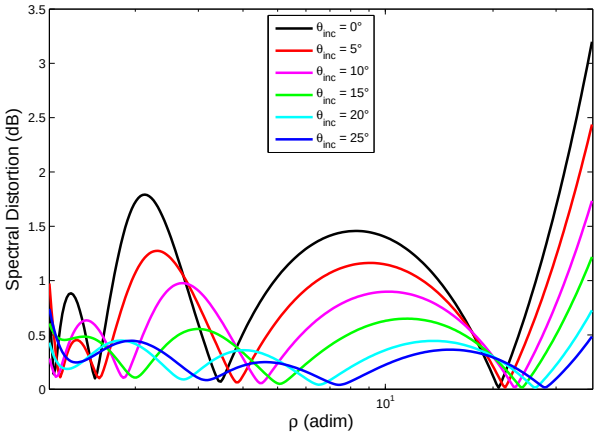


Figure 5.10: *Reconstructed NFTFs as functions of distance and incidence angle.*

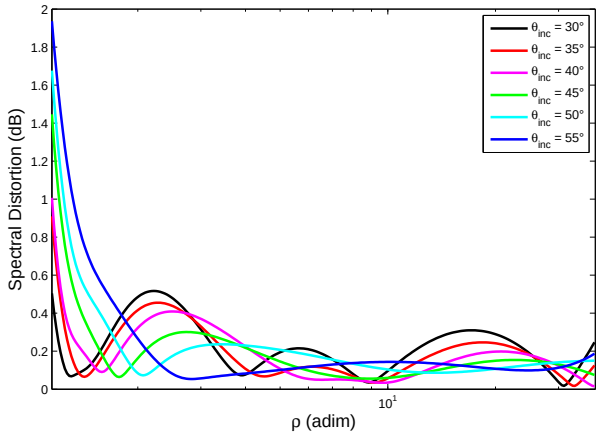
representation of distance and interpolation over incidence angles of the key parameters. Specifically, the magnitude of H_{dist} for a 8.75-cm spherical head was computed through Eq. (5.19) for the usual 250 distances, this time at 5-degree angular steps ($\theta_{\text{inc}} = 0°, 5°, 10°, \ldots, 180°$), and compared to the magnitude response of the corresponding original NFTFs up to 15 kHz. Distance-dependent spectral distortion plots for the 37 considered incidence angles are all shown in Fig. 5.11.

Notice that the overall fitness of the approximation is excellent for contralateral sources, being the SD lower than 1 dB in almost all of the considered source locations except for the very nearest ones around $\theta_{\text{inc}} = 90°$. Concerning ipsilateral positions, the biggest discrepancies for very close distances appear in the middle range, i.e. from $\theta_{\text{inc}} = 45°$ to $\theta_{\text{inc}} = 65°$. These are well explained by the already mentioned passage from a shelving to a lowpass frequency behaviour, that doesn't find a smooth correspondence in the exponential functions.

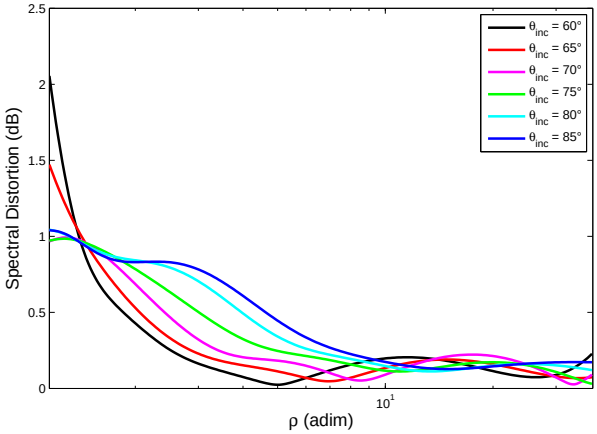
Then, as distance increases up to $\rho = 10$, SD tends to decrease except for a small peak centered around $\rho = 3 - 4$ for contralateral sources, and a series of peaks for $\theta_{\text{inc}} < 40°$ that are in most cases weak except for those at the smallest angular positions, $\theta_{\text{inc}} = 0°$ and $\theta_{\text{inc}} = 5°$. By taking a look to Table 5.1 one can realize that the latter behaviour is due to the non-perfect fit provided by the first two exponential functions G_0^{app}. These are responsible for the sudden SD rise at the farthest distances ($\rho > 20$) observed for the smallest incidence angles that was also seen in the last plot of Fig. 5.10, being asymptotically not tending to 0 dB with distance



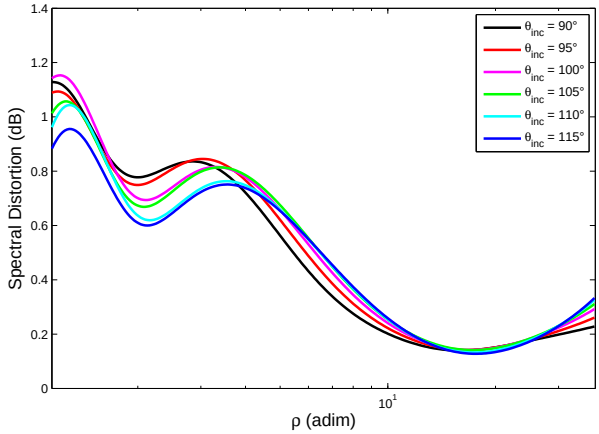
(a) $0° \leq \theta_{\text{inc}} \leq 25°$.



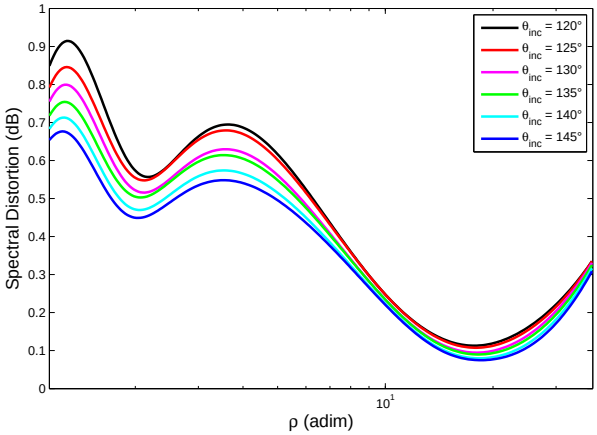
(b) $30° \leq \theta_{\text{inc}} \leq 55°$.



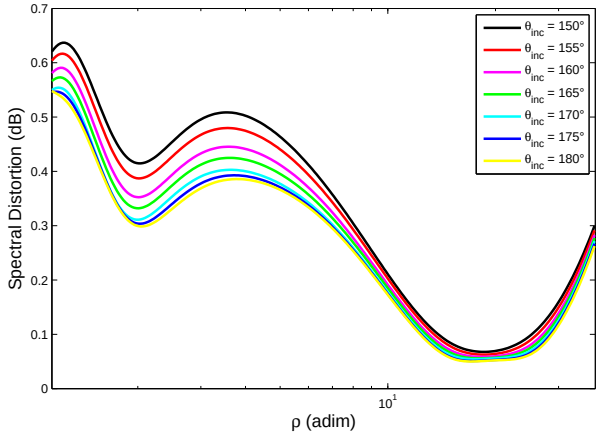
(c) $60° \leq \theta_{\text{inc}} \leq 85°$.



(d) $90° \leq \theta_{\text{inc}} \leq 115°$.



(e) $120° \leq \theta_{\text{inc}} \leq 145°$.



(f) $150° \leq \theta_{\text{inc}} \leq 180°$.

Figure 5.11: *Spectral distortion introduced by model H_{dist}.*

yet falling under this threshold. However, the problem can be easily be overcome by forcing the exponential function G_0^{app} to always tend to 0 dB (possibly at the cost of worsening the approximations at close distances) or, even better, by letting the contribution of the whole filter structure H_{dist} smoothly dissolve as the source enters the far field, which is definitely the case for $\rho > 20$.

Finally, note how there is no evident relative SD increase between reconstructed NFTFs for angles that were considered in the analysis process and other angles. To be more precise, SD for angle θ_{inc} remains approximately halfway the two nearest analyzed angles $\theta'_{\text{inc}} = \theta_{\text{inc}} - 5°$ and $\theta''_{\text{inc}} = \theta_{\text{inc}} + 5°$ for almost all the considered distances. As a consequence, linear interpolations of the key coefficients are already effective as they are, not needing to be improved through higher-order interpolations and/or a denser sampling along the angular axis.

5.3 Conclusions

An extremely low-order model for distance rendering, thought for real-time applications, was proposed in this Chapter. The main purpose of the model is to cheaply simulate the impact that source distance has on the sound waves arriving at the ears in the near field, a region where the relation between sound pressure and distance is both highly frequency-dependent and nonlinear. The reference for the model was based on an analytical description of a spherical head response, appropriately filtered out so as to include distance-dependent patterns only. With respect to such reference, the model was objectively seen to provide a very good fit in almost the whole near field, despite its simplicity.

It could be questioned whether analytical near-field transfer functions really reflect distance-dependent patterns in measured HRTFs, and if a weak customization (onto the head radius only) may be enough to account for differences among subjects. In particular, when a source is very close to the human head the finer details of the subject's anthropometry, such as the shape of the head or the presence of the nose, could prominently come into play. Unfortunately, most HRTF measurements are performed in the far field or in its vicinity at one single given distance: collecting HRTFs for more than one distance would intolerably multiplicate the required measurement time. Numerical simulations (such as the BEM) are thus needed to address such a question.

Further work in this direction should take into consideration alternative filter structures to the single, first-order shelving filter, such as a higher-order shelving filter or a lowpass filter realization allowing slope control for contralateral positions, in order to better approximate normalized NFTFs. Also, if one remains within the assumption that ITD does not change with distance, the design of an all-pass section counterbalancing the effect that the shelving/lowpass filter's phase response has on ITD needs to be carried out.

Last but not least, the choice of the far-field head filter to be coupled with the distance rendering model will turn out to be pivotal for a good STF approximation. Brown and Duda's first-order

filter, although replicating with some degree of approximation the mean magnitude characteristics of the far-field STF, does not simulate the rippled behaviour seen for contralateral sources. Experimental evaluations on the psychoacoustical importance of these ripples or an alternative head model are thus needed.

Chapter 6

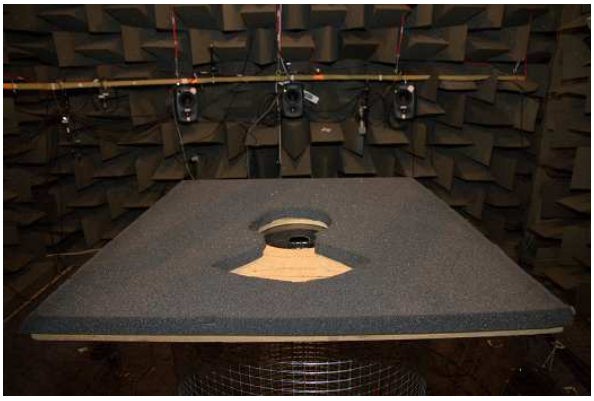
Pinna-Related Transfer Functions: Estimation Methods and Analysis

Pinna-Related Transfer Functions (PRTFs) reflect the modifications undergone by an acoustic signal as it interacts with the listener's outer ear. These can be seen as the pinna contribution to the HRTF. Although perceptually dominated by head motion cues, pinna effects on incident sound waves are of great importance in sound spatialization. Several experiments have shown that, contrarily to azimuth effects which are dominated by diffraction around the listener's head and may be reduced to simple and intuitive binaural quantities, elevation cues are basically monaural and heavily depend on the listener's pinna shape, being the result of a superposition of scattering waves influenced by a number of resonant modes inside pinna cavities. Within this framework, it is crucial to find a suitable model for representing PRTFs; linking the model parameters to simple anthropometric measurements on the user's pinnae represents the ultimate challenge in this direction. Once this model is available, cascading it to a Head-and-Torso (HAT) model [6] would yield a complete structural HRTF representation.

This chapter, along with the following two, considers the problem of modeling PRTFs for 3-D sound rendering. Following a structural *modus operandi*, two approaches for PRTF derivation, either by direct measurement (Section 6.1) or by HRIR processing (Section 6.2.1), and an approach for the decomposition of PRTFs into ear resonances and frequency notches (Section 6.2.2) that will allow separate control of the evolution of each physical phenomenon in the final PRTF model, are presented. Results of PRTF decomposition will be finally discussed and further elaborated in Section 6.3.

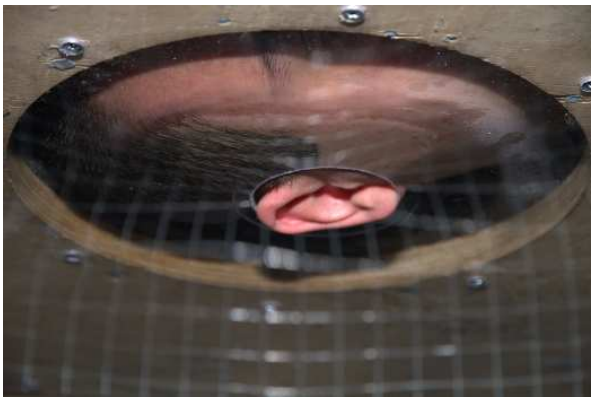
The work presented in this Chapter was published in papers [168] (Section 6.1) and [57] (Sections 6.2- 6.3).

(a) The isolation device, configuration 1.



(b) The isolation device, configuration 2.

(c) Close-up of the pinna hole.



(d) Pinna isolation.

Figure 6.1: *Isolation of the pinna through an ad hoc device.*

6.1 PRTF measurement

In this first Section I will describe a database of PRTFs collected from measurements performed at the Department of Signal Processing and Acoustics, Aalto University, Finland, primarily focusing on the choices and tools through which the final responses were collected (i.e., experimental setup, measurement procedure, and polishing operations applied to obtain the final PRTFs from the measured responses). The database, accompanied by detailed photographs of the subjects' pinnae and the measurement setup, consists of median-plane PRIRs measured at 61 different elevation angles from 25 subjects and is publicly downloadable as a .zip archive from my website at `http://www.dei.unipd.it/~spagnols/PRTF_db.zip`.

6.1.1 Measurement procedure and apparatus

In an ideal situation, the PRTF is the response of the pinna mounted on an infinite plane [5]. For the actual measurements, an ad hoc pinna isolation device that approximates the ideal case,

(a) The measurement setup. On bottom right, the boom-controlled loudspeaker used for sweep reproduction.

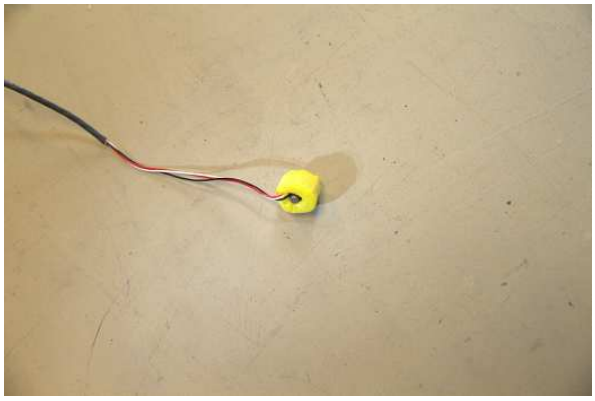
(b) Subject position during the measurements.

Figure 6.2: *Measurement setup and subject position.*

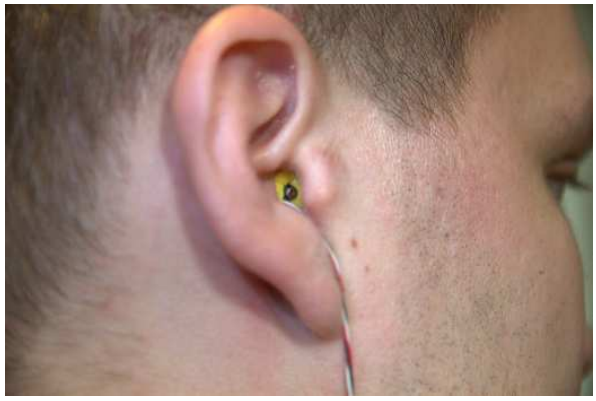
pictured in Fig. 6.1(a), was built and used. The test subjects' torso and shoulders were isolated by a 1-m $\times$ 1-m, 15-mm thick wooden board having a 24-cm-diameter circular hole in the middle of it that approximately fits the size of the human head. A polycarbonate sheet with grinded edges and a 6-cm-diameter circular hole in the middle was fixed with a dozen flat head screws to the board in order to completely cover the hole for the head while letting the subjects' right pinna come out of the other side of it (see Fig. 6.1(c)-(d)). Furthermore, a thick layer of foam with a head-profile-shaped cut in the middle was glued to the upper side of the board with the purpose of adding comfort to the subjects. A piece of such layer could be taken off accordingly with the specific subject's build (Fig. 6.1(b)).

The isolation device was brought right in the middle of an anechoic chamber and placed over an acoustically transparent, one meter high cylindrical metallic fence having 1.75mm thread width in order to avoid reflections from prospective table legs. A controlled boom mounted on the room's ceiling had the purpose of moving the sound source (a Genelec 8030A loudspeaker) along a circumference centered in the pinna hole and laying on the plane parallel to the isolation device. The loudspeaker was positioned upside down, so that the woofer was at the level of the forementioned plane while the tweeter was under it, allowing high-frequency components to directly join the pinna hole without reflecting on the border of the isolation device. Fig. 6.2(a) reports a global view of this experimental setup.

Furthermore, the distance between the loudspeaker and the pinna hole was approximately 1.6 m, so that the incident wave can be assumed plane for frequencies above 3 kHz (the loudspeaker's crossover frequency). This assumption may not be guaranteed below 3 kHz, yet the relative little importance of pinna features below this threshold makes this problem negligible. Since the boom was not acoustically transparent and other loudspeakers were fixed to the chamber's walls, the environment should be labeled as low-echoic rather than anechoic. In spite of this, all the data



(a) Knowles FG-23329 microphone stuffed inside a hollow earplug.



(b) Placement inside the ear canal.

Figure 6.3: *The microphone used for HRTF acquisition.*

will be adequately windowed so as to discard late reflections occurring on the room's equipment.

25 subjects (18 men and 7 women), mostly students and staff of Aalto University, participated to the measurements, which were performed using the blocked-ear-canal technique [63]: a Knowles FG-23329 microphone carefully stuffed in the middle of a hollow earplug was placed right at the entrance of the right ear canal of each subject in turn, as pictured in Fig. 6.3. Then, the subject was asked to stand in front of one side of the panel (eventually with the help of a pedestal to let his waist reach the level of the isolation device), bend 90 degrees forwards and lay his head on the right side in order to let his pinna pass the hole (see Fig. 6.2(b)). The required $90°$ head-neck rotation could be reached thanks to the thick layer of foam which allowed the right shoulder to sink at a lower level than the left. This way, the plane spanned by the loudspeaker's rotation approximately corresponded to the subject's median plane. The pinna position was then adjusted both by instructing the subject on how to move his head and by manual intervention through a big hole in the fence. Finally, vertical orientation was adjusted by manually rotating the subject's head to let his ear axis point at a precise mark on one of the chamber's walls. Subjects were told to remain as still as possible, yet their movements were not monitored during the actual measurement session.

The responses were measured via the logarithmic sweep (or logsweep) method [121]. The used sine sweep had 48 kHz sampling frequency, 1 s duration, and exponentially spanned the frequency range from 20 Hz to 22 kHz. By controlling the boom rotation and sweep reproduction from a Max/MSP patch running on a workstation just outside the anechoic chamber, sweep responses for 61 different elevation angles were recorded at 48 kHz sampling frequency in approximately six minutes' time per subject. The selected elevation angles, considering the interaural polar coordinate system, spanned the range from $\phi = -60°$ to $\phi = 240°$ (i.e. $\phi = -120°$) at 5-degree steps. The boom constantly rotated during the measurements, hence high frequencies

(a) Subject 02. (b) Subject 08. (c) Subject 13. (d) Subject 18.

Figure 6.4: *Right pinnae of four participating subjects.*

were measured from a slightly different elevation angle than low frequencies. However, since the angular speed was almost constantly less that one degree per second, its impact on measurements looks negligible.

In addition, free-field responses were taken by placing the microphone-stuffed earplug inside a small foam cut, positioning it in the middle of the pinna hole of the isolation device, and repeating the measurement procedure in the same way as for the test subjects.

Pictures of the subjects' right pinnae were also taken before or after the measurements (see Fig. 6.4). The distance and orientation of the camera with respect to the pinna was kept as constant among subjects as possible through the help of a tripod. Also, each subject's pinna height (variable d_5 in Fig. 4.9) was measured and tracked down for resizing purposes. This information, along with each subject's sex and evidenced anomalies in the experiment with respect to the optimal situation, can be found in Table 6.1. As for anomalies, Subject 06's pinna did not completely pass the hole, Subject 13 had a piercing on the helix which could not be taken off, and Subject 18 had the earplug slightly displaced at the end of the measurements.

6.1.2 Data post-processing

According to the logsweep method, inverse filtering was performed on the measured sweeps (including free-field sweeps) in order to obtain the corresponding impulse responses. Specifically, the inverse response of the excitation signal was first computed and then low-passed and high-passed with fifth-order digital Butterworth filters to compensate for the original zero sound pressure level below 20 Hz and above 22 kHz in the sweep signal. Since the pinna has no effect below 3 kHz and sounds above 15-20 kHz are hardly perceptible by humans, the high-pass and low-pass Butterworth filters' cutoff frequency was kept loose, that is 1.2 and 21.6 kHz respectively. Hence, each impulse response was calculated by convolving such band-passed inverse

subject	sex	pinna height	anomalies
01	F	5.6 cm	no
02	M	6.5 cm	no
03	M	6.5 cm	no
04	M	6.5 cm	no
05	F	5.9 cm	no
06	M	6.9 cm	yes
07	M	6.2 cm	no
08	M	6.3 cm	no
09	F	6.0 cm	no
10	M	6.1 cm	no
11	M	6.7 cm	no
12	M	6.3 cm	no
13	F	6.2 cm	yes
14	M	6.3 cm	no
15	M	5.8 cm	no
16	M	6.5 cm	no
17	F	5.9 cm	no
18	M	6.7 cm	yes
19	M	7.2 cm	no
20	M	5.6 cm	no
21	F	5.8 cm	no
22	M	6.4 cm	no
23	M	6.3 cm	no
24	M	6.0 cm	no
25	F	5.6 cm	no

Table 6.1: *PRTF database: subjects' information.*

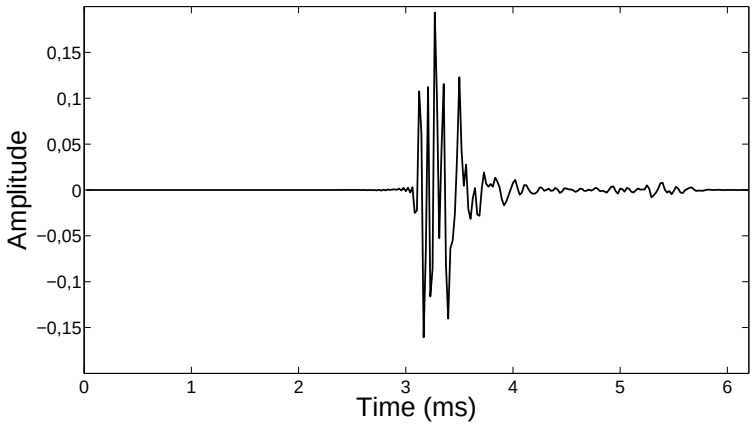


Figure 6.5: *PRIR for elevation* $\phi = 0°$*, subject* 04*.*

filter with the measured sweep.

Subsequently, a 300-sample Hann window was applied to each impulse response with the aim of cutting off late reflections possibly occurring on the subject's legs, the pedestal, or the room equipment. The window was centered in the first positive peak p exceeding a heuristic amplitude threshold in the impulse response, so that the windowed impulse lasted approximately 3 ms from p.

Finally, free-field compensation of the subjects' impulse responses had to be performed. To this end, for each elevation ϕ, a 10^{th}-order minimum-phase IIR filter which approximates the magnitude of the inverse free-field response at source elevation ϕ was designed through the least-squares fit procedure provided by the Yule-Walker method of ARMA spectral estimation [51]. As expected, all of the free-field responses had similar and almost flat - except for a ripple around 2.5 kHz probably due to the loudspeaker's crossover frequency - magnitude plots, with no tangible diffraction occurring on the wooden board. This result certifies the transparency of the measurement setup. Straightforward filtering of the subject's impulse response at elevation ϕ through the so built IIR filter gave the free-field compensated, final pinna-related impulse response (PRIR) that is currently stored in the online database, an example of which can be seen in Fig. 6.5.

Fig. 6.6 shows the magnitude plots of an original recorded sweep and the corresponding post-processed PRTF. It can be clearly seen how the general notch/resonance structure of the typical PRTF is preserved, excluding the very upper and lower frequency ranges which are, however, not of interest in this context.

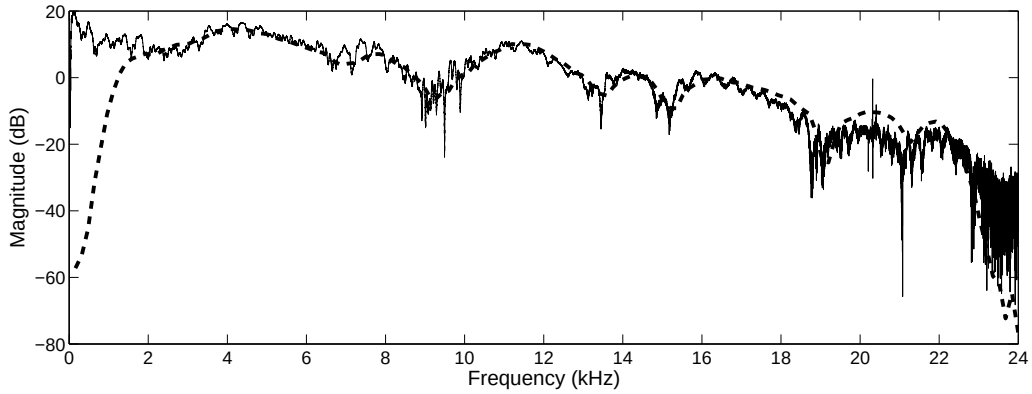


Figure 6.6: *Original sweep magnitude response (solid line) and post-processed PRTF magnitude (thick dashed line).*

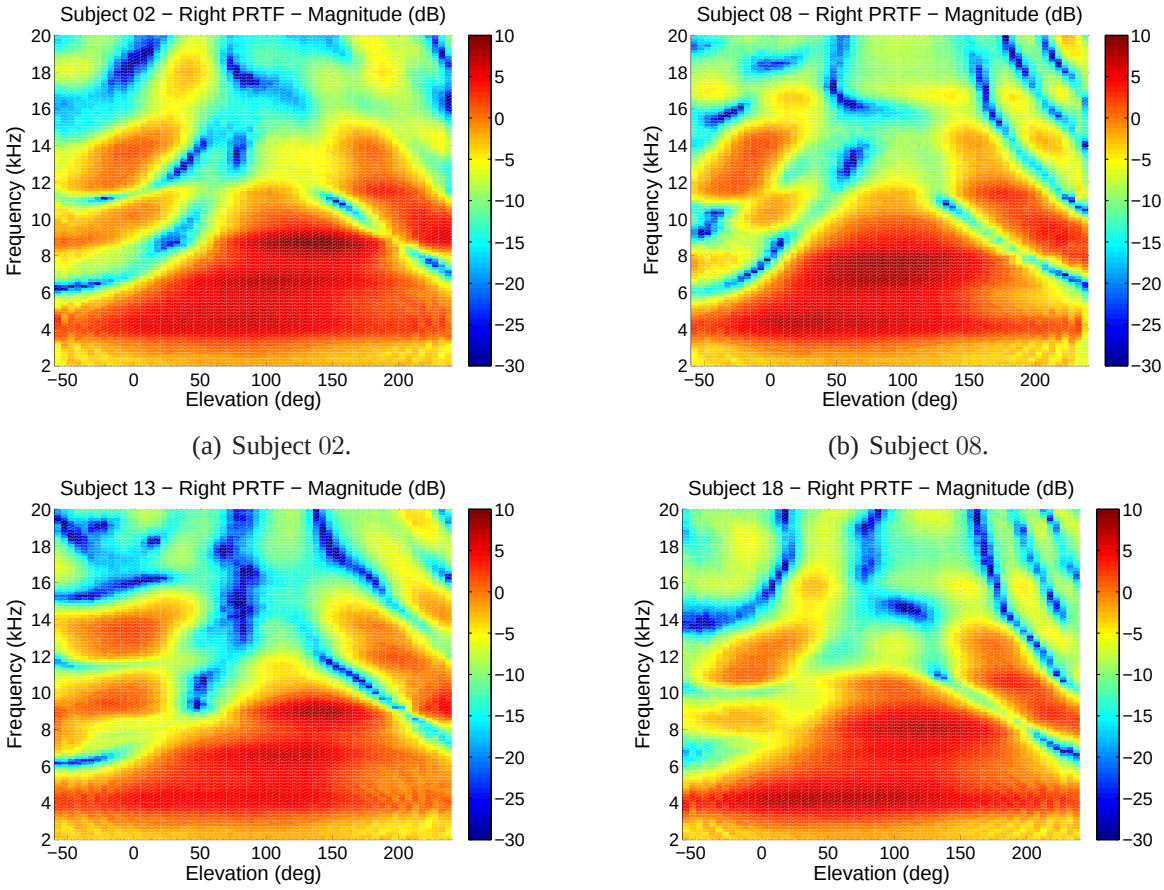
6.1.3 Early results

Through direct inspection of the PRTF magnitude plots of all 25 subjects, a couple of observations can be made. First, when the source is ahead of the frontal plane (in this case when $-60° \leq \phi < 90°$), the PRTF behaviour is quite complex and greatly varies from subject to subject. However, commonly known features evidenced in previous works on PRTFs [159, 81], such as the 4-kHz omnidirectional resonance mode and the notch whose frequency ($6 - 10$ kHz) increases with elevation, appear in the vast majority of subjects, as can be seen in Fig. 6.7). In some cases (e.g. Subject 18), however, the reflection structure is unclear, the magnitude plot presenting valleys which happen to be excessively shallow.

Secondly, while all PRTFs greatly differ among subjects when the source is ahead of the frontal plane, their behaviour is similar for all other elevations. Specifically, allowing some degree of approximation:

- for $90° \leq \phi \leq 125°$ the majority of PRTFs show a descending magnitude plot with one major resonance around 7 kHz and no evident notches;
- from about $\phi = 130°$ one frequency notch appears at around 10 kHz, eventually followed by others at higher frequencies when the source is about to cross the horizontal plane at $\phi = 180°$ (this notch was found in [81] too);
- PRTFs for the last elevation angles, especially $\phi = 240°$, show a more complex magnitude structure with 3 or more notches below 15 kHz (also reported in [81]).

These features can all be detected in Fig. 6.7. The absence of evident notches when the source is above the listener may easily be attributed to the presence of the helix which "masks" the concha, evading direct reflections on it. Conversely, the presence of complicated patterns at $\phi = 240°$



(a) Subject 02.

(b) Subject 08.

(c) Subject 13.

(d) Subject 18.

Figure 6.7: *PRTF magnitude plots of four subjects at all available elevations.*

comparable to those for sources ahead of the frontal plane may be both attributed to reflections on different pinna contours such as the upper part of the helix, the tragus or the crus helias, or to possible unwindowed reflections on the subject's legs.

Finally, even after post-processing some PRTFs still present a "noisy" spectrum. This artifact may likely be associated to slight movements of the subjects during the sweep reproduction or to a rattling noise coming from the metallic fence which was reported by a few subjects right after their measurement session. However, besides being isolated cases only, the main features of PRTFs remain preserved. The early results and assumptions traced above will be further investigated later in this thesis, especially for what concerns PRTF behaviour in the elevation range up to $45°$ where pinna modifications happen in greater number.

6.2 PRTF extraction and separation

The procedure for PRTF measurement described in the previous Section requires, just as typical HRTF measurements, specific (expensive) equipment, hence it is hardly replicable. In absence of such facilities, the most straightforward way of obtaining PRTFs it to extract them through signal processing techniques from existing HRTF databases. This Section describes such an approach, where the initial data was chosen to be a set of measured HRIRs taken from the CIPIC database [7], a public-domain database of high spatial resolution HRIR measured at 1250 directions for 45 different subjects. Once a PRTF is derived, a method for separating the contribution of reflections to that due to resonances is described in detail.

6.2.1 Data collection and pre-processing

Extraction of PRTF features first requires an analysis step. Taking as reference system the interaural polar coordinate system previously sketched in Fig. 4.1(a), the focus is placed on median-plane (azimuth angle $\theta = 0°$) HRIRs, with the elevation angle ϕ varying from $\phi = -45°$ to $\phi = 90°$ at 5.625-degree steps (25 HRIRs per subject).

The first problem that needs to be addressed is how to extract the PRTF from the corresponding HRIR: basically, the head, torso and shoulders contributions need to be discarded from the response. Knowing that pinna reflection delays usually range between 100 and 300μs in the median plane [12], the HRIR is shortened by applying a 1-ms Hann window starting from the HRIR onset [137]. In this way spectral effects due to reflections caused by shoulders and torso are removed from the response, while those due to the pinna are preserved.

Concerning head diffraction compensation, if the pinnaless head is treated as a sphere, then the ear canal lies around $\theta = \pm 90° - 100°$.[1] Assuming CIPIC HRTF measurements, taken at 1-meter distance, to be comparable to far-field measurements, it can be directly seen from Fig. 4.10(a) that the corresponding responses of spherical diffraction for a source in the frontal side of the median plane are approximately flat. Further evidence of such "flatness" is found in [113], where the authors show that the mean spectral distance between measured responses on a complete KEMAR head and FDTD-simulated responses on its pinna alone is 2.3 dB only.

As a consequence, no further preprocessing step is applied to the windowed and zero-padded HRIR, whose FFT, calculated on a 512-sample window size, yields the estimated PRTF. Fig. 6.8 reports an example of extracted PRTF, where spectral notches and resonances can be easily detected.

Both in the PRTF database described in Section 6.1 and in the post-processed CIPIC HRTF responses, median-plane data was considered. Therefore, it ought to be mentioned that since the data was collected for a single azimuth value only there is no guarantee that integrating a future

[1] Since human ears typically lie slightly behind and below the x axis [2], the source-ear angular distance is certainly greater than $90°$ for sources between $\phi = 0°$ and $\phi = 45°$ at least.

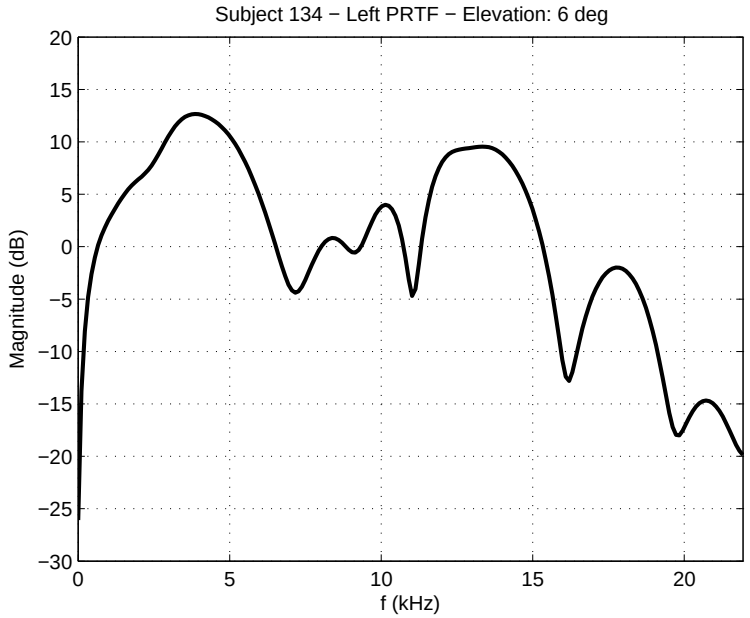


Figure 6.8: *PRTF extracted from a CIPIC database HRIR.*

pinna model based on these responses in a complete structural model would give an appropriate representation of the HRTF. In other words, the PRTF for elevation ϕ and azimuth $\theta = 0°$ may have a totally different look than the PRTF for elevation ϕ and e.g. azimuth $\theta = 60°$. However, relative azimuthal variations up to at least $\Delta\theta = 30°$ at fixed elevation cause very slight spectral changes in the PRTF [114, 96, 137]. This observation was supported by an informal personal inspection of different PRTF sets too. Hence, under the assumption that the source moves in the vicinity of the median plane, pinna effects can be thought of solely depending on source elevation.

6.2.2 The separation algorithm

The following issue concerns feature extraction from the obtained PRTF, with the constraint that reflections and resonances must be treated as two separated phenomena. To this end, the PRTF can be split into a "resonant" and a "reflective" component by means of an ad-hoc designed algorithm which I will refer to as *separation algorithm* and describe in detail. The idea that drives the algorithm is the iterative compensation of the PRTF magnitude spectrum through a sequence of synthetic multi-notch filters until no local notches above a given amplitude threshold are left. Each multi-notch filter is fitted to the shape of the PRTF spectrum at the current iteration with its spectral envelope removed and subtracted to it, giving the spectrum for the next iteration. Eventually, when convergence is reached (say at iteration $\underline{i}$), the final spectrum $H_{res}^{(\underline{i})}$ contains the resonant component, while the reflective component is given by direct combination of all the

calculated n multi-notch filters.

Fig. 6.9 reports the complete flow chart of the separation algorithm. The algorithm's initial conditions heavily influence the final result; indeed, three parameters have to be chosen:

- N_{ceps}, the number of cepstral coefficients used for estimating the PRTF spectral envelope at each iteration;
- D_{min}, the minimum dB depth threshold for notches to be considered;
- ξ, the reduction factor for every notch filter bandwidth, whose purpose will be discussed below.

Before entering the core of the algorithm, let $H_{res}^{(1)}$ match the PRTF and set $H_{refl}^{(1)}$ to 1. These two frequency responses will be updated at each iteration, resulting in $H_{res}^{(i)}$ and $H_{refl}^{(i)}$ at the beginning of the i-th iteration. If $N_{nch}^{(i)}$ is the number of "valid" notches algebraically identified at the end of it, the algorithm will terminate at iteration $\underline{i}$ if $N_{nch}^{(\underline{i})} = 0$, while $H_{res}^{(\underline{i})}$ and $H_{refl}^{(\underline{i})}$ will respectively contain the resonant and reflective components of the PRTF. As one may expect, both the number of iterations and the quality of the decomposition strongly rely on a good choice of the above parameters. For instance, choosing D_{min} too close to zero may lead to an unacceptable number of iterations; conversely, a high value of D_{min} could result in a number of uncompensated notches in the resonant part of the PRTF. In the following, the step-by-step analysis procedure on $H_{res}^{(i)}$ is presented, assuming that $N_{nch}^{(i-1)} > 0$. For the sake of simplicity, in the following the apex (i) indicating iteration number is dropped from all notation.

First, in order to properly extract the local minima due to pinna notches in the PRTF, the resonant component of the spectrum must be compensated for. To this end, the real cepstrum of H_{res} is calculated; then, by liftering the cepstrum with the first N_{ceps} cepstral coefficients and performing the FFT, an estimate C_{res} of the spectral envelope of H_{res} is obtained.

The parameter N_{ceps} must be chosen adequately, since it is crucial in determining the degree of detail of the spectral envelope. As N_{ceps} increases, the notches' contribution is reduced both in magnitude and in passband while the resonance plot becomes more and more detailed. The optimal number of coefficients that capture the resonant structure of the PRTF while leaving all the notches out of the spectral envelope was experimentally found to be $N_{ceps} = 4$. This number also matches the maximum number of modes identified by Shaw [158] which appear at one specific spatial location: for elevations close to zero, modes 1, 4, 5, and 6 are excited. Once C_{res} is computed, it is subtracted from the dB magnitude of H_{res} yielding the residue E_{res}.

At this point E_{res} should present an almost flat spectrum with a certain number of notches. Parameter N_{nch} is first set to the number of local minima in E_{res} deeper than D_{min}, extracted by a simple notch picking algorithm. The aim here is to compensate each notch with a second-order notch filter, defined by three parameters: central frequency f_C, 3-dB bandwidth f_B, and notch depth D.

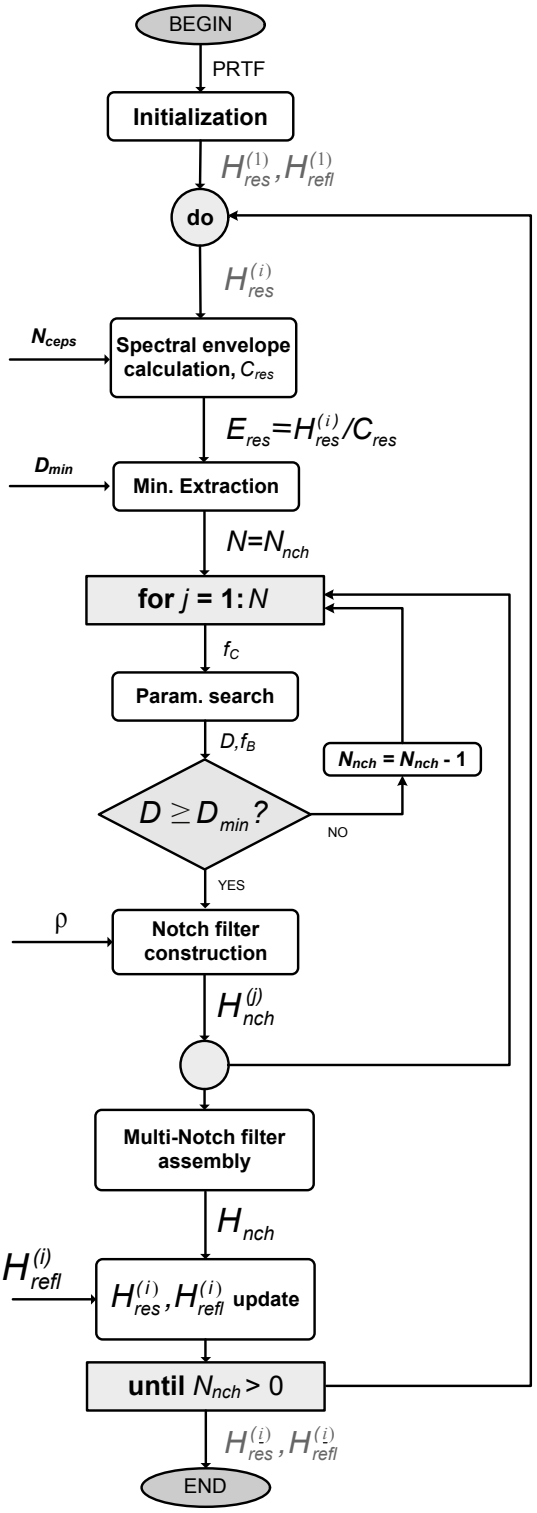


Figure 6.9: *Flow chart of the separation algorithm.*

Consider the j-th local minimum. The central frequency of the corresponding notch f_C is immediately determined, while notch depth is found as $D = |E_{res}(f_C)|$. Computation of f_B is less straightforward. Indeed, f_B is calculated as the standard 3-dB bandwidth, i.e. $f_B = f_r - f_l$, where f_l and f_r are respectively the left and right +3 dB level points relative to f_C in E_{res}, except for the following situations:

1. if $D < 3$ dB, the 3-dB bandwidth is not defined. Then f_r and f_l are placed at an intermediate dB level, halfway between 0 and $-D$ in a linear scale;

2. if the local maximum of E_{res} immediately preceding (following) f_C does not lie above the 0-dB line while the local maximum immediately following (preceding) does, f_B is calculated as twice the half-bandwidth between f_C and f_r (f_l);

3. if both local maxima do not lie above the 0-dB line, E_{res} is vertically shifted until the 0-dB level meets the closest of the two. Then, f_B is calculated as before except if the new notch depth is smaller than D_{min} in the shifted residue plot, in which case the parameter search procedure for the current notch is aborted and N_{nch} is decreased by one.

Note that case 1 may occur simultaneously with respect to case 2 or 3: in this situation, both corresponding effects are considered when calculating f_B.

The so found parameters f_C, D, and f_B need to uniquely define a filter structure. To this end, a second-order notch filter implementation of the form [192]

$$H_{nch}^{(j)}(z) = \frac{1 + (1+k)\frac{H_0}{2} + l(1-k)z^{-1} + (-k-(1+k)\frac{H_0}{2})z^{-2}}{1 + l(1-k)z^{-1} - kz^{-2}}, \tag{6.1}$$

is used, where

$$k = \frac{\tan(\pi\frac{f_B}{f_s}) - V_0}{\tan(\pi\frac{f_B}{f_s}) + V_0}, \tag{6.2}$$

$$l = -\cos(2\pi\frac{f_C}{f_s}), \tag{6.3}$$

$$V_0 = 10^{\frac{D}{20}}, \tag{6.4}$$

$$H_0 = V_0 - 1, \tag{6.5}$$

and f_s is the sampling frequency. Using such an implementation allows to directly fit the parameters to the filter structure. Clearly, not every combination of the three parameters is accurately approximated by the second-order filter: if the notch to be compensated is particularly deep and sharp, the filter will produce a shallower and broader notch, having a center frequency which is slightly less than f_C.

Although moderate frequency shift and attenuation is not detrimental to the estimation algorithm (an underestimated notch will be fully compensated through the following iterations),

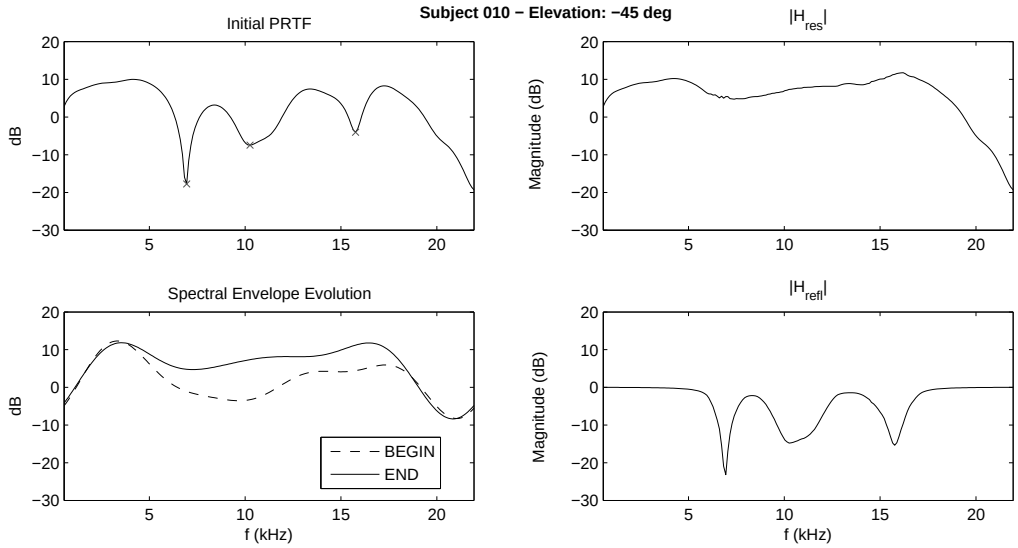


Figure 6.10: *An example of the separation algorithm's evolution. The PRTF magnitude in the top left panel is decomposed into resonances (top right panel) and frequency notches (bottom right panel). The bottom left panel shows the evolution of the PRTF spectral envelope from the first iteration to convergence.*

an excessive notch bandwidth could lead to undesired artifacts in the final resonance spectrum. Here is where parameter ξ comes into play: if f_B is divided by $\xi > 1$, the new bandwidth specification will produce a filter whose notch amplitude will be further reduced, allowing to reach a smaller bandwidth. Typically, in order to achieve a satisfactory trade-off between the size of the reduction factor and the number of iterations, ξ is set to 2.

Consequently, the parameters to be fed to the filter are $(f_C, D, f_B/\xi)$, yielding coefficients vectors $\boldsymbol{b}^{(j)}$ and $\boldsymbol{a}^{(j)}$ for $H_{nch}^{(j)}$. The parameter search and notch filter construction procedures are repeated for all N_{nch} notches. In order to build the complete multi-notch filter H_{nch},

$$H_{nch}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}} = \prod_{j=1}^{N_{nch}} H_{nch}^{(j)}(z), \tag{6.6}$$

it is now sufficient to convolve all the coefficient vectors computed during iteration i:

$$\boldsymbol{b} = [b_0, b_1, b_2] = \boldsymbol{b}^{(1)} * \boldsymbol{b}^{(2)} * \cdots * \boldsymbol{b}^{(N_{nch})}, \tag{6.7}$$

$$\boldsymbol{a} = [a_0, a_1, a_2] = \boldsymbol{a}^{(1)} * \boldsymbol{a}^{(2)} * \cdots * \boldsymbol{a}^{(N_{nch})}. \tag{6.8}$$

Finally, before considering the next iteration, the global multi-notch filter $H_{refl}^{(i+1)} = H_{refl}^{(i)} \cdot H_{nch}$ must be updated and the PRTF compensated by applying $H_{res}^{(i+1)} = H_{res}^{(i)} / H_{nch}$.

Fig. 6.10 illustrates the algorithm's evolution for a particular PRTF. The specific choice of the initial parameters was $N_{ceps} = 4$, $D_{min} = 0.1$ dB, and $\xi = 2$. The top left panel illustrates

Subject 010's PRTF for an elevation of -45 degrees. The bottom left panel reports the spectral envelope evolution, where we can see how interfering spectral notches negatively influence the initial estimate. The panels on the right represent the resonant (H_{res}) and reflective (H_{refl}) parts of the PRTF at the end of the algorithm.

Consider the range where acoustic effects of the pinna are relevant, i.e. the range from 4 to 16 kHz approximately [68]. Fig. 6.10 shows that inside such range the algorithm has produced a realistic decomposition: the gain of the reflective component is unitary outside the notch regions, while the peaks appearing in the resonant component reveal a good correspondence to Shaw's modes (this point is further discussed in the next Section). Outside the relevant range for the pinna, there is a sharp gain decrease in the resonant part and further imperfections that appear for different subjects and elevations. Nevertheless, this is not a problem as long as the pinna contribution to the HRTF is considered alone. The behavior exemplified in Fig. 6.10 is observed for different elevations and subjects too.

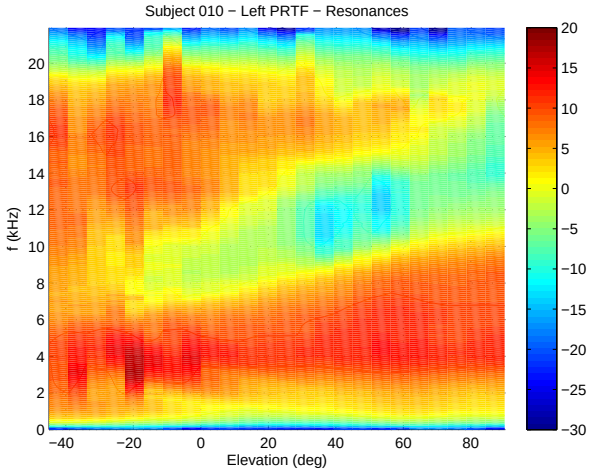
6.3 PRTF analysis: results

PRTF features identified through the decomposition carried out by the separation algorithm are now discussed. The most general result that will be highlighted is that while the resonant component is in broad terms similar among different subjects, the reflective component comes along critically subject-dependent. In order to facilitate comparison with previous works, most of the following plots report results for the same CIPIC subjects that appear in [137] and [149], specifically Subjects 010, 027, 048, 134, and 165 (KEMAR head with small pinnae).

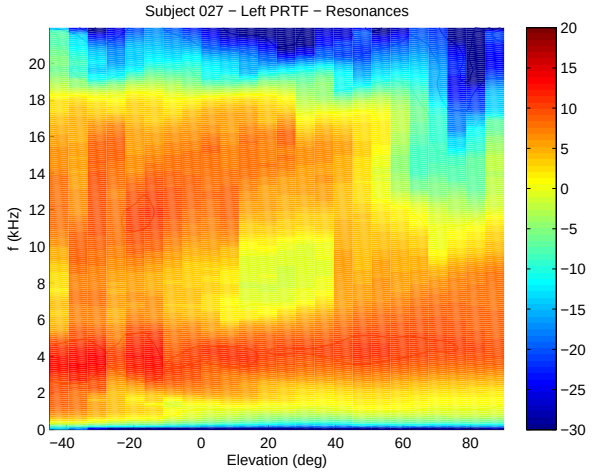
6.3.1 The resonant component

The variation in the contribution of pinna resonances to the PRTF throughout the considered elevation range can be studied by examining the 3-D plots in Fig. 6.11. Two major hot-colored areas can be easily identified in these plots. The first one, centered around 4 kHz, appears to be very similar amongst subjects since it spans all elevations. One may immediately notice that this area includes Shaw's omnidirectional mode. The resonance's bandwidth appears to increase with elevation; however, knowledge of pinna modes implies that a second resonance is likely to interfere within this frequency range at higher elevations, specifically Shaw's vertical mode 2 (centered around 7 kHz with a magnitude of 10 dB).

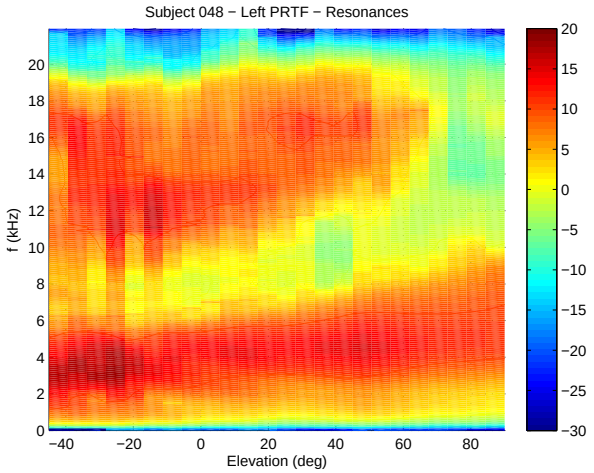
On the other hand, the second hot-colored area differs both in shape and shade amongst subjects. Still it is most prominent at low elevations between 12 and 18 kHz, a frequency range which is in general agreement with Shaw's horizontal modes 4, 5, and 6, and smoothly dissolves as the elevation angle increases up above the horizontal plane. Note that this higher resonance area and Shaw's vertical mode 2 seem to be excited in mutually exclusive elevation ranges. This



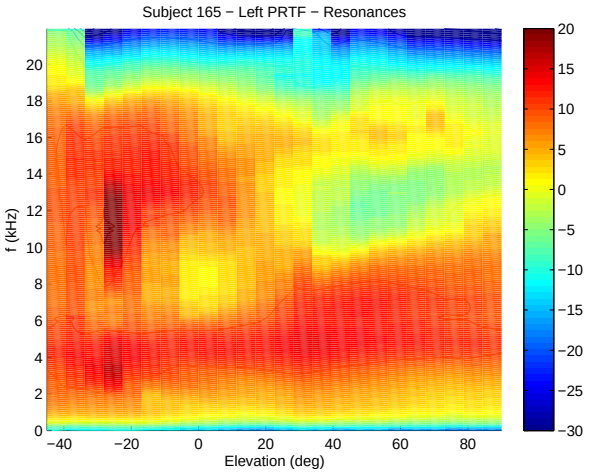
(a) Subject 010.



(b) Subject 027.



(c) Subject 048.

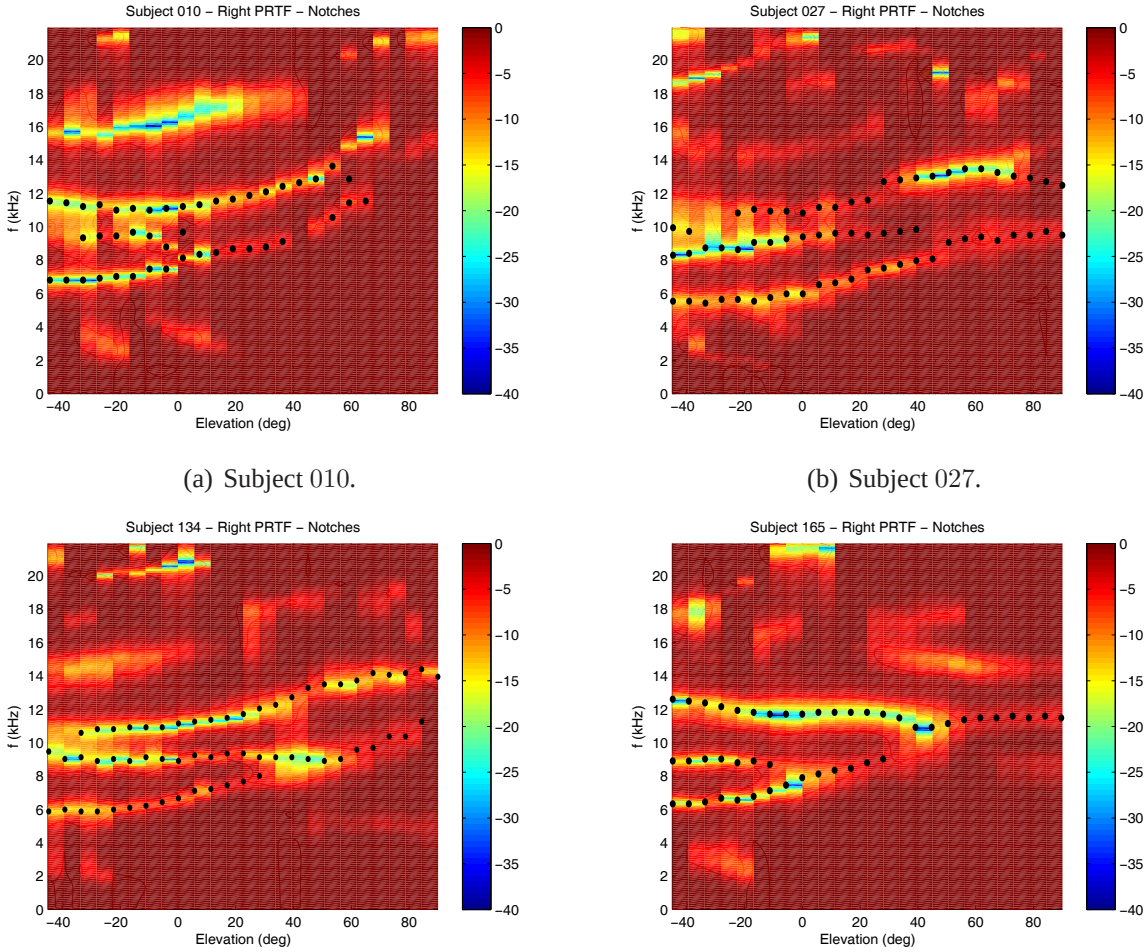


(d) Subject 165.

Figure 6.11: *Resonant component of four subjects' left pinnae for* $-45° \leq \phi \leq 90°$.

effect, which appears for all the analyzed subjects and is especially evident for Subject 165, gives the impression of a smooth transition from one resonance to the other, and allows to look forward to a double-resonance model for the pinna.

Finally, note that the magnitude response around 12 kHz occasionally takes low negative dB values at high elevations, especially in Subject 010's plot: such an incongruity may be explained by phenomena other than reflections or resonances, e.g. diffraction around the pinna. One may advance the same observation for very low and very high frequency zones; nevertheless, these effects are due to the pre-processing step and lie whatever fairly outside the frequency range that interests the pinna.



(a) Subject 010.

(b) Subject 027.

(c) Subject 134.

(d) Subject 165.

Figure 6.12: *Reflective component of four subjects' right pinnae for* $-45° \leq \phi \leq 90°$.

6.3.2 The reflective component

As already mentioned, reflection patterns strongly depend on elevation and pinna shape. Nevertheless, a number of common trends can be acknowledged here too. In general it can be stated that, while PRTFs generally exhibit poor notch structures when the source is above the head, as soon as elevation decreases the number and depth of frequency notches grows to an extent that varies between subjects. These remarks can be immediately verified in Fig. 6.12, where the spectral notches' contribution for four different pinnae are reported. In particular, Subjects 134 and 165 exhibit at low elevations a clear reflection structure with three prominent notches. Still, robust techniques are required in order to track the effective notch patterns along elevation, discarding the weak ones together with those which appear occasionally, and to have a consistent labeling along subsequent elevation angles.

As an adequate solution to this problem, a widely used analysis tool in the field of sinusoidal modeling, the McAulay-Quatieri partial tracking algorithm [105] – originally used to group sinusoidal partials along consecutive temporal windows according to their spectral location – can be inherited and fitted to this context. As a matter of fact, the very algorithm can be exploited to track the most marked notch patterns along elevation, as long as temporal evolution is conceptually replaced by elevation dependency and spectral notches take the role of partials: for this reason, let me refer to it as "notch tracking" algorithm. With respect to its original formulation, suffice it to add that the notch detection (originally "peak detection") step trivially locates all of the local minima in the reflective component's spectrum, and that the matching interval for the notch tracking procedure is set to $\Delta = 3$ kHz.

Since it is preferable to restrict the attention to the frequency range where reflections due to the pinna alone are most likely seen, and ignore notches which are overall feeble hence not likely to be associated with a major reflection, two post-processing steps are performed on the obtained tracks:

- keep only those tracks which remain inside the range $4 - 16$ kHz, where pinna cues are most likely to be detected;
- delete the tracks that do not present a notch deeper than 5 dB.

The dotted tracks superimposed on the plots in Fig. 6.12 represent the outputs of the notch tracking algorithm. Results are definitely akin to the findings by Raykar (Fig. 11(a) in [137]) obtained through the use of the labyrinthine DSP-based algorithm depicted back in Fig. 4.17. In particular, three main tracks are seen for all four subjects, whereas the shorter tracks in the plots of Subject 010 and Subject 027 very probably represent the continuation of the missing track at those specific elevations. Realistically, gaps between tracks may be caused by the algorithm's unlikelihood of locating proper minima due to uncontrollable events such as the superposition of two different notches or the presence of shallow valleys in the considered region of the magnitude plot. Nonetheless, the three aforementioned main tracks indicate that congruous reflection patterns appear in different PRTFs.

As a further step, notch patterns for 20 different CIPIC subjects were analyzed in the elevation range $-45° \leq \phi \leq 45°$ where one notch at least appears at a specific elevation. The majority of them exhibits three notch tracks at a given elevation: only two subjects (Subjects 019 and 020) lack of one track, the lowest and the highest in frequency respectively. Average notch frequencies for the three tracks at each available elevation are reported in Table 6.2: frequencies in the first two tracks (T_1 and T_2) monotonically grow with elevation, while frequencies in the third track (T_3) remain almost constant up to $\phi = -11.25°$, then grow until $\phi = 28.125°$, and decrease at higher elevations on average. These trends were seen to be consistent across subjects. Not reported in the table is the number of subjects that exhibit a notch for each track/elevation coordinate: for the sake of brevity, suffice it to mention that all tracks begin at $-45°$ except for

ϕ	T_1	T_2	T_3
$-45°$	6.10 kHz	8.90 kHz	12.17 kHz
$-39.375°$	6.10 kHz	8.91 kHz	12.14 kHz
$-33.75°$	6.14 kHz	8.96 kHz	12.12 kHz
$-28.125°$	6.31 kHz	8.96 kHz	12.10 kHz
$-22.5°$	6.34 kHz	8.97 kHz	12.18 kHz
$-16.875°$	6.56 kHz	9.10 kHz	12.21 kHz
$-11.25°$	6.75 kHz	9.12 kHz	12.17 kHz
$-5.625°$	6.89 kHz	9.26 kHz	12.37 kHz
$0°$	7.19 kHz	9.42 kHz	12.39 kHz
$5.625°$	7.34 kHz	9.60 kHz	12.49 kHz
$11.25°$	7.71 kHz	9.71 kHz	12.67 kHz
$16.875°$	8.12 kHz	9.81 kHz	12.72 kHz
$22.5°$	8.21 kHz	9.89 kHz	13.00 kHz
$28.125°$	8.24 kHz	10.09 kHz	13.38 kHz
$33.75°$	8.44 kHz	10.27 kHz	13.16 kHz
$39.375°$	8.76 kHz	10.69 kHz	12.92 kHz
$45°$	9.21 kHz	10.84 kHz	12.64 kHz

Table 6.2: *Notch frequencies averaged across* 20 *subjects per elevation and track.*

three cases only, that T_1 terminates earlier than T_2 on average, and the same applies to T_2 with respect to T_3.

6.4 Conclusions

Analysis of the PRTF resonant component in different CIPIC subjects revealed common trends with respect to elevation: in particular, two prominent peaks at quasi-steady central frequencies can be distinctly identified in the considered frequency range. Similarly, analysis of the PRTF reflective component with notch tracking along elevation angles highlighted the presence, in the vast majority of subjects, of three main (and apparently continuous) notch tracks between 5 and 15 kHz approximately, whose evolution will be directly related to the location of reflection points over pinna surfaces. These two results suggest that Satarzadeh's filter model (see again Fig. 4.18) can be generalized through consideration of multiple reflection paths, and extended to a wider frontal space. This can be done by construction of three different notch filters, each tuned to a specific anthropometric measure, as a replacement to the simpler comb filter. Fig. 6.13 describes such an extension, that will be widely analyzed in Chapter 8.

Nevertheless, since robust common trends cannot be identified at a first glance in the evolu-

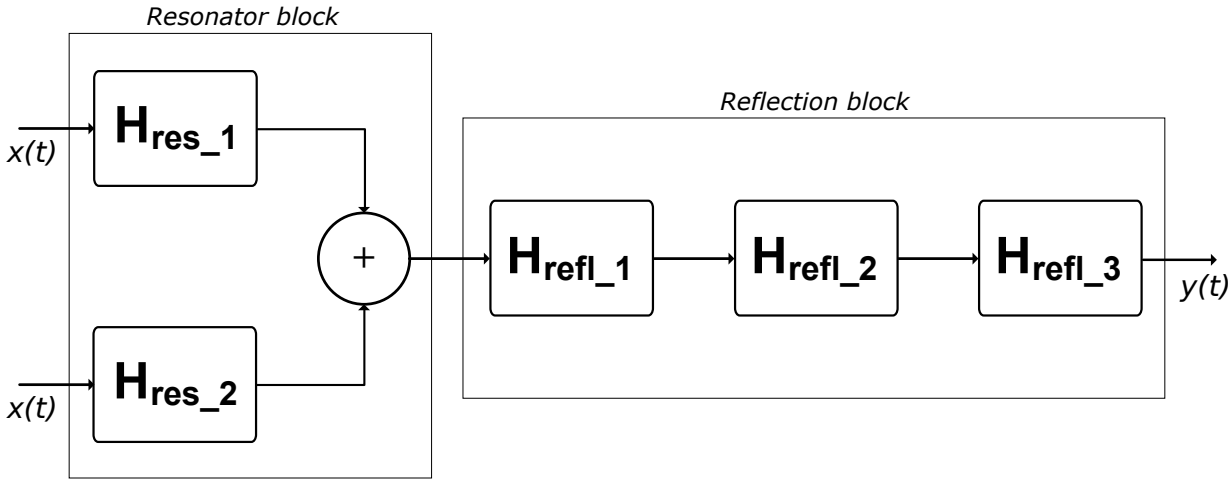


Figure 6.13: *General model for PRTF reconstruction.*

tion of spectral notches, and following the common idea that notches are of major relevance for elevation detection in the frontal region [75, 118, 189, 68], attention in the next chapter will be focused onto the reflective component.

Chapter 7

Pinna-Related Transfer Functions: Relation to Anthropometry

There is no doubt that the greatest dissimilarities among different people's HRTFs for a same spatial location are due to the massive subject-to-subject pinna shape diversity, which reflects itself onto different resonance and reflection patterns in the corresponding PRTF. This chapter is dedicated to investigation of the anthropometric mapping lying behind the frequency location of notches in the PRTF spectrum. After an informal ray-tracing analysis on four subjects similar to the one described in [137] and sketched in Section 7.1, a formal analysis of the optimal mapping between contours extracted from a pinna picture and PRTF frequency notches on twenty CIPIC database subjects is described in Section 7.2. Results are presented and discussed in Sections 7.3 and 7.4, respectively.

Similarly to [137], each notch is associated to its own reflection path. Note that since notch tracks in PRTFs are pairwise in non-harmonic relationship, both on average (see again Table 6.2) and for every single analyzed subject, a single reflection path cannot be assigned to any pair of tracks. Hence the assumption that each notch in the considered frequency range is the result of a distinct reflection path is well-grounded.

Also, similarly to previous works on reflection modeling [137, 149], central frequency is considered as the most relevant notch feature. Inspection of different PRTF plots reveals that the notch moves continuously along the frequency axis depending on the elevation angle [159, 68] to an extent that can definitely be detected by the human auditory system [118]. Conversely, changes in notch bandwidth and amplitude along elevation are seen to be far less systematic (this point will be further discussed in the next chapter), and their perceptual relevance is little understood in previous literature.

The work presented in this Chapter was published in papers [164], [165], and [166] (Section 7.1) and has been submitted for publication in [167] (Sections 7.2- 7.4).

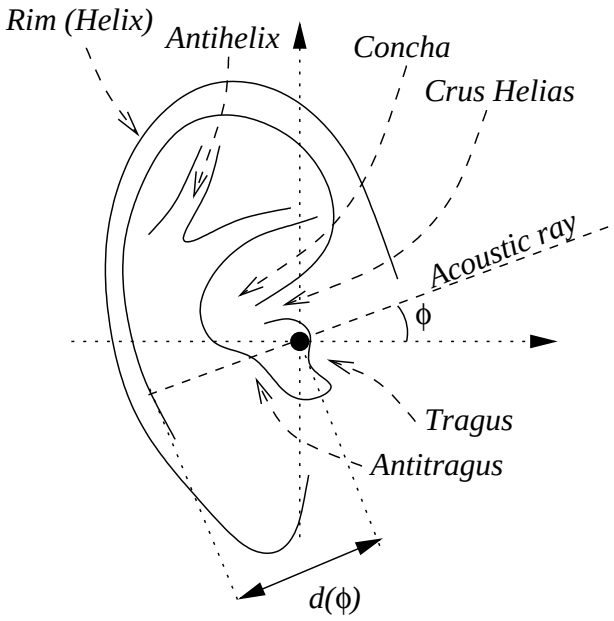


Figure 7.1: *Reflection ray-tracing on the pinna.*

7.1 Reflections and ray tracing

Ray-tracing reflection models [68] are based on a very simple and intuitive observation: the elevation-dependent temporal delay $t_d(\phi)$ between the direct and the reflected wave projects the point of reflection at distance

$$d_c(\phi) = \frac{ct_d(\phi)}{2} \tag{7.1}$$

from the ear canal (where c is again the speed of sound), as can be seen from the raw model reported in Fig. 7.1. Knowing the simple law described by Eq. (7.1), and assuming the reflection coefficient to be positive (which is the typical case), then destructive interference (i.e., a notch) will appear at all those frequencies where the reflection's phase shift equals π:

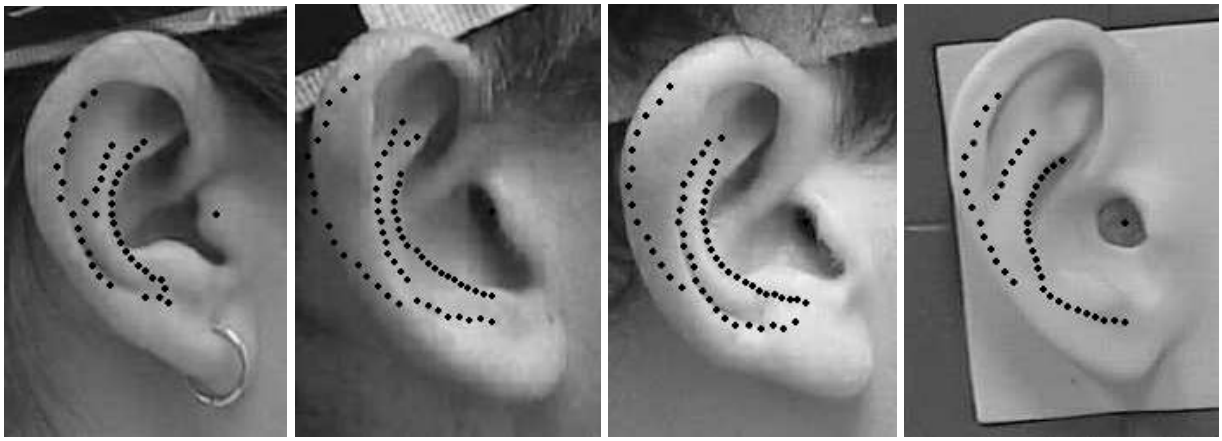
$$f_n(\phi) = \frac{2n+1}{2t_d(\phi)} = \frac{c(2n+1)}{4d_c(\phi)}, \quad n = 0, 1, \ldots. \tag{7.2}$$

Hence the first notch falls at frequency

$$f_0(\phi) = \frac{c}{4d_c(\phi)}. \tag{7.3}$$

The positive reflection assumption was also adopted by Raykar [137] when tracing reflection points over pinna images based on the extracted notch frequencies.

Nevertheless, Satarzadeh [148] drew attention to the fact that almost 80% out of a test bed of 20 CIPIC subjects exhibit a clear negative reflection in their HRIRs. With the help of a simple



(a) Subject 010. (b) Subject 027. (c) Subject 134. (d) Subject 165.

Figure 7.2: *Ray-traced reflection points on four CIPIC subjects' right pinnae.*

physical model of the pinna he argued that, since the impedance of the pinna is greater than that of air, there may be a boundary created by an impedance discontinuity which could produce its own reflection and ultimately reverse the phase of the wave. In case of negative reflection destructive interference would not appear at half-wavelength delays anymore, but at full-wavelength delays. Eqs. (7.2) and (7.3) would then become

$$f_n(\phi) = \frac{n+1}{t_d(\phi)} = \frac{c(n+1)}{2d_c(\phi)}, \quad n = 0, 1, \ldots \tag{7.4}$$

and

$$f_0(\phi) = \frac{c}{2d_c(\phi)}. \tag{7.5}$$

Following Satarzadeh's hypothesis, this last assumption is now exploited in a simple ray-tracing procedure over pinna pictures of four CIPIC subjects: Subjects 010, 027, 134, and 165, whose PRTFs were analyzed in [137] too. Right pinna images are first uniformly rescaled in order to match parameters d_5 (pinna height) and d_6 (pinna width) in Fig. 4.9. Each notch frequency f_0 is then extracted as described in the previous chapter and associated to a single reflection point. Finally, the distance of each reflection point with respect to the entrance of the ear canal is calculated by reversing Eq. (7.5) and, considering the 2-D polar coordinate system illustrated in Fig. 7.1 having the right ear canal entrance as origin, each notch is mapped to the point $(d(\phi), \pi + \phi)$. Clearly, the negative reflection coefficient assumption causes distances to be doubled with respect to those calculated in [137].

Results are reported in Fig. 7.2. For all the subjects, the so-obtained mapping shows a high degree of correspondence between computed reflection points and pinna geometry. One can immediately notice that the track nearest to the ear canal very closely follows the concha wall of each subject for all elevations, except for a couple of cases:

Figure 7.3: *The* 20 *pinna pictures used in the contour matching procedure.*

- at low elevations, displacement of points may be caused by the little extra distance needed by the wave to pass over the crus helias;
- Subject 010's track disappears at around $\phi = 60°$ probably because of the insufficient space between tragus and antitragus that causes the incoming wave to reflect outside the concha.

The intermediate track falls upon the area between concha and rim, with variable length among subjects:

- in the case of subjects 010 and 165 the track is faint and probably due to the antihelix;
- conversely, subjects 027 and 134 present a longer and deeper track, that is visually associated to a reflection on the rim's edge.

Finally, the furthest track follows the shape of the rim and stops in the vicinity of the point where the rim terminates, hence it is likely to be associated to a reflection in the inner wall of it, except for Subject 010 whose reflection occurs at the rim's edge.

Such an attempt towards the explanation of the pinna reflection process resulting in the most important spectral notches in the PRTF provided visually convincing results. However, in order to fully justify these preliminary findings, a rigorous analysis using a vast test bed of subjects is required. This problem is addressed in the following section.

7.2 The contour matching procedure

Attention is now restricted to the elevation range $-45° \leq \phi \leq 45°$. The upper elevation limit is chosen because of the high degree of uncertainty in elevation judgement for sources from $\phi > 45°$ [17, 119] and the general lack of deep spectral notches in PRTFs in this region [81,

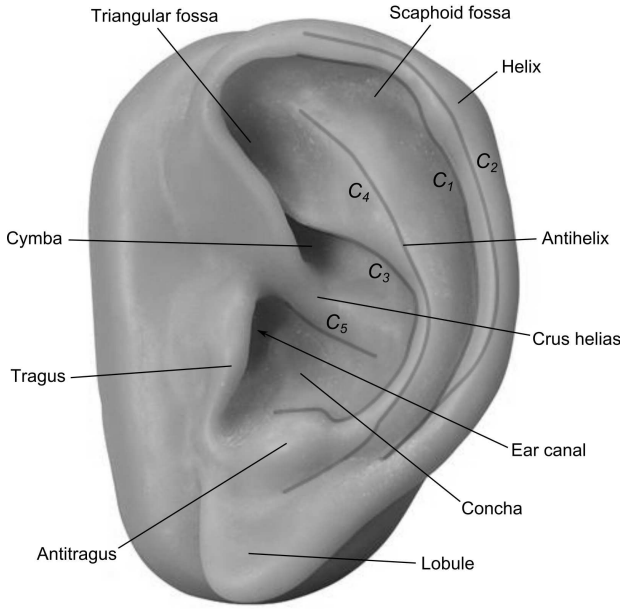


Figure 7.4: *Pinna anatomy and the five chosen contours for the matching procedure.* C_1*: helix border;* C_2*: helix wall;* C_3*: concha border;* C_4*: antihelix and concha wall;* C_5*: crus helias.*

137], which may besides be two faces of the same coin. Again as in the previous informal analysis, each extracted notch frequency is treated as the f_0 of its respective reflection. Based on these choices, the correspondence between pinna anatomy and theoretical reflection points under different reflection sign conditions on a wide morphological variety of CIPIC subjects' pinnae, a glimpse of which is retained by Fig. 7.3, is now investigated. Since this work involves the anthropometry of subjects in the form of a picture of their left or right pinna, attention is restricted to the 20 of them for which the corresponding photograph is available [1]: subjects 003, 008, 009, 010, 011, 012, 015, 017, 019, 020, 021 (KEMAR with large pinna), 027, 028, 033, 040, 044, 048, 050, 134, and 165 (KEMAR with small pinna).

7.2.1 Pinna contour extraction

The basic assumption that drives the analysis procedure is that each notch track is associated to a distinct reflection surface on the subject's pinna. Since the available data for each subject is a side-view of his/her head showing the left or right pinna, extraction of the "candidate" reflection surfaces must be reduced to a two-dimensional basis. I chose to investigate as possible reflection surfaces a set of three contours directly recognizeable from the pinna photograph, together with two hidden surfaces approximating the real inner back walls of the concha and helix. Specifically, as Fig. 7.4 depicts, the following contours are considered:

1. helix border (C_1), visible on picture;
2. helix inner wall (C_2), following the jutting light surface at the helix approximately halfway between the rim border and the rim outer wall;
3. concha outer border (C_3), visible on picture;
4. antihelix and concha inner wall (C_4), following the jutting light surface just behind the concha outer border up to the shaded area below the antitragus;
5. crus helias inferior surface (C_5), visible on picture.

Since automatic contour extraction is hard to obtain because of both the low resolution of the pictures and the presence of hidden contours, the extraction procedure was performed by manual tracing through a pen tablet. Photographs were accurately resized to match a $1:1$ scale based on the quantitative pinna height parameter (d_5 in Fig. 4.9) available from the HRTF database's anthropometric data [7], or based on the measuring tape pictured in the photograph close to the pinna in those cases where d_5 was not defined. Right pinna photographs were horizontally mirrored so that all pinnae headed left, and contours were drawn and stored as sequences of pixels in the post-processed image. Of all the contours, C_4 was the hardest to recognize due to the forementioned low resolution; it is therefore necessary to point out that in some cases the lower part of this contour was almost blindly traced.

7.2.2 Contour matching algorithm

Before describing the contour matching procedure, some useful definitions are formally stated.

- the *focus* $\psi = (\psi_x, \psi_y)$ is the reference point where the direct and reflected waves meet, usually set at the entrance of the ear canal where the microphone is assumed to have been placed during HRTF measurements;
- the rotation ρ is a tolerance on elevation that counterbalances possible angular mismatches between the actual orientation of the subject's ear and the picture's x-axis;
- a *reflection sign configuration* $\boldsymbol{s} = [s_1, s_2, s_3]$ (with $s_j = \{0, 1\}$), abbreviated as *configuration*, is the combination of reflection coefficient signs attributed to the three notch tracks $\{T_1, T_2, T_3\}$. Here s_j takes 0 value if a negative sign is attributed to T_j and 1 otherwise;
- the *distance* $d(p, C_i)$ *between a point* p *and a contour* C_i is defined as the Euclidean distance between p and the nearest point of C_i lying in the 5-degree elevation range centered in p with the focus as reference point.

The main goal of this analysis is to discover which of the 8 configurations is the most likely to hold according to an error measure between extracted contours and ray-traced notch tracks.

First, in order to perform ray tracing for each configuration $\boldsymbol{s} = [s_1, s_2, s_3]$ the focus needs to be known. Unfortunately, no documentation on the exact microphone position is provided with the CIPIC database; hence, in order to avoid blind focus fixing, an optimization procedure is run pixelwise over a rectangular search area A of the pinna picture covering the whole ear canal entrance. Also, a rotation tolerance $\rho \in I = [-5°, 5°]$ at 1-degree steps is considered. More in detail, for each track T_j the corresponding notch frequencies $f_0^j(\phi)$, $j = \{1, 2, 3\}$, are firstly translated into Euclidean distances (in pixels) through a sign-dependent combination of Eqs. (7.3) and (7.5),

$$d_c^j(\phi) = \frac{c}{2(s_j + 1)f_0^j(\phi)}, \tag{7.6}$$

and subsequently projected onto the point

$$p_{\psi,\rho}^j(\phi) = (\psi_x + d_c^j(\phi)\cos(\phi + \rho), \psi_y + d_c^j(\phi)\sin(\phi + \rho)) \tag{7.7}$$

on the pinna image. The optimal focus and rotation of the configuration, $(\psi_{\mathbf{s}}^{\text{opt}}, \rho_{\mathbf{s}}^{\text{opt}})$, are then defined as those sastisfying the following minimization problem:

$$\min_{\psi \in A, \rho \in I} \sum_{j=1}^{3} \min_i d_{\psi,\rho}(T_j, C_i)^2, \tag{7.8}$$

where $d_{\psi,\rho}(T_j, C_i)$ is the distance between track T_j and contour C_i, which is defined as the average of distances $d(p_{\psi,\rho}^j(\phi), C_i)$ across all the track points. Obviously, for those subjects that lack a notch track the mean is computed onto two tracks only.

Having fixed the eight optimal foci and rotations, one per configuration, a simple scoring function is now used to indicate the *fitness* of each configuration. This is defined as

$$F(\boldsymbol{s}) = \frac{1}{3} \sum_{j=1}^{3} \min_i \frac{d_{\psi_{\mathbf{s}}^{\text{opt}}, \rho_{\mathbf{s}}^{\text{opt}}}(T_j, C_i)}{2 - s_j}, \tag{7.9}$$

that is, the mean of all the (linear) distances between each ray-traced track T_j, $j = 1, 2, 3$, and its nearest contour C_i, $i = 1, \ldots, 5$. Note that the innermost quantity in Eq. (7.9) is scaled by a 2 factor if the reflection sign is negative; this factor takes into account the halvened resolution of the ray-traced negative reflection with respect to a positive reflection. The smaller the fitness value, the larger the fit, clearly.

7.3 Contour matching procedure: results

The above contour matching procedure was run for all the 20 considered CIPIC subjects. Table 7.1 summarizes the final scores (fitness values) for all possible configurations, while Table 7.2

Subject	$F(0,0,0)$	$F(0,0,1)$	$F(0,1,0)$	$F(0,1,1)$	$F(1,0,0)$	$F(1,0,1)$	$F(1,1,0)$	$F(1,1,1)$
003	4.03	9.19	9.27	13.78	7.83	12.45	13.03	17.54
008	2.95	4.86	5.33	7.30	3.69	7.89	5.58	10.64
009	2.55	5.18	4.79	7.02	2.95	5.08	2.94	5.01
010	1.88	5.18	2.26	6.02	3.57	5.69	4.46	6.70
011	2.62	5.10	5.60	9.53	3.16	5.79	4.97	9.25
012	2.08	4.21	4.76	7.30	2.70	5.32	3.20	6.78
015	4.99	9.92	6.14	10.59	3.02	6.70	3.39	3.19
017	2.81	6.35	4.53	8.12	2.99	5.02	5.63	6.79
019	1.64	6.64	4.85	8.00	1.64	6.64	4.85	8.00
020	1.15	1.15	5.27	5.27	1.85	1.85	5.45	5.45
021	2.90	6.40	4.06	8.44	3.30	8.97	6.25	11.54
027	2.07	6.53	5.04	8.56	2.32	5.27	2.80	4.25
028	1.71	3.54	4.21	5.57	3.79	4.02	5.62	6.10
033	2.51	4.73	6.66	6.61	3.42	7.68	9.08	9.98
040	1.74	5.48	2.59	5.35	2.57	5.86	3.30	5.96
044	1.88	2.84	5.33	4.81	2.86	2.49	4.13	3.74
048	2.02	5.33	5.45	7.86	3.70	5.06	5.27	6.97
050	3.25	6.29	7.68	10.52	4.37	7.59	7.57	11.23
134	1.64	6.11	5.18	8.56	3.38	6.31	4.56	7.37
165	1.09	5.35	3.08	5.93	3.43	3.89	3.00	2.99

Table 7.1: *Contour matching procedure: fitness scores.*

reports the resulting "best" configuration $\boldsymbol{s}^{\text{opt}}$ for each subject along with the corresponding best matching contours and optimal rotation $\rho_{\mathbf{s}}^{\text{opt}}$. For subjects with two tracks only the missing track's reflection sign is conventionally labeled with "$*$". As an example, Fig. 7.5 shows the optimal ray-traced tracks for Subjects 040 and 134.

One can immediately notice that configuration $\boldsymbol{s} = [0, 0, 0]$, i.e. negative coefficient sign for all reflections, obtains the best score in all cases except for Subject 015. However, it was seen that for both this subject and Subject 009 the optimal focus of the winning configuration is located well outside the ear canal area, even when the search area A is widened. Closer inspection of the corresponding pinna pictures revealed that they were taken from an angle which is far from being approximately aligned to the interaural axis, resulting much displaced towards the back of the head. As an effect, the pinna image is stretched with respect to all other cases. Consequently, as no consistent matching can be defined on these two pinna pictures, in the following Subject 009 and Subject 015 are regarded as outliers.

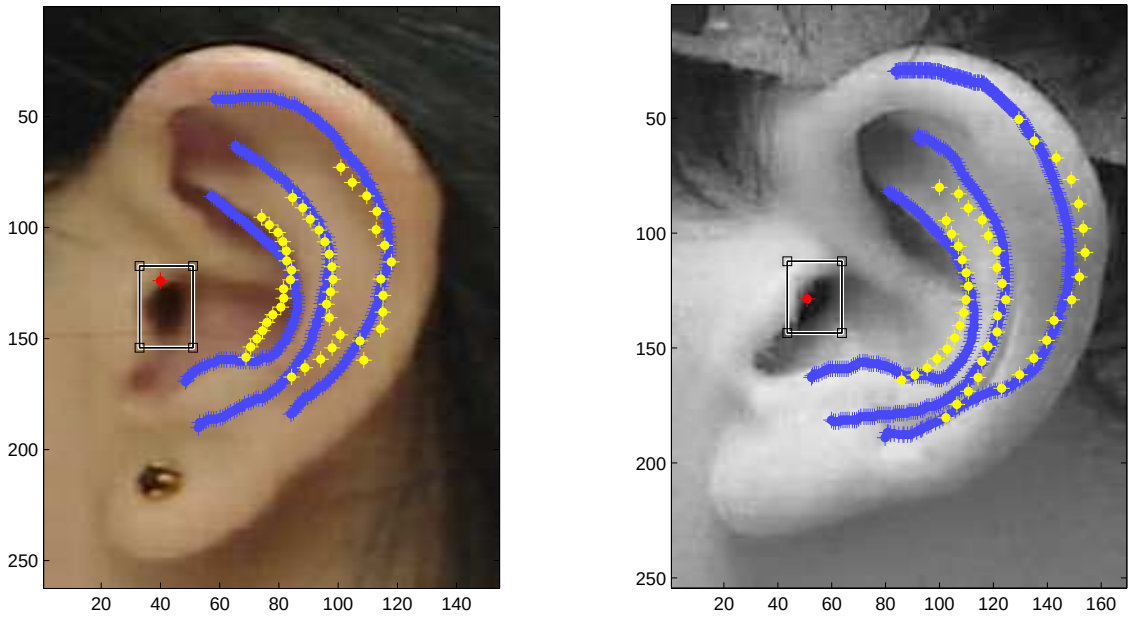
All the remaining subjects exhibit $\boldsymbol{s}^{\text{opt}} = [0, 0, 0]$ as the winning configuration. Quantitative correspondence between tracks and contours varies from subject to subject, e.g. assigning a much

Subject	$\boldsymbol{s}^{\text{opt}}$	$\rho_{\text{s}}^{\text{opt}}$	Nearest contours
003	$[0,0,0]$	$5°$	$1,4,3$
008	$[0,0,0]$	$-3°$	$1,4,3$
009	$[0,0,0]$	$-5°$	$2,4,4$
010	$[0,0,0]$	$-1°$	$1,4,3$
011	$[0,0,0]$	$-5°$	$1,4,3$
012	$[0,0,0]$	$-5°$	$2,4,3$
015	$[1,0,0]$	$-3°$	$3,1,4$
017	$[0,0,0]$	$5°$	$1,4,3$
019	$[*,0,0]$	$-5°$	$-,4,3$
020	$[0,0,*]$	$-5°$	$2,4,-$
021	$[0,0,0]$	$-5°$	$2,4,3$
027	$[0,0,0]$	$-5°$	$2,4,3$
028	$[0,0,0]$	$-5°$	$2,4,3$
033	$[0,0,0]$	$0°$	$1,4,3$
040	$[0,0,0]$	$-5°$	$1,4,3$
044	$[0,0,0]$	$2°$	$2,4,3$
048	$[0,0,0]$	$5°$	$1,4,3$
050	$[0,0,0]$	$-5°$	$2,4,3$
134	$[0,0,0]$	$0°$	$2,4,3$
165	$[0,0,0]$	$-5°$	$2,4,3$

Table 7.2: *Contour matching procedure: winning configurations.*

lower score to Subject 165 with respect to Subject 003; still, scores were defined as above with the aim to give an indication of the probability of a configuration for a series of subjects rather than an intersubjective fitness measure. Interestingly, in all cases except one, scores for $\boldsymbol{s} = [1,1,1]$ are more than doubled with respect to the complementary configuration $\boldsymbol{s} = [0,0,0]$, a result which catalogues the hypothesis of an overall positive reflection sign as unlikely. Also, note that the second best configuration is generally $\boldsymbol{s} = [1,0,0]$. Moreover, tracks T_2 and T_3 always best match with C_4 and C_3, respectively, while T_1 matches best with C_1 in 47% of subjects and with C_2 in 53% of subjects. These results enforce the hypothesis of negative reflection sign for T_2 and T_3 while leaving a halo of uncertainty on T_1's actual reflection sign.

Nevertheless, the optimality of $\boldsymbol{s}^{\text{opt}} = [0,0,0]$ is further supported by the following observations. First, if $s_1 = 1$, T_1 would fall near to contour C_3 just like T_3 (see e.g. Fig. 7.5 for graphical evidence), hence the hypothesis of two different signs for reflections onto the same surface seems unlikely. Second, as mentioned in the previous chapter, T_1 terminates on average earlier than T_2 and T_3. This indicates that for elevations approaching $\phi = 45°$ the incoming wave hardly finds a perpendicular reflection surface, and this is compatible with a reflection on the helix, which



(a) Subject 040. (b) Subject 134.

Figure 7.5: *Optimal ray-tracing for two CIPIC subjects. The red point surrounded by the search area A is the optimal focus of the winning configuration $\boldsymbol{s}^{opt} = [0, 0, 0]$. Yellow points indicate the three projected tracks, and blue points the hand-traced nearest contours to the tracks.*

normally ends just below the eye level. Last but not least, if $s_1 = 0$, T_1 falls near C_2 for all those subjects having a protruding ear; this would mean that reflections are most likely to happen on the wide helix wall rather than the border C_1, which conversely is the significant reflector for subjects with a narrow helix.

Another quantitative result that deserves to be commented is the score per track, averaged on the 18 "good" subjects: 2.37 for T_1, 1.84 for T_2, and 2.57 for T_3. Surprisingly, the best score is obtained for C_4, which was harder to trace in the preprocessing phase. By contrast, one of the clearest contours, C_3, is also the one that exhibits the greatest mismatch with respect to its relative track. This is mainly due to a number of track points around elevation $\phi = 0°$ projected nearer to the ear canal than C_3 on the pinna image, a common trend that is observed in 11 subjects over 18 and is clearly detectable in Fig. 7.5, especially for Subject 040. This point is further discussed next.

Finally, note that in many cases the optimal rotation ρ_s^{opt} equals $-5°$ or $5°$, i.e. the extremes of the search interval I. This result suggests that the real optimal rotation could lie outside the current search interval, and indeed by widening I this happened for some of the considered

subjects. However, the difference in fitness values is really small, hence the choice of the correct extremes for I does not represent a big deal.

7.4 Discussion and conclusions

The above results numerically give credit to Satarzadeh's negative reflection hypothesis. Three main notches apparently due to three different reflections on the concha border, antihelix/concha wall, and helix are seen in most HRTFs. One may think of the pinna seen from the median plane as a sequence of three protruding borders: concha border, antihelix, and helix border. These are regarded by Satarzadeh as boundaries between skin and air, that in a mechanical wave transmission analogy would introduce an impedance discontinuity $Z_1/Z_2 < 1$ at the reflection point [148]. Thus, a part of the wave would follow a straight path while another with diminished amplitude and inverted phase would be reflected back to the ear canal. Despite the clever intuition, there is no substance to the fact that waves are only reflected at borders and not onto inner pinna walls.

A recent study by Takemoto *et al.* on pressure distribution patterns in median-plane PRTFs [171] reveals through FDTD simulations on four different subjects' pinnae the existence of vast negative pressure anti-nodes inside pinna cavities at the first notch frequency. Specifically, when the source is below the horizontal plane the cymba, triangular fossa, and scaphoid fossa resonate in the same phase which is reverse to that of the incoming wave, while when the source is placed in the anterosuperior direction the same phenomenon appears at the back of the concha and the lower part of the helix. The authors then observe that these negative pressure zones cancel the wave and, as a consequence, a pressure node appears at the ear canal entrance. The observed negative pressure zones approximately cover both contours C_1 and C_2, hence one could speculate about the following generation mechanism for notches in track T_1, all of which are referred to as N_1: a given frequency component of the incoming sound wave encounters a negative pressure area in the vicinity of the helix wall or border, reflects back with inverted phase, and meets the direct wave at the ear canal entrance after a full period delay canceling that frequency component. Unfortunately, similar pressure distribution patterns for notches in T_2 and T_3 (respectively N_2 and N_3) have not been studied in [171]; still we can think of analogous generation mechanisms for these tracks too.

Shifting the focus to actual pinna contours that are responsible for spectral notches, one further clue confirms contour C_3 as most likely associated to track T_3. The observed "anticipation" of contour C_3 exhibited by T_3 at elevations close to $\phi = 0°$ (see Fig. 7.5) may be regarded as a delay that affects the direct wave alone due to diffraction across the tragus. Evidence of this phenomenon is also conjectured in [115]. Concerning track T_1, the above findings seem to conflict with the common idea that N_1 is due to a reflection on the concha wall [68, 96, 137]. In two works by Mokhtari *et al.* [114, 115], micro-perturbations to pinna surface geometry in the form

of 2-mm voxels are introduced at each possible point on a simulated KEMAR pinna. The authors observe that perturbations across the whole area of the pinna, helix included, introduce positive or negative shifts in the center frequency of N_1, especially at elevations between $\phi = -45°$ and $\phi = 0°$ in the median plane. Such shifts do not appear if voxels are introduced over the helix area in higher order notches, whose center frequency sensitively varies for perturbations introduced within the concha, cymba and triangular fossa only. This result clearly indicates that the reflection path responsible for N_1 crosses the whole pinna area, calling into question the above common belief and giving credit to the results of this work instead.

Admittedly, as [115] points out, the last result also suggests that ray-tracing models are based on a wrong assumption, i.e. that a single path is responsible for a notch. The dependence of N_1 on the whole pinna surface clearly indicates that multiple reflection paths concur in the determination of the notch distinctive parameters. However, even if multiple paths are responsible for the exact frequency location of the notch, thanks to the concave shape of the considered contours one may think of a specific time delay for which the greatest portion of reflections counteract the direct wave as an approximation to a single, direct ray.

Another objectionable point of the described approach is the adequateness of using a 2-D representation for contour extraction. As a matter of fact, since in most cases the pinna structure does not lie on a parallel plane with respect to the head's median plane, especially in subjects with protruding ears, a 3-D model of the pinna would allow to investigate its horizontal section, in such a way that projections on the side-view images could take into account the displacement caused by the flare angle of the pinna. Beside the unavalaibility of such kind of reconstruction for the considered subjects, my original aim was to keep the contour extraction procedure as low-cost and accessible as possible; furthermore, additional results in the following chapter will confirm that the 2-D approximation is, on a theoretical basis at least, already satisfactory.

To conclude, such an analysis has revealed a convincing correspondence between computed reflection points and reflective structures over the pinna. This opens the door for a very attractive approach to the parametrization of a PRTF model based on individual anthropometry. Indeed, given a 2-D image or a 3-D reconstruction of the user's pinna, one can easily trace the contours of the concha wall, antihelix and helix, compute each contour's distance with respect to the ear canal for all elevations, and extrapolate the notch frequencies by reversing Eq. 7.5. Obviously, since notch depth strongly varies within subjects and elevations, the reflection coefficient must also be estimated for each point. This problem theoretically requires strong physical arguments; alternatively, psychoacoustical criteria could be used in order to evaluate the perceptual relevance of notch depth, and potentially simplify the fitting procedure.

Chapter 8

A Personalized Head-Related Transfer Function Model

In this last Chapter an extension of Satarzadeh's structural pinna filter model [149] is proposed. Satarzadeh's work has evidenced how a very simple pinna model can incorporate the gross magnitude characteristics of the PRTF, by straightforward parametrization on two physical measures. However, besides solely considering the frontal direction of the sound wave, taking into account a single reflection seems to be a limiting factor: as an example, PRTFs with a poor notch structure do not exhibit a clear reflection in the impulse response.

Instead, the information collected from the outputs of the separation algorithm allows to look forward to a structural model which models two resonances and three reflections in the PRTF. Although including both contributions of the head and pinna into two separate structures, the motivation that led to the building of such model is the definition of a pinna model that can be easily merged with the several solutions proposed in literature regarding the head, torso, and shoulders' contributions. Indeed, the proposed model was designed so as to avoid expensive computational and temporal steps such as HRTF interpolation on different spatial locations, best fitting non-individual HRTFs, or the addition of further artificial localization cues, allowing future implementation and evaluation in a real-time environment. The filter structure of the model is presented in Section 8.1; two instances (one per ear) of such model, appropriately synchronized through ITD estimation methods, allow for real-time binaural rendering.

In light of the previously discussed approximate invariance of PRTFs in the vicinity of the median plane a fundamental assumption is introduced, i.e. elevation and azimuth cues are handled orthogonally throughout the considered frontal workspace, where azimuth ranges from -30° to 30° and elevation ranges from -45° to 45°. The angular range of validity of the presented model will thus be at least as broad as the shaded area depicted in Fig. 8.1. This way, it is possible to define customized elevation and azimuth cues that maintain their average behaviour throughout a significant slice of the front hemisphere. At this preliminary stage, the model can thus be exploited to simulate a real 3-D representation of a sound source in a number of "frontal"

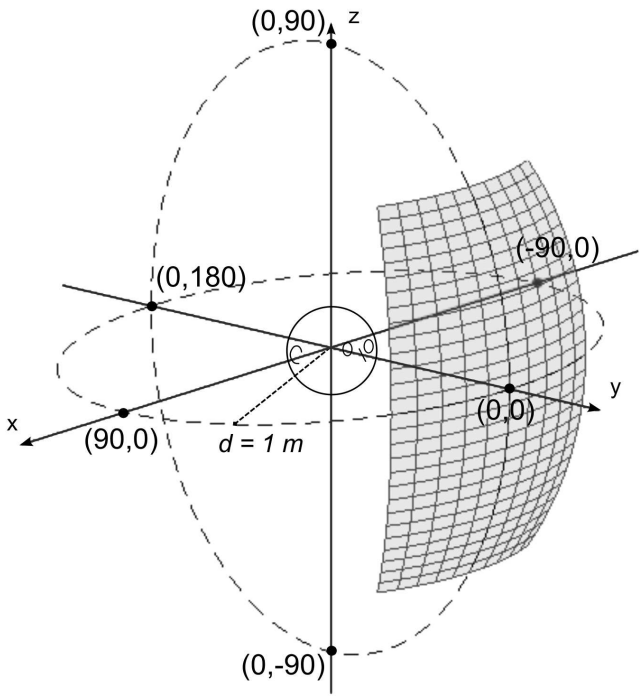


Figure 8.1: *Spatial range of validity of the structural HRTF model.*

applications, e.g. a sonified screen.

Vertical control is associated with the acoustic effects of the pinna while the horizontal one is delegated to head diffraction. No modeling for the shoulders and torso is considered, even though it is known that their presence would generally add low-frequency secondary HRTF cues for elevation perception [2]. Furthermore, dependence on source distance is negligible in the pinna model but critical in the head counterpart in the near field [42]; however, distance information is by now not integrated in the structural model. Although the inclusion of the model proposed in Chapter 5 would be straightforward in this context, no objective evaluations can be performed against real, measured responses such as CIPIC HRTFs, that were taken for a constant 1-meter source distance. Furthermore, distance information in real HRTFs would probably always be retained in the resonant component extracted by the separation algorithm because of its "frequency smoothness", and thus already be included by the pinna model. As a consequence, the overall structure is assumed to be valid only for sources at 1 m from the center of the head or farther.

As Section 8.2 will describe, parameters to be fed to the model are both derived from spectral features directly extracted from PRTF analysis or averaged in a collection of measured HRTFs, and anthropometric features of the specific subject (some of which are taken from a photograph of his/her outer ear), hence allowing customization onto a specific listener. Objective evaluation of the model against measured HRTFs of a number of CIPIC database subjects will be finally

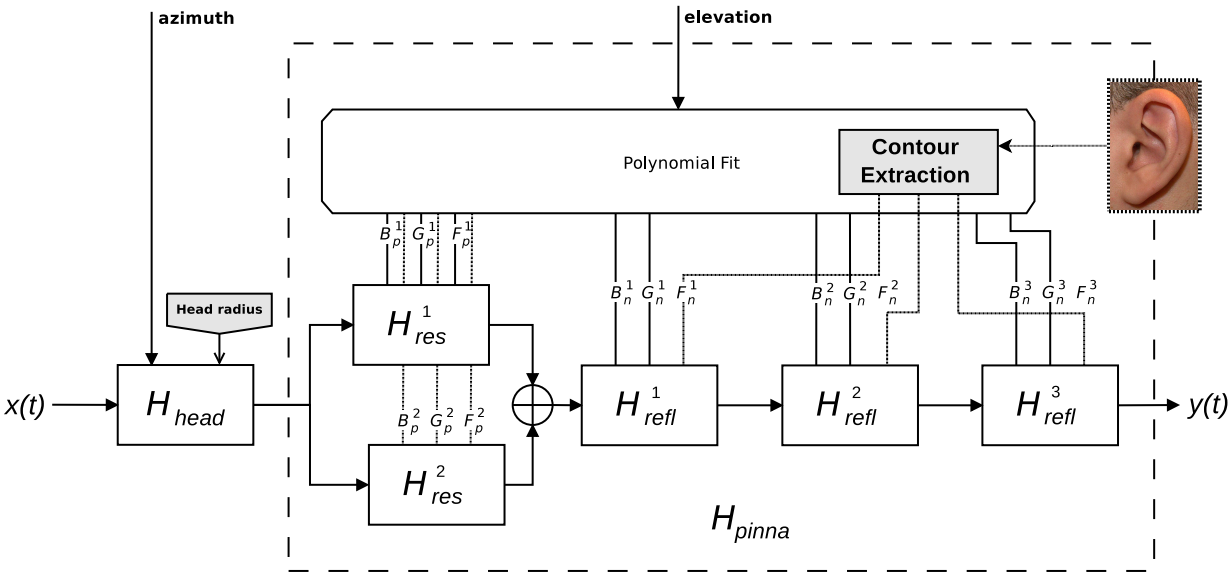


Figure 8.2: *The structural HRTF model. Customization is performed through parameter extraction from anthropometric measurements and a pinna photograph.*

carried out in Section 8.3.

Results of the work presented in this Chapter have been progressively published in papers [57], [165], [166], [58], and [59], and have been submitted for publication in [167].

8.1 Filter model

Fig. 8.2 reports a global view of the model. From left to right, the first block is the head model. Different possible existing models can be exploited here; in order to keep the overall structure as cheap as possible, the digital counterpart of the head shadow analog filter described in Eq. (4.4) [20], obtained through the bilinear transform, was chosen:

$$H_{\text{head}}(z) = \frac{\frac{\beta+\alpha f_s}{\beta+f_s} + \frac{\beta-\alpha f_s}{\beta+f_s} z^{-1}}{1 + \frac{\beta-f_s}{\beta+f_s} z^{-1}}, \tag{8.1}$$

where f_s is the sampling frequency, β depends on the head radius parameter a as $\beta = c/a$, and α is defined as in Eq. (4.6). In the calculation of the last parameter, θ_{inc} relates to azimuth θ as $\theta_{\text{inc}} = 90° - \theta$ for the right ear and $\theta_{\text{inc}} = 90° + \theta$ for the left ear, assuming the interaural axis to coincide with the x axis for sake of brevity. A reasonably good approximation of real diffraction curves in the considered range of interest for the azimuth angle $-30° < \theta < 30°$ is euristically found for parameters $\alpha_{\text{min}} = 0.1$ and $\theta_{\text{min}} = 180°$.

Coming to the pinna block, the only independent parameter used here is source elevation ϕ, which drives the evolution of resonances' center frequency $F_p^i(\phi)$, 3dB bandwidth $B_p^i(\phi)$,

and gain $G_p^i(\phi)$, $i = 1, 2$, and of the corresponding notch parameters ($F_n^j(\phi)$, $B_n^j(\phi)$, $G_n^j(\phi)$, $j = 1, 2, 3$). The resonant part of the pinna model is represented as a parallel of two different second-order peak filters. The first peak ($i = 1$) has the form [192]

$$H_{\text{res}}^{(1)}(z) = \frac{1 + (1+k)\frac{H_0}{2} + l(1-k)z^{-1} + (-k-(1+k)\frac{H_0}{2})z^{-2}}{1 + l(1-k)z^{-1} - kz^{-2}}, \tag{8.2}$$

where

$$k = \frac{\tan\left(\pi \frac{B_p^1(\phi)}{f_s}\right) - 1}{\tan\left(\pi \frac{B_p^1(\phi)}{f_s}\right) + 1}, \tag{8.3}$$

$$l = -\cos\left(2\pi \frac{F_p^1(\phi)}{f_s}\right), \tag{8.4}$$

$$V_0 = 10^{\frac{G_p^1(\phi)}{20}}, \tag{8.5}$$

$$H_0 = V_0 - 1, \tag{8.6}$$

and f_s is the usual sampling frequency. The second peak ($i = 2$) is implemented as in [127],

$$H_{\text{res}}^{(2)}(z) = \frac{V_0(1-h)(1-z^{-2})}{1 + 2lhz^{-1} + (2h-1)z^{-2}}, \tag{8.7}$$

where

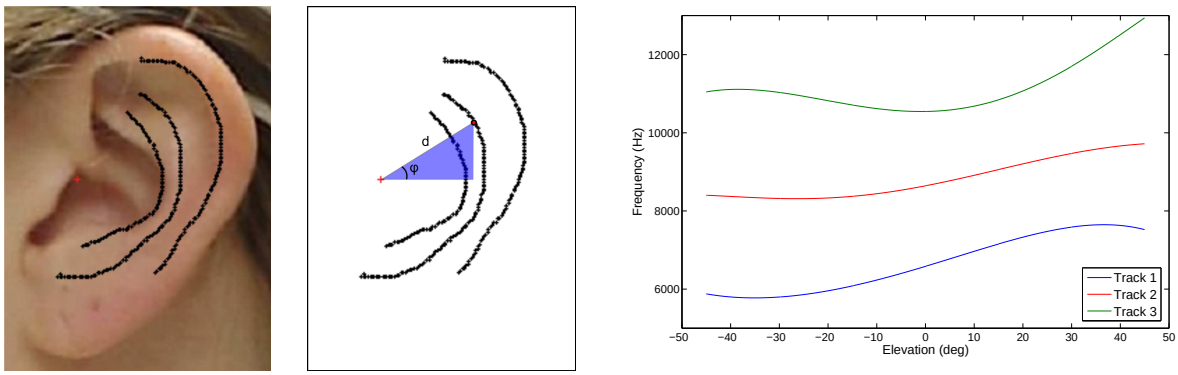
$$h = \frac{1}{1 + \tan\left(\pi \frac{B_p^2(\phi)}{f_s}\right)}, \tag{8.8}$$

while l and V_0 are defined as in Eqs. (8.4) and (8.5) with polynomial index $i = 2$. The reason for this distinction lays on the low-frequency behaviour that needs to be modeled: the former implementation has unitary gain at low frequencies so as to preserve such characteristic in the parallel filter structure, while the latter has a negative dB magnitude in the same frequency range. In this way, the all-round pinna filter does not alter low-frequency components in the signal forwarded by the head shadow filter.

The notch filter implementation is of the same form as peak filter $H_{\text{res}}^{(1)}$ with the only differences in the parameters' description. In order to keep notation correct, polynomials $\mathcal{P}_p^1$ must be substituted by the corresponding notch counterparts $\mathcal{P}_n^j$, $j = 1, 2, 3$, and parameter k defined in Eq. (8.3) replaced by its "cut" version

$$k = \frac{\tan\left(\pi \frac{B_n^j(\phi)}{f_s}\right) - V_0}{\tan\left(\pi \frac{B_n^j(\phi)}{f_s}\right) + V_0}. \tag{8.9}$$

Note that the notch filter implementation is identical to the one used in the separation algorithm, reported in Eq. (6.1). The three notch filters are placed in series and cascaded to the parallel of the two peak filters, resulting in an eighth-order global pinna filter.



(a) Contour tracing. (b) Polar coordinates. (c) Notch track extraction and approximation.

Figure 8.3: *Notch frequency extraction from a picture of the pinna.*

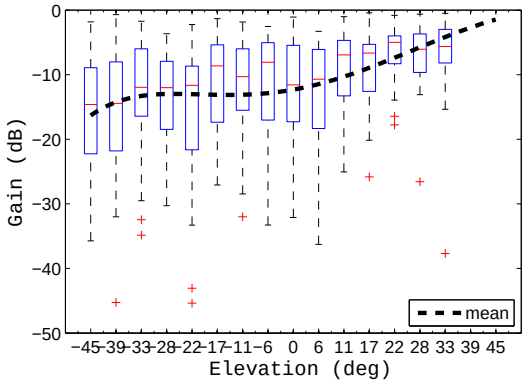
8.2 Parametric model fitting

Clearly, in order to reach complete control of the model, a mapping of anthropometric parameters, where available, and/or average or extracted PRTF features onto filter parameters is needed. As for the head filter, the radius parameter a, whose value influences the cutoff frequency of head shadowing, is defined by the weighted sum of the subject's head dimensions in Eq. (4.10) using the optimal weights obtained reported in Eq. (4.11) [3].

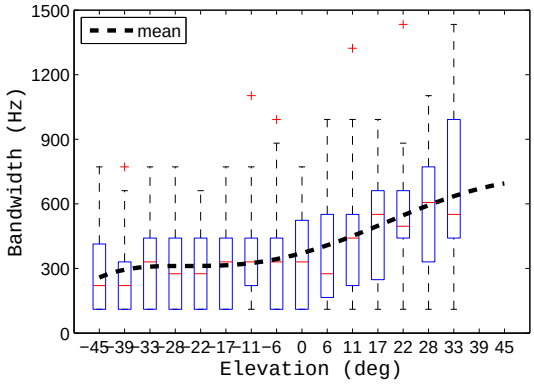
For what concerns the pinna block, the original peak and notch parameters of each subject can be derived as follows. First, they are estimated from the separated resonant or reflective (i.e. notch tracks) component of median-plane PRTFs for all the available ϕ values. An identification system based on a sixth-order ARMA model [46] that extracts for every ϕ the two highest maxima of the resonant component allows straightforward computation of the gain, center frequency, and 3dB bandwidth of each resonance peak, while required notch features are taken from the already available notch tracks computed as in Section 6.3.2.[1] Second, a fifth order polynomial $\mathcal{P}_p^i$ or $\mathcal{P}_n^j$, where $\mathcal{P} \in \{F, B, G\}$, is best fitted to the corresponding sequence of parameter values, yielding a complete functional parametrization of the filters. These functions are used in the structural model of Fig. 8.2 in order to continuously control the evolution of the resonant and reflective components when the sound source is moving along elevation. Obviously, all the polynomials must be computed offline previous to the rendering process.

However, following the important findings of the previous chapter, functions $F_n^j(\phi)$ can alternatively be extracted from the subject's anthropometry. Fig. 8.3 sums up the procedure needed to extract notch frequencies from a representation of the pinna contours. Specifically, first a picture of the subject's pinna is resized to match the real dimensions according to his/her anthropomet-

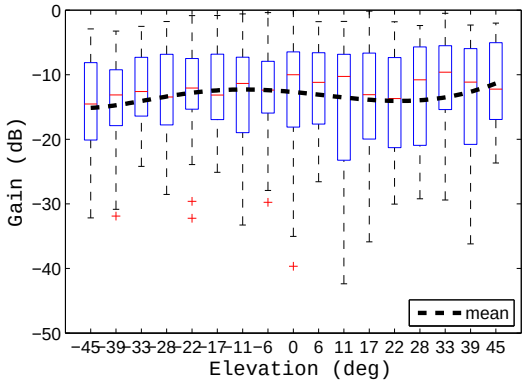
[1] In order to avoid bad outcomes in the design of notch filters, gaps in notch tracks are assigned a gain equal to 0 dB while bandwidth and center frequency are given the value of the previous notch feature in the track.



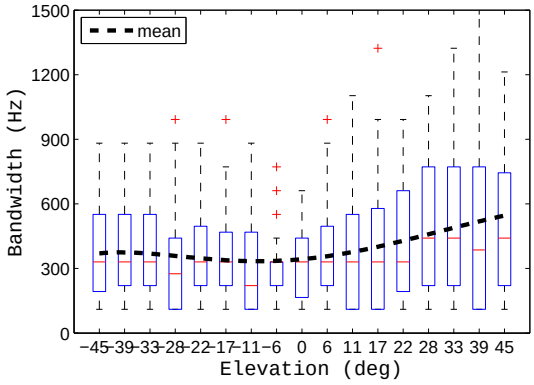
(a) Track T_1, gain.



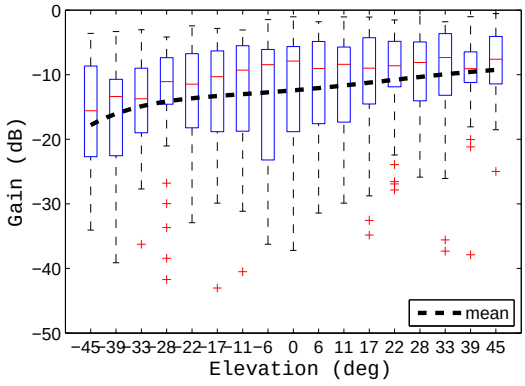
(b) Track T_1, bandwidth.



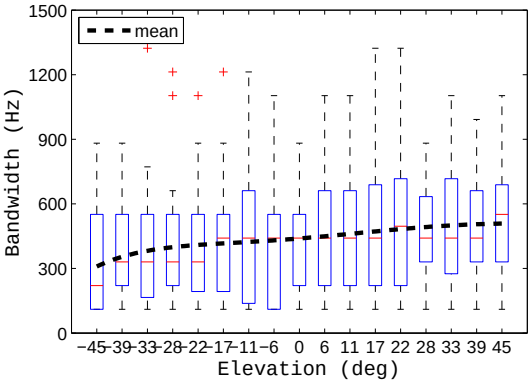
(c) Track T_2, gain.



(d) Track T_2, bandwidth.



(e) Track T_3, gain.



(f) Track T_3, bandwidth.

Figure 8.4: *Box plot and mean of notch tracks gain and bandwidth values among CIPIC subjects.*

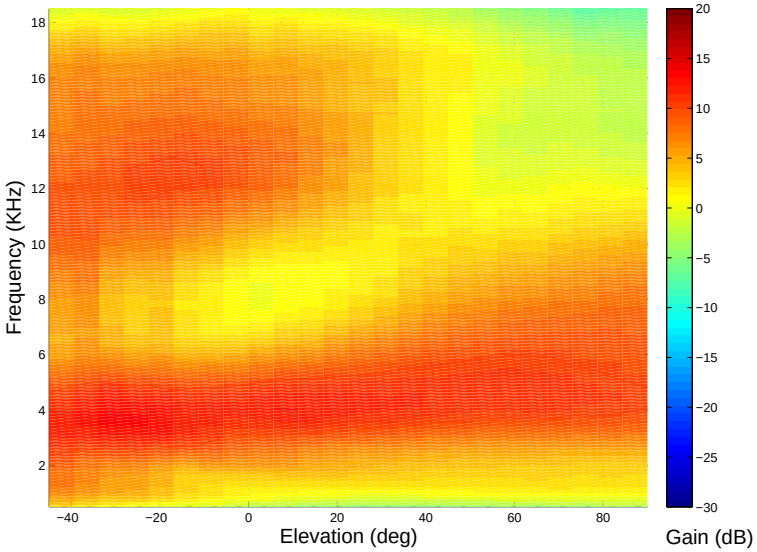


Figure 8.5: *Mean magnitude spectrum of the pinna resonant component, averaged on all* 45 *CIPIC subjects' left-ear responses.*

ric data. Then, contours C_2 or C_1 (depending if the subject's ear is respectively protruding or not), C_4, and C_3 are traced and stored as a sequence of pixels. These are translated into couples of polar coordinates (d, ϕ), with respect to a fixed point within the ear canal, through simple trigonometric computations. Finally, distances d are translated into sequences of frequencies through Eq. (7.5), thus assuming overall negative reflection coefficients. Again, a fifth order polynomial is best fitted to these sequences, resulting in functions $F_n^j(\phi)$, $j = 1, 2, 3$.

For what concerns the other two parameters defining a notch, i.e. gain and 3dB bandwidth, there is still no evidence of correspondence with anthropometric quantities. A first-order statistical analysis on the depth and bandwidth of notches, subdivided by notch track and elevation, was performed among all 45 left pinnae of CIPIC subjects. Such an analysis revealed a high variance of these values within each track and elevation, and mean values which lie approximately constant apart from a slight decrease in notch depth and a slight increase in bandwidth as the elevation increases up to $\phi = 45°$. These trends can be clearly seen in the various plots of Fig. 8.4, where box plots are absent at a specific elevation if the number of subjects presenting a notch in that track is less than 20. In absence of clear elevation-dependent patterns, the mean of both gains and bandwidths for all tracks and elevations ϕ among all subjects can be computed, and again a fifth-order polynomial dependent on elevation fitted to each of these sequences of points, yielding functions $G_n^j(\phi)$ and $B_n^j(\phi)$, $j = 1, 2, 3$ which can be feeded to the structural model as an alternative to the corresponding functions derived from subject-specific analysis.

Finally, given that resonances have a similar behaviour in all of the analyzed PRTFs, cus-

tomization of this component for the model may be overlooked. The mean magnitude spectrum of the resonant component for the 45 left pinnae of CIPIC subjects (shown in Fig. 8.5) may be instead calculated and substituted to the specific listener's resonant component for resynthesis via the usual polynomial fitting procedure onto the three distinctive parameters of its two peaks.

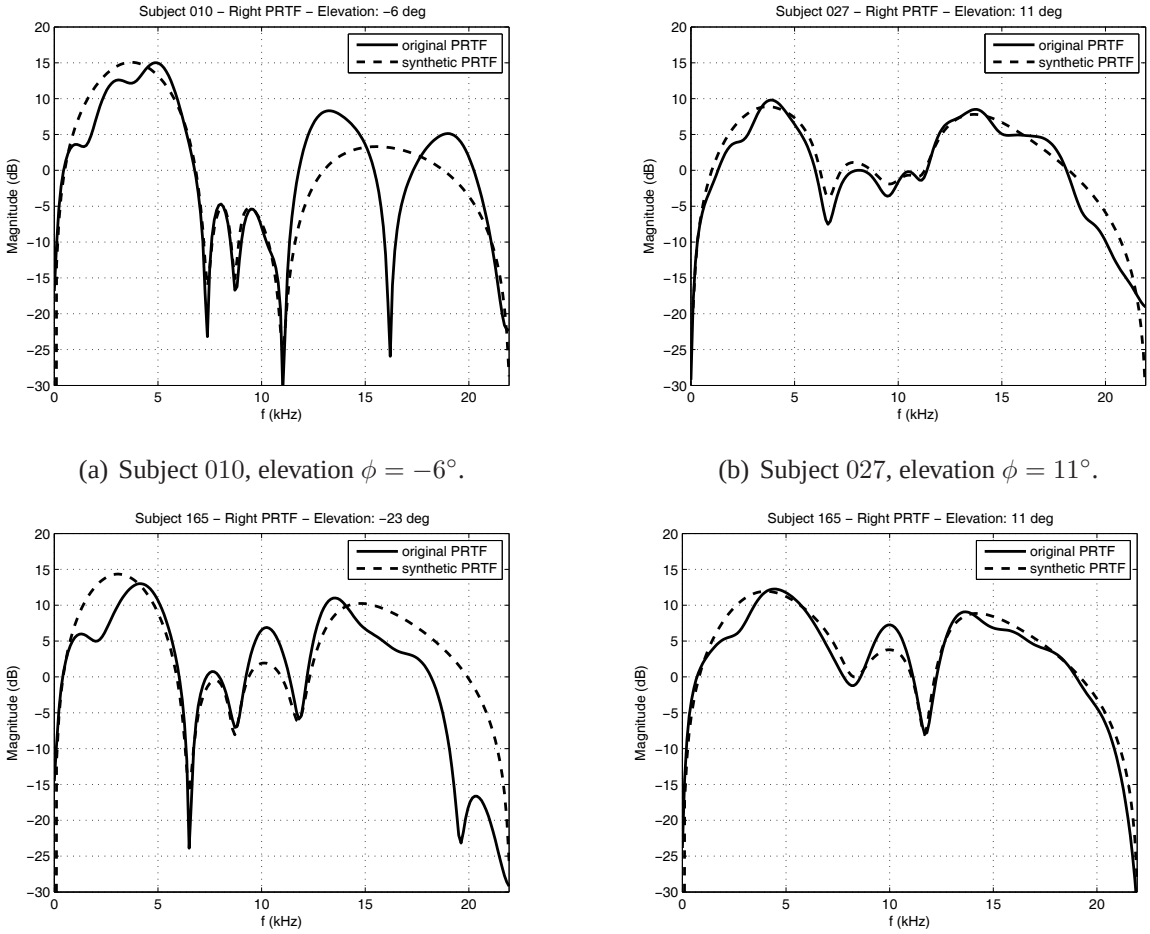
In the remainder of this chapter let me refer to HRTFs given by the fully resynthesized model (without contour extraction nor averaged peak and notch features) as H^r; to HRTFs resulting from the contour-parameterized model as H^c; and to HRTFs built through the fully synthesized model (contour extraction plus averaged peak and notch features) as H^s. In the following section I'm going to discuss the effectiveness of all the introduced approximations through objective comparison of the synthesized responses to the original HRTFs for a bunch of subjects.

8.3 Results and discussion

Fig. 8.6 reports the comparison between original and re-synthesized PRTF magnitudes through the pinna block of the structural model for three distinct subjects at four different elevation angles. Adherence rate to the original PRTFs is overall satisfactory in the frequency range up to 14 kHz. Still, several types of imperfections stand out: as a first example, deep frequency notches occasionally complicate the notch filter design procedure. In point of fact, if the notch to be approximated is particularly deep and sharp, the second-order filter will produce a shallower and broader notch whose bandwidth may interfere with adjacent notches, resulting in underestimation of the PRTF magnitude response in the frequency interval between them: Fig. 8.6(c) exhibits this behaviour around 10 kHz. Using a filter design procedure which forces to respect the notch bandwidth specification during re-synthesis would grant a better rendering of resonances, at the expense of worsening notch depth accuracy.

The absence of modeled notches over the upper frequency threshold is another cause of imprecision. For instance, Fig. 8.6(a) presents an evident mismatch between original and modeled PRTF just after the 11-kHz notch, due to the cut of the following frequency notch at 16 kHz. This problem may be corrected by increasing the upper frequency threshold of the notch tracking algorithm in order to take into account a higher number of notches. However, being the psychoacoustic relevance of this frequency range not effectively known, the overall weight of such mismatch could bear little significance.

Last but not least, resonance modeling may bring approximation errors too. In particular, the possible presence of non-modeled interfering resonances represents a limitation to the resynthesis procedure. Furthermore, center frequencies extracted by the ARMA identification method mentioned in the previous section do not always coincide with peaks in the PRTF. Thus, a stronger criterion for extracting the main parameters of each resonance could be needed. Nevertheless, the approximation error seems to be negligible in all those cases where resonances, just as well as notches, are distinctly identifiable in the PRTF.



(a) Subject 010, elevation $\phi = -6°$.

(b) Subject 027, elevation $\phi = 11°$.

(c) Subject 165, elevation $\phi = -23°$.

(d) Subject 165, elevation $\phi = 11°$.

Figure 8.6: *Original versus synthetic PRTF magnitude plots.*

However, in order to objectively evaluate the whole structural model against the original measured HRTFs in the CIPIC database an error measure needs to be introduced. A measure definitely suitable to the purpose is the already defined spectral distortion, see Eq. (5.15). In this case, the considered frequency range is limited between 500 Hz and 16 kHz.

Fig. 8.7 reports SD values, averaged across the 18 non-outlier CIPIC subjects examined in the previous chapter, of five different median-plane reconstructed responses:

1. the all-round response of the contour-parameterized model, H^c_{tot};

2. the reflective component of the contour-parameterized model given by notch filters, H^c_{refl};

3. the resonant component of the model (either contour-parameterized or resynthesized) given by peak filters, H_{res};

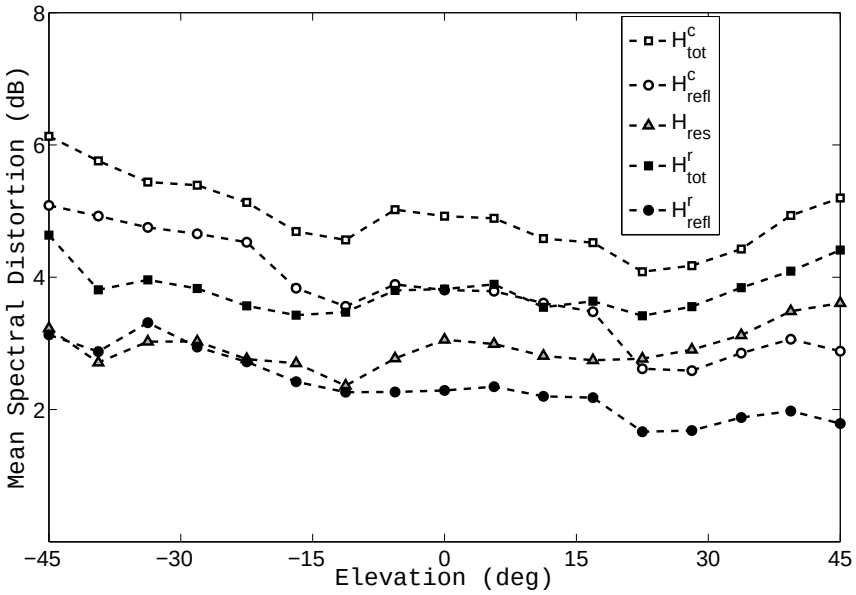


Figure 8.7: *Mean spectral distortion between reconstructed and measured median-plane HRTFs, averaged among* 18 *CIPIC subjects. Square markers represent total responses, circular markers represent reflective components, and triangular markers represent resonant components. Black-filled markers refer to fully resynthesized responses, while white-filled markers refer to contour-parameterized responses.*

4. the all-round response of the fully resynthesized model, H^r_{tot};

5. the reflective component of the fully resynthesized model given by notch filters, H^r_{refl}.

Resonant and reflective components are obviously compared to their counterparts extracted by the separation algorithm.

As expected, H^c_{tot} is the response with the highest average SD among the five considered responses. As a matter of fact, errors in the resynthesized resonant (H_{res}) and contour-parameterized reflective (H^c_{refl}) components combine together yielding the SD for H^c_{tot}, which ranges from 4 to 6 dB and is worse at low elevations. This fact can be explained by the occurrence of very deep notches at low elevations, that causes large errors in the SD when a notch extracted from a contour is not perfectly reconstructed at its proper frequency.

In proof of this note that, as notches become fainter and fainter with increasing elevation, the SD of H^c_{tot} tends to decrease apart from a new rise at the last elevation angles, which is conversely due to greater errors in the resonant component H_{res}. Inspection of resonant components at higher elevations reveals indeed that the second modeled high-frequency peak (horizontal mode) disappears, gradually letting non-modeled lower-frequency vertical modes in. As a further confirmation of the criticality of the exact notch frequency location in SD computation, note that when frequencies are extracted from real HRTFs the SD of the reflective component H^r_{refl}

Subject	$m(T_1, C_1)$	$m(T_1, C_2)$	$m(T_2, C_4)$	$m(T_3, C_3)$
003	11.42%	–	12.02%	18.25%
008	8.98%	–	8.69%	14.07%
010	4.80%	–	2.90%	18.74%
011	8.75%	–	7.77%	12.20%
012	–	5.57%	8.98%	8.69%
017	7.80%	–	3.44%	17.97%
019	–	–	4.48%	5.92%
020	–	5.50%	4.27%	–
021	–	9.18%	10.16%	11.73%
027	–	8.14%	2.09%	7.63%
028	–	7.39%	8.05%	14.79%
033	4.52%	–	3.55%	16.44%
040	2.98%	–	5.50%	12.92%
044	–	9.63%	6.49%	8.10%
048	4.01%	–	3.18%	16.19%
050	–	8.62%	7.28%	18.95%
134	–	2.59%	5.10%	10.13%
165	–	3.91%	4.11%	6.44%

Table 8.1: *Notch frequency mismatch between tracks and contours.*

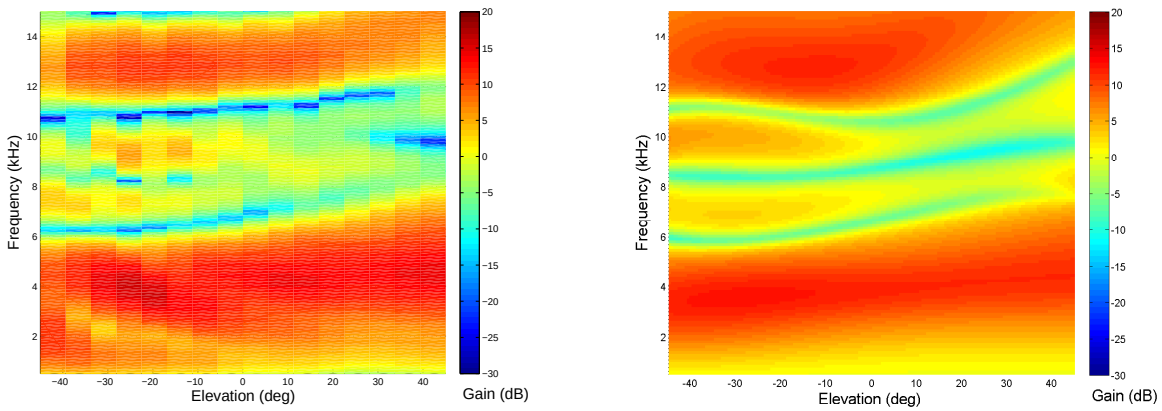
distinctly decreases to 3 dB or less, resulting in a sensibly lower SD (about 4 dB) in the total response H^r_{tot}.

Another error measure is now introduced to show that, even if contour-extracted notch frequencies are not exactly correspondent to their measured counterparts, the effective frequency shift is almost everywhere not likely to result in a perceptual difference. Specifically, we define the *mismatch* between a computed notch track T_j and its associated contour C_i as the percentual ratio between the forementioned frequency shift and the measured notch frequency, averaged on all elevations where the notch is present:

$$m(T_j, C_i) = \frac{1}{n(T_j)} \sum_{\phi} \frac{|f_0^j(\phi) - F_n^j(\phi)|}{f_0^j(\phi)} \cdot 100\%, \tag{8.10}$$

where $n(T_j)$ is the number of available notch frequencies in track T_j and $F_n^j(\phi)$ is extracted from the associated contour C_i as described in the previous section.

Table 8.1 shows frequency mismatches computed for the same 18 CIPIC subjects. These results can be directly compared to the findings by Moore *et al.* included in Experiment V in [118]: two steady notches in the high-frequency range (around 8 kHz) differing just in central frequency are not distinguishable on average if the mismatch is less than approximately 9%, regardless of



(a) Original plot.

(b) Synthetic plot H^s_{tot}.

Figure 8.8: *Original and synthetic HRTF magnitude plots for Subject* 048.

notch bandwidth. Although these results were found for just one high-frequency location, mismatches of T_1 and T_2 may be informally compared with the 9%-threshold, concluding that only 4 tracks over 35 exhibit a mismatch greater than the threshold and suggesting that the frequency shift caused by contour extraction is not perceptually relevant on average.

Conversely, track T_3 shows much greater mismatches, mostly due to the "contour anticipation" effect discussed in Section 7.4. Beside possible improvements that may take into account such an effect while extracting contour C_3 and lower the mismatch, no results are available in the literature about notch perception in the region between 10 and 15 kHz. However, as already mentioned in Section 4.1, the third notch is of lesser importance than the first two in elevation perception [75], hence psychoacoustical criticality of its center frequency is somehow questionable.

To sum up, assuming that the forementioned mismatches are in most cases not perceptually relevant, we can consider the mean SD of 4 dB in H^r_{tot} as a satisfactory result, being comparable to SD values found in similar works that deal with HRTF resynthesis by means of HRIR decomposition [48] or anthropometric parametrization through multiple regression analysis on HRTF decomposition [123]. What's more, the structural model is composed by first- and second-order filters only: given that many responses exhibit sharp notches whose slope cannot be reached by a second-order filter, increasing the order of notch filters in particular would further improve the SD score. However, low-order filters allow cheap and fast real-time simulation, which is a valuable merit of the model.

Finally, let's have a more informal look at the completely synthetic reconstruction given by the H^s version of the model. As an example, the magnitude plot for Subject 048, compared to her original HRTF, is provided in Fig. 8.8. The use of mean, non-customized values for peak parameters and notches' gain and bandwidth specifications negatively influences the SD score

as expected. However, beside the (desired) different elevation resolution in the original and synthetic HRTF plots, similar features can be observed:

1. the first resonance, being omnidirectional and having an almost common behaviour in all subjects, is well approximated;

2. as already pointed out for the H^c version of the model the extracted notch tracks closely follow the measured patterns, although being smoother in frequency, gain, and bandwidth than the original ones;

3. gains, even in the intermediate frequency areas between notches and resonances, are overall preserved.

Coming to subtler differences, comparing Subject 048's pinna picture (see Fig. 8.3) with the original HRTF plots we can note a relationship between the shorter antihelix and concha wall reflection surfaces and two distinct notch tracks, the first located around 8 kHz at negative elevations and the second around 10 kHz at positive elevations. Since the initial choice was to model three contours only, these two notches are collapsed into one continuous track. A further notch appears around 15 kHz, yet it is likely associated with a mild pinna contour. Furthermore, the second resonance is clearly overestimated and its shape doesn't find a strong visual correspondence. Such mismatch highlights a complex spectrum evolution due the presence of two or more resonances interacting in the upper frequency range for elevations in proximity of the horizontal plane [158]. However, following the choice of limiting the number of resonances to two, and assuming the first resonance to be omnipresent, the second synthetic resonance has to cover multiple contributions.

Further analysis is required toward a detailed model that takes into account the individual differences among subjects and their psychoacoustical relevance besides the observed objective dissimilarities. Synthetic notches bear a smoother magnitude and bandwidth evolution compared to the original ones; in particular, magnitude irregolarities in the original notches could arise from superposition of multiple reflections and, in addition, from a strong sensitivity of the subject's spatial position during the HRTF recording session. Psychoacoustical evaluations into virtual environments are definitely needed to reveal the appreciation degree of such an approach together with the real perceived weight of such homogeneous notch and peak shapes.

Chapter 9

Conclusions and Future Work

This thesis has developed and discussed techniques for an effective, customized rendering of spatial audio, that is nowadays one of the most challenging and interesting research areas for virtual and augmented reality. The final application area of the studied techniques is that of technology-assisted motor rehabilitation, a field in which the consistent use of auditory feedback is largely underestimated yet where the use of even simple forms of auditory feedback can enhance performance and learning of a rehabilitative task, as the upper limb rehabilitation experiments and the pilot study on the gait training experiment reported in this thesis have pointed out.

As a matter of fact, these experiments corroborated the initial hypothesis that continuous sound feedback can be successfully employed during motor training to provide the subject with additional and/or substitutive information on task and/or error. In particular, it was found that rendering task-related information through sound helps subjects to increase performance; that a visuomotor transformation can be learned through a consistent auditory feedback; and that sound spatialization can further enhance performance. Thus, in this context, the aware use of spatial sound is expected to bring even more advantages, such as positive effects on patient engagement and effort during movement training, and help in performing and hopefully relearning complex functional movements. Future works in this direction will investigate how and to what extent spatialized auditory feedback can improve learning and motor recovery, possibly in post-stroke subjects.

Coming to spatial sound rendering, two novel personalizable models, one for distance rendering and one for pinna-related transfer functions simulation, were introduced and objectively evaluated. The main purpose of the distance model was to minimize magnitude differences with respect to the distance-dependent part of an analytical spherical head model through a low-order filter structure. This was done through direct fitting of three parameters easily extracted from the analytical responses to a number of exponential functions and through the use of a first-order shelving filter. The approximation was found to be appropriate; however, more work is needed in terms of further improving the model's accuracy and correctly tuning its phase response in order to grant a correct ITD estimation when using two of such models in a real-time listening

scenario.

The structural PRTF model was obtained through a definitely more complex analysis. An algorithm that separates the resonant and reflective parts of the PRTF spectrum was implemented, and the resulting decomposition drove the design of a low-order model consisting of two peak filters and three notch filters. Moreover, an analysis of real HRTF data was performed in order to study the relationship between PRTF features and anthropometry in the frontal median plane, the findings supporting the hypothesis that reflections occurring on pinna surfaces can be reduced for the sake of design to three main contributions, each carrying a negative reflection coefficient. Based on this observation, the PRTF model was parameterized onto anthropometric features of the listener extracted from a picture of his/her pinna. Spectral distortion and notch frequency mismatch measures indicated that the approximation is objectively satisfactory.

Further investigations on the correspondence between anthropometry of the human pinna and PRTF features will be soon carried out using the new PRTF database as analysis material. As a matter of fact, the PRTF database can claim the following advantages compared to the CIPIC HRTF database, thus representing a better alternative as the starting point for future work on this topic:

- the absence of the head and torso's contributions in the HRIRs, which cannot be fully eliminated a posteriori in CIPIC HRIRs;
- a denser and more extended sampling on the median plane (61 elevation angles against 50);
- the availability of highly detailed photographs of the subjects' pinnae, that allows an easier contour extraction;
- the public unavailability of all the CIPIC subjects' pinna pictures.

Clearly, in order to replicate the contour matching procedure on the new data, first the separation algorithm will have to be adapted to the present database in order to analyze the resonant and reflective components separately. In other words, the algorithm needs to be rendered format-independent. The results will be expected to provide better understanding of the behaviour of the pinna for source positions behind, above, and almost below the listener, possibly allowing extension of the pinna model to a wider spatial range, including the upper and back side of the sagittal plane.

More open issues include:

- improvements in the separation algorithm, in particular through the use of a more effective multi-notch filter design;
- automatic extraction through image processing techniques of pinna contours from 2-D pictures, that need to be stored in a format which allows computation of distances between different contours and observation points;

- understanding the influence of notch depth and bandwidth in elevation perception along with the relation between the resonant component of the PRTF and the shape of pinna cavities, both required to have a complete anthropometric parametrization of the pinna model;
- study of the PRTF behaviour outside the median plane: the model should need to be readjusted for more lateral sound sources accordingly with the examined reflection/resonance pattern variation, the psychoacoustical relevance of azimuth effects on the pinna also needing to be verified through listening tests;
- integration of the pinna model with the near-field distance model and a suitable far-field head model;
- possibly, the study of a model to be built from scratch for representing the shoulders and torso contributions. The model would add further reflection patterns and shadowing effects to the structural HRTF model, in particular for sources below the listener.

These future studies will ultimately allow a complete representation of the auditory scene surrounding the listener, offering him or her a full, surrounding binaural experience. However, subjective evaluations of the model in each of its development phases are required in order to attest its effectiveness in binaural hearing. This point requires new HRTF measurements as reference and model reconstruction onto a number of physical subjects, as well as a novel experimental setup for both static and dynamic listening tests. Admittedly, the setup is already available and is now briefly depicted.

9.1 A real-time system for 3-D audio evaluation

The experimental setup is schematically represented in Fig. 9.1. The user is wearing a pair of wireless headphones with three LED markers positioned on them, one on the left earphone (green marker), one on the right (red), and one on top of the headphones (blue). A specific area is delimited by 8 high-resolution cameras coordinated by the PhaseSpace Impulse motion tracking system, featuring a data capture rate of 480 Hz (frames/s). The square walking area is 3.0 m $\times$ 3.5 m, while the eight cameras are placed on 3 horizontal bars (parallel to the square perimeter) placed 2.5 m high. Such redundant camera placement easily allows tracking of the three markers even in certain occlusion cases.

Two informations are required to properly compute the two audio channels: the listener's head position in the tracked region and the relative direction of the head with respect to one or more simulated sound sources. In this case we can focus on a single source scenario because it's the simplest experiment not involving any acoustic effects due the mutual interaction of possible multiple sound sources. The sound source movements could be defined in different

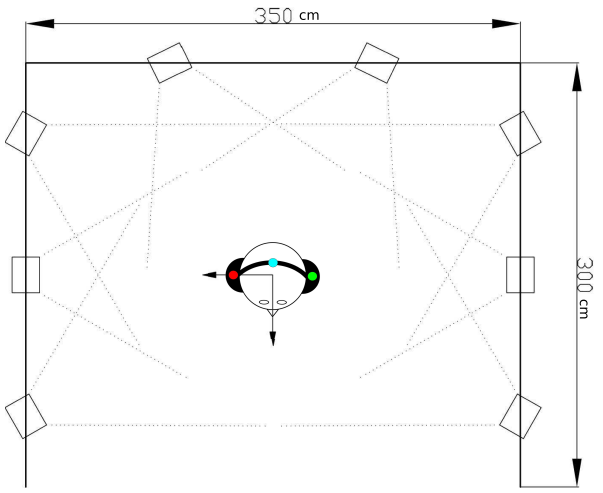


Figure 9.1: *A real-time experimental setup for evaluating the structural HRTF model.*

ways, depending on the evaluation techniques involved. In sake of simplicity, the sound source can be assumed to move in front of the listener following precomputed trajectories.

The PhaseSpace Impulse routes each marker position via the OSC protocol towards another workstation responsible for computationally carrying out the sound rendering task. Currently, sound spatialization is provided by the already cited `earplug~` Pure Data external that, to be more precise than in Chapter 3, convolves the incoming audio signals with the HRTF corresponding to the angular direction of one or more virtual sources, interpolated in real time from a set of measured KEMAR HRTFs. Additionally, source distance is rendered in the whole spatial range by varying the signal's loudness proportionally to the inverse of the square distance, with air absorption and reverberation ready to be added as auxiliary distance cues. The exact current spatial location is computed through a C-developed external module that performs some simple three-dimensional geometric calculations in order to convert the markers' absolute positions in the interaural polar coordinates of the sound source with respect to the listener's head; this way, the 3-D head position is continuously kept updated.

Azimuth, elevation and distance of the sound source thus become the input parameters on each of the two audio channels, left and right, processed in an omnicomprehensive Pure Data patch and correctly synchronized by means of a delay block taking into account for the ITD cue as in Eq. (4.9). Implementation in Pure Data of the structural model presented in this thesis, or parts of it, will allow to replace the non-customized `earplug~` external and the simplistic distance mapping in the near field with a customized filter model tuned to the listener's anthropometric parameters to be subjectively evaluated. Heavier computations can be of course delegated to external modules developed in C/C++ language.

Thinking of possible experimental protocols, each subject's anthropometric quantities will first be gathered and feeded to the customized HRTF model. Then, three different experimental

conditions for the model validation will be considered.

1. In the *static* condition, single sounds will be presented to the user through his/her customized HRTF model at a specific point around him/her. The user will remain still, listen to the sound and report the perceived source position.

2. In the *semi-static* condition the user still won't be moving, yet the sound will simulate a source which moves in the space around him/her following a precomputed trajectory. The trajectory will continuously define azimuth, elevation, and distance parameters to be feeded to the HRTF model, and the user will follow the perceived trajectory by virtually drawing lines in the 3D space, possibly with the help of a head-mounted display.

3. In the *dynamic* condition, the subject will be free to move his/her head and body and judge the position of a virtual sound source which remains fixed at a specific point while he/she is moving. After a free exploration of the space around him/her, the user will have to indicate the exact position of the simulated sound source.

Obviously, choice of the most suitable condition for evaluation and future exploitation of the model will depend on the development phase of the model itself and on the final possible virtual rehabilitation application.

9.2 Publications

The work presented in this thesis has produced the following publications.

9.2.1 International Journals (submitted for publication)

♢ S. Spagnol, M. Geronazzo, and F. Avanzini. *On the relation between pinna reflection patterns and head-related transfer function features.* IEEE Transactions on Audio, Speech, and Language Processing (IEEE TASLP).

♢ D. Zanotto, G. Rosati, S. Spagnol, P. Stegall, and S. K. Agrawal. *Effects of auditory feedback in robot-assisted lower extremity motor adaptation.* IEEE Transactions on Neural Systems and Rehabilitation Engineering (IEEE TNSRE).

♢ G. Rosati, F. Oscari, S. Spagnol, F. Avanzini, and S. Masiero. *Effect of task-related continuous auditory feedback during learning of tracking motion exercises.* Journal of NeuroEngineering and Rehabilitation (JNER).

9.2.2 International Conferences

- 2011:
  - ♢ M. Geronazzo, S. Spagnol, and F. Avanzini. *A head-related transfer function model for real-time customized 3-D sound rendering.* In Proc. INTERPRET Workshop, SITIS 2011 Conference, pages 174-179, Dijon, November-December 2011.
  - ♢ S. Spagnol, M. Hiipakka, and V. Pulkki. *A single-azimuth pinna-related transfer function database.* In Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11), Paris, September 2011.
  - ♢ G. Rosati, F. Oscari, D. J. Reinkensmeyer, R. Secoli, F. Avanzini, S. Spagnol, and S. Masiero. *Improving robotics for neurorehabilitation: enhancing engagement, performance, and learning with auditory feedback.* In Proc. IEEE 12th Int. Conf. on Rehabilitation Robotics (ICORR2011), pages 341-346, Zurich, June-July 2011. **Best Poster Award finalist.**
- 2010:
  - ♢ S. Spagnol, M. Geronazzo, and F. Avanzini. *Fitting pinna-related transfer functions to anthropometry for binaural sound rendering.* In Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP'10), pages 194-199, Saint-Malo, October 2010. **Top 10% Paper Award winner.**
  - ♢ M. Geronazzo, S. Spagnol, and F. Avanzini. *Estimation and modeling of pinna-related transfer functions.* In Proc. 13th Int. Conf. on Digital Audio Effects (DAFx-10), Graz, September 2010.
  - ♢ S. Spagnol, M. Geronazzo, and F. Avanzini. *Structural modeling of pinna-related transfer functions.* In Proc. 7th Int. Conf. on Sound and Music Computing (SMC 2010), pages 422-428, Barcelona, July 2010.
- 2009:
  - ♢ F. Avanzini, A. De Götzen, S. Spagnol, and A. Rodà. *Integrating auditory feedback in motor rehabilitation systems.* In Proc. Int. Conf. on Multimodal Interfaces for Skills Transfer (SKILLS09), Bilbao, December 2009.
  - ♢ F. Avanzini, L. Mion, and S. Spagnol. *Personalized 3D sound rendering for content creation, delivery, and presentation.* NEM Summit 2009, pages 12-16, Saint-Malo, September 2009.
  - ♢ S. Spagnol and F. Avanzini. *Real-time binaural audio rendering in the near field.* In Proc. 6th Int. Conf. on Sound and Music Computing (SMC09), pages 201-206, Porto, July 2009.

9.2.3 National Conferences

- 2011:
  - ♢ M. Geronazzo, S. Spagnol, and F. Avanzini. *Customized 3D sound for innovative interaction design.* In Proc. SMC-HCI Workshop, CHItaly 2011 Conference, Alghero, September 2011.
- 2010:
  - ♢ S. Spagnol, M. Geronazzo, and F. Avanzini. *Structural modeling of pinna-related transfer functions for 3-D sound rendering.* In Proc. XVIII Colloquio di Informatica Musicale (XVIII CIM), Torino, October 2010.

9.2.4 Book Chapters

- 2012 (expected):
  - ♢ F. Avanzini, S. Spagnol, A. De Götzen, and A. Rodà. Designing interactive sound for neurorehabilitation systems. Chapter in *Sonic Interaction Design* book, edited by K. Franinovic and S. Serafin, MIT Press. *Accepted for publication*.

Bibliography

[1] V. R. Algazi. Private communications, 2010.

[2] V. R. Algazi, C. Avendano, and R. O. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *J. Acoust. Soc. Am.*, 109(3):1110–1122, March 2001.

[3] V. R. Algazi, C. Avendano, and R. O. Duda. Estimation of a spherical-head model from anthropometry. *J. Audio Eng. Soc.*, 49(6):472–479, 2001.

[4] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Am.*, 112(5):2053–2064, November 2002.

[5] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson. Structural composition and decomposition of HRTFs. In *IEEE Work. Appl. Signal Process., Audio, Acoust.*, pages 103–106, New Paltz, New York, USA, 2001.

[6] V. R. Algazi, R. O. Duda, and D. M. Thompson. The use of head-and-torso models for improved spatial sound synthesis. In *Proc. 113th Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, October 5-8 2002.

[7] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF database. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, pages 1–4, New Paltz, New York, USA, 2001.

[8] F. Asano, Y. Suzuki, and T. Sone. Role of spectral cues in median plane localization. *J. Acoust. Soc. Am.*, 88(1):159–168, July 1990.

[9] F. Avanzini, A. De Götzen, S. Spagnol, and A. Rodá. Integrating auditory feedback in motor rehabilitation systems. In *Proc. Int. Conf. Multimodal Interfaces for Skills Transfer (SKILLS09)*, Bilbao, Spain, December 2009.

[10] F. Avanzini, D. Rocchesso, and S. Serafin. A toolkit for interactive sonification. In *Proc. 10th Int. Conf. Auditory Display (ICAD 2004)*, Sydney, Australia, 2004.

[11] F. Avanzini, S. Spagnol, A. De Götzen, and A. Rodá. Designing interactive sound for neurorehabilitation systems. In *Sonic Interaction Design*. K. Franinovic and S. Serafin, MIT Press, Cambridge, MA, USA, 2012. Accepted for publication.

[12] D. W. Batteau. The role of the pinna in human localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168(1011):158–180, August 1967.

[13] F. L. Bedford. Can a space-perception conflict be solved with three sense modalities? *Perception*, 36:508–515, 2007.

[14] D. R. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[15] D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *J. Audio Eng. Soc.*, 49(10):904–916, October 2001.

[16] A. J. Berkhout. A holographic approach to acoustic control. *J. Audio Eng. Soc.*, 36(12):977–995, December 1988.

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

[18] R. F. Boian, J. E. Deutsch, C. S. Lee, G. C. Burdea, and J. Lewis. Haptic effects for virtual reality-based post-stroke rehabilitation. *HAPTICS*, page 247, 2003.

[19] A. W. Bronkhorst. Localization of real and virtual sound sources. *J. Acoust. Soc. Am.*, 98(5):2542–2553, 1995.

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

[21] D. S. Brungart. Auditory localization of nearby sources. II. Stimulus effects. *J. Acoust. Soc. Am.*, 106(6):3589–3602, December 1999.

[22] D. S. Brungart. Near-field virtual audio displays. *Presence*, 11(1):93–106, February 2002.

[23] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz. Auditory localization of nearby sources. II. Localization of a broadband source. *J. Acoust. Soc. Am.*, 106(4):1956–1968, October 1999.

[24] D. S. Brungart and W. M. Rabinowitz. Auditory localization of nearby sources. Head-related transfer functions. *J. Acoust. Soc. Am.*, 106(3):1465–1479, September 1999.

[25] M. S. Cameirao, S. Bermuidez i Badia, L. Zimmerli, E. Duarte Oller, and P. F. M. J. Verschure. The rehabilitation gaming system: a virtual reality based system for the evaluation and rehabilitation of motor deficits. In *Proc. IEEE Virtual Rehab. Conf.*, pages 29–33, 27-29 September 2007.

[26] A. Camurri, G. Volpe, H. Vinet, R. Bresin, E. Maestre, J. Llop, J. Kleimola, S. Oksanen, V. Välimäki, and J. Seppanen. User-centric context-aware mobile applications for embodied music listening. In *Proc. 1st Int. ICST Conf. User Centric Media*, Venice, Italy, 2009.

[27] G. Castellano, R. Bresin, A. Camurri, and G. Volpe. Expressive control of music and visual media by full-body movement. In *Proc. 7th Int. Conf. New Interf. Music. Expr. (NIME '07)*, pages 390–391, New York, NY, USA, 2007.

[28] S. Cenci, G. Rosati, D. Zanotto, F. Oscari, and A. Rossi. First test results of a haptic tele-operation system to enhance stability of telescopic handlers. In *Proc. 10th Conf. Eng. Syst. Des. Anal. (ESDA2010)*, Istanbul, Turkey, July 12-14 2010.

[29] J. Chen, B. D. Van Veen, and K. E. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *J. Acoust. Soc. Am.*, 97(1):439–452, January 1995.

[30] C. I. Cheng and G. H. Wakefield. Introduction to head-related transfer functions (HRTFs): Representations of hrtfs in time, frequency, and space. *J. Audio Eng. Soc.*, 49(4):231–249, April 2001.

[31] M. C. Cirstea and M. F. Levin. Compensatory strategies for reaching in stroke. *Brain*, 123:940–953, 2000.

[32] R. Colombo, F. Pisano, S. Micera, A. Mazzone, C. Delconte, M. C. Carrozza, P. Dario, and G. Minuco. Robotic techniques for upper limb evaluation and rehabilitation of stroke patients. *IEEE Trans. Neural Syst. Rehab. Eng.*, 13(3):311–324, September 2005.

[33] R. Colombo, F. Pisano, S. Micera, A. Mazzone, C. Delconte, M. C. Carrozza, P. Dario, and G. Minuco. Assessing mechanisms of recovery during robot-aided neurorehabilitation of the upper limb. *Neurorehabil. Neural Repair*, 22:50–63, 2008.

[34] B. B. Connor, A. M. Wing, G. W. Humphreys, R. M. Bracewell, and D. A. Harvey. Errorless learning using haptic guidance: research in cognitive rehabilitation following stroke. In *Proc. 4th Int. Conf. Disability, Virtual Reality & Assoc. Tech.*, pages 77–84, Veszprém, Hungary, 2002.

[35] A. G. D. Correa, G. A. de Assis, M. do Nascimento, I. Ficheman, and R. de Deus Lopes. GenVirtual: An augmented reality musical game for cognitive and motor rehabilitation. In *Proc. IEEE Virtual Rehab. Conf.*, pages 1–6, 27-29 September 2007.

[36] A. de Götzen and D. Rocchesso. The speed accuracy trade-off through tuning tasks. In *Proc. 4th Int. Conf. Enactive Interfaces (Enactive 07)*, pages 81–84, November 2007.

[37] J. A. Deutsch, J. A. Lewis, E. Whitworth, R. Boian, G. Burdea, and M. Tremaine. Formative evaluation and preliminary findings of a virtual reality telerehabilitation system for the lower extremity. *Presence: Teleoperators and Virtual Environments*, 14(2):198–213, April 2005.

[38] A. di Lauro, L. Pellegrino, G. Savastano, C. Ferraro, M. Fusco, F. Balzarano, M. M. Franco, L. G. Biancardi, and A. Grasso. A randomized trial on the efficacy of intensive rehabilitation in the acute phase of ischemic stroke. *J. Neurology*, 10(250):1206–1208, 2003.

[39] A. W. Dromerick, D. F. Edwards, and M. Hahn. Does the application of constraint-induced movement therapy during acute rehabilitation reduce arm impairment after ischemic stroke? *Stroke*, 31:2984–2988, 2000.

[40] R. O. Duda, C. Avendano, and V. R. Algazi. An adaptable ellipsoidal head model for the interaural time difference. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'99)*, pages 965–968, Phoenix, AZ, USA, March 1999.

[41] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification: Second Edition*. John Wiley & Sons, New York, NY, USA, 2001.

[42] R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5):3048–3058, November 1998.

[43] E. C. Durant and G. H. Wakefield. Efficient model fitting using a genetic algorithm: pole-zero approximations of HRTFs. *IEEE Trans. Speech Audio Process.*, 10(1):18–27, 2002.

[44] P. Ellis and L. van Leeuwen. Confronting the transition: Improving quality of life for the elderly with an interactive multisensory environment - a case study. In Constantine Stephanidis, editor, *Universal Access in Human-Computer Interaction. Addressing Diversity*, volume 5614 of *Lecture Notes in Computer Science*, pages 210–219. Springer Berlin / Heidelberg, 2009.

[45] J. L. Emken, R. Benitez, A. Sideris, J. E. Bobrow, and D. J. Reinkensmeyer. Motor adaptation as a greedy optimization of error and effort. *J. Neurophysiol.*, 97:3997–4006, 2007.

[46] P. A. A. Esquef, M. Karjalainen, and V. Välimäki. Frequency-zooming ARMA modeling for analysis of noisy string instrument tones. *EURASIP J. Appl. Signal Process.*, 10:953–967, 2003.

[47] M. J. Evans, J. A. S. Angus, and A. I. Tew. Analyzing head-related transfer function measurements using surface spherical harmonics. *J. Acoust. Soc. Am.*, 104(4):2400–2411, October 1998.

[48] K. J. Faller II, A. Barreto, and M. Adjouadi. Augmented Hankel total least-squares decomposition of head-related transfer functions. *J. Audio Eng. Soc.*, 58(1/2):3–21, January/February 2010.

[49] K. J. Faller II, A. Barreto, N. Gupta, and N. Rishe. Time and frequency decomposition of head-related impulse responses for the development of customizable spatial audio models. *WSEAS Trans. Signal Proc.*, 2(11):1465–1472, 2006.

[50] T. Flash and N. Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *J. Neurosc.*, 5(7):1688–703, 1985.

[51] B. Friedlander and B. Porat. The modified Yule-Walker method of ARMA spectral estimation. *IEEE Trans. Aerosp. Electron. Syst.*, AES-20(2):158–173, March 1984.

[52] M. B. Gardner. Distance estimation of 0° or apparent 0°-oriented speech signals in anechoic space. *J. Acoust. Soc. Am.*, 45(1):47–53, 1969.

[53] M. B. Gardner and R. S. Gardner. Problem of localization in the median plane: Effect of pinnae cavity occlusion. *J. Acoust. Soc. Am.*, 53(2):400–408, 1973.

[54] W. G. Gardner. *3-D audio using loudspeakers*. Kluwer Acad. Publ., Boston u.a., 1998.

[55] G. Gatzsche, B. Michel, J. Delvaux, and L. Altmann. Beyond DCI: The integration of object oriented 3D sound into the Digital Cinema. In *Proc. 2008 NEM Summit*, pages 247–251, Saint-Malo, France, October 2008.

[56] T. Geneva. The World Health Report 2006: working together for health. Technical report, World Health Organization, 2006.

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

[58] M. Geronazzo, S. Spagnol, and F. Avanzini. Customized 3D sound for innovative interaction design. In *Proc. SMC-HCI Work., CHItaly 2011 Conf.*, Alghero, Italy, September 2011.

[59] M. Geronazzo, S. Spagnol, and F. Avanzini. A head-related transfer function model for real-time customized 3-D sound rendering. In *Proc. INTERPRET Work., SITIS 2011 Conf.*, pages 174–179, Dijon, France, November-December 2011.

[60] M. A. Gerzon. Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.*, 33(11):859–871, November 1985.

[61] J. A. Gil, M. Alcafiiz, J. Montesa, M. Ferrer, J. Chirivella, E. Noe, C. Colomer, and J. Ferri. Low-cost virtual motor rehabilitation system for standing exercises. In *Proc. IEEE Virtual Rehab. Conf.*, pages 34–38, September 2007.

[62] N. Gupta, A. Barreto, and M. Choudhury. Modeling head-related transfer functions based on pinna anthropometry. In *Proc. Int. Lat. Am. Carib. Conf. Eng. Tech. (LACCEI)*, Miami, FL, USA, 2004.

[63] D. Hammershøi and H. Møller. Sound transmission to and within the human ear canal. *J. Acoust. Soc. Am.*, 100(1):408–427, July 1996.

[64] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho. Augmented reality audio for mobile and wearable appliances. *J. Audio Eng. Soc.*, 52(6):618–639, 2004.

[65] R. V. L. Hartley and T. C. Fry. The binaural localization of pure tones. *Phys. Rev.*, 18:431–442, 1921.

[66] W. S. Harwin, J. L. Patton, and V. R. Edgerton. Challenges and opportunities for robot-mediated neurorehabilitation. *Proc. IEEE*, 94(9):1717–1726, September 2006.

[67] J. Hebrank and D. Wright. Are two ears necessary for localization of sound sources on the median plane? *J. Acoust. Soc. Am.*, 56(3):935–938, September 1974.

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

[69] C. Hendrix and W. Barfield. Presence in virtual environments as a function of visual and auditory cues. In *Proc. IEEE Virtual Reality Annual Int. Symp.*, page 74, Washington, DC, USA, 1995.

[70] J. Hidler, D. Nichols, M. Pelliccio, K. Brady, D. D. Campbell, J. H. Kahn, and T. G. Hornby. Multicenter randomized clinical trial evaluating the effectiveness of the Lokomat in subacute stroke. *Neurorehab. Neural Repair*, 23:5–13, 2009.

[71] D. Hilton, S. Cobb, T. Pridmore, and J. Gladman. Virtual reality and stroke rehabilitation: a tangible interface to an every day task. In *Proc. 4th Int. Conf. Disability, Virtual Reality & Assoc. Tech.*, pages 63–70, Veszprm, Hungary, 2002.

[72] M. K. Holden. Virtual environments for motor rehabilitation: Review. *Cyberpsychology & Behavior*, 8(3):187–219, 2005.

[73] X. L. Hu, K.-Y. Tong, R. Song, X. J. Zheng, and W. W. F. Leung. A comparison between electromyography-driven robot and passive motion device on wrist rehabilitation for chronic stroke. *Neurorehab. Neural Repair*, 23(8):837–846, 2009.

[74] S. Hwang, Y. Park, and Y. Park. Modeling and customization of head-related impulse responses based on general basis functions in time domain. *Acta Acustica united with Acustica*, 94(6):965–980, November 2008.

[75] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto. Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Appl. Acoust.*, 68:835–850, 2007.

[76] J. F. Israel, D. D. Campbell, J. H. Kahn, and T. G. Hornby. Metabolic costs and muscle activity patterns during robotic and therapist-assisted treadmill walking in individuals with incomplete spinal cord injury. *Phys. Ther.*, 86(1):1466–1478, 2006.

[77] H. Jo, Y. Park, and Y. Park. Approximation of head related transfer function using prolate spheroidal head model. In *Proc. 15th Int. Congr. Sound Vibr. (ICSV15)*, pages 2963–2970, Daejeon, Korea, July 6-10 2008.

[78] H. Jo, Y. Park, and Y. Park. Optimization of spherical and spheroidal head model for head related transfer function customization: Magnitude comparison. In *Proc. Int. Conf. Control, Automat., Syst.*, pages 251–254, Seoul, Korea, October 14-17 2008.

[79] M. Johnson, H. V. der Loos, C. Burgar, P. Shor, and L. Leifer. Design and evaluation of driver's seat: A car steering simulation environment for upper limb stroke therapy. *Robotica*, 21(1):13–23, January 2003.

[80] M. J. Johnson, K. J. Wisneski, J. Anderson, D. Nathan, and R. O. Smith. Development of ADLER: The activities of daily living exercise robot. *Biom. Robotics and Biomechatronics*, pages 881–886, 2006.

[81] Y. Kahana and P. A. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *J. Sound Vibr.*, 300(3-5):552–579, 2007.

[82] T. Kanade, B. Davies, and C. N. Riviere. Special issue on medical robotics. *Proc. IEEE*, 94(9):1649–1651, September 2006.

[83] B. Kapralos, M. R. Jenkin, and E. Milios. Virtual audio systems. *Presence: Teleoper. Virtual Environ.*, 17:527–549, December 2008.

[84] C. I. Karageorghis and P. C. Terry. The psycho-physical effects of music in sport and exercise: A review. *J. Sport Behavior*, 20:54–68, 1997.

[85] B. F. G. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *J. Acoust. Soc. Am.*, 110(5):2440–2448, November 2001.

[86] S. H. Kim, S. K. Banala, E. A. Brackbill, S. K. Agrawal, V. Krishnamoorthy, and J. P. Scholz. Robot-assisted modifications of gait in healthy individuals. *Exp. Brain Res.*, 202:809–824, 2010.

[87] O. Kirkeby, E. T. Seppälä, A. Kärkkäinen, L. Kärkkäinen, and T. Huttunen. Some effects of the torso on head-related transfer functions. In *Proc. 122nd Conv. Audio Eng. Soc.*, Vienna, Austria, May 5-8 2007.

[88] D. J. Kistler and F. L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647, 1992.

[89] P. L. Kolominsky-Rabas, P. U. Heuschmann, D. Marschall, M. Emmert, N. Baltzer, B. Neundörfer, O. Schöffski, and K. J. Krobot. Lifetime cost of ischemic stroke in Germany: results and national projections from a population-based stroke registry: the Erlangen Stroke Project. *Stroke*, 37(5):1179–1183, 2006.

[90] S. Kousidou, N. G. Tsagarakis, C. Smith, and D. G. Caldwell. Task-orientated biofeedback system for the rehabilitation of the upper limb. In *Proc. 10th IEEE Int. Conf. Rehab. Robotics (ICORR 2007)*, pages 376–384, June 2007.

[91] H. I. Krebs and N. Hogan. Therapeutic robotics: A technology push. *Proc. IEEE*, 94(9):1727–1738, September 2006.

[92] H. I. Krebs, N. Hogan, M. L. Aisen, and B. T. Volpe. Robot-aided neurorehabilitation. *IEEE Trans. Rehab. Eng.*, 6(1):75–87, March 1998.

[93] G. F. Kuhn. Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Am.*, 62(1):157–167, July 1977.

[94] G. Kwakkel, B. J. Kollen, and H. I. Krebs. Effects of robot-assisted therapy on upper limb recovery after stroke: A systematic review. *Neurorehab. Neural Repair*, 22(2):111–121, 2008.

[95] D. Lloyd-Jones, R. J. Adams, T. M. Brown, M. Carnethon, S. Dai, G. De Simone, T. B. Ferguson, E. Ford, K. Furie, C. Gillespie, A. Go, K. Greenlund, N. Haase, S. Hailpern, P. M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lackland, L. Lisabeth, A. Marelli, M. M. McDermott, J. Meigs, D. Mozaffarian, M. Mussolino, G. Nichol, V. L. Roger, W. Rosamond, R. Sacco, P. Sorlie, V. L. Roger, R. Stafford, T. Thom, S. Wasserthiel-Smoller, N. D. Wong, and J. Wylie-Rosett. Heart disease and stroke statistics - 2010 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*, 121(7):e46–e215, 2010.

[96] E. A. Lopez-Poveda and R. Meddis. A physical model of sound diffraction and reflections in the human concha. *J. Acoust. Soc. Am.*, 100(5):3248–3259, November 1996.

[97] R. Loureiro, F. Amirabdollahian, M. Topping, B. Driessen, and W. Harwin. Upper limb robot mediated stroke therapy-GENTLE/s approach. *Autonomous Robots*, 15:35–51, 2003.

[98] I. S. MacKenzie, A. Sellen, and W. Buxton. A comparison of input devices in elemental pointing and dragging tasks. In *CHI91*, pages 161–166, New York, 1991.

[99] L. Marchal-Crespo and D. J. Reinkensmeyer. Review of control strategies for robotic movement training after neurologic injury. *J. Neuroeng. Rehab.*, 6:20, 2008.

[100] S. Masiero, M. Armani, and G. Rosati. Upper extremity robot-assisted therapy in rehabilitation of acute stroke patients: focused review and results of a new randomized controlled trial. *J. Rehab. Res. Develop.*, 48(4), 2011.

[101] S. Masiero, A. Celia, M. Armani, G. Rosati, B. Tavolato, and C. Ferraro. Robot-aided intensive training in post-stroke recovery. *Aging Clin. Exp. Res.*, 18:261–265, 2006.

[102] S. Masiero, A. Celia, G. Rosati, and M. Armani. Robotic-assisted rehabilitation of the upper limb after acute stroke. *Arch. Phys. Med. Rehab.*, 88:142–149, 2007.

[103] M. Massimino. Improved force perception through sensory substitution. *Control Eng. Practice*, 3(2):215–222, February 1995.

[104] R. A. Maulucci and R. H. Eckhouse. Retraining reaching in chronic stroke with real-time auditory feedback. *NeuroRehab.*, 16:171–182, 2001.

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

[106] D. K. McGookin and S. A. Brewster. Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition. *ACM Trans. Applied Perceptions*, 1(2):130–155, October 2004.

[107] M. McLaughlin, R. Zimmermann, L. S. Liu, Y. Jung, W. Peng, S. A. Jin, J. Stewart, S. C. Yeh, W. Zhu, and B. Seo. Integrated voice and haptic support for tele-rehabilitation. *Pervasive Comp. Comm. Work.*, pages 1–4, March 2006.

[108] J. Mehrholz, T. Platz, J. Kugler, and M. Pohl. Electromechanical and robot-assisted arm training for improving arm function and activities of daily living after stroke (review). *Cochrane Database of Systematic Reviews*, 4, 2008.

[109] P. B. Meijer. An experimental system for auditory image representations. *IEEE Trans. Biomed. Eng.*, 39(2):112–121, February 1992.

[110] D. H. Mershon and J. N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8(3):311–322, 1979.

[111] A. W. Mills. Lateralization of high-frequency tones. *J. Acoust. Soc. Am.*, 32(1):132–135, January 1960.

[112] A. Misra, G. Essl, and M. Rohs. Microphone as sensor in mobile phone performance. In *Proc. 8th Int. Conf. New Interf. Music. Expr. (NIME '08)*, Genova, Italy, 2008.

[113] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Acoustic simulation of KEMAR's HRTFs: Verification with measurements and the effects of modifying head shape and pinna concavity. In *Proc. Int. Work. Princ. Appl. Spatial Hearing (IWPASH)*, Zao, Miyagi, Japan, November 2009.

[114] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Acoustic sensitivity to micro-perturbations of KEMAR's pinna surface geometry. In *Proc. 20th Int. Congr. Acoust. (ICA 2010)*, Sydney, Australia, August 2010.

[115] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2011)*, Prague, Czech Republic, May 2011.

[116] B. I. Molier, E. H. F. van Asseldonk, G. B. Prange, and J. H. Buurke. Influence of reaching direction on visuomotor adaptation: an explorative study. In *Proc. IEEE 12th Int. Conf. Rehab. Rob. (ICORR2011)*, Zurich, CH, June 29 - July 1 2011.

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

[118] B. C. J. Moore, S. R. Oldfield, and G. J. Dooley. Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *J. Acoust. Soc. Am.*, 85(2):820–836, February 1989.

[119] M. Morimoto. The contribution of two ears to the perception of vertical angle in sagittal planes. *J. Acoust. Soc. Am.*, 109(4):1596–1603, April 2001.

[120] D. Morris, H. Tan, F. Barbagli, T. Chang, and K. Salisbury. Haptic feedback enhances force skill learning. In *EuroHaptics Conf. and Symp. Haptic Interf. Virt. Env. Teleop. Syst.*, pages 21 – 26, 2007.

[121] S. Müller and P. Massarani. Transfer-function measurement with sweeps. *J. Audio Eng. Soc.*, 49(6):443–471, June 2001.

[122] T. Nef, M. Mihelj, G. Colombo, and R. Riener. ARMin - robot for rehabilitation of the upper extremities. In *IEEE Int. Conf. Robot. Autom. (ICRA 2006)*, pages 3152–3157, Orlando, FL, USA, 2006.

[123] T. Nishino, N. Inoue, K. Takeda, and F. Itakura. Estimation of HRTFs on the horizontal plane using physical features. *Acoust. Science Technol.*, 68:897–908, 2007.

[124] R. W. Novy. Characterizing elevation effects of a prolate spheroidal HRTF model. Master's thesis, San Jose State University, 1998.

[125] R. J. Nudo. Postinfarct cortical plasticity and behavioral recovery. *Stroke*, 38(2):840–845, 2007.

[126] R. J. Nudo, B. M. Wise, F. SiFuentes, and G. W. Milliken. Neural substrates for the effects of rehabilitative training on motor recovery after ischemic infarct. *Science*, 272:1791–1794, 1996.

[127] S. J. Orfanidis, editor. *Introduction To Signal Processing*. Prentice Hall, 1996.

[128] N. Orio, N. Schnell, and M. Wanderley. Input devices for musical expression: borrowing tools from HCI. In *Proc. 1st Int. Work. New Interf. Music. Expr. (NIME'01)*, pages 1–4, Seattle, WA, USA, 2001.

[129] J. Perry. *Gait Analysis: Normal and Pathological Function*. Slack Incorporated, Thorofare, NJ, USA, 1992.

[130] A. Pirhonen, S. Brewster, and C. Holguin. Gestural and audio metaphors as a means of control for mobile devices. In *CHI2002*, pages 291–298, Minneapolis, MN, USA, 2002.

[131] G. Prange, M. Jannink, C. Groothuis-Oudshoorn, H. Hermens, and M. Ijzerman. Systematic review of the effect of robot-aided therapy on recovery of the hemiparetic arm after stroke. *J. Rehab. Res. Develop.*, 43(2):171–183, 2006.

[132] M. Puckette. Pure Data: another integrated computer music environment. In *Proc. Int. Computer Music Conf.*, pages 37–41, 1996.

[133] T. Qu, Z. Xiao, M. Gong, Y. Huang, X. Li, and X. Wu. Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. *IEEE Trans. Audio, Speech, Lang. Process.*, 17(6):1124–1132, August 2009.

[134] W. M. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei. Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere. *Presence*, 2:125–129, 1993.

[135] M. Rath and D. Rocchesso. Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69, April-June 2005.

[136] M. Rauterberg and E. Styger. Positive effects of sound feedback during the operation of a plant simulator. *Lecture Notes Comp. Science*, 876:35 – 44, 1994.

[137] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *J. Acoust. Soc. Am.*, 118(1):364–374, July 2005.

[138] A. Reben, M. Laibowitz, and J. Paradiso. Responsive music interfaces for performance. In *Proc. 9th Int. Conf. New Interf. Music. Expr. (NIME '09)*, Pittsburgh, PA, USA, 2009.

[139] D. J. Reinkensmeyer and J. Galvez. Some key problems for robot-assisted movement therapy research: A perspective from the university of California at Irvine. In *IEEE 10th Int. Conf. Rehab. Rob. (ICORR 2007)*, pages 1009–1015, 2007.

[140] J. V. G. Robertson, T. Hoellinger, P. Lindberg, D. Bensmail, S. Hanneton, and A. Roby-Brami. Effect of auditory feedback differs according to side of hemiparesis: a comparative pilot study. *J. Neuroeng. Rehab.*, 6:45, 2009.

[141] R. Ronsse, R. C. Miall, and S. P. Swinnen. Multisensory integration in dynamical behaviors: maximum likelihood estimation across bimanual skill learning. *J. Neurosc.*, 29(26):8419–28, 2009.

[142] G. Rosati. The place of robotics in post-stroke rehabilitation. *Exp. Rev. Med. Devices*, 7(6):753–758, 2010.

[143] G. Rosati, S. Masiero, E. Carraro, P. Gallina, M. Ortolani, and A. Rossi. Robot-aided upper limb rehabilitation in the acute phase. In *Proc. IEEE Virtual Rehab. Conf.*, 27-29 September 2007.

[144] G. Rosati, F. Oscari, D. J. Reinkensmeyer, R. Secoli, F. Avanzini, S. Spagnol, and S. Masiero. Improving robotics for neurorehabilitation: Enhancing engagement, performance, and learning with auditory feedback. In *Proc. IEEE 12th Int. Conf. Rehab. Rob. (ICORR2011)*, pages 341–346, Zurich, Switzerland, June-July 2011.

[145] G. Rosati, F. Oscari, S. Spagnol, F. Avanzini, and S. Masiero. Effect of task-related continuous auditory feedback during learning of tracking motion exercises. *J. Neuroeng. Rehab.*, 2012 (expected). Submitted for publication.

[146] D. Rubine. *The automatic recognition of gesture*. PhD thesis, Carnegie-Mellon University, School of Computer Science, Pittsburgh, PA, USA, 1991.

[147] H. Sadeghi, P. Allard, F. Prince, and H. Labelle. Symmetry and limb dominance in able-bodied gait: a review. *Gait & posture*, 12(1):34–45, September 2000.

[148] P. Satarzadeh. A study of physical and circuit models of the human pinnae. Master's thesis, University of California Davis, 2006.

[149] P. Satarzadeh, R. V. Algazi, and R. O. Duda. Physical and filter pinna models based on anthropometry. In *Proc. 122nd Conv. Audio Eng. Soc.*, Vienna, Austria, May 5-8 2007.

[150] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Creating interactive virtual acoustic environments. *J. Audio Eng. Soc.*, 47(9):675–705, September 1999.

[151] C. Scaletti. *Auditory Display: Sonification, Audification, and Auditory Interfaces*, volume 1, chapter Sound synthesis algorithms for auditory data representations, pages 223–251. Reading, MA: Addison Wesley, 1994.

[152] A. Schaufelberger, J. Zitzewitz, and R. Riener. Evaluation of visual and auditory feedback in virtual obstacle walking. *Presence*, 17(5):512–524, 2008.

[153] R. A. Scheidt, M. A. Conditt, E. L. Secco, and F. A. Mussa-Ivaldi. Interaction of visual and proprioceptive feedback during adaptation of human reaching movements. *J. Neurophysiol.*, 93(6):3200–13, 2005.

[154] M. Scholz. *Approaches to Analyse and Interpret Biological Profile Data*. PhD thesis, University of Potsdam, Germany, 2006.

[155] R. Secoli, M.-H. Milot, G. Rosati, and D. J. Reinkensmeyer. Effect of visual distraction and auditory feedback on patient effort during robot-assisted movement training after stroke. *J. NeuroEng. Rehab.*, 8(21), 2011.

[156] T. Seizova-Cajic and R. Azzi. A visual distracter task during adaptation reduces the proprioceptive movement aftereffect. *Exp. Brain Res.*, 203:213–219, 2010.

[157] E. A. G. Shaw. The acoustics of the external ear. In *Acoustical Factors Affecting Hearing Aid Performance*. G. A. Studebaker and I. Hochberg, University Park Press, Baltimore, MD, USA, 1980.

[158] E. A. G. Shaw. Acoustical features of human ear. In *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 25–47. R. H. Gilkey and T. R. Anderson, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1997.

[159] E. A. G. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *J. Acoust. Soc. Am.*, 44(1):240–249, 1968.

[160] C. Y. Shing, C. P. Fung, T. Y. Chuang, I. W. Penn, and J. L. Doong. The study of auditory and haptic signals in a virtual reality-based hand rehabilitation system. *Robotica*, 21(2):211–218, 2003.

[161] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *J. Acoust. Soc. Am.*, 107(3):1627–1636, March 2000.

[162] S. J. Sober and P. N. Sabes. Multisensory integration during motor planning. *J. Neurosc.*, 23(18):6982–92, 2003.

[163] S. Spagnol and F. Avanzini. Real-time binaural audio rendering in the near field. In *Proc. 6th Int. Conf. Sound and Music Computing (SMC09)*, pages 201–206, Porto, Portugal, July 2009.

[164] S. Spagnol, M. Geronazzo, and F. Avanzini. Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP'10)*, pages 194–199, Saint-Malo, France, October 2010.

[165] S. Spagnol, M. Geronazzo, and F. Avanzini. Structural modeling of pinna-related transfer functions. In *Proc. 7th Int. Conf. Sound and Music Computing (SMC10)*, pages 422–428, Barcelona, Spain, July 2010.

[166] S. Spagnol, M. Geronazzo, and F. Avanzini. Structural modeling of pinna-related transfer functions for 3-D sound rendering. In *Proc. XVIII Colloquio di Informatica Musicale (XVIII CIM)*, Torino, Italy, October 2010.

[167] S. Spagnol, M. Geronazzo, and F. Avanzini. On the relation between pinna reflection patterns and head-related transfer function features. *IEEE Trans. Audio, Speech, Lang. Process.*, 2012 (expected). Submitted for publication.

[168] S. Spagnol, M. Hiipakka, and V. Pulkki. A single-azimuth pinna-related transfer function database. In *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, Paris, France, September 2011.

[169] J. W. Strutt. On our perception of sound direction. *Phil. Mag.*, 13:214–232, 1907.

[170] H. Sveistrup. Motor rehabilitation using virtual reality. *J. NeuroEng. Rehab.*, 1:10, 2004.

[171] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida. Pressure distribution patterns on the pinna at spectral peak and notch frequencies of head-related transfer functions in the median plane. In *Proc. Int. Work. Princ. Appl. Spatial Hearing (IWPASH)*, Zao, Miyagi, Japan, November 2009.

[172] H. Takemoto, P. Mokhtari, H. Kato, R. Nishimura, and K. Iida. A simple pinna model for generating head-related transfer functions in the median plane. In *Proc. 20th Int. Congr. Acoust. (ICA 2010)*, Sydney, Australia, August 2010.

[173] J. A. Taylor and K. A. Thoroughman. Divided attention impairs human motor adaptation but not feedback control. *J. Neurophysiol.*, page 10, April 2007.

[174] R. Teranishi and E. A. G. Shaw. External-ear acoustic models with simple geometry. *J. Acoust. Soc. Am.*, 44(1):257–263, 1968.

[175] K. A. Thoroughman and R. Shadmehr. Learning of action through adaptive combination of motor primitives. *Stroke*, 407:742–7, 2000.

[176] W. R. Thurlow and P. S. Runge. Effect of induced head movements on localization of direction of sounds. *J. Acoust. Soc. Am.*, 42(2):480–488, August 1967.

[177] A. A. A. Timmermans, H. A. M. Seelen, R. D. Willmann, and H. Kingma. Technology-assisted training of arm-hand skills in stroke: concepts on reacquisition of motor control and therapist guidelines for rehabilitation technology design. *J. NeuroEng. Rehab.*, 6, 2009.

[178] R. Vertegaal. *An Evaluation of input devices for timbre space navigation*. PhD thesis, University of Bradford, Bradford, UK, 1994.

[179] K. Vogt, D. Pirró, I. Kobenz, R. Höldrich, and G. Eckel. Physiosonic - movement sonification as auditory feedback. In *Proc. 15th Int. Conf. Auditory Display (ICAD 2009)*, Copenhagen, Denmark, 18-21 May 2009.

[180] R. G. von Wettschureck. Die absoluten Unterschiedsschwellen der Richtungswahrnehmung in der Medianebene beim natürlichen Hören, sowie beim Hören über ein Kunstkopf-Übertragungssystem (The absolute difference limen of directional perception

in the median plane under conditions of both, natural hearing and hearing with artificial-head-system). *Acustica*, 28:197–208, 1973.

[181] A. J. Watkins. Psychoacoustical aspects of synthesized vertical locale cues. *J. Acoust. Soc. Am.*, 63(4):1152–1165, April 1978.

[182] M. Wellner, A. Schaufelberger, and R. Riener. A study on sound feedback in a virtual environment for gait rehabilitation. In *Proc. Virtual Rehab. Conf.*, pages 53–56, 2007.

[183] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94(1):111–123, 1993.

[184] F. L. Wightman and D. J. Kistler. Resolution of front-back ambiguity in spatial hearing by listener and source movement. *J. Acoust. Soc. Am.*, 105(5):2841–2853, May 1999.

[185] A. Wilska. *Studies on Directional Hearing*. English translation, Aalto University School of Science and Technology, Department of Signal Processing and Acoustics, 2010. PhD thesis originally published in German as "Untersuchungen über das Richtungshören", University of Helsinki, 1938.

[186] K. N. Winfree, P. Stegall, and S. K. Agrawal. Design of a minimally constraining, passively supported gait training exoskeleton: ALEX II. In *Proc. IEEE 12th Int. Conf. Rehab. Rob. (ICORR2011)*, pages 1053–1058, Zurich, CH, June 29 - July 1 2011.

[187] E. T. Wolbrecht, V. Chan, D. J. Reinkensmeyer, and J. E. Bobrow. Optimizing compliant, model-based robotic assistance to promote neurorehabilitation. *IEEE Trans. Neur. Sys. Rehab. Eng.*, 16(3):286–297, 2008.

[188] R. S. Woodworth and H. Schlosberg. *Experimental Psychology*. Holt, Rinehard and Winston, NY, USA, 1954.

[189] D. Wright, J. H. Hebrank, and B. Wilson. Pinna reflections as cues for localization. *J. Acoust. Soc. Am.*, 56(3):957–962, September 1974.

[190] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: a summary of past and present research. *Acta Acustica united with Acustica*, 91:409–420, 2005.

[191] D. Zanotto, G. Rosati, S. Spagnol, P. Stegall, and S. K. Agrawal. Effects of auditory feedback in robot-assisted lower extremity motor adaptation. *IEEE Trans. Neur. Syst. Rehab. Eng.*, 2012 (expected). Submitted for publication.

[192] U. Zölzer, editor. *Digital Audio Effects*. J. Wiley & Sons, New York, NY, USA, 2002.