UNIVERSITÀ DEGLI STUDI PADOVA

Scuola di Dottorato in Ingegneria dellInformazione
Indirizzo: Scienza e Tecnologia dellInformazione

CICLO XXVI

MIXED STRUCTURAL MODELS FOR 3D AUDIO IN VIRTUAL ENVIRONMENTS

Direttore della Scuola: Ch.mo Prof. Matteo Bertocco

Supervisore: Ch.mo Prof. Federico Avanzini

Dottorando: Michele Geronazzo

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Ingegneria dell'informazione

Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione

Indirizzo: Scienza e Tecnologia dell'Informazione

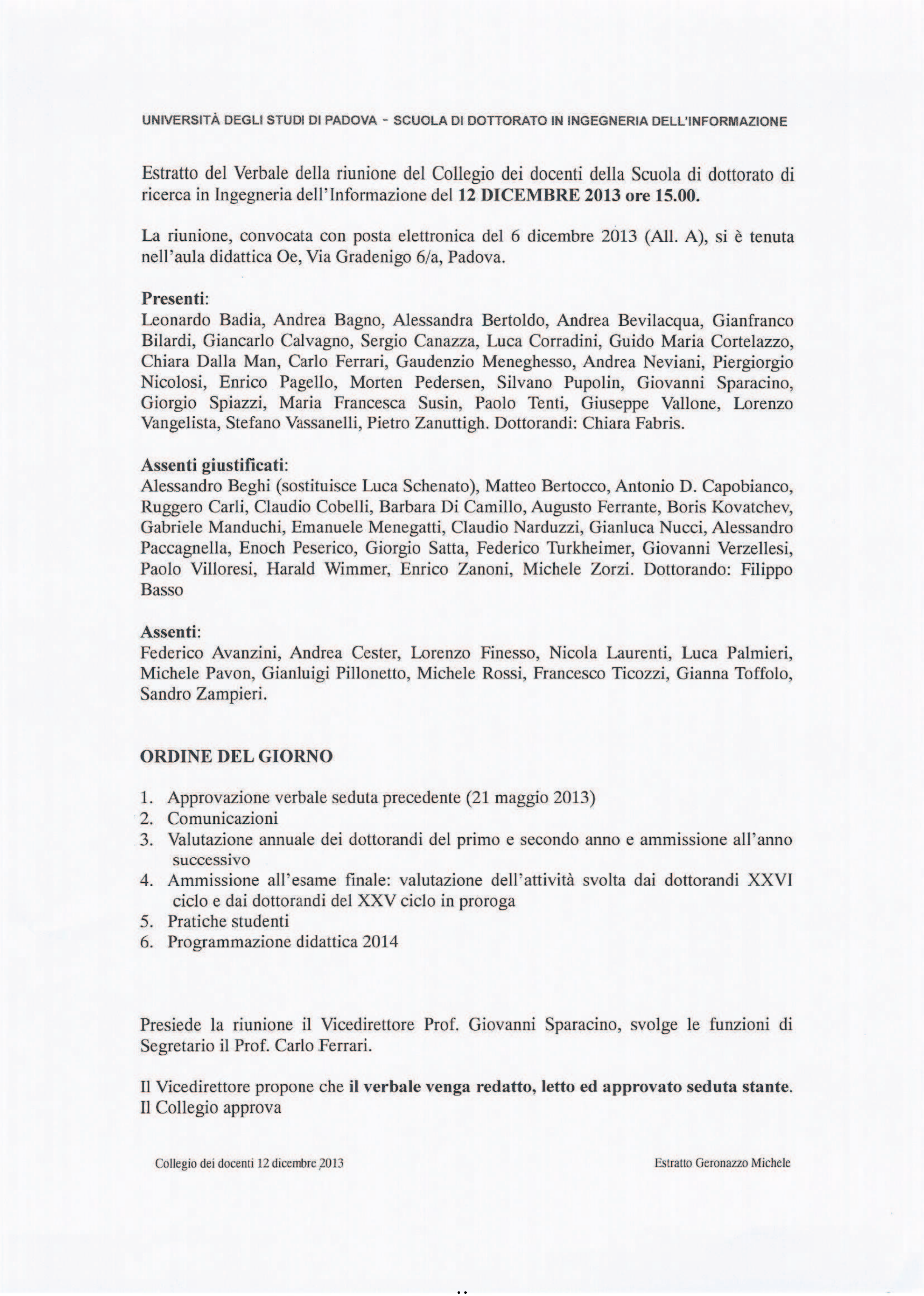
Ciclo XXVI

MIXED STRUCTURAL MODELS FOR 3D AUDIO IN VIRTUAL ENVIRONMENTS

Direttore della Scuola : Ch.mo Prof. Matteo BERTOCCO

Supervisore : Ch.mo Prof. Federico AVANZINI

Dottorando : Michele GERONAZZO



UNIVERSITÀ DEGLI STUDI DI PADOVA - SCUOLA DI DOTTORATO IN INGEGNERIA DELL'INFORMAZIONE

Estratto del Verbale della riunione del Collegio dei docenti della Scuola di dottorato di ricerca in Ingegneria dell'Informazione del **12 DICEMBRE 2013 ore 15.00.**

La riunione, convocata con posta elettronica del 6 dicembre 2013 (All. A), si è tenuta nell'aula didattica Oe, Via Gradenigo 6/a, Padova.

Presenti:
Leonardo Badia, Andrea Bagno, Alessandra Bertoldo, Andrea Bevilacqua, Gianfranco Bilardi, Giancarlo Calvagno, Sergio Canazza, Luca Corradini, Guido Maria Cortelazzo, Chiara Dalla Man, Carlo Ferrari, Gaudenzio Meneghesso, Andrea Neviani, Piergiorgio Nicolosi, Enrico Pagello, Morten Pedersen, Silvano Pupolin, Giovanni Sparacino, Giorgio Spiazzi, Maria Francesca Susin, Paolo Tenti, Giuseppe Vallone, Lorenzo Vangelista, Stefano Vassanelli, Pietro Zanuttigh. Dottorandi: Chiara Fabris.

Assenti giustificati:
Alessandro Beghi (sostituisce Luca Schenato), Matteo Bertocco, Antonio D. Capobianco, Ruggero Carli, Claudio Cobelli, Barbara Di Camillo, Augusto Ferrante, Boris Kovatchev, Gabriele Manduchi, Emanuele Menegatti, Claudio Narduzzi, Gianluca Nucci, Alessandro Paccagnella, Enoch Peserico, Giorgio Satta, Federico Turkheimer, Giovanni Verzellesi, Paolo Villoresi, Harald Wimmer, Enrico Zanoni, Michele Zorzi. Dottorando: Filippo Basso

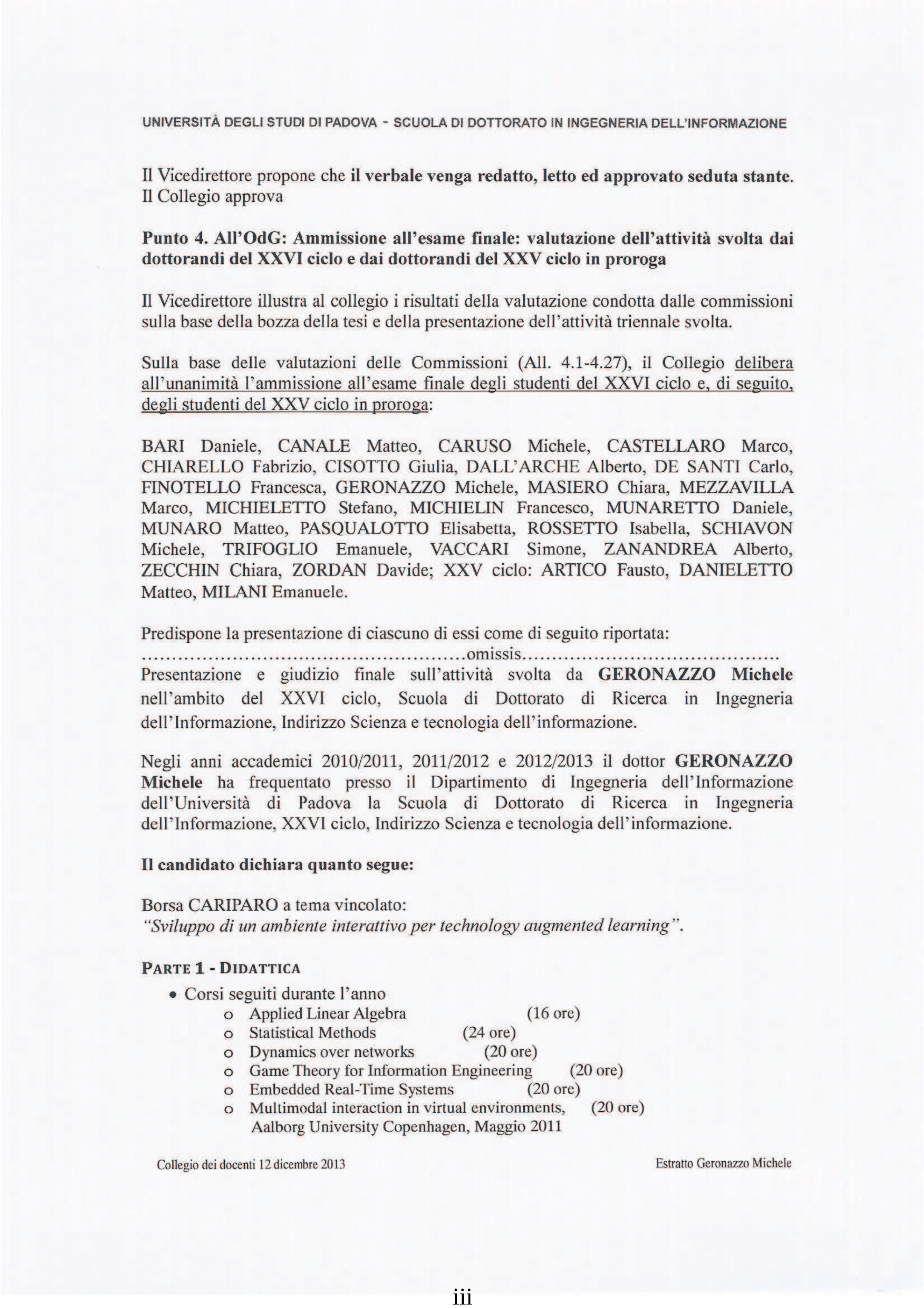
Assenti:
Federico Avanzini, Andrea Cester, Lorenzo Finesso, Nicola Laurenti, Luca Palmieri, Michele Pavon, Gianluigi Pillonetto, Michele Rossi, Francesco Ticozzi, Gianna Toffolo, Sandro Zampieri.

ORDINE DEL GIORNO

1. Approvazione verbale seduta precedente (21 maggio 2013)
2. Comunicazioni
3. Valutazione annuale dei dottorandi del primo e secondo anno e ammissione all'anno successivo
4. Ammissione all'esame finale: valutazione dell'attività svolta dai dottorandi XXVI ciclo e dai dottorandi del XXV ciclo in proroga
5. Pratiche studenti
6. Programmazione didattica 2014

Presiede la riunione il Vicedirettore Prof. Giovanni Sparacino, svolge le funzioni di Segretario il Prof. Carlo Ferrari.

Il Vicedirettore propone che **il verbale venga redatto, letto ed approvato seduta stante**. Il Collegio approva



Il Vicedirettore propone che **il verbale venga redatto, letto ed approvato seduta stante.**
Il Collegio approva

Punto 4. All'OdG: Ammissione all'esame finale: valutazione dell'attività svolta dai dottorandi del XXVI ciclo e dai dottorandi del XXV ciclo in proroga

Il Vicedirettore illustra al collegio i risultati della valutazione condotta dalle commissioni sulla base della bozza della tesi e della presentazione dell'attività triennale svolta.

Sulla base delle valutazioni delle Commissioni (All. 4.1-4.27), il Collegio delibera all'unanimità l'ammissione all'esame finale degli studenti del XXVI ciclo e, di seguito, degli studenti del XXV ciclo in proroga:

BARI Daniele, CANALE Matteo, CARUSO Michele, CASTELLARO Marco, CHIARELLO Fabrizio, CISOTTO Giulia, DALL'ARCHE Alberto, DE SANTI Carlo, FINOTELLO Francesca, GERONAZZO Michele, MASIERO Chiara, MEZZAVILLA Marco, MICHIELETTO Stefano, MICHIELIN Francesco, MUNARETTO Daniele, MUNARO Matteo, PASQUALOTTO Elisabetta, ROSSETTO Isabella, SCHIAVON Michele, TRIFOGLIO Emanuele, VACCARI Simone, ZANANDREA Alberto, ZECCHIN Chiara, ZORDAN Davide; XXV ciclo: ARTICO Fausto, DANIELETTO Matteo, MILANI Emanuele.

Predispone la presentazione di ciascuno di essi come di seguito riportata:
....omissis..
Presentazione e giudizio finale sull'attività svolta da **GERONAZZO Michele** nell'ambito del XXVI ciclo, Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione, Indirizzo Scienza e tecnologia dell'informazione.

Negli anni accademici 2010/2011, 2011/2012 e 2012/2013 il dottor **GERONAZZO Michele** ha frequentato presso il Dipartimento di Ingegneria dell'Informazione dell'Università di Padova la Scuola di Dottorato di Ricerca in Ingegneria dell'Informazione, XXVI ciclo, Indirizzo Scienza e tecnologia dell'informazione.

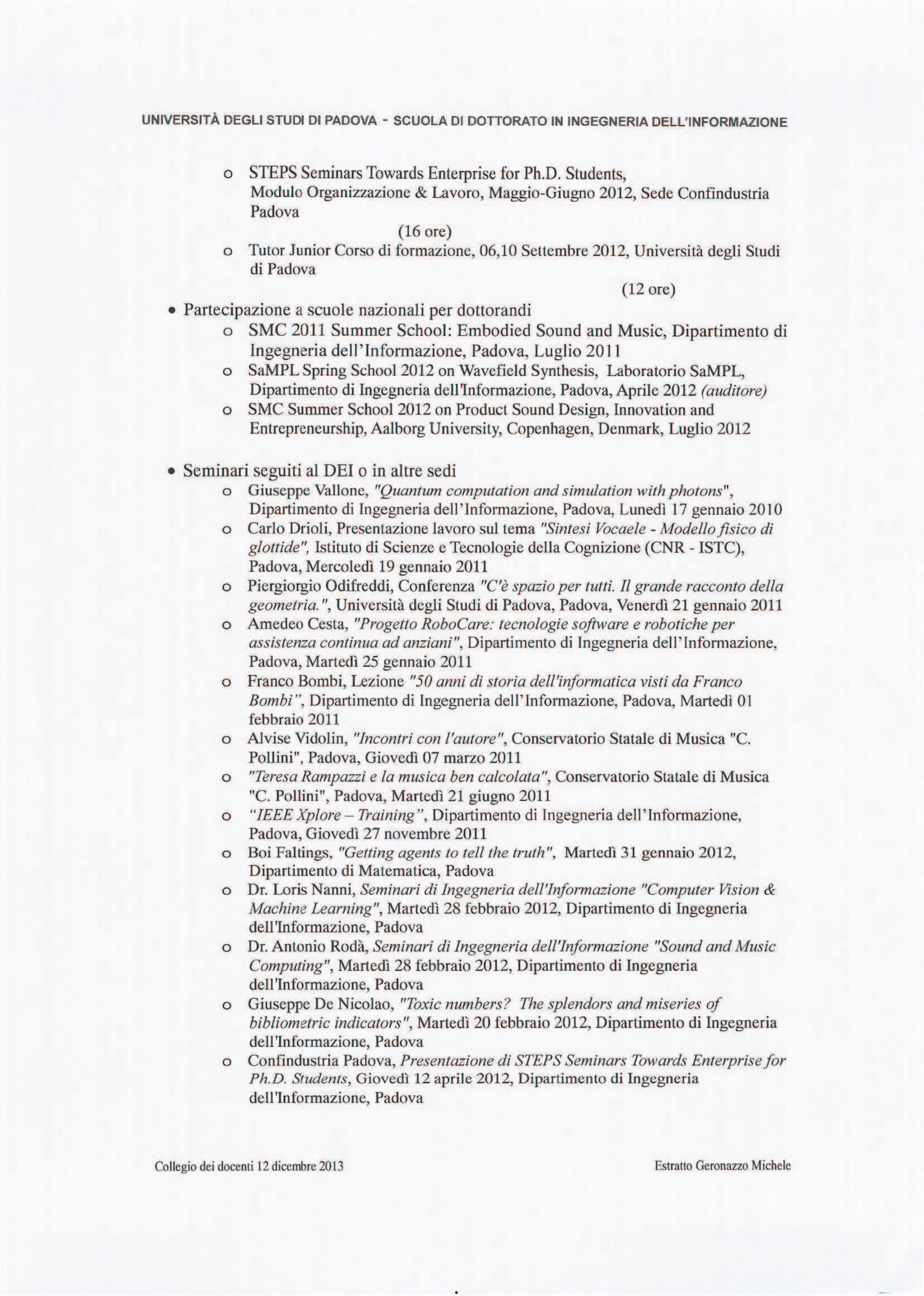
Il candidato dichiara quanto segue:

Borsa CARIPARO a tema vincolato:
"Sviluppo di un ambiente interattivo per technology augmented learning".

PARTE 1 - DIDATTICA

- Corsi seguiti durante l'anno

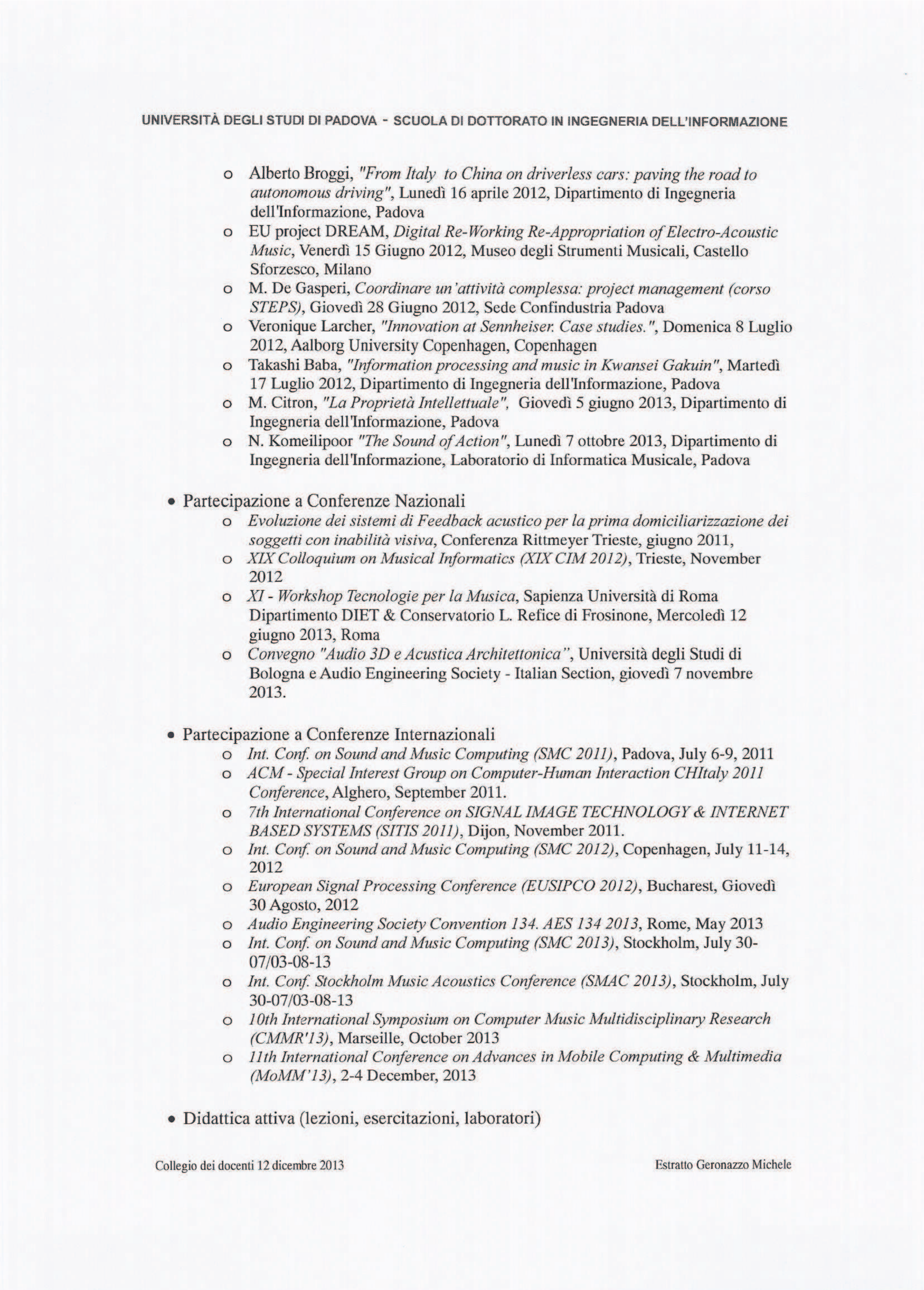
Applied Linear Algebra	(16 ore)
Statistical Methods	(24 ore)
Dynamics over networks	(20 ore)
Game Theory for Information Engineering	(20 ore)
Embedded Real-Time Systems	(20 ore)
Multimodal interaction in virtual environments, Aalborg University Copenhagen, Maggio 2011	(20 ore)



  - STEPS Seminars Towards Enterprise for Ph.D. Students, Modulo Organizzazione & Lavoro, Maggio-Giugno 2012, Sede Confindustria Padova (16 ore)
  - Tutor Junior Corso di formazione, 06,10 Settembre 2012, Università degli Studi di Padova (12 ore)

- Partecipazione a scuole nazionali per dottorandi
  - SMC 2011 Summer School: Embodied Sound and Music, Dipartimento di Ingegneria dell'Informazione, Padova, Luglio 2011
  - SaMPL Spring School 2012 on Wavefield Synthesis, Laboratorio SaMPL, Dipartimento di Ingegneria dell'Informazione, Padova, Aprile 2012 *(auditore)*
  - SMC Summer School 2012 on Product Sound Design, Innovation and Entrepreneurship, Aalborg University, Copenhagen, Denmark, Luglio 2012

- Seminari seguiti al DEI o in altre sedi
  - Giuseppe Vallone, *"Quantum computation and simulation with photons"*, Dipartimento di Ingegneria dell'Informazione, Padova, Lunedì 17 gennaio 2010
  - Carlo Drioli, Presentazione lavoro sul tema *"Sintesi Vocaele - Modello fisico di glottide"*, Istituto di Scienze e Tecnologie della Cognizione (CNR - ISTC), Padova, Mercoledì 19 gennaio 2011
  - Piergiorgio Odifreddi, Conferenza *"C'è spazio per tutti. Il grande racconto della geometria."*, Università degli Studi di Padova, Padova, Venerdì 21 gennaio 2011
  - Amedeo Cesta, *"Progetto RoboCare: tecnologie software e robotiche per assistenza continua ad anziani"*, Dipartimento di Ingegneria dell'Informazione, Padova, Martedì 25 gennaio 2011
  - Franco Bombi, Lezione *"50 anni di storia dell'informatica visti da Franco Bombi"*, Dipartimento di Ingegneria dell'Informazione, Padova, Martedì 01 febbraio 2011
  - Alvise Vidolin, *"Incontri con l'autore"*, Conservatorio Statale di Musica "C. Pollini", Padova, Giovedì 07 marzo 2011
  - *"Teresa Rampazzi e la musica ben calcolata"*, Conservatorio Statale di Musica "C. Pollini", Padova, Martedì 21 giugno 2011
  - *"IEEE Xplore – Training"*, Dipartimento di Ingegneria dell'Informazione, Padova, Giovedì 27 novembre 2011
  - Boi Faltings, *"Getting agents to tell the truth"*, Martedì 31 gennaio 2012, Dipartimento di Matematica, Padova
  - Dr. Loris Nanni, *Seminari di Ingegneria dell'Informazione "Computer Vision & Machine Learning"*, Martedì 28 febbraio 2012, Dipartimento di Ingegneria dell'Informazione, Padova
  - Dr. Antonio Rodà, *Seminari di Ingegneria dell'Informazione "Sound and Music Computing"*, Martedì 28 febbraio 2012, Dipartimento di Ingegneria dell'Informazione, Padova
  - Giuseppe De Nicolao, *"Toxic numbers? The splendors and miseries of bibliometric indicators"*, Martedì 20 febbraio 2012, Dipartimento di Ingegneria dell'Informazione, Padova
  - Confindustria Padova, *Presentazione di STEPS Seminars Towards Enterprise for Ph.D. Students*, Giovedì 12 aprile 2012, Dipartimento di Ingegneria dell'Informazione, Padova

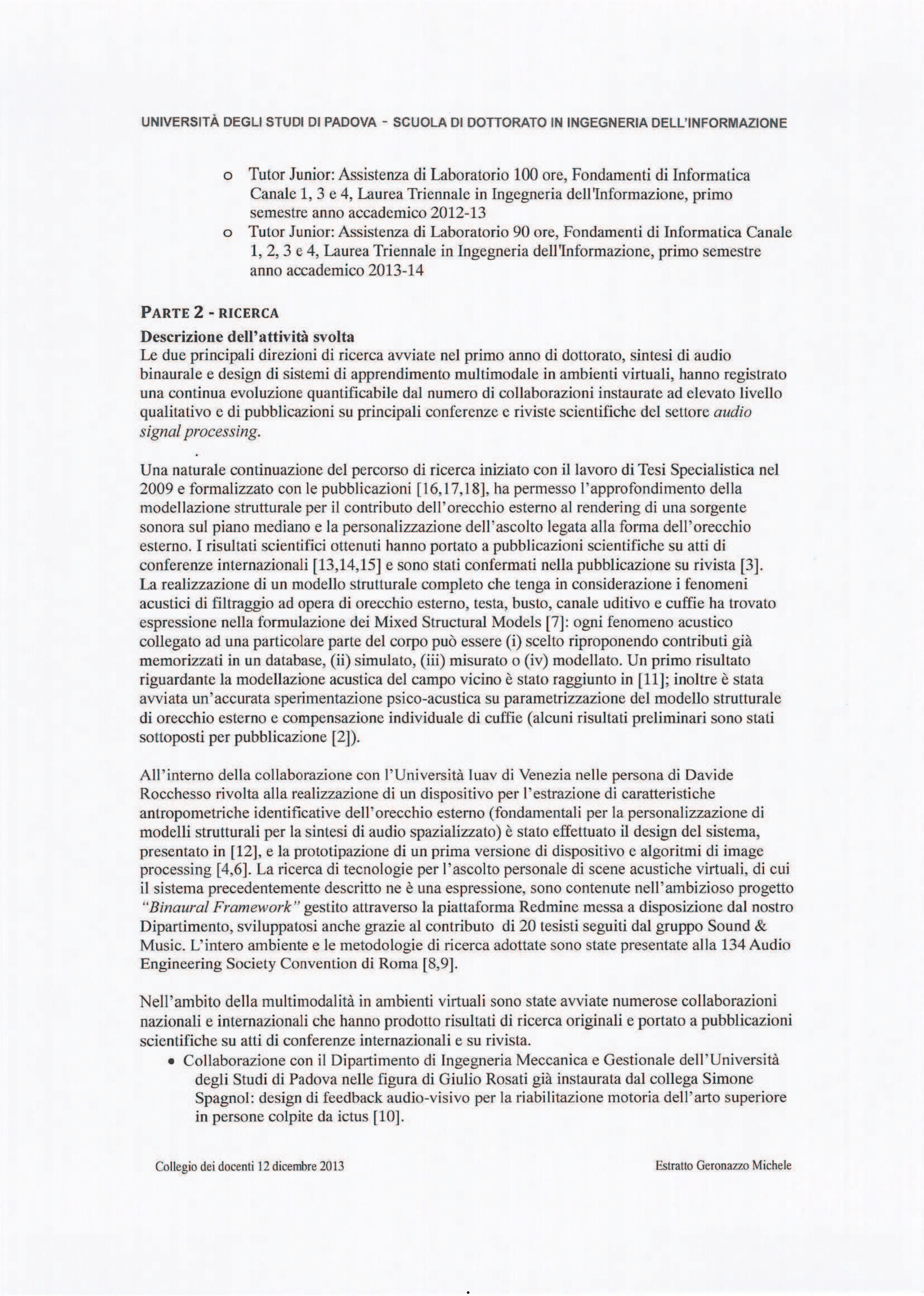


- Alberto Broggi, *"From Italy to China on driverless cars: paving the road to autonomous driving"*, Lunedì 16 aprile 2012, Dipartimento di Ingegneria dell'Informazione, Padova
- EU project DREAM, *Digital Re-Working Re-Appropriation of Electro-Acoustic Music*, Venerdì 15 Giugno 2012, Museo degli Strumenti Musicali, Castello Sforzesco, Milano
- M. De Gasperi, *Coordinare un'attività complessa: project management (corso STEPS)*, Giovedì 28 Giugno 2012, Sede Confindustria Padova
- Veronique Larcher, *"Innovation at Sennheiser. Case studies."*, Domenica 8 Luglio 2012, Aalborg University Copenhagen, Copenhagen
- Takashi Baba, *"Information processing and music in Kwansei Gakuin"*, Martedì 17 Luglio 2012, Dipartimento di Ingegneria dell'Informazione, Padova
- M. Citron, *"La Proprietà Intellettuale"*, Giovedì 5 giugno 2013, Dipartimento di Ingegneria dell'Informazione, Padova
- N. Komeilipoor *"The Sound of Action"*, Lunedì 7 ottobre 2013, Dipartimento di Ingegneria dell'Informazione, Laboratorio di Informatica Musicale, Padova

- Partecipazione a Conferenze Nazionali
  - *Evoluzione dei sistemi di Feedback acustico per la prima domiciliarizzazione dei soggetti con inabilità visiva*, Conferenza Rittmeyer Trieste, giugno 2011,
  - *XIX Colloquium on Musical Informatics (XIX CIM 2012)*, Trieste, November 2012
  - *XI - Workshop Tecnologie per la Musica*, Sapienza Università di Roma Dipartimento DIET & Conservatorio L. Refice di Frosinone, Mercoledì 12 giugno 2013, Roma
  - *Convegno "Audio 3D e Acustica Architettonica"*, Università degli Studi di Bologna e Audio Engineering Society - Italian Section, giovedì 7 novembre 2013.

- Partecipazione a Conferenze Internazionali
  - *Int. Conf. on Sound and Music Computing (SMC 2011)*, Padova, July 6-9, 2011
  - *ACM - Special Interest Group on Computer-Human Interaction CHItaly 2011 Conference*, Alghero, September 2011.
  - *7th International Conference on SIGNAL IMAGE TECHNOLOGY & INTERNET BASED SYSTEMS (SITIS 2011)*, Dijon, November 2011.
  - *Int. Conf. on Sound and Music Computing (SMC 2012)*, Copenhagen, July 11-14, 2012
  - *European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Giovedì 30 Agosto, 2012
  - *Audio Engineering Society Convention 134. AES 134 2013*, Rome, May 2013
  - *Int. Conf. on Sound and Music Computing (SMC 2013)*, Stockholm, July 30-07/03-08-13
  - *Int. Conf. Stockholm Music Acoustics Conference (SMAC 2013)*, Stockholm, July 30-07/03-08-13
  - *10th International Symposium on Computer Music Multidisciplinary Research (CMMR'13)*, Marseille, October 2013
  - *11th International Conference on Advances in Mobile Computing & Multimedia (MoMM'13)*, 2-4 December, 2013

- Didattica attiva (lezioni, esercitazioni, laboratori)



- Tutor Junior: Assistenza di Laboratorio 100 ore, Fondamenti di Informatica Canale 1, 3 e 4, Laurea Triennale in Ingegneria dell'Informazione, primo semestre anno accademico 2012-13
- Tutor Junior: Assistenza di Laboratorio 90 ore, Fondamenti di Informatica Canale 1, 2, 3 e 4, Laurea Triennale in Ingegneria dell'Informazione, primo semestre anno accademico 2013-14

PARTE 2 - RICERCA

Descrizione dell'attività svolta

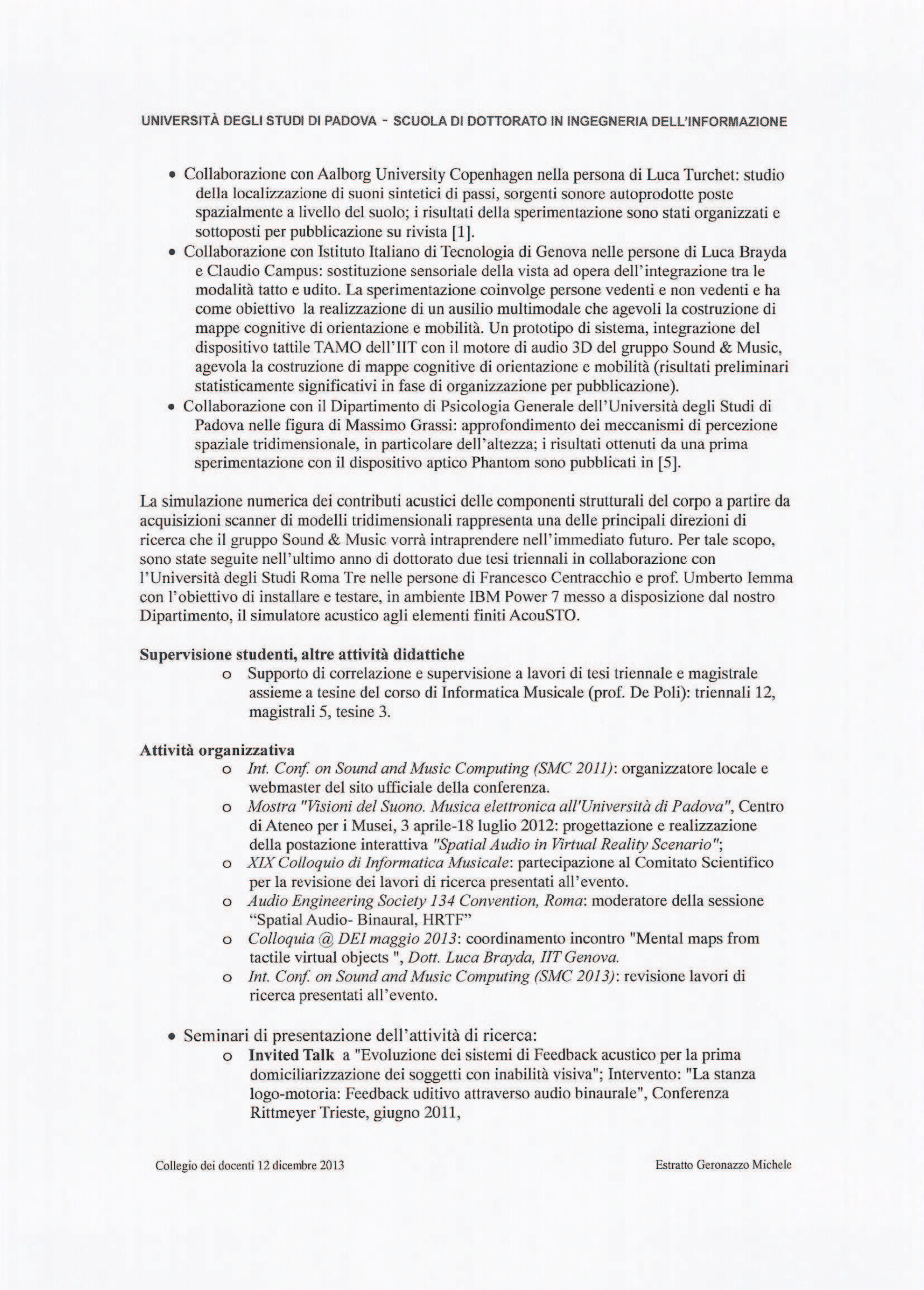
Le due principali direzioni di ricerca avviate nel primo anno di dottorato, sintesi di audio binaurale e design di sistemi di apprendimento multimodale in ambienti virtuali, hanno registrato una continua evoluzione quantificabile dal numero di collaborazioni instaurate ad elevato livello qualitativo e di pubblicazioni su principali conferenze e riviste scientifiche del settore *audio signal processing*.

Una naturale continuazione del percorso di ricerca iniziato con il lavoro di Tesi Specialistica nel 2009 e formalizzato con le pubblicazioni [16,17,18], ha permesso l'approfondimento della modellazione strutturale per il contributo dell'orecchio esterno al rendering di una sorgente sonora sul piano mediano e la personalizzazione dell'ascolto legata alla forma dell'orecchio esterno. I risultati scientifici ottenuti hanno portato a pubblicazioni scientifiche su atti di conferenze internazionali [13,14,15] e sono stati confermati nella pubblicazione su rivista [3]. La realizzazione di un modello strutturale completo che tenga in considerazione i fenomeni acustici di filtraggio ad opera di orecchio esterno, testa, busto, canale uditivo e cuffie ha trovato espressione nella formulazione dei Mixed Structural Models [7]: ogni fenomeno acustico collegato ad una particolare parte del corpo può essere (i) scelto riproponendo contributi già memorizzati in un database, (ii) simulato, (iii) misurato o (iv) modellato. Un primo risultato riguardante la modellazione acustica del campo vicino è stato raggiunto in [11]; inoltre è stata avviata un'accurata sperimentazione psico-acustica su parametrizzazione del modello strutturale di orecchio esterno e compensazione individuale di cuffie (alcuni risultati preliminari sono stati sottoposti per pubblicazione [2]).

All'interno della collaborazione con l'Università Iuav di Venezia nelle persona di Davide Rocchesso rivolta alla realizzazione di un dispositivo per l'estrazione di caratteristiche antropometriche identificative dell'orecchio esterno (fondamentali per la personalizzazione di modelli strutturali per la sintesi di audio spazializzato) è stato effettuato il design del sistema, presentato in [12], e la prototipazione di un prima versione di dispositivo e algoritmi di image processing [4,6]. La ricerca di tecnologie per l'ascolto personale di scene acustiche virtuali, di cui il sistema precedentemente descritto ne è una espressione, sono contenute nell'ambizioso progetto *"Binaural Framework"* gestito attraverso la piattaforma Redmine messa a disposizione dal nostro Dipartimento, sviluppatosi anche grazie al contributo di 20 tesisti seguiti dal gruppo Sound & Music. L'intero ambiente e le metodologie di ricerca adottate sono state presentate alla 134 Audio Engineering Society Convention di Roma [8,9].

Nell'ambito della multimodalità in ambienti virtuali sono state avviate numerose collaborazioni nazionali e internazionali che hanno prodotto risultati di ricerca originali e portato a pubblicazioni scientifiche su atti di conferenze internazionali e su rivista.

- Collaborazione con il Dipartimento di Ingegneria Meccanica e Gestionale dell'Università degli Studi di Padova nelle figura di Giulio Rosati già instaurata dal collega Simone Spagnol: design di feedback audio-visivo per la riabilitazione motoria dell'arto superiore in persone colpite da ictus [10].



- Collaborazione con Aalborg University Copenhagen nella persona di Luca Turchet: studio della localizzazione di suoni sintetici di passi, sorgenti sonore autoprodotte poste spazialmente a livello del suolo; i risultati della sperimentazione sono stati organizzati e sottoposti per pubblicazione su rivista [1].
- Collaborazione con Istituto Italiano di Tecnologia di Genova nelle persone di Luca Brayda e Claudio Campus: sostituzione sensoriale della vista ad opera dell'integrazione tra le modalità tatto e udito. La sperimentazione coinvolge persone vedenti e non vedenti e ha come obiettivo la realizzazione di un ausilio multimodale che agevoli la costruzione di mappe cognitive di orientazione e mobilità. Un prototipo di sistema, integrazione del dispositivo tattile TAMO dell'IIT con il motore di audio 3D del gruppo Sound & Music, agevola la costruzione di mappe cognitive di orientazione e mobilità (risultati preliminari statisticamente significativi in fase di organizzazione per pubblicazione).
- Collaborazione con il Dipartimento di Psicologia Generale dell'Università degli Studi di Padova nelle figura di Massimo Grassi: approfondimento dei meccanismi di percezione spaziale tridimensionale, in particolare dell'altezza; i risultati ottenuti da una prima sperimentazione con il dispositivo aptico Phantom sono pubblicati in [5].

La simulazione numerica dei contributi acustici delle componenti strutturali del corpo a partire da acquisizioni scanner di modelli tridimensionali rappresenta una delle principali direzioni di ricerca che il gruppo Sound & Music vorrà intraprendere nell'immediato futuro. Per tale scopo, sono state seguite nell'ultimo anno di dottorato due tesi triennali in collaborazione con l'Università degli Studi Roma Tre nelle persone di Francesco Centracchio e prof. Umberto Iemma con l'obiettivo di installare e testare, in ambiente IBM Power 7 messo a disposizione dal nostro Dipartimento, il simulatore acustico agli elementi finiti AcouSTO.

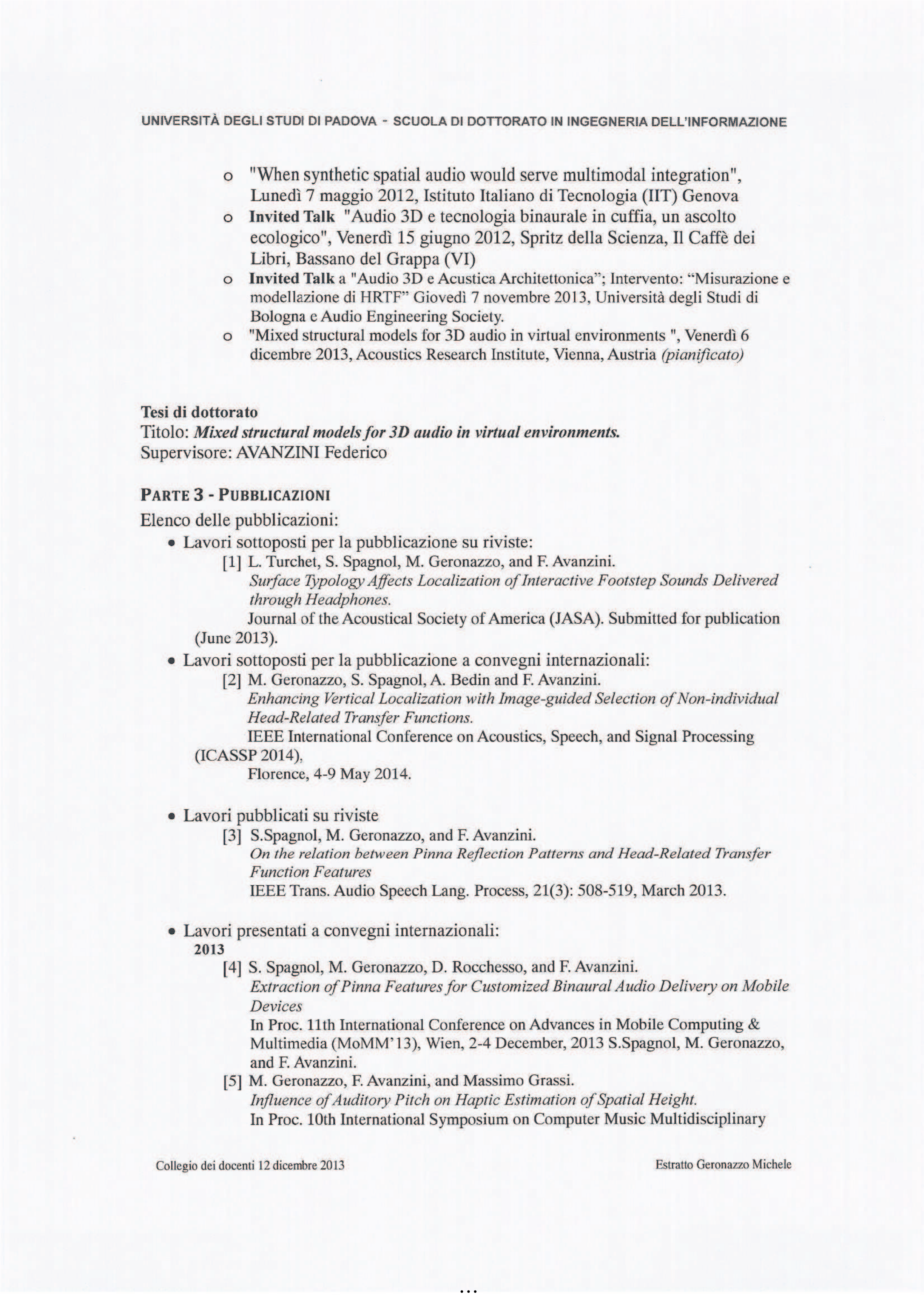
Supervisione studenti, altre attività didattiche

- o Supporto di correlazione e supervisione a lavori di tesi triennale e magistrale assieme a tesine del corso di Informatica Musicale (prof. De Poli): triennali 12, magistrali 5, tesine 3.

Attività organizzativa

- o *Int. Conf. on Sound and Music Computing (SMC 2011)*: organizzatore locale e webmaster del sito ufficiale della conferenza.
- o *Mostra "Visioni del Suono. Musica elettronica all'Università di Padova"*, Centro di Ateneo per i Musei, 3 aprile-18 luglio 2012: progettazione e realizzazione della postazione interattiva *"Spatial Audio in Virtual Reality Scenario"*;
- o *XIX Colloquio di Informatica Musicale*: partecipazione al Comitato Scientifico per la revisione dei lavori di ricerca presentati all'evento.
- o *Audio Engineering Society 134 Convention, Roma*: moderatore della sessione "Spatial Audio- Binaural, HRTF"
- o *Colloquia @ DEI maggio 2013*: coordinamento incontro "Mental maps from tactile virtual objects ", *Dott. Luca Brayda, IIT Genova.*
- o *Int. Conf. on Sound and Music Computing (SMC 2013)*: revisione lavori di ricerca presentati all'evento.

- Seminari di presentazione dell'attività di ricerca:
  - o **Invited Talk** a "Evoluzione dei sistemi di Feedback acustico per la prima domiciliarizzazione dei soggetti con inabilità visiva"; Intervento: "La stanza logo-motoria: Feedback uditivo attraverso audio binaurale", Conferenza Rittmeyer Trieste, giugno 2011,



- o "When synthetic spatial audio would serve multimodal integration", Lunedì 7 maggio 2012, Istituto Italiano di Tecnologia (IIT) Genova
- o **Invited Talk** "Audio 3D e tecnologia binaurale in cuffia, un ascolto ecologico", Venerdì 15 giugno 2012, Spritz della Scienza, Il Caffè dei Libri, Bassano del Grappa (VI)
- o **Invited Talk** a "Audio 3D e Acustica Architettonica"; Intervento: "Misurazione e modellazione di HRTF" Giovedì 7 novembre 2013, Università degli Studi di Bologna e Audio Engineering Society.
- o "Mixed structural models for 3D audio in virtual environments ", Venerdì 6 dicembre 2013, Acoustics Research Institute, Vienna, Austria *(pianificato)*

Tesi di dottorato
Titolo: ***Mixed structural models for 3D audio in virtual environments.***
Supervisore: AVANZINI Federico

PARTE 3 - PUBBLICAZIONI

Elenco delle pubblicazioni:

- Lavori sottoposti per la pubblicazione su riviste:

  [1] L. Turchet, S. Spagnol, M. Geronazzo, and F. Avanzini.
  Surface Typology Affects Localization of Interactive Footstep Sounds Delivered through Headphones.
  Journal of the Acoustical Society of America (JASA). Submitted for publication (June 2013).

- Lavori sottoposti per la pubblicazione a convegni internazionali:

  [2] M. Geronazzo, S. Spagnol, A. Bedin and F. Avanzini.
  Enhancing Vertical Localization with Image-guided Selection of Non-individual Head-Related Transfer Functions.
  IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014),
  Florence, 4-9 May 2014.

- Lavori pubblicati su riviste

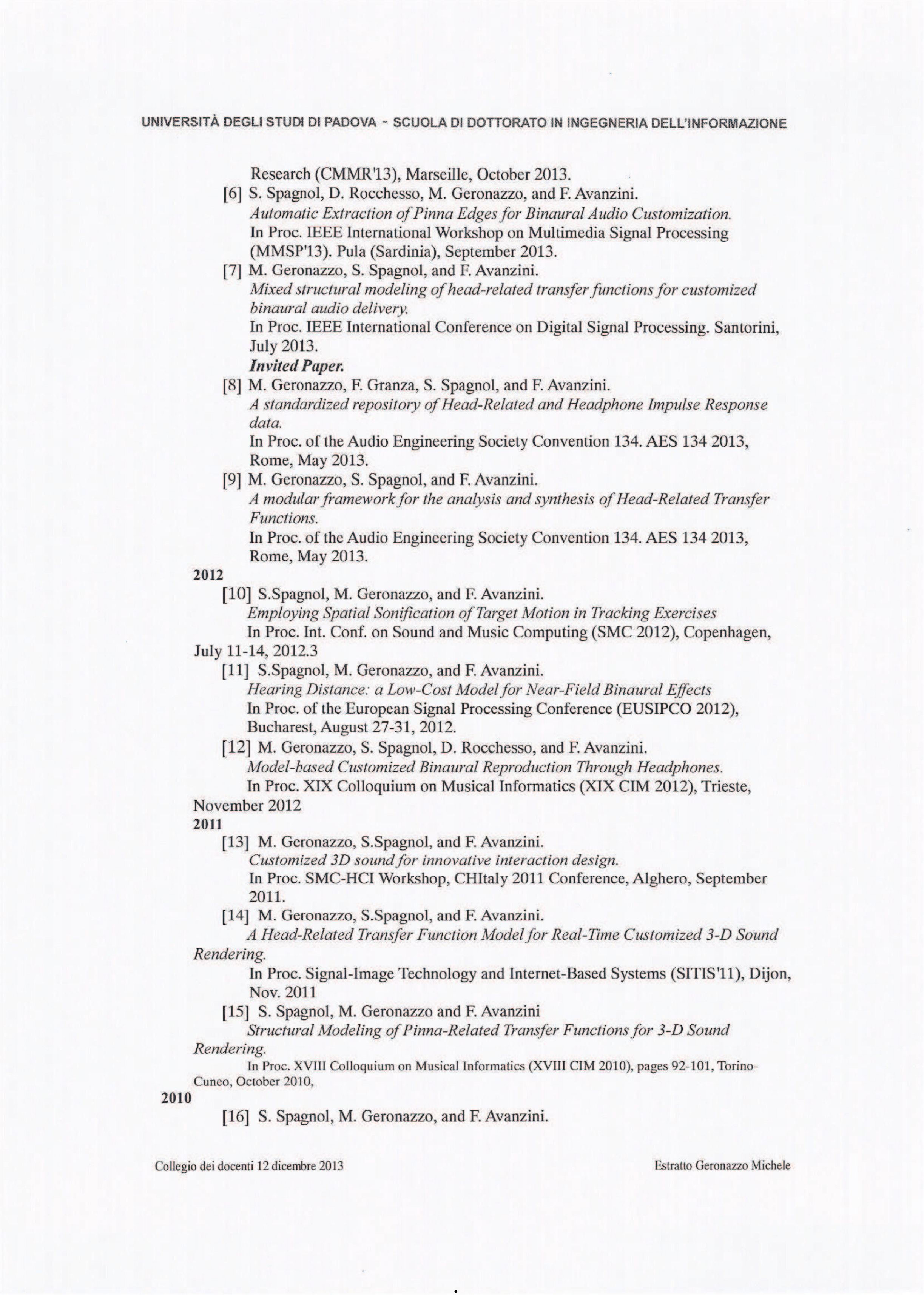
  [3] S.Spagnol, M. Geronazzo, and F. Avanzini.
  On the relation between Pinna Reflection Patterns and Head-Related Transfer Function Features
  IEEE Trans. Audio Speech Lang. Process, 21(3): 508-519, March 2013.

- Lavori presentati a convegni internazionali:

  2013

  [4] S. Spagnol, M. Geronazzo, D. Rocchesso, and F. Avanzini.
  Extraction of Pinna Features for Customized Binaural Audio Delivery on Mobile Devices
  In Proc. 11th International Conference on Advances in Mobile Computing & Multimedia (MoMM'13), Wien, 2-4 December, 2013 S.Spagnol, M. Geronazzo, and F. Avanzini.

  [5] M. Geronazzo, F. Avanzini, and Massimo Grassi.
  Influence of Auditory Pitch on Haptic Estimation of Spatial Height.
  In Proc. 10th International Symposium on Computer Music Multidisciplinary



Research (CMMR'13), Marseille, October 2013.

[6] S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini.
Automatic Extraction of Pinna Edges for Binaural Audio Customization.
In Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP'13). Pula (Sardinia), September 2013.

[7] M. Geronazzo, S. Spagnol, and F. Avanzini.
Mixed structural modeling of head-related transfer functions for customized binaural audio delivery.
In Proc. IEEE International Conference on Digital Signal Processing. Santorini, July 2013.
Invited Paper.

[8] M. Geronazzo, F. Granza, S. Spagnol, and F. Avanzini.
A standardized repository of Head-Related and Headphone Impulse Response data.
In Proc. of the Audio Engineering Society Convention 134. AES 134 2013, Rome, May 2013.

[9] M. Geronazzo, S. Spagnol, and F. Avanzini.
A modular framework for the analysis and synthesis of Head-Related Transfer Functions.
In Proc. of the Audio Engineering Society Convention 134. AES 134 2013, Rome, May 2013.

2012

[10] S.Spagnol, M. Geronazzo, and F. Avanzini.
Employing Spatial Sonification of Target Motion in Tracking Exercises
In Proc. Int. Conf. on Sound and Music Computing (SMC 2012), Copenhagen, July 11-14, 2012.3

[11] S.Spagnol, M. Geronazzo, and F. Avanzini.
Hearing Distance: a Low-Cost Model for Near-Field Binaural Effects
In Proc. of the European Signal Processing Conference (EUSIPCO 2012), Bucharest, August 27-31, 2012.

[12] M. Geronazzo, S. Spagnol, D. Rocchesso, and F. Avanzini.
Model-based Customized Binaural Reproduction Through Headphones.
In Proc. XIX Colloquium on Musical Informatics (XIX CIM 2012), Trieste, November 2012

2011

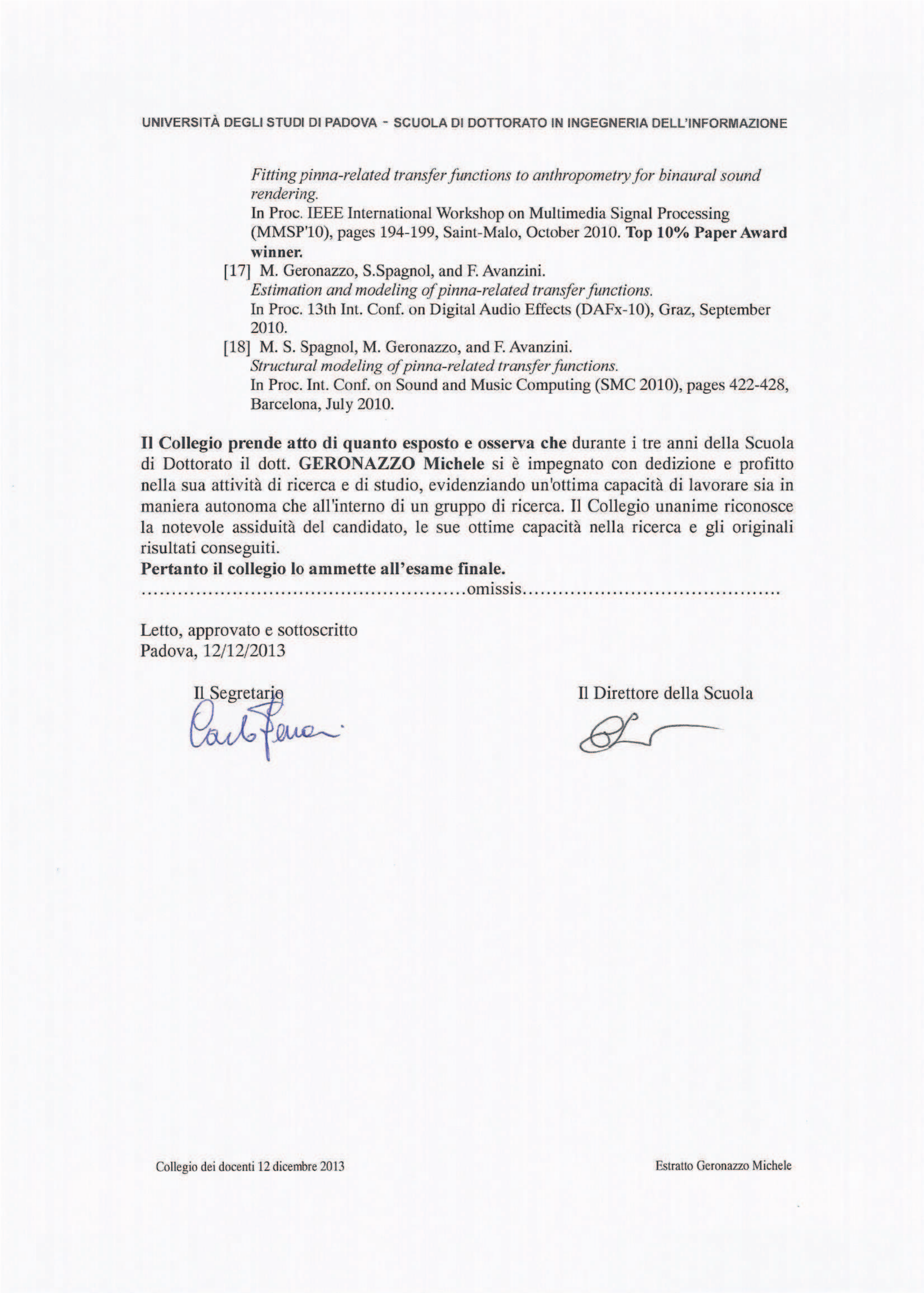
[13] M. Geronazzo, S.Spagnol, and F. Avanzini.
Customized 3D sound for innovative interaction design.
In Proc. SMC-HCI Workshop, CHItaly 2011 Conference, Alghero, September 2011.

[14] M. Geronazzo, S.Spagnol, and F. Avanzini.
A Head-Related Transfer Function Model for Real-Time Customized 3-D Sound Rendering.
In Proc. Signal-Image Technology and Internet-Based Systems (SITIS'11), Dijon, Nov. 2011

[15] S. Spagnol, M. Geronazzo and F. Avanzini
Structural Modeling of Pinna-Related Transfer Functions for 3-D Sound Rendering.
In Proc. XVIII Colloquium on Musical Informatics (XVIII CIM 2010), pages 92-101, Torino-Cuneo, October 2010,

2010

[16] 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.**

[17] 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.

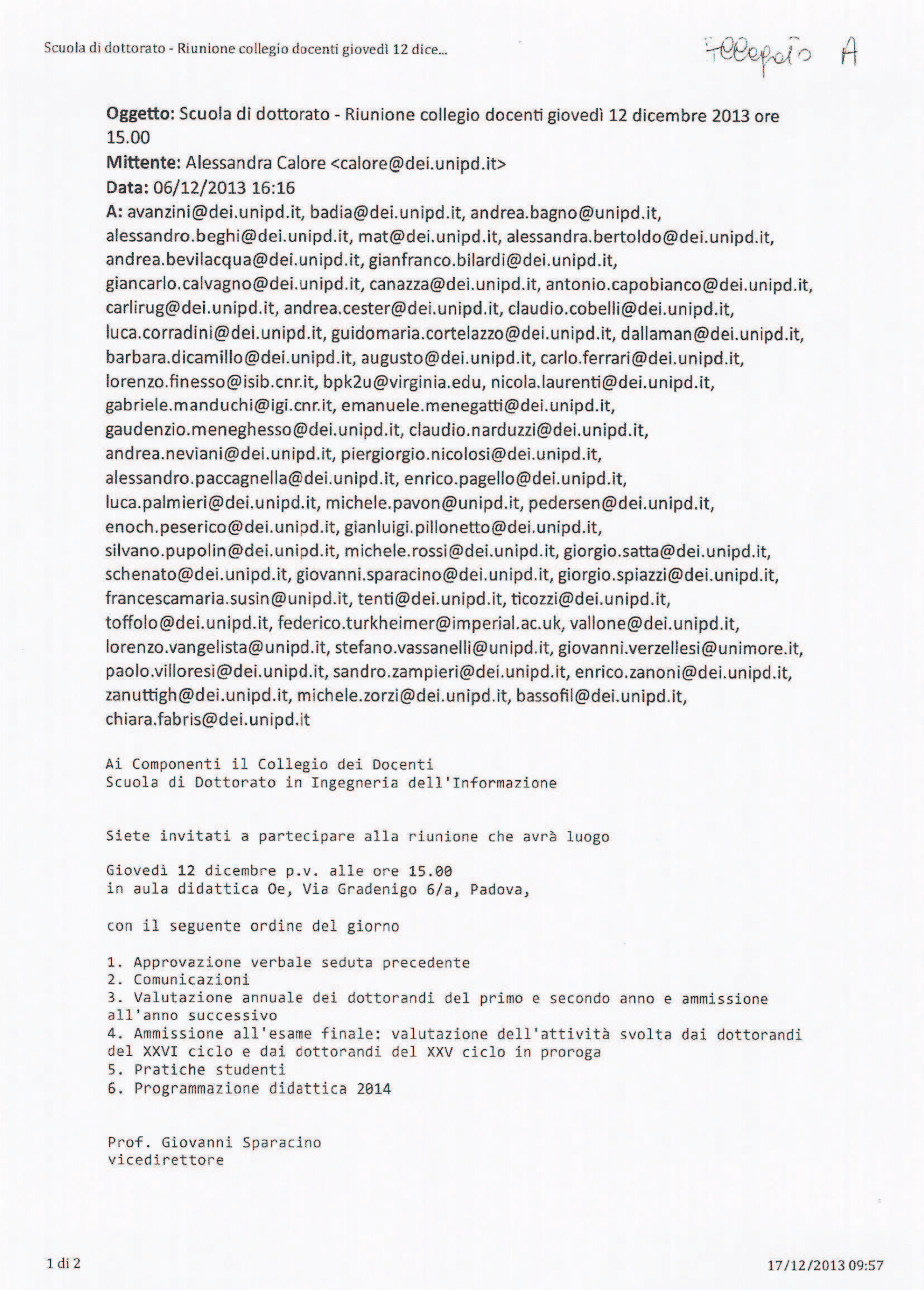
[18] M. S. Spagnol, M. Geronazzo, and F. Avanzini.
Structural modeling of pinna-related transfer functions.
In Proc. Int. Conf. on Sound and Music Computing (SMC 2010), pages 422-428, Barcelona, July 2010.

Il Collegio prende atto di quanto esposto e osserva che durante i tre anni della Scuola di Dottorato il dott. **GERONAZZO Michele** si è impegnato con dedizione e profitto nella sua attività di ricerca e di studio, evidenziando un'ottima capacità di lavorare sia in maniera autonoma che all'interno di un gruppo di ricerca. Il Collegio unanime riconosce la notevole assiduità del candidato, le sue ottime capacità nella ricerca e gli originali risultati conseguiti.
Pertanto il collegio lo ammette all'esame finale.
....omissis.......................................

Letto, approvato e sottoscritto
Padova, 12/12/2013

Il Segretario
Carlo Ferrari

Il Direttore della Scuola



Scuola di dottorato - Riunione collegio docenti giovedì 12 dice...

Allegato A

Oggetto: Scuola di dottorato - Riunione collegio docenti giovedì 12 dicembre 2013 ore 15.00
Mittente: Alessandra Calore <calore@dei.unipd.it>
Data: 06/12/2013 16:16
A: avanzini@dei.unipd.it, badia@dei.unipd.it, andrea.bagno@unipd.it, alessandro.beghi@dei.unipd.it, mat@dei.unipd.it, alessandra.bertoldo@dei.unipd.it, andrea.bevilacqua@dei.unipd.it, gianfranco.bilardi@dei.unipd.it, giancarlo.calvagno@dei.unipd.it, canazza@dei.unipd.it, antonio.capobianco@dei.unipd.it, carlirug@dei.unipd.it, andrea.cester@dei.unipd.it, claudio.cobelli@dei.unipd.it, luca.corradini@dei.unipd.it, guidomaria.cortelazzo@dei.unipd.it, dallaman@dei.unipd.it, barbara.dicamillo@dei.unipd.it, augusto@dei.unipd.it, carlo.ferrari@dei.unipd.it, lorenzo.finesso@isib.cnr.it, bpk2u@virginia.edu, nicola.laurenti@dei.unipd.it, gabriele.manduchi@igi.cnr.it, emanuele.menegatti@dei.unipd.it, gaudenzio.meneghesso@dei.unipd.it, claudio.narduzzi@dei.unipd.it, andrea.neviani@dei.unipd.it, piergiorgio.nicolosi@dei.unipd.it, alessandro.paccagnella@dei.unipd.it, enrico.pagello@dei.unipd.it, luca.palmieri@dei.unipd.it, michele.pavon@unipd.it, pedersen@dei.unipd.it, enoch.peserico@dei.unipd.it, gianluigi.pillonetto@dei.unipd.it, silvano.pupolin@dei.unipd.it, michele.rossi@dei.unipd.it, giorgio.satta@dei.unipd.it, schenato@dei.unipd.it, giovanni.sparacino@dei.unipd.it, giorgio.spiazzi@dei.unipd.it, francescamaria.susin@unipd.it, tenti@dei.unipd.it, ticozzi@dei.unipd.it, toffolo@dei.unipd.it, federico.turkheimer@imperial.ac.uk, vallone@dei.unipd.it, lorenzo.vangelista@unipd.it, stefano.vassanelli@unipd.it, giovanni.verzellesi@unimore.it, paolo.villoresi@dei.unipd.it, sandro.zampieri@dei.unipd.it, enrico.zanoni@dei.unipd.it, zanuttigh@dei.unipd.it, michele.zorzi@dei.unipd.it, bassofil@dei.unipd.it, chiara.fabris@dei.unipd.it

Ai Componenti il Collegio dei Docenti
Scuola di Dottorato in Ingegneria dell'Informazione

Siete invitati a partecipare alla riunione che avrà luogo

Giovedì 12 dicembre p.v. alle ore 15.00
in aula didattica Oe, Via Gradenigo 6/a, Padova,

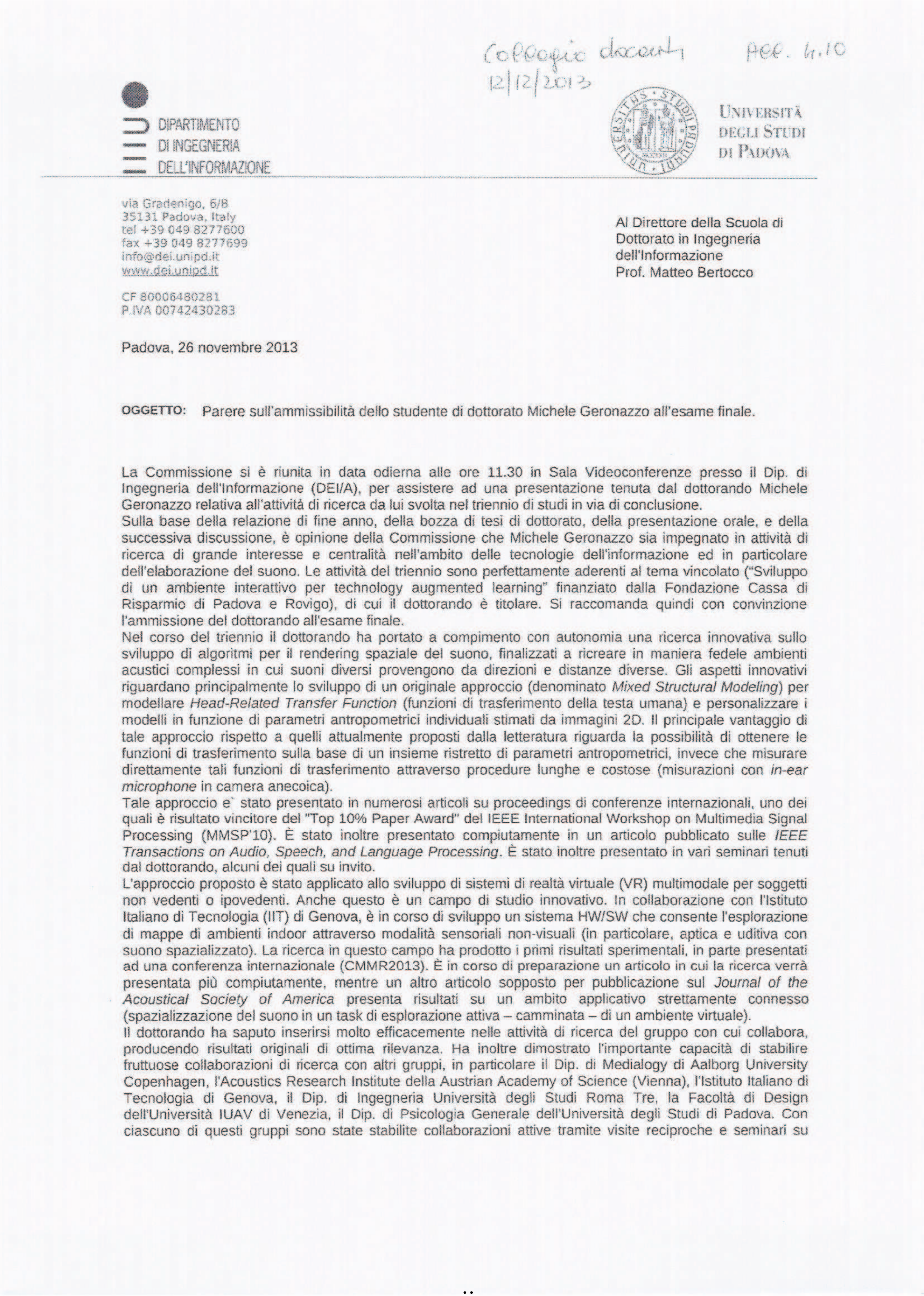
con il seguente ordine del giorno

1. Approvazione verbale seduta precedente
2. Comunicazioni
3. Valutazione annuale dei dottorandi del primo e secondo anno e ammissione all'anno successivo
4. Ammissione all'esame finale: valutazione dell'attività svolta dai dottorandi del XXVI ciclo e dai dottorandi del XXV ciclo in proroga
5. Pratiche studenti
6. Programmazione didattica 2014

Prof. Giovanni Sparacino
vicedirettore

1 di 2

17/12/2013 09:57



Collegio docenti
12/12/2013

All. 4.10

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

via Gradenigo, 6/B
35131 Padova, Italy
tel +39 049 8277600
fax +39 049 8277699
info@dei.unipd.it
www.dei.unipd.it

CF 80006480281
P.IVA 00742430283

Al Direttore della Scuola di
Dottorato in Ingegneria
dell'Informazione
Prof. Matteo Bertocco

Padova, 26 novembre 2013

OGGETTO: Parere sull'ammissibilità dello studente di dottorato Michele Geronazzo all'esame finale.

La Commissione si è riunita in data odierna alle ore 11.30 in Sala Videoconferenze presso il Dip. di Ingegneria dell'Informazione (DEI/A), per assistere ad una presentazione tenuta dal dottorando Michele Geronazzo relativa all'attività di ricerca da lui svolta nel triennio di studi in via di conclusione.

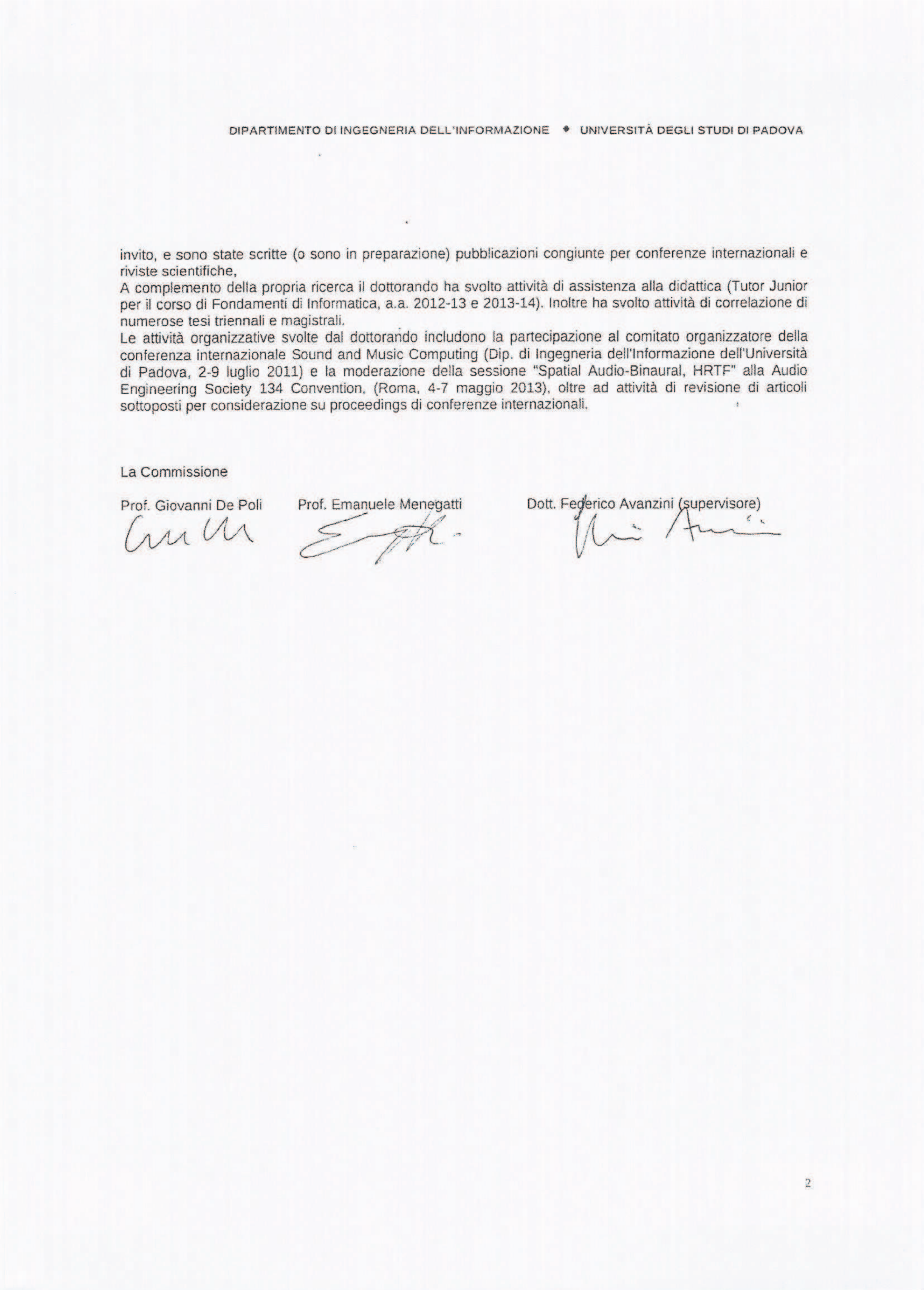
Sulla base della relazione di fine anno, della bozza di tesi di dottorato, della presentazione orale, e della successiva discussione, è opinione della Commissione che Michele Geronazzo sia impegnato in attività di ricerca di grande interesse e centralità nell'ambito delle tecnologie dell'informazione ed in particolare dell'elaborazione del suono. Le attività del triennio sono perfettamente aderenti al tema vincolato ("Sviluppo di un ambiente interattivo per technology augmented learning" finanziato dalla Fondazione Cassa di Risparmio di Padova e Rovigo), di cui il dottorando è titolare. Si raccomanda quindi con convinzione l'ammissione del dottorando all'esame finale.

Nel corso del triennio il dottorando ha portato a compimento con autonomia una ricerca innovativa sullo sviluppo di algoritmi per il rendering spaziale del suono, finalizzati a ricreare in maniera fedele ambienti acustici complessi in cui suoni diversi provengono da direzioni e distanze diverse. Gli aspetti innovativi riguardano principalmente lo sviluppo di un originale approccio (denominato *Mixed Structural Modeling*) per modellare *Head-Related Transfer Function* (funzioni di trasferimento della testa umana) e personalizzare i modelli in funzione di parametri antropometrici individuali stimati da immagini 2D. Il principale vantaggio di tale approccio rispetto a quelli attualmente proposti dalla letteratura riguarda la possibilità di ottenere le funzioni di trasferimento sulla base di un insieme ristretto di parametri antropometrici, invece che misurare direttamente tali funzioni di trasferimento attraverso procedure lunghe e costose (misurazioni con *in-ear microphone* in camera anecoica).

Tale approccio e` stato presentato in numerosi articoli su proceedings di conferenze internazionali, uno dei quali è risultato vincitore del "Top 10% Paper Award" del IEEE International Workshop on Multimedia Signal Processing (MMSP'10). È stato inoltre presentato compiutamente in un articolo pubblicato sulle *IEEE Transactions on Audio, Speech, and Language Processing*. È stato inoltre presentato in vari seminari tenuti dal dottorando, alcuni dei quali su invito.

L'approccio proposto è stato applicato allo sviluppo di sistemi di realtà virtuale (VR) multimodale per soggetti non vedenti o ipovedenti. Anche questo è un campo di studio innovativo. In collaborazione con l'Istituto Italiano di Tecnologia (IIT) di Genova, è in corso di sviluppo un sistema HW/SW che consente l'esplorazione di mappe di ambienti indoor attraverso modalità sensoriali non-visuali (in particolare, aptica e uditiva con suono spazializzato). La ricerca in questo campo ha prodotto i primi risultati sperimentali, in parte presentati ad una conferenza internazionale (CMMR2013). È in corso di preparazione un articolo in cui la ricerca verrà presentata più compiutamente, mentre un altro articolo sopposto per pubblicazione sul *Journal of the Acoustical Society of America* presenta risultati su un ambito applicativo strettamente connesso (spazializzazione del suono in un task di esplorazione attiva – camminata – di un ambiente virtuale).

Il dottorando ha saputo inserirsi molto efficacemente nelle attività di ricerca del gruppo con cui collabora, producendo risultati originali di ottima rilevanza. Ha inoltre dimostrato l'importante capacità di stabilire fruttuose collaborazioni di ricerca con altri gruppi, in particolare il Dip. di Medialogy di Aalborg University Copenhagen, l'Acoustics Research Institute della Austrian Academy of Science (Vienna), l'Istituto Italiano di Tecnologia di Genova, il Dip. di Ingegneria Università degli Studi Roma Tre, la Facoltà di Design dell'Università IUAV di Venezia, il Dip. di Psicologia Generale dell'Università degli Studi di Padova. Con ciascuno di questi gruppi sono state stabilite collaborazioni attive tramite visite reciproche e seminari su



invito, e sono state scritte (o sono in preparazione) pubblicazioni congiunte per conferenze internazionali e riviste scientifiche,

A complemento della propria ricerca il dottorando ha svolto attività di assistenza alla didattica (Tutor Junior per il corso di Fondamenti di Informatica, a.a. 2012-13 e 2013-14). Inoltre ha svolto attività di correlazione di numerose tesi triennali e magistrali.

Le attività organizzative svolte dal dottorando includono la partecipazione al comitato organizzatore della conferenza internazionale Sound and Music Computing (Dip. di Ingegneria dell'Informazione dell'Università di Padova, 2-9 luglio 2011) e la moderazione della sessione "Spatial Audio-Binaural, HRTF" alla Audio Engineering Society 134 Convention, (Roma, 4-7 maggio 2013), oltre ad attività di revisione di articoli sottoposti per considerazione su proceedings di conferenze internazionali.

La Commissione

Prof. Giovanni De Poli Prof. Emanuele Menegatti Dott. Federico Avanzini (supervisore)

Prefazione

Il settore dell' *Information and Communications Technology*(ICT) sta investendo in strategie di innovazione e sviluppo sempre piú rivolte ad applicazioni capaci di interazione complesse grazie alla rappresentazione spaziale in ambienti virtuali multimodali capaci di rispettare i vincoli di tempo reale. Una delle principali sfide da affrontare riguarda la centralitá dell'utente, che si riflette, ad esempio, sullo sviluppo di servizi la cui complessitá tecnologica viene nascosta al destinatario, e la cui offerta di servizi sia personalizzabile dallutente e per lutente. Per queste ragioni , le interfacce multimodali rappresentano un elemento chiave per consentire un uso diffuso di queste nuove tecnologie. Per raggiungere questo obiettivo necessario ottenere dei modelli multimodali realistici che siano capaci di descrivere lambiente circostante, e in particolare modelli che sappiano rappresentare accuratamente l'acustica dell'ambiente e la trasmissione di informazione attraverso la modalitá uditiva. Alcuni esempi di aree applicative e direzioni di ricerca attive nella comunitá scientifica internazionale includono 3DTV e internet del futuro , codifica, trasmissione e ricostruzione della scena 3D video e audio e sistemi di teleconferenza , per citarne solo alcuni.

La presenza concomitante di piú modalitá sensoriali e la loro integrazione rendono gli ambienti virtuali multimodali potenzialmente flessibili e adattabili, permettendo agli utenti di passare dalluna allaltra modalitá in base alle necessitá dettata dalle mutevoli condizioni di utilizzo di tali sistemi. Modalitá sensoriali aumentata attraverso altri sensi e tecniche di sostituzione sensoriale sono elementi essenziali per la veicolazione dellinformazioni non visivamente, quando, ad esempio, il canale visivo sovraccaricato, quando i dati sono visivamente ostruiti, o quando il canale visivo non disponibile per l'utente (ad esempio, per le persone non vedenti). I sistemi multimodali per la rappresentazione delle informazioni spaziali beneficano sicuramente della realizzazione di motori audio che possiedano una conoscenza approfondita degli aspetti legati alla percezione spaziale e allacustica virtuale. I modelli per il rendering di audio spazializzato sono in grado di fornire accurate informazioni dinamiche sulla relazione tra la sorgente sonora e l'ambiente circostante , compresa l'interazione del corpo dellascoltatore che agisce da ulteriore filtraggio acustico. Queste informazioni non possono essere sostituite da altre modalitá (ad esempio quella visiva o tattile). Tuttavia , la rappresentazione spaziale del suono nei feedback acustici tende ad essere, al giorno doggi, semplicistica e con scarse capacitá di interazione, questo perché i sistemi multimediali attualmente si focalizzano per lo piú sullelaborazione grafica, e si accontentano di semplici tecnologie stereofoniche o surround multicanale per il rendering del suono.

Il rendering binaurale riprodotto in cuffia rappresenta un approccio avveniristico, tenendo conto che i possibili svantaggi (es. invasivitá , risposte in frequenza non piane) possono essere man mano gestiti e controbilanciati da una serie di desiderabili caratteristiche. Questi sistemi sono caratterizzati dalla possibilitá di controllare e/o eliminare il riverbero e altri effetti acustici dello spazio di ascolto circostante, di ridurre il rumore di fondo e fornire dei display audio adattabili e portatili, tutti aspetti rilevanti soprattutto in contesti di innovazione.

La maggior parte delle tecniche di rendering binaurale impiegate oggigiorno in ricerca si basano sull'uso di *Head Related Transfer Functions* (HRTFs), vale a dire di filtri particolari che catturano gli effetti acustici di testa, busto e orecchie dellascoltatore. Le HRTF permettono una simulazione fedele del segnale audio che si presenta all'ingresso del canale uditivo in funzione della posizione spaziale della sorgente sonora. I filtri basati su HRTF sono generalmente presentati sotto forma di segnali acustici misurati a partire da una testa di manichino costruito secondo misurazioni antropometriche medie. Tuttavia, le caratteristiche antropometriche individuali hanno un ruolo fondamentale nel determinare le HRTF: diversi studi hanno riscontrato come lascolto di audio binaurale non individuale produce errori di localizzazione evidenti . D'altra parte , le misurazioni individuali di HRTF su un numero significativo di soggetti richiedono un impiego di risorse e tempo non trascurabili.

Sono state proposte negli ultimi due decenni diverse tecniche per il design di HRTF sintetiche e tra le piú promettente vi quella che utilizza i modelli strutturali di HRTF. In questo approccio rivoluzionario, gli effetti piú importanti coinvolti nella percezione spaziale del suono (i ritardi acustici e le ombre acustiche ad opera della diffrazione attorno alla testa, le riflessioni sui contorni dellorecchio esterno e sulle spalle, le risonanze all'interno delle cavitá dellorecchio) sono isolati e modellati separatamente nellelemento filtrante corrispondente. La selezione di HRTF non individuali e queste procedure di modellazione possono essere entrambe analizzate con una interpretazione fisica: i parametri di ogni blocco di rendering o i criteri di selezione possono venir stimati dalla relazione tra dati reali e simulati e antropometria dellascoltatore. La realizzazione di efficaci display uditivi personali rappresenta un notevole passo in avanti per numerose applicazioni; lapproccio strutturale consente una intrinseca scalabilitá a seconda delle risorse computazionali o della larghezza di banda disponibili. Scene altamente realistiche con piú oggetti audiovisivi riescono ad essere gestite sfruttando il parallelismo della Graphics Processing Unit (GPU) sempre piú onnipresenti. Ottenere un equalizzazione individuale delle cuffie con tecniche di filtraggio inverso che siano percettivamente robuste costituisce un passo fondamentale verso la creazione di *display uditivi virtuali* personali. A titolo desempio, vengono di seguito riportate alcune aree applicative che possono trarre beneficio da queste considerazioni: riproduzione multi canale in cuffia, rendering spaziale del suono in dispositivi mobile, motori di rendering per computer-game e standard audio binaurali individuali per film e produzione musicale.

Questa tesi presenta una famiglia di approcci in grado di superare gli attuali limiti dei sistemi di audio 3D in cuffia, con lobiettivo di realizzare display uditivi personali attraverso modelli strutturali per laudio binaurale volti ad una riproduzione immersiva del suono. I modelli che ne derivano permettono adattamento e personalizzazione di contenuti, grazie alla gestione dei parametri relativi allantropometria dell'utente oltre a quelli relativi alle sorgenti sonore nell'ambiente .

Le direzioni di ricerca intraprese convergono verso una metodologia per la progettazione e personalizzazione di HRTF sintetiche che unisce il paradigma di modellazione strutturale con altre tecniche di selezione per HRTF (ispirate a procedure di selezione non-individuali di HRTF) e rappresenta il principale contributo di questa tesi: l approccio a *modellazione strutturale mista*(MSM) che considera la HRTF globale come una combinazione di elementi strutturali, che possono essere scelti tra componenti sia sintetiche che registrate. In entrambi i casi, la personalizzazione si basa su dati antropometrici individuali, utilizzati per adattare sia i parametri del modello sia per selezionare un componente simulato o misurato, tra un insieme di risposte allimpulso disponibili.

La definizione e la validazione sperimentale dell'approccio a MSM affronta alcune questioni cruciali riguarda l'acquisizione e il rendering di scene acustiche binaurali, definendo alcune linee guida di progettazione per ambienti virtuali personali che utilizzano laudio 3D e che possiedono nuove forme di comunicazione su misura e di interazione con contenuti sonori e musicali.

In questa tesi viene anche presentato un sistema interattivo multimodale utilizzato per condurre test soggettivi sullintegrazione multisensoriale in ambienti virtuali. Vengono proposti quattro scenari sperimentali al fine di testare le funzionalitá di un feedback sonoro integrato a modalitá tattili o visive. (i) Un feedback con audio 3D legato ai movimenti dell'utente durante una semplice attivitá di inseguimento di un bersaglio viene presentato come un esempio applicativo di sistema riabilitativo audiovisivo. (ii) La percezione della direzione sonora dei passi interattivamente generati in cuffia durante la camminata evidenzia come l'informazione spaziale sia in grado di mettere in luce la congruenza semantica tra movimento e feedback multimodale. (iii) Un sistema audio tattile interattivo e real-time sintetizza l'informazione spaziale di mappe virtuali per leducazione allorientamento e alla mobilitá (O&M) rivolta a persone non vedenti. (iv) Un ultimo esperimento analizza la stima tattile delle dimensioni di un oggetto virtuale 3D (un gradino), mentre l'esplorazione accompagnata da un feedback sonoro generato in tempo reale i cui parametri variano in funzione dellaltezza del punto di interazione aptico.

I dati raccolti da questi esperimenti suggeriscono che feedback multimodali che sfruttano correttamente modelli di audio 3D, possono essere utilizzati per migliorare la navigazione nella realtá virtuale, lorientamento e lapprendimento di azioni motorie complesse, grazie all'alto livello di attenzione, impegno e immersivitá fornito all'utente. La metodologia di ricerca, basata sull'approccio a MSM, rappresenta un importante strumento di valutazione per determinare progressivamente i principali attributi spaziali del suono in relazione a ciascun dominio applicativo. In questa prospettiva, tali studi rappresentano una novitá nella letteratura scientifica corrente che ha come principale argomento di indagine la realtá virtuale e aumentata, soprattutto per quanto riguarda l'uso di tecniche di sonicazione legate alla cognizione spaziale e alla rappresentazione multisensoriale interna del corpo .

Questa tesi organizzata come segue. Unintroduzione e una panoramica sulla percezione spaziale del suono e sulle tecnologie binaurali in cuffia sono fornite nel **Capitolo 1**. Il **Capitolo 2** dedicato al formalismo sulla modellazione strutturale mista e sua corrispondente filosofia di ricerca. Nel **Capitolo 3** vengono presentati i modelli strutturali relativi ad ogni parte del corpo, risultanti da precedenti ricerche. Due nuove proposte di modello di testa e orecchio approfondiscono rispettivamente la dipendenza dalla distanza nel near-field e le informazioni spettrali fornite dallorecchio esterno per la localizzazione verticale del suono. Il **Capitolo 4** si occupa di un caso di studio completo riguardante l'approccio a modellazione strutturale mista, fornendo degli ap-

profondimenti riguardanti i principali aspetti innovativi di tale *modus operandi*. Il **Capitolo 5** fornisce una panoramica di strumenti sviluppati per l'analisi e la sintesi di HRTF. Inoltre linee guida per il design di ambienti di realtá virtuale vengono discussi in termini di problematiche riguardanti vincoli di tempo reali, requisiti per la mobilitá e personalizzazione del segnale audio. Nel **Capitolo 6**, attraverso due casi di studio viene approfondita l'importanza dell'attributo spaziale del suono nel comportamento dellascoltatore e come la continua interazione in ambienti virtuali possa utilizzare con successo algoritmi per laudio spaziale. Il **Capitolo 7** descrive una serie di esperimenti volti a valutare il contributo dellaudio binaurale in cuffia in processi di apprendimento di mappe cognitive spaziali e nell'esplorazione di oggetti virtuali. Infine, il **Capitolo 8** apre a nuovi orizzonti per futuri lavori di ricerca.

Parole chiave: audio 3D, head-related transfer function, tecnologia binaurale, percezione spaziale del suono, elaborazione digitale del segnale audio, display uditivi, ambienti virtuali, design dellinterazione sonora, multimodalitá, conoscenza spaziale.

Ringraziamenti

Sono passati in fretta questi tre anni di ricerca scientifica. Un periodo della mia vita ricco di cambiamenti e sfide continue. Non ci si costruisce da soli, sono fortunato ad avere incontrato lungo la mia strada le persone giuste.

Un grazie di cuore a Federico Avanzini perché se sono giunto a questo traguardo, molto é dovuto alla sintonia che abbiamo costruito. Professionalitá e umanitá coesistono in un equilibrio che va sempre cercato, desiderato e sostenuto. La tua esperienza, gentilezza ed entusiasmo sono per me motivo di speranza.

Grazie caro collega e caro amico Simone Spagnol per essere un “fratello di ricerca” maggiore, per le sempre pi numerose vittorie a fronte di qualche sconfitta in pi, che male non fa; per aver corretto miei innumerevoli orrori linguistici e per la condivisione di gioie e dolori dei laboratori e del CSC.

Un grazie che attraversa le Alpi va ad Enrico Marchetto, guida saggia ed equilibrata; nei primi anni del mio dottorato abbiamo condiviso un doloroso trasloco di laboratorio e molte pause pranzo di sostegno vicendevole.

Ringrazio il gruppo Sound & Music Computing group del Dipartimento di Ingegneria dell’Informazione dell’Universitá di Padova e in particolare il prof. De Poli, per i suoi saggi consigli e perché la sua passione e dedizione alla disciplina dell’Informatica Musicale hanno permesso che io iniziassi questo percorso di ricerca a Padova.

Un grazie particolare alle donne del SMC International group, Stefania Serafin e Amalia De Götzen, per la loro testimonianza di tenacia e coraggio, nonché per le occasioni di formazione presso la Aalborg University Denmark e presso il Conservatorio Pollini di Padova.

Non posso dimenticare Stefano Delle Monache e Davide Rocchesso dellUniversitá IUAV di Venezia che in occasioni diverse hanno saputo consigliarmi con estro e simpatia, Grazie!

Ringrazio tutti i co-autori dei miei lavori per aver messo a dura prova le mie capacitá di lavorare in gruppo.

Un grazie per avermi sopportato ai numerosi tesisti binaurali che ha seguito in lavori sempre sovradimensionati, ma portati avanti con passione e spirito di iniziativa. In particolare e rigorosamente in ordine alfabetico coloro che hanno condiviso pi attivamente i miei interessi di ricerca: Alberto Bedin, Fabrizio Granza, Lorenzo Monni, Giovanni Schiesaro e Alberto Ziliotto.

Infine una menzione speciale all’Audio Engineering Society Italia per le occasioni di crescita professionale fornitemi, con l’augurio che possa continuare una collaborazione attiva per la crescita di questa realtá.

Preface

In the world of ICT, strategies for innovation and development are increasingly focusing on applications that require spatial representation and real-time interaction with and within 3D media environments. One of the major challenges that such applications have to address is user-centricity, reflecting e.g. on developing complexity-hiding services so that people can personalize their own delivery of services. In these terms, multimodal interfaces represent a key factor for enabling an inclusive use of the new technology by everyone. In order to achieve this, multimodal realistic models that describe our environment are needed, and in particular models that accurately describe the acoustics of the environment and communication through the auditory modality. Examples of currently active research directions and application areas include 3DTV and future internet, 3D visual-sound scene coding, transmission and reconstruction and teleconferencing systems, to name but a few.

The concurrent presence of multimodal senses and activities make multimodal virtual environments potentially flexible and adaptive, allowing users to switch between modalities as needed during the continuously changing conditions of use situation. Augmentation through additional modalities and sensory substitution techniques are compelling ingredients for presenting information non-visually, when the visual bandwidth is overloaded, when data are visually occluded, or when the visual channel is not available to the user (e.g., for visually impaired people). Multimodal systems for the representation of spatial information will largely benefit from the implementation of audio engines that have extensive knowledge of spatial hearing and virtual acoustics. Models for spatial audio can provide accurate dynamic information about the relation between the sound source and the surrounding environment, including the listener and his/her body which acts as an additional filter. Indeed, this information cannot be substituted by any other modality (i.e., visual or tactile). Nevertheless, today's spatial representation of audio within sonification tends to be simplistic and with poor interaction capabilities, being multimedia systems currently focused on graphics processing mostly, and integrated with simple stereo or multi-channel surround-sound.

On a much different level lie binaural rendering approaches based on headphone reproduction, taking into account that possible disadvantages (e.g. invasiveness, non-flat frequency responses) are counterbalanced by a number of desirable features. Indeed, these systems might control and/or eliminate reverberation and other acoustic effects of the real listening space, reduce background noise, and provide adaptable and portable audio displays, which are all relevant aspects especially in enhanced contexts.

Most of the binaural sound rendering techniques currently exploited in research rely on the use of *Head-Related Transfer Functions* (HRTFs), i.e. peculiar filters that capture the acoustic effects of the human head and ears. HRTFs allow loyal simulation of the audio signal that arrives at the entrance of the ear canal as a function of the sound source's spatial position. HRTF filters are usually presented under the form of acoustic signals acquired on dummy heads built according to mean anthropometric measurements. Nevertheless, anthropometric features of the human body have a key role in HRTF shaping: several studies have attested how listening to non-individual binaural sounds results in evident localization errors. On the other hand, individual HRTF measurements on a significant number of subjects result both time- and resource-expensive.

Several techniques for synthetic HRTF design have been proposed during the last two decades and the most promising one relies on structural HRTF models. In this revolutionary approach, the most important effects involved in spatial sound perception (acoustic delays and shadowing due to head diffraction, reflections on pinna contours and shoulders, resonances inside the ear cavities) are isolated and modeled separately with a corresponding filtering element. HRTF selection and modeling procedures can be determined by physical interpretation: parameters of each rendering blocks or selection criteria can be estimated from real and simulated data and related to anthropometric geometries.

Effective personal auditory displays represent an innovative breakthrough for a plethora of applications and structural approach can also allow for effective scalability depending on the available computational resources or bandwidth. Scenes with multiple highly realistic audiovisual objects are easily managed exploiting parallelism of increasingly ubiquitous GPUs (Graphics Processing Units). Building individual headphone equalization with perceptually robust inverse filtering techniques represents a fundamental step towards the creation of personal *virtual auditory displays* (VADs). To this regard, several examples might benefit from these considerations: multi-channel downmix over headphones, personal cinema, spatial audio rendering in mobile devices, computer-game engines and individual binaural audio standards for movie and music production.

This thesis presents a family of approaches that overcome the current limitations of headphone-based 3D audio systems, aiming at building personal auditory displays through structural binaural audio models for an immersive sound reproduction. The resulting models allow for an interesting form of content adaptation and personalization, since they include parameters related to the user's anthropometry in addition to those related to the sound sources and the environment.

The covered research directions converge to a novel framework for synthetic HRTF design and customization that combines the structural modeling paradigm with other HRTF selection techniques (inspired by non-individualized HRTF selection procedures) and represents the main novel contribution of this thesis: the *Mixed Structural Modeling* (MSM) approach considers the global HRTF as a combination of structural components, which can be chosen to be either synthetic or recorded components. In both cases, customization is based on individual anthropometric data, which are used to either fit the model parameters or to select a measured/simulated component within a set of available responses.

The definition and experimental validation of the MSM approach addresses several pivotal issues towards the acquisition and delivery of binaural sound scenes and designing guidelines for personalized 3D audio virtual environments holding the potential of novel forms of customized communication and interaction with sound and music content.

The thesis also presents a multimodal interactive system which is used to conduct subjective test on multi-sensory integration in virtual environments. Four experimental scenarios are proposed in order to test the capabilities of auditory feedback jointly to tactile or visual modalities. 3D audio feedback related to users movements during simple target following tasks is tested as an applicative example of audio-visual rehabilitation system. Perception of direction of footstep sounds interactively generated during walking and provided through headphones highlights how spatial information can clarify the semantic congruence between movement and multimodal feedback. A real time, physically informed audio-tactile interactive system encodes spatial information in the context of virtual map presentation with particular attention to orientation and mobility (O&M) learning processes addressed to visually impaired people. Finally, an experiment analyzes the haptic estimation of size of a virtual 3D object (a stair-step) whereas the exploration is accompanied by a real-time generated auditory feedback whose parameters vary as a function of the height of the interaction point.

The collected data from these experiments suggest that well-designed multimodal feedback, exploiting 3D audio models, can definitely be used to improve performance in virtual reality and learning processes in orientation and complex motor tasks, thanks to the high level of attention, engagement, and presence provided to the user. The research framework, based on the MSM approach, serves as an important evaluation tool with the aim of progressively determining the relevant spatial attributes of sound for each application domain. In this perspective, such studies represent a novelty in the current literature on virtual and augmented reality, especially concerning the use of sonification techniques in several aspects of spatial cognition and internal multisensory representation of the body.

This thesis is organized as follows. An overview of spatial hearing and binaural technology through headphones is given in **Chapter 1**. **Chapter 2** is devoted to the Mixed Structural Modeling formalism and philosophy. In **Chapter 3**, topics in structural modeling for each body component are studied, previous research and two new models, i.e. near-field distance dependency and external-ear spectral cue, are presented. **Chapter 4** deals with a complete case study of the mixed structural modeling approach and provides insights about the main innovative aspects of such *modus operandi*. **Chapter 5** gives an overview of number of a number of proposed tools for the analysis and synthesis of HRTFs. System architectural guidelines and constraints are discussed in terms of real-time issues, mobility requirements and customized audio delivery. In **Chapter 6**, two case studies investigate the behavioral importance of spatial attribute of sound and how continuous interaction with virtual environments can benefit from using spatial audio algorithms. **Chapter 7** describes a set of experiments aimed at assessing the contribution of binaural audio through headphones in learning processes of spatial cognitive maps and exploration of virtual objects. Finally, conclusions are drawn and new research horizons for further work are exposed in **Chapter 8**.

Keywords: 3D audio, head-related transfer function, binaural technology, spatial hearing, audio signal processing, auditory displays, virtual environments, sonic interaction design, multimodality, spatial cognition.

Acknowledgments

These years of scientific researches are quickly gone. This time of my life is characterized by continuous changes and challenges. No one is a self made man and I am very lucky to have met along my path the right people.

I would like to begin by thanking Federico Avanzini, my supervisor; our mutual understanding has allowed to reach this goal. Professional skills and human abilities live together in an always sought for, desired and supported balance. His experience, kindness and enthusiasm give me hope.

It is a pleasure for me to thank dear colleague and friend Simone Spagnol, for being an elder "brother in reaserch", for the more and more victories despite some defeats which did not hurt though. He has often corrected my horrible language style and shared joys and sorrows in the laboratories and in the CSC.

My gratitude goes beyond the Alpes to Enrico Marchetto who was a wise and balanced mentor for me; in my first years of PhD, we shared an unexpected lab moving and several lunch breaks supporting each other.

I would like to thank all the Sound and Music Computing Group at Department of Information Engineering in the University of Padova, particularly Professor De Poli, for his wise advices and his dedication to computer music discipline which encouraged me to undertake this research course in Padova.

I am grateful to some of the Sound and Music Computing International Group ladies, Stefania Serafin and Amalia De Gotzen, for their constancy and courage, furthermore for the training courses at Aalborg University in Denmark e at Pollini conservatory in Padua.

I could not absolutely forget Stefano Delle Monache and Davide Rocchesso of the University IUAV of Venice who have on several occasions advised me with inspiration and congeniality. Thanks, really!

Thanks to all my co-authors for having put a strain my capability of team working. Thanks to all my B.S and M.A thesis students who have always tolerated me with works of considerable dimensions carried out enthusiastically and with personal initiative. I would like to mention (in alphabetic order) those who mostly have shared my research interests: Alberto Bedin, Fabrizio Granza, Lorenzo Monni, Giovanni Schiesaro and Alberto Ziliotto.

Finally I would like express special thanks to Audio Engineering Society Italian Section for the opportunities they gave me for my professional growth. I hope I can continue actively cooperating to improve this scientific organization.

Treviso, 28th January, 2014

Michele

Contents

List of Figures

List of Tables

List of Symbols and Acronyms

Symbol	*Quantity*	*Unit*
BEM	Boundary element method	
BRIR	Binaural room impulse response	
CSD	Coordinate system deviation	
DOA	Direction-of-arrival	
DI	Degree of immersion	
ECTF	Ear canal transfer function	
EPM	Extent of presence metaphor	
EWK	Extent of world knowledge	
FEC	Free-air equivalent coupling	
HATS	Head and torso simulator	
HpIR	Headphone impulse response	
HpTF	Headphone transfer function	
HRIR	Head-related impulse response	
HRTF	Head-related transfer function	
IID	Interaural intensity difference	

ILD	Interaural level difference	
ITD	Interaural time difference	[s]
ITD$_{env}$	Envelope interaural time difference	
ITD$_p$	Interaural phase delay difference	
IPD	Interaural phase difference	[rad]
JND	Just noticeable difference	
KEMAR	Knowles electronics mannikin for acoustic research	
MAA	Minimal audible angle (localizatio blur)	[deg]
MSM	Mixed structural modeling	
O&M	Orientation and mobility	
PDR	Pressure division ratio	
PRTF	Pinna-related transfer function	
pHRIR	Partial head-related impulse response	
pHRTF	Partial head-related transfer function	
RF	Reproduction fidelity	
RIR	Room impulse response	
SPL	Sound pressure level	[dB$_{SPL}$]
TAMO	TActile MOuse	
TOA	Time-of-arrival	[s]
VAD	Virtual auditory displays	

Chapter 1

Binaural Technologies

The ability of the human auditory system to estimate the spatial location of sound sources in acoustic environments 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 the information provided by the visual channel: as an example, it can provide localization information on targets that are out of sight.

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 (Begault, 1994), particularly in cases where visual interface is limited in extension and/or resolution, as in mobile devices (Härmä et al., 2004), or is useless if users are visually-impaired people (Afonso et al., 2010); it improves the sense of presence in augmented/virtual reality systems, and adds engagement to computer games.

According to Morimoto (2002), human subjective evaluation of virtual/real sound environments takes the form of a multiple stage flowchart (see Fig. 1.1) divided into a physical and psychological space.

The sound waves produced by everyday sound sources are subjected to diverse transformations along their path towards the listener's eardrums. The approach that best the acoustic information at the eardrum, involves the use of individual *binaural room impulse responses* (BRIRs). A sounding object radiates an acoustic signal which undergoes temporal and spectral modifications by the environment and the listener body. Environmental properties are contained in the *room impulse response* (RIR) while *head-related impulse response* (HRIR) incorporates listener acoustic contribution. To this regard, BRIRs holds both RIRs and HRIRs properties, being the signature of the room response for a particular sound source and with respect to a the human receiver (Kleiner et al., 1993). Fig. 1.2 offers a different visualization and at the same time groups relevant physical information for this thesis.

As mentioned above, one critical transformation is provided by the listener himself: as a matter of fact, sound waves are influenced by the active role of the listener's body, thanks to which he/she can collect salient information on sound source attributes. Listener's perception of auditory events spans three main groups of perceptual attributes, or *elemental senses* (Morimoto, 2002):

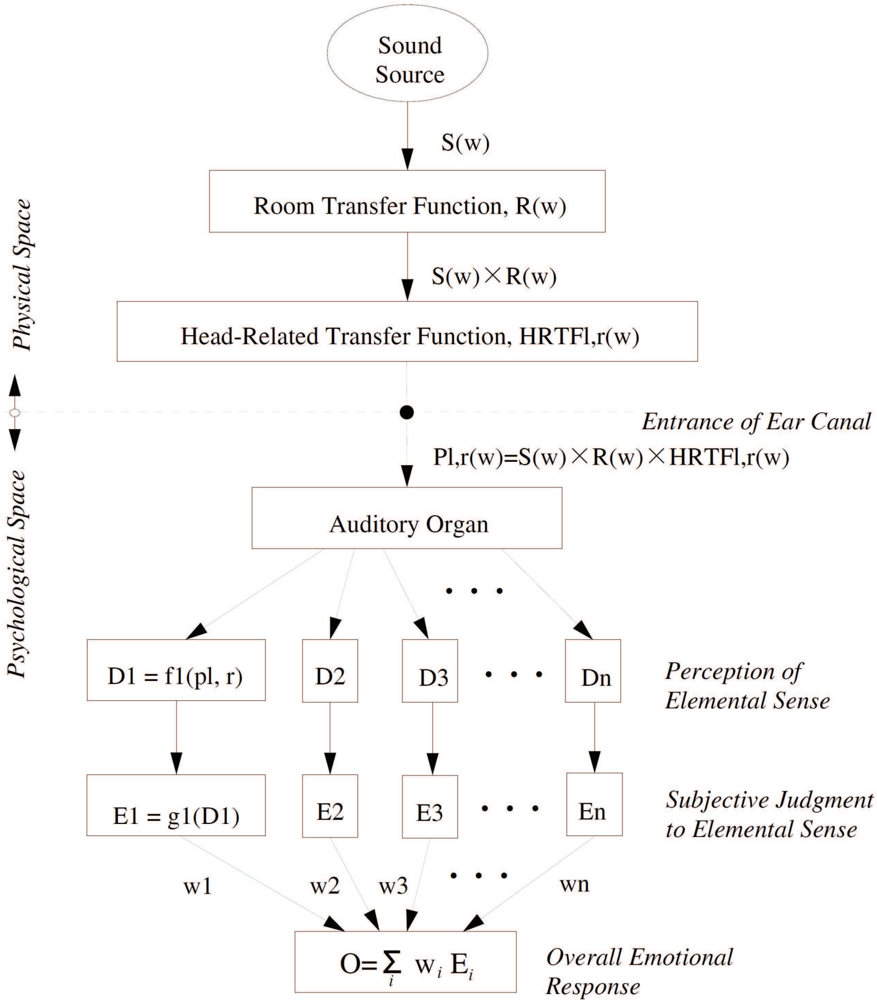

Sound Source
S(w)
Room Transfer Function, R(w)
S(w)×R(w)
Head-Related Transfer Function, HRTFl,r(w)
Physical Space
Psychological Space
Entrance of Ear Canal
Pl,r(w)=S(w)×R(w)×HRTFl,r(w)
Auditory Organ
D1 = f1(pl, r)
D2
D3
Dn
Perception of Elemental Sense
E1 = g1(D1)
E2
E3
En
Subjective Judgment to Elemental Sense
w1
w2
w3
wn
O=Σ_i w_i E_i
Overall Emotional Response

Figure 1.1: *Subjective evaluation system of sound environment (figure reproduced from (Morimoto, 2002)).*

1. temporal attributes: rhythm, durability, reverberance, etc.
2. spatial attributes: direction, distance, spatial impression, etc.
3. quality attributes: loudness, pitch, timbre, etc.

The listener assigns subjective judgments for each elemental sense, being nevertheless influenced by his/her personal emotional state.

In this monograph, particular attention is given to the second group of elemental senses: auditory cues produced by the human body 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 external ear. All these features are summarized into the so-called *Head-Related Transfer Functions (HRTFs)* (Cheng and Wakefield, 2001), i.e. the free-field compensated frequency- and space-dependent acoustic transfer functions between the sound source and

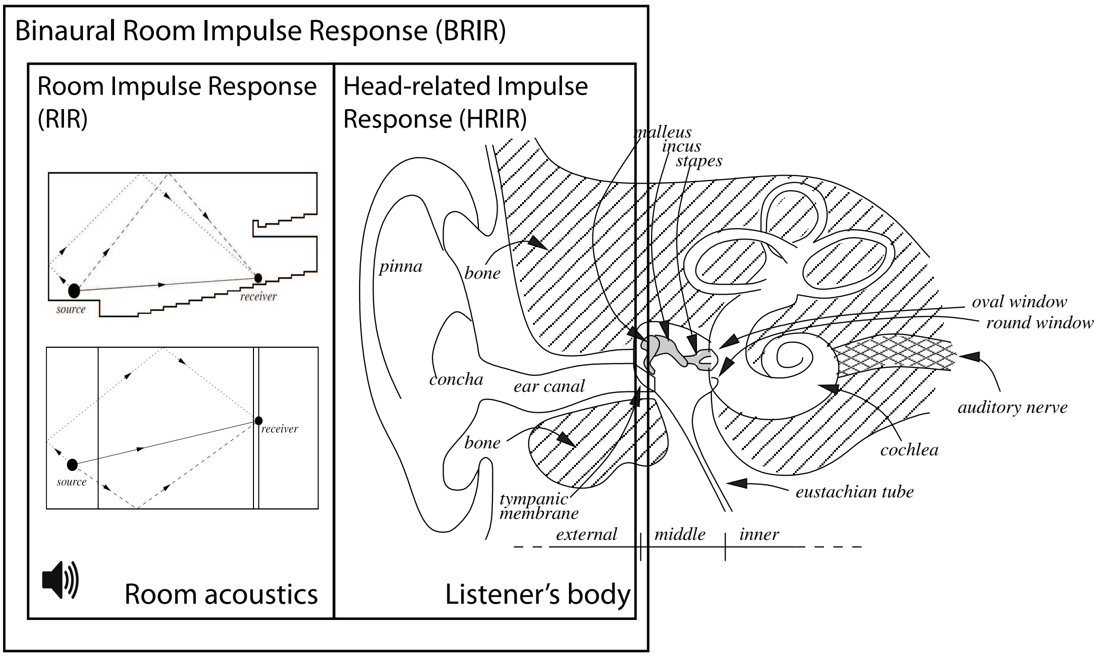


Figure 1.2: *The room acoustic information and the structure of the ear.*

the eardrum. Binaural spatial sound can be synthesized by convolving an anechoic sound signal with the corresponding left and right HRTFs, and presented through a pair of suitably compensated playback device.

It has to be noticed that binaural technologies relies on individual anthropometric and perceptual features having a key role in characterizing and modeling HRTFs. However, personal measured HRTF data for a vast number of listeners is currently time- and resource-expensive. Therefore, non-individualized HRTF sets are preferred in practice and they are typically recorded using “dummy heads”, i.e. mannequins constructed from averaged anthropometric measures, and represent a cheap and straightforward mean of providing 3-D rendering in headphone reproduction. However, they are known to produce evident sound localization errors (Wenzel et al., 1993), including incorrect perception of elevation, front-back reversals, and lack of externalization (Møller et al., 1996), especially when head tracking is not utilized in the reproduction (Thurlow et al., 1967).

1.1 Spatial hearing

Researches from several disciplines form the knowledge on spatial hearing. Physics, physiology, psychology and signal processing have jointly undertaken a wide spectrum of studies in many base aspects and application domains. Localization of a single source or multiple auditory events, subjective spatial perception in different real and virtual environments are some representative topics studied by internationally renowned scientists as Blauert (1983) and Xie (2013).

Human hearing incorporates perception of loudness, pitch, timbre and spatial attributes of

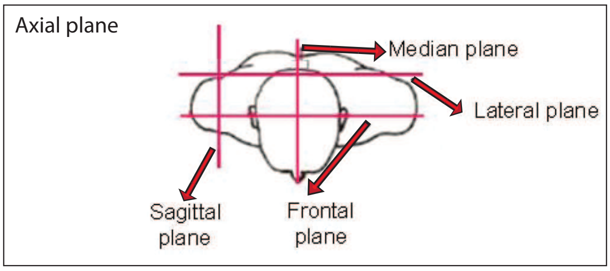


Figure 1.3: *Main planes of symmetry for human body. The axial plane is parallel to the page (figure adapted from (Strumillo, 2011)).*

sound. In particular, the auditory system is capable of estimating the location of sound sources in terms of direction and distance, as well as the spatial impression from the surrounding acoustic space. The position and movement of the listener inside the environment plays a key role in the perception of the characteristics of the listening space and in the identification of primary and early reflections coming from reflecting surfaces. Then, multisensory integration is all the more vital in spatial perception where auditory cues assist visual attention for seeking potential dangers around primates and mammals (Holmes and Spence, 2005; Heffner, 2004), apprehending location and movements of a potential prey, and communicating between peers (Gridi-Papp and Narins, 2008).

According to Strumillo (2011), auditory spatial perception might be categorized in four basic elements whose level of expertise and abilities vary across individuals:

- horizontal localization, azimuth;
- vertical localization, elevation;
- distance estimation;
- perception of space properties, spaciousness and externalization.

Their final evaluation is investigated through with psycho-acoustic experiments, usually based upon two basic types of localization judgments:

- relative localization in discrimination task;
- absolute localization in identification task.

The first two elements of spatial perception belong to *direction-of-arrival* (DOA) judgments in the corresponding planes, i.e horizontal and vertical for front/back, up/down and left/right discrimination. The main planes of symmetry usually take as references for the human body are depicted in Fig. 1.3

A distinction between localization and *lateralization* has to be immediately made. The former refers to estimating source position in an external three-dimensional space, while the latter might be seen as a special form of one-dimensional localization judgments where inside-the-head

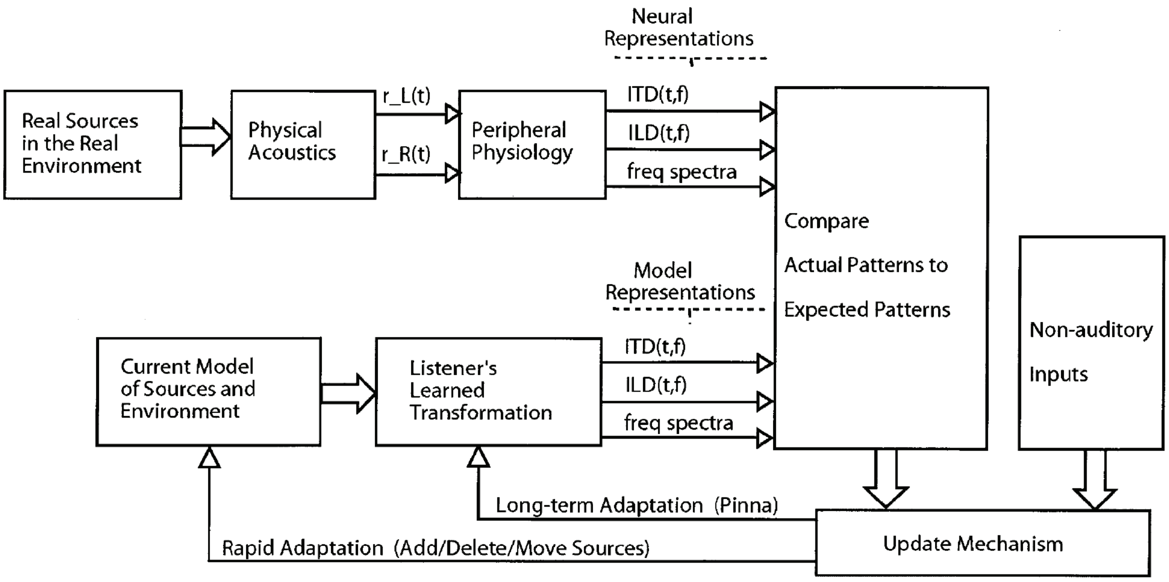


Figure 1.4: *General model of human sound localization. (figure reproduced from (Popper and Fay, 2005)).*

virtual phantom sound sources are located along the inter-aural axis (Yost and Hafter, 1987), e.g. during stereophonic presentation over headphones.

The third and fourth elements, i.e. distance estimation and space properties, mostly depend on prior knowledge of source properties and acoustic environments. The study of these abilities has received increasing scientific interest especially due to a rapid development of immersive virtual auditory displays, but are not well understood because a comprehensive analysis of multiple cues is still necessary. Auditory depth judgments (Zahorik et al., 2005), categorical judgments of spatial impression (Morimoto, 2002) and degree of externalization (Sakamoto et al., 1976; Hartmann and Wittenberg, 1996), are some examples of multidimensional phenomena without well defined dimensions.

A general model of human sound localization introduced by Colburn and Kulkarni (2005) can be seen in Fig. 1.4. The upper sequence of blocks depicts physical and physiological processes that give shape to neural representations of stimulus attributes such as ITD, ILD, monoaural and interaural spectral differences. On the other hand, the lower sequence of blocks represents how the listener creates his/her own representation of sound sources and the environment, continuously controlled and refined via update mechanisms based on the interplay between expectation and dynamic input data. Relative distances between source and receiver at previous time frames (e.g. head movements or dynamic sources), and prediction of source location expected from non-auditory modalities or by behavioral experience, highly change listener's awareness of the acoustic scene and his/her adaptation to previous stimuli.

Unfortunately, in this kind of auditory model, many elements related to behavioral and cognitive aspects are not sufficiently developed mainly in complex perceptual situation where several concurrent factors are included. The following subsections review the most notable perceptual mechanisms in order to identify quantitative criteria in the analysis and modeling of HRTFs.

1.1.1 Horizontal localization

Human ability for detecting changes in the direction of a sound source, i.e. relative localization performance, along the horizontal plane, known as "localization blur" or "minimal audible angle" (MAA), is characterized by the *just noticeable difference* (JND) for azimuth, $\Delta\theta$, of approximately $1^\circ - 2^\circ$ for a vast class of sounds (Blauert, 1983; Perrott and Saberi, 1990). This value grows to about $6^\circ - 8^\circ$ at lateral positions.

At the beginning of the last century, Lord Rayleigh studies on the scattering of sound waves around a rigid sphere gave birth to the field of 3D audio. In the context of his Duplex Theory of Localization (Strutt, 1904), the most relevant parameters that play a key role in azimuth perception are:

- *interaural time/phase difference (ITD/IPD)*: sound waves travel an extra distance before reaching the farthest ear, thus introducing a temporal delays, a time difference between the *time-of-arrival* (TOA) of a sound source at the two ears;
- *interaural level/intensity difference (ILD/IID)*: the head introduces an acoustic shadow (especially at high frequencies) for the farthest ear, and the ratio between the instantaneous level/intensity of the signals at the two ears defines this frequency- and direction- dependent quantity.

Further psychoacoustic experiments showed that ITD dominates horizontal localization for stimuli containing low-frequency components and, only for high-pass filtered stimuli, ILD plays a determinant role in localization performance (Wightman and Kistler, 1992) and (Macpherson and Middlebrooks, 2002).

Interaural time difference

ITD is a relative binaural quantity related to acoustic scene with a single sound source. In a real listening scenario with multiple sources, inter-source time differences and ITDs are extracted from each monoaural timing information, i.e. absolute delay related to sound wave propagation of each source to listener's ear. This generalization is described by the so called *time-of-arrival* (TOA) that inherits the same issues from ITD.

As a first approximation, the human head can be treated as a sphere. In the horizontal plane, this simplified scenario (see Fig. 1.5) allows to consider plane waves generated by sound sources positioned in the far field.[1] The acoustic wave is required to travel an extra distance before reaching the farthest ear and, in literature, the Woodworth's formula well approximates ITD (Woodworth and Schlosberg, 1954):

$$\text{ITD} = \frac{a(\sin\theta + \theta)}{c}, 0 \leq \theta \leq \pi/2, \tag{1.1}$$

[1] At leat 1 m far from the center of the head, see Sec. 1.1.3

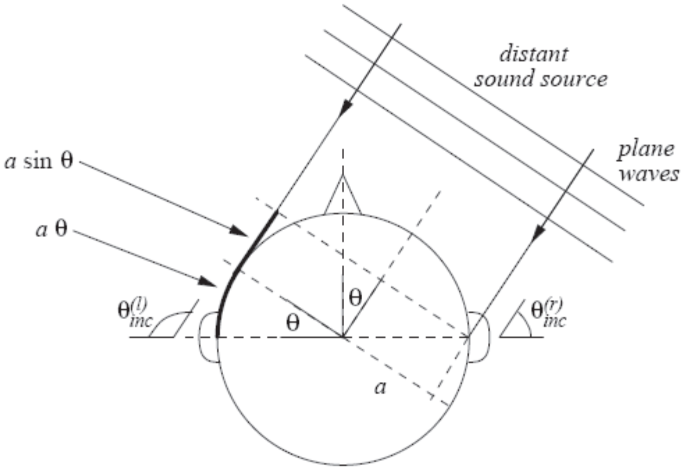


Figure 1.5: *ITD estimation of a spherical head; a plane wave is generated by a sound source in the far field.*

where *c* is the speed of sounds, *a* is the sphere radius and θ is the azimuthal angle. In the median plane ($\theta = 0$), ITD is equal to zero and becomes non-zero when the sound source deviates from that plane, i.e. in sagittal planes, reaching its maximum value where the sound source is located directly in front of one ear ($\theta = \pi/2$).

ITD is strictly dependent on the anatomical dimension of the listener head. As an example, a maximum value of approximately 0.6 ms is obtained for a spherical head radius of $a = 8.5$ cm. This approximation considers the ITD as independent from frequency, and effective for horizontal localization only for wavelengths comparable to individual head dimension, between 0.7 kHz and 1.5 kHz.[2]

However, ITD information are associated with *interaural phase delay difference*, ITD$_p$ which is defined as a directional- and frequency- dependent function:

$$\text{ITD}_p(\theta, \phi, f) = \frac{\Delta\Phi(\theta, \phi, f)}{2\pi f} = -\frac{\Phi_L(\theta, \phi, f) - \Phi_R(\theta, \phi, f)}{2\pi f}, \tag{1.2}$$

where $\Phi_L(\theta, \phi, f)$ and $\Phi_R(\theta, \phi, f)$ are the left and right directional- and frequency- dependent phase delay, respectively. The auditory system can also extract *envelope ITD*, ITD$_{env}$, from the slowly-varying temporal envelope of complex sounds at higher-frequencies containing multiple frequency components. The latter cue was shown to contribute to some extent in the localization of high-pass filtered sounds when amplitude modulations are introduced (Macpherson and Middlebrooks, 2002) and (Majdak and Laback, 2009).

Usually, ITD$_p$ and ITD$_{env}$ are not easily analyzed due to frequency and source signal dependency, respectively. To this regard, many methods and definitions have been proposed in the

[2] Interaural time difference provides ambiguous localization cues outside the frequency range related to head dimension: (i) when the wavelength is less than half of head dimension, roughly at 0.7 kHz, pressures at both ears for lateral sources are out of phase and (ii) when head dimension is larger than the wavelength, roughly above 1.5 kHz, phase difference exceed 2π (Blauert, 1983)

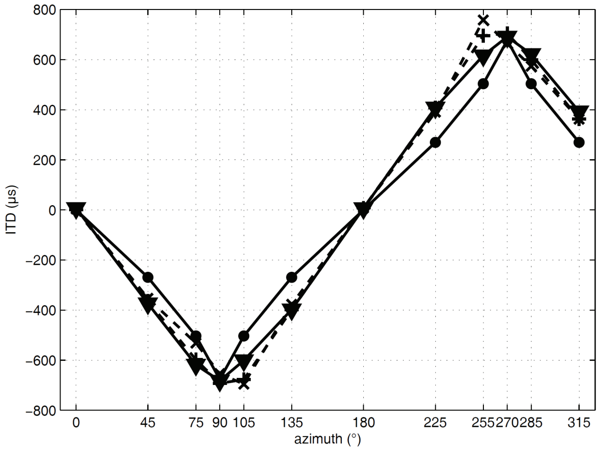


Figure 1.6: *Various ITD estimation methods averaging on 21 subjects. The individualized Woodworth's method (Algazi et al., 2001a) (solid line with circles), the threshold method based on 50 % maximum (solid line with triangles), the linear phase method (dotted line with crosses) and the inter-aural cross correlation method (dotted line with plus marks). (figure reproduced from (Busson et al., 2005))*

literature in order to estimate ITD/TOA (for an extensive review of this topic see (Minnaar et al., 2000; Nam et al., 2008a; Xie, 2013)) and mainly derived from the HRTF decomposition in pure delay (frequency independent TAO) and minimum-phase system (see Sec. 1.2.1). Figure 1.6 shows the variability in estimation for different kinds of ITD predictors. Usually, ITD$_p$ and ITD$_{env}$ are not easily analyzed due to frequency and source signal dependency, respectively. To this regard, many methods and definitions have been proposed in the literature in order to estimate ITD/TOA (for an extensive review of this topic see (Minnaar et al., 2000; Nam et al., 2008a; Xie, 2013)) and mainly derived from the HRTF decomposition 1.2.1 in pure delay (frequency independent TAO) and minimum-phase system. Figure 1.6 exhibits the variability in estimation for different kind of ITD predictors.

Interaural level difference

ILD is a relevant localization cue combined with ITD information and is much more salient at high frequencies where the wavelength is smaller than the physical dimensions of the head (Blauert, 1983). As a matter of fact, the sound pressure at contralateral side is attenuated at high frequencies by the shadowing effect of the head and boosted at the ipsilateral side. This acoustic information is formally described by the following equation:

$$\mathrm{ILD}(\theta,\phi,r,f) = 20\log_{10}\left|\frac{P_R(\theta,\phi,r,f)}{P_L(\theta,\phi,r,f)})\right|(d), \tag{1.3}$$

where $P_{L/R}$ indicates the sound pressure in the frequency domain generated by a sound source in a fixed $3-D$ position at the left/right ear. Figure 1.7 shows the average ILD (in dB) measured

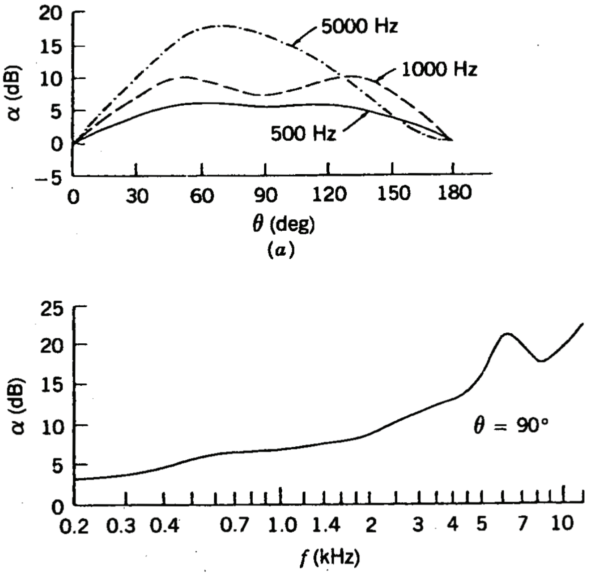


Figure 1.7: *ILD (in dB) as a function of the azimuth angle for three different frequencies, $500Hz$ – $1kHz$ – $5kHz$ (upper panel), and as a function of the frequency for a fixed azimuth angle, $\theta = 90^{\circ}$ (lower panel). (figure reproduced from (Hafter and Trahiotis, 1997))*

in a number of historical studies as a function of azimuth for three different frequencies (upper panel) and as a function of the frequency for a fixed azimuth, $\theta = 90^{\circ}$ (lower panel).

Since human head is not perfectly spherical or ellipsoidal, nor perfectly symmetric with respect to the median plane, individual irregularities from fine structures, such as the external ears (as discussed in Sec. 1.1.2), have to be taken into account determining localization performance. Therefore, individual ILD and its relationships with frequency and DOA are complicated to estimate and evaluate.

Head movements

It is well explained by Wallach's hypotesis (Wallach, 1940) and by (Wightman and Kistler, 1999) that head movements, primarily rotation movements about a vertical axis (turning left or right directions) (Thurlow et al., 1967), resolve front-back confusions where ambiguities of interaural differences arise especially outside the horizontal plane. If the listener is unable to move their head, azimuth localization errors increase in the proximity of the so called *cone/torus of confusion* for sound sources positioned in the far/near field[3] respectively, or in the median plane (a limit case). These regions are symmetrical around the listeners interaural axis and they contain those spatial points for which ITD and ILD values are constant. Fig. 1.8 visually describes these critical areas in the simplified spherical head approximation.

More recently, several works have investigated the distribution of spontaneous head movements during realistic listening activities by means of motion capture systems and head tracker

[3] See Sec. 1.1.3 for a precise far/near field distinction.

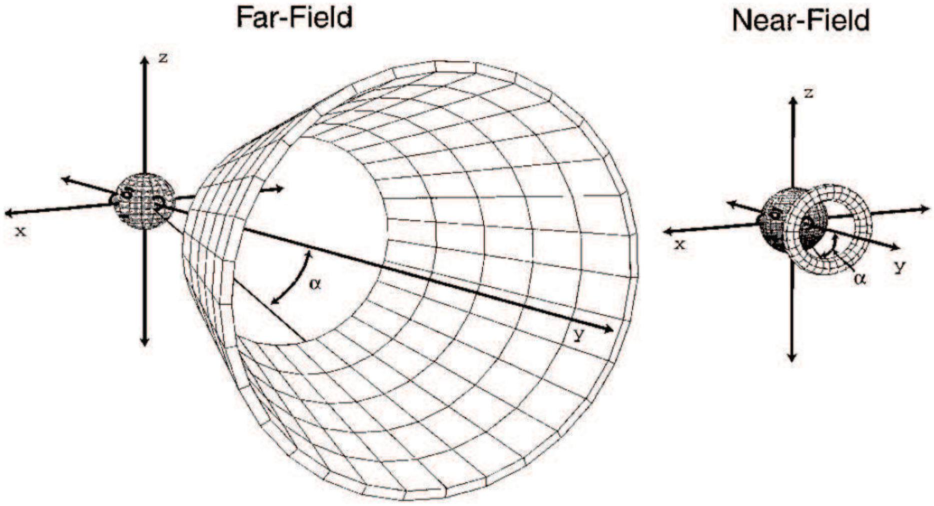


Figure 1.8: *Front-back critical area, cone of confusion and torus of confusion (figure reproduced from (Brungart, 2002))*

data (see (Hess, 2012) for an updated review of such devices). Kim et al. (2013) observed that the majority of head orientations of the listener is confined around the initial forward-facing direction, and a rotational movement is sometimes employed to localize sound sources, depending on whether or not listeners vision is usable (Nojima et al., 2013). The listeners move their heads to larger extents while judging source width or envelopment or deciding which sound source radiates an acoustic signal (in scenarios with multiple sound sources) rather than while localizing.

Furthermore, time and frequency features of sound stimuli also influence the effectiveness of head motions (Macpherson, 2011). This observation is supported by slow responsiveness of the auditory system to dynamic changes of binaural parameters (Grantham and Wightman, 1978; Grantham, 1984), even though this behavior is often related to a persistence mechanism in maintaining interaural differences.

1.1.2 Vertical localization

Directional hearing in the median vertical plane has long been known to have a coarser resolution compared with the horizontal plane (Wilska, 2010). The threshold for detecting changes in the direction of a sound source along the median plane was found to be never less than $4°$, reaching a much larger threshold ($\approx 17°$) for unfamiliar sounds (e.g. foreign languages). Such a poor resolution is motivated by two basic observations:

- the almost nonexistent interaural differences (ITD and ILD) between the signals arriving at the left and right ear, which conversely play a primary role in horizontal perception;
- the need of high-frequency content (above $4 - 5$ kHz) for accurate vertical localization (Hebrank and Wright, 1974b; Moore et al., 1989; Asano et al., 1990).

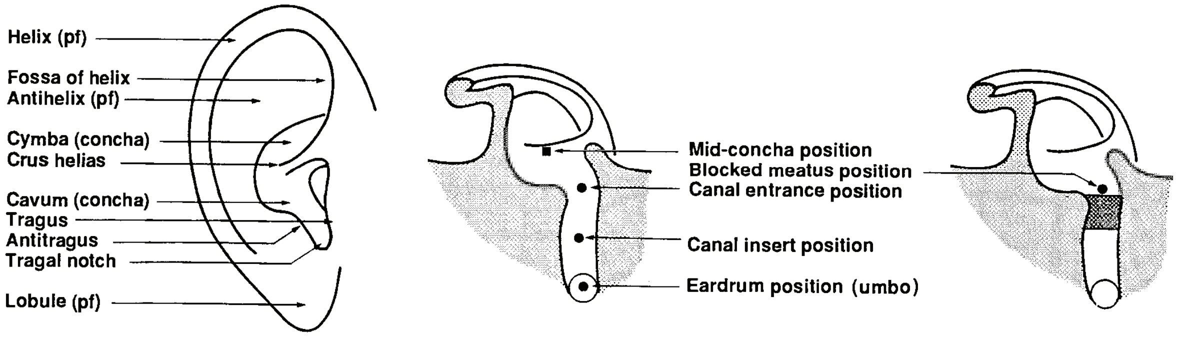


Figure 1.9: *Frontal and horizontal section of the human outer ear, showing five typical measurements positions. (figure reproduced from (Shaw, 1997))*

It is undisputed that vertical localization ability is brought by the presence of the pinnae (Gardner and Gardner, 1973) (see Fig. 1.9 for a detailed anatomical description of the pinna). Even though localization in any plane involves pinna cavities of both ears (Morimoto, 2001), determination of the perceived vertical angle of a sound source in the median plane is essentially a monaural process (Hebrank and Wright, 1974a). 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 depend strongly on elevation angle (Shaw and Teranishi, 1968), to a remarkably minor extent on azimuth (Lopez-Poveda and Meddis, 1996), and are almost independent on distance between source and listener beyond a few centimeters from the ear (Brungart and Rabinowitz, 1999).

Spectral cue: high-frequencies

Following two historical theories of localization, the pinna can be seen both as a filter in the frequency domain (Blauert, 1983) and a delay-and-add reflection system in the time domain (Batteau, 1967) as long as typical pinna reflection delays, detectable by the human hearing apparatus (Wright et al., 1974), produce spectral notches in the high-frequency range.

The evolution of notches in the median plane was studied by Raykar *et al.* (Raykar et al., 2005). Robust digital signal processing techniques based on the residual of a linear prediction model were applied to measured head-related impulse responses in order to extract the frequencies of those spectral notches caused by the presence of the pinna. The authors exploited a simple ray-tracing law (borrowed from (Hebrank and Wright, 1974b)) 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 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.

In addition to reflections, pinna resonances and diffraction inside the concha also contribute to the HRTF spectral shape. Shaw (1997) identified five resonant modes, schematized in Fig. 1.10, of the pinna excited at different directions which clearly produce the most prominent HRTF

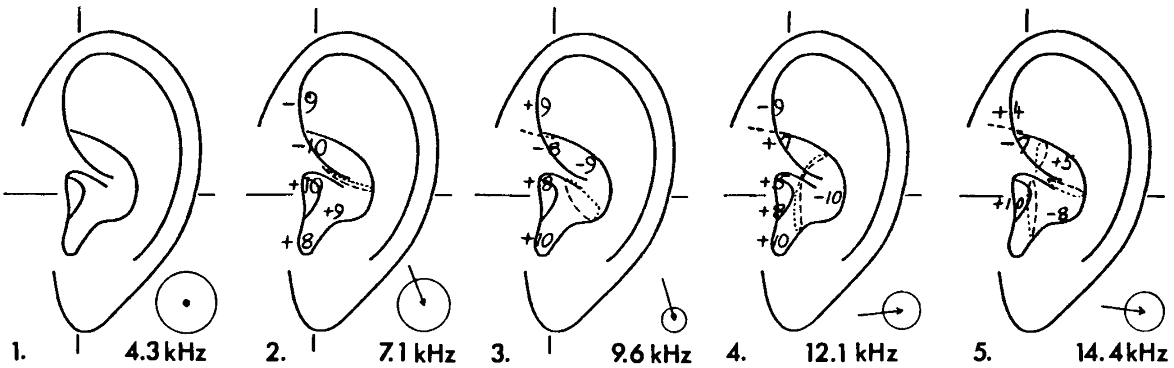


Figure 1.10: *The five pinna resonance modes identified by Shaw. For each resonance: average transverse pressure distribution, resonance frequency, nodal surface (broken lines), relative pressure (numerals) and direction of excitation (circles with arrows). (figure reproduced from (Shaw, 1997))*

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 two horizontal resonances at 12.2 and 14.4 kHz (modes 4 and 5).[4] These results are confirmed by a more recent study by Kahana *et al.* (Kahana and Nelson, 2007) on numerical simulation of PRTFs using BEM over baffled pinnae.

Concerning diffraction effects, Lopez-Poveda and Meddis (1996) 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. (2010, 2011).

Nevertheless, the relative importance of major peaks and notches in elevation perception has been disputed over the past years.[5] A recent study (Iida et al., 2007) showed a parametric HRTF 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 (Moore et al., 1989), 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 (Hebrank and Wright, 1974b) 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 direction. In general, hence, both peaks and notches seem to play an important function in vertical localization of a sound source.

[4] 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.

[5] 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 (Moore et al., 1989), which is the case for most measured HRTFs.

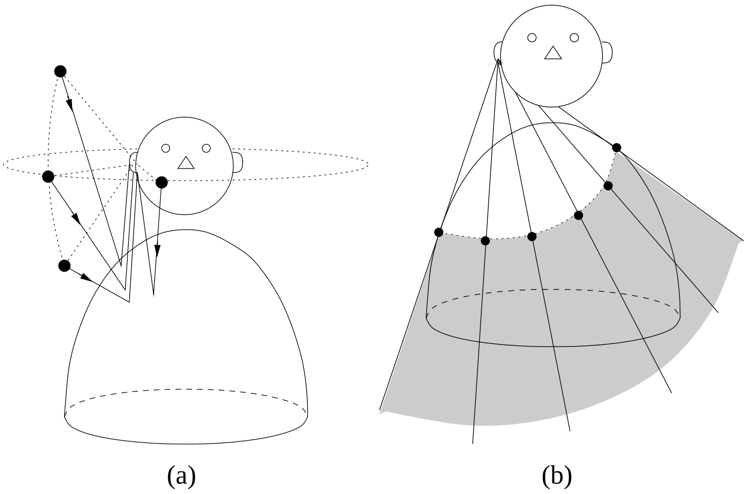


Figure 1.11: *Torso effects: shoulder reflections (a) and shadowing (b).*

Spectral cue: low-frequencies

It is generally considered that a sound source has to contain substantial energy in the high-frequency range for accurate judgment 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.

However, in absence of mid-high frequencies, subjects are still able to estimate elevation with good accuracy (Algazi et al., 2001b). These findings suggest that head diffraction together with torso and shoulders shadow and reflections provide to some extent a contribution to elevation cues, although remaining relatively weak effects if compared to those due to the pinnae. Experiments to establish the perceptual importance of each cue have in general produced mixed results (Brown and Duda, 1998; Asano et al., 1990; Algazi et al., 2001b).

Torso introduces a shadowing effect for sound waves coming from below. Complementarily at low frequencies, shoulders disturb incident sound waves coming from all directions other than below the listener. In particular, when the sound source is directly above the listener, shoulders provide a major reflection whose delay is proportional to the ear-shoulder distance (Huttunen et al., 2007). Fig. 1.11 sketches these two acoustic effects.[6]

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 (Huttunen et al., 2007). However, as Algazi *et al.* remarked in (Algazi et al., 2001b), when a signal is low-passed below 3 kHz elevation judgment is very poor in the sagittal plane if compared to a broadband source, but proportionally improves as the source is progressively moves 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 significantly away from the median plane.

[6] More details about the head and torso model are exposed in Sec. 3.2

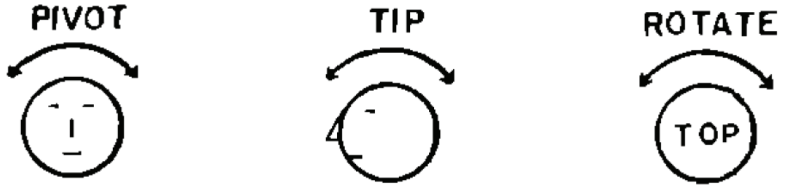


Figure 1.12: *A schematic illustration of head movements in each rotational axis. (figure reproduced from (Thurlow and Runge, 1967))*

Head movements

Perrett and Noble (1997) investigated elevation detection under distorted pinna functions and employing signals without high-frequency energy. Head rotations along the vertical axis lead to variations of low-frequency interaural time/phase differences, thus representing an important dynamic localization cue for elevation especially in the front median vertical plane.

Further studies (Thurlow and Runge, 1967; Rao, 2005) investigated which head movements provide information for localization in the median plane (Fig. 1.12 sketches the three-dimensional head movements): rotation and also rotate-pivot movement.

It is also still true, as in the horizontal localization, that improvement in performances varies with respect to time- and spectral- information of the stimuli (Morikawa et al., 2011).

1.1.3 Distance estimation

Under anechoic conditions, perception of sound source distance is correlated to the signal's intensity (the $1/r$ law).[7] On the other hand, if the environment is reverberant, a sensation of distance changing occurs if the overall intensity is constant but the proportion of the *reflected-to-direct* energy, the so-called *R/D ratio,* is altered (Begault, 1994; Bronkhorst and Houtgast, 1999; Kolarik et al., 2013). Accordingly, a number of other cues contribute to a correct perception of distance, such as distance-dependent spectral effects, familiarity of the involved sounds (Coleman, 1962), and dynamic cues (Zahorik et al., 2005).

It is well known from the literature (Mershon and Bowers, 1979; Loomis et al., 1999) that humans systematically underestimate the distance for sources in the *far field* (i.e. more than approximately $1 - 1.5$ m from the center of the listener's head) and overestimate the distance in *near field* (i.e. when the source is within $1 - 1.5$ m from the center of the head). [8]

In the literature, several models have been proposed to predict distance estimation by listeners. Bronkhorst and Houtgast (1999) relate the perceived auditory distance, d_s (in m), to the R/D

[7] Sound pressure level decreases by 6 dB for each doubling of the distance, r , between source and listener/receiver

[8] From a physical standpoint, the near field os the region where incoming wavefront cannot be assumed to be planar anymore. This definition is strongly related to the concept of peripersonal space defined as region where the subject is able to grasp objects with hand reaching movements. This further definition is useful for spatial exploration purposes in Ch. 7.

ratio:

$$d_s = Ar_h\left(\frac{\hat{E}_r}{\hat{E}_d}\right)^j, \tag{1.4}$$

where A and j are constants, $\hat{E}_d$ and $\hat{E}_r$ are the energies of the direct and reverberant sound respectively and r_h is the reverberation radius or critical distance. [9] The calculation of $\hat{E}_d$ and $\hat{E}_r$ is performed within the following integration window containing two parameters, i.e. time duration, t_w, and slope s: [10]

$$W = \begin{cases} 1 & \text{for } 0 < t \le t_w - \frac{\pi}{2s} \\ 0.5 - 0.5\sin[s(t - t_w)] & \text{for } t_w - \frac{\pi}{2s} < t < t_w + \frac{\pi}{2s} \\ 0 & \text{for } t > t_w + \frac{\pi}{2s} \end{cases}, \tag{1.5}$$

Afterward, Zahorik (2002) suggested a power function fit analysis based upon a variety of previous studies (84 data sets), in order to map the physical distance, r, to the perceived auditory distance, r', in reverberant environments:

$$r' = kr^a, \tag{1.6}$$

where the average value of constant k is 1.32 and the power-law exponent, a, exhibits an average value of 0.4 being influenced by experimental conditions and subject sensitivity/expertise.

When the source is in the far field, directional cues are roughly distance-independent. By gradually approaching the near field, it is known that whereas ITD remains almost independent from distance, ILD is boosted across the whole spectrum and in particular at low frequencies. Since distance dependence must then be taken into account in the near field, a prompt characterization of the head's response in such region has to be studied (Liu and Xie, 2013). For the sake of simplicity, the head of the listener can be treated as a rigid sphere (Duda and Martens, 1998).

Switching from a static to a dynamic environment where the source moves with respect to the listener and/or *vice versa*, slight translations of the listeners head on the horizontal plane can help discriminating near source distance thanks to the motion parallax effect (Speigle and Loomis, 1993; Brungart et al., 1999).

1.1.4 Externalization using headphones

Throughout this thesis, the majority of examples and applications are related to headphones delivery of spatial audio contents, thus the reference scenario employs *dichotic* stimuli, i.e. conditions in which different signals are presented to the left and right channels of the earphones.[11]

[9] Reverberation radius defines the distance between the listener and a sound source at which the energy of the direct sound becomes equal to the reflected sound energy. Sound information deteriorate as the listener moves away from this critical distance. Reverberation radius has many formal descriptions depending on the properties of room and sound source, for more detail see (Kuttruff, 2000).

[10] Constant parameters A and j, and slope parameter, s, were estimated with an iterative least-squares fit on the corresponding room impulse response.

[11] Signal presented at one single headphone is called *monotic*, while identical signals presented at both headphones are called *diotic*

Producing realistic virtual acoustic scenarios over headphones with particular attention to space properties and externalization issues remains one major challenge due to the interconnections of all the above mentioned localization cues. For this reason, this section well summarizes all the perceptual requirements for a virtual auditory display.

The term externalization is usually used to indicate whether stimuli are heard inside (being *lateralized*, subjected to the so called *in-head localization* (IHD)) or outside the head (being *localized*). The contributing factors to such phenomenon can be grouped in four categories (Hafter and Trahiotis, 1997; Loomis et al., 1999):

- *spectral cues*: availability of binaural signals appropriate to head and pinna spectral filtering; and in particular, appropriate compensation for headphones acoustic distortion on delivered signals (Wightman and Kistler, 1989a,b) and correction of the acoustic coupling between headphones and ear canal (Hiipakka, 2008);

- *room acoustics*: availability of virtual room acoustics resembling real-world reverberation (Plenge, 1974), therefore containing early reflections which are related to spatial impression and perception of the acoustic space (Barron, 1971; Sakamoto et al., 1976); moreover, realistic spatial impression must be provided conforming to ambience, source width and listener envelopment (Bradley and Soulodre, 1995; Morimoto, 2002).

- *head movements*: the presence of head movements, as in the real acoustic world, produces dynamic changes in interaural cues, but their role in the externalization of signals is not unanimously recognized due to the lacking of quantitative data (Brimijoin et al., 2013);

- *ergonomic delivery system*: if the listener is not aware of the actual of sound emitting source, i.e. the headphones transducer, he/she is more likely to externalize stimuli, especially with low invasiveness of headphones strap or cups; two example of *transparent* devices are the tube-phone apparatus by Kulkarni and Colburn (2000) and the bone conducted headphones by Walker and Lindsay (2005), illustrated in Fig. 1.13.

The following sections mainly deal with the first group of factors with particular attention to how binaural signals are measured, modeled and reproduced. It has to be mentioned that the modular approach described in Chapters 2 and 5 permits future formal investigation on which factors dominate the perception of out-of-the-head stimuli localization.

1.2 Spatial audio over headphones

In order to enable authentic auditory experiences, the correct sound pressure level (SPL) due to one or more acoustic sources positioned in a virtual space shall be reproduced at the eardrums of a listener by a sound delivery system, be it a pair of headphones, stereo loudspeakers or multiple loudspeaker arrays.

In the literature, sound transmission from the headphone to the eardrum is often represented through an analogue circuit model (Møller, 1992), reported in Fig. 1.14. With reference to such model, the goal of the research reported in this thesis is the reproduction of the sound pressure P_6

(a)

(b)

Figure 1.13: *Etymotic ER-*1 *insert headphones (http://www.etymotic.com/pro/er1.aspx) and Teac Filltune HP-F*100 *bone-conduction transducers (http://teac-ipd.com/) (devices used by Stanley (2009) in his studies).*

that would be guaranteed at the entrance of the blocked ear canal of the listener by a sound source placed around him/her, even though nothing prevents to extend these studies to circuit points closer to the eardrum impedance as soon as proper tools are available (e.g., HRTF measurements at the eardrum, ear canal models, etc.). The analysis and synthesis of P_6 requires the collection of HRIR data and *headphone impulse responses* (HpIRs). The former are usually recorded onto a dummy head and/or a significative number of human subjects by varying the position of the sound source with respect to the head, while the latter lead the equalization process of several types and models of headphones.

By convolving a desired monophonic and anechoic sound signal with a set of personal HRIRs adequately compensated for headphone-induced spectral coloration or for loudspeaker dynamic crosstalk cancellation (Gardner, 1999; Song et al., 2010), one can reach almost the same localization accuracy as in free-field listening conditions (Bronkhorst, 1995), especially when head motion, artificial reverberation (Valimaki et al., 2012) and RIR prediction programs are considered (Wightman and Kistler, 1999; Begault et al., 2001).

1.2.1 Head-related impulse response data

Auditory cues discussed in Sec. 1.1 are summarized into the so-called *Head Related Transfer Functions*, i.e. the frequency- and space-dependent acoustic transfer functions between the sound source and the eardrum. More formally, the HRTF at one ear is the frequency-dependent ratio between the sound pressure level (SPL) $P_L(\theta, \phi, r, \omega)$ and $P_R(\theta, \phi, r, \omega)$ at the left and right eardrum respectively, and the free-field SPL at the center of the head $P_f(r, \omega)$ as if the listener

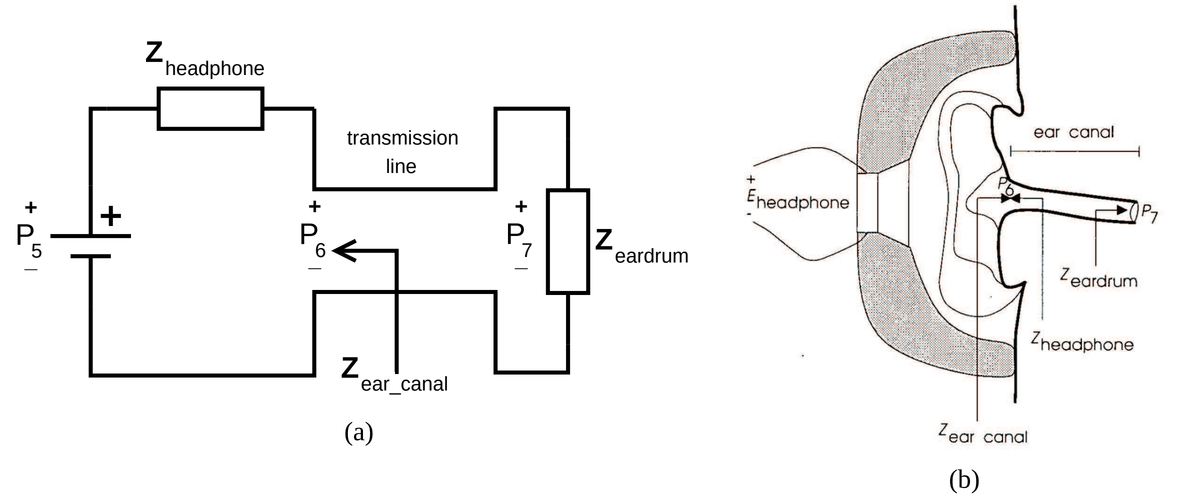


Figure 1.14: *Sound transmission from the circumaural headphone to the eardrum; (a) circuit model (b) listener's anatomy (after (Møller, 1992)).*

were absent:

$$
\begin{aligned}
H_L(\theta,\phi,r,\omega) &= \frac{P_L(\theta,\phi,r,\omega)}{P_f(r,\omega)}, \\
H_R(\theta,\phi,r,\omega) &= \frac{P_R(\theta,\phi,r,\omega)}{P_f(r,\omega)},
\end{aligned}
\tag{1.7}
$$

where θ and ϕ indicate the angular position of the source relative to the listener, r expresses the listener-to-source distance, and ω is the angular frequency. Binaural anechoic spatial sound can be synthesized by convolving an anechoic sound signal with the corresponding left and right HRTFs. Moreover, the minimum-phase characteristic of HRTF (both in the far and near field) allows its decomposition in a pure delay, $\tau(\bullet)$, followed by a minimum-phase system, $H_{min}(\bullet)$:

$$
H(\theta,\phi,r,\omega) = H_{min}(\theta,\phi,r,\omega)\exp[-j2\pi f\tau(\theta,\phi,r,\omega)]. \tag{1.8}
$$

Several studies (see Kulkarni et al. (1995, 1999), Nam et al. (2008b), and Yu et al. (2012)) confirmed high degree of similarity between a measured HRTF and its minimum-phase counterpart, though pairwise analysis of the cross-coherence over azimuth and elevation. HRTFs and their minimum-phase sequences among all spatial locations exhibit cross-coherence values of approximately ≈0.9 (the maximum cross-coherence would be 1.0), thus signal-processing procedures largely benefit from this HRTF decomposition. As a matter of fact, the extracted pure delay or time-shift is related to the monaural TOA; therefore, as presented in Sec. 1.1.1, ITDs can be estimated as the difference between the left and right ear TOA.

Obtaining personal measured HRTF data for a vast number of users is currently unpracticable because specific hardware, anechoic spaces, and long collection times are strictly required (Brown and Duda, 1998). This is the main reason why non-individual HRTFs, usually measured on anthropomorphic mannequins (Burkhard and Sachs, 1975), are often preferred in practice. The drawback with non-individual HRTFs is that they rarely match with the listener's

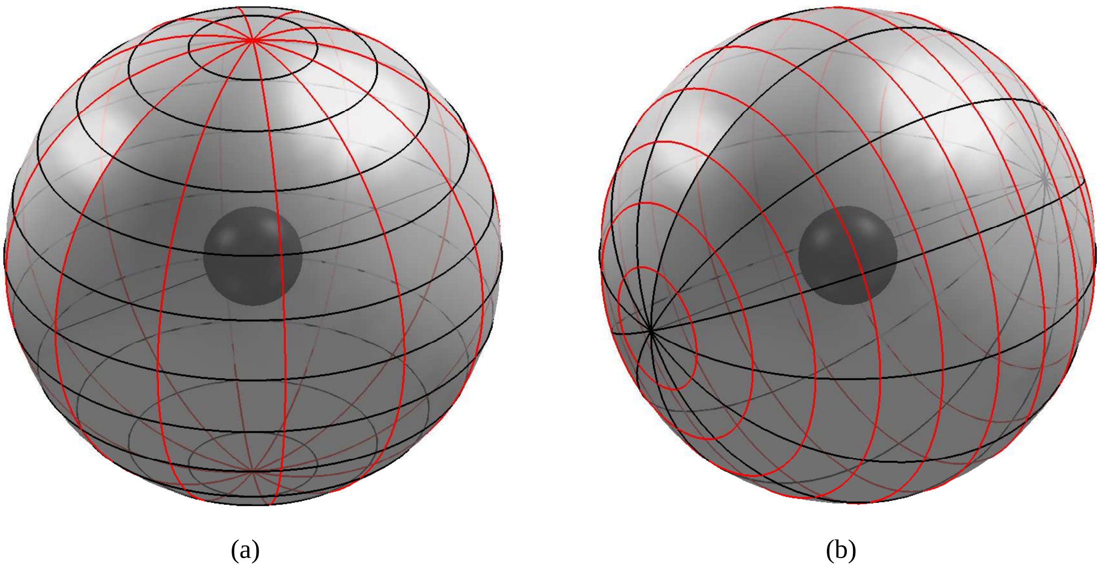


Figure 1.15: *Coordinate systems: vertical polar (a) and interaural polar (b). The inner grey sphere represents the human head, with the interaural axis evidenced by a thin grey line. Points along a red curve on the outer sphere have constant azimuth angle, points along a black curve have constant elevation angle.*

unique anthropometry, and especially his/her outer ear (Abaza et al., 2010), resulting in frequent localization errors such as front/back reversals (Wenzel et al., 1993), elevation angle misperception (Møller et al., 1996), and inside-the-head localization (Plenge, 1974; Brungart, 2002). Since recording individual HRTFs is both time- and resource-expensive, obtaining reliable HRTFs for a particular subject in different and more convenient ways is desirable.

Measurements

HRIRs are usually measured in anechoic environment for a discrete spherical spatial grid of sound source locations around a listener, while BRIRs are measured in an echoic environment for a discrete number of head orientations of a dummy head (Lindau and Brinkmann, 2010) or a human listener or a spherical microphone array (Algazi et al., 2004; O'Donovan et al., 2008; Zotkin et al., 2010).

Several research groups provided public-domain HRIR databases. Among these, the CIPIC HRTF database (Algazi et al., 2001a)[12] and the LISTEN HRIR database[13] are today the most known and exploited. The intrinsic differences among HRIR databases can be mainly summarized referring to the recording process and their storage format, some aspects of which are now discussed.

First of all, the discrete spatial grid where the responses are taken has to be interpreted rela-

[12] http://interface.cipic.ucdavis.edu/

[13] http://recherche.ircam.fr/equipes/salles/listen/

Database	CIPIC[a]	LISTEN	FIU[b]	MIT[c]	AALTO[d]	ARI	PKU&IOA[e]
Sampling frequency	44100 Hz	44100 Hz	96000 Hz	44100 Hz	48000 Hz	48000 Hz	65536 Hz
HRIR length	200pts	512pts	256pts	512pts	8192pts	256pts	1024pts
Coordinate system	IP	VP	VP	VP	VP	VP	VP
Spatial grid	25A,50E	24A,10E	12A,6E	72A,14E	72A,19E	90A,22E	72A,14E
No. of directions	1250	187	72	710	240	1550	793
No. of distances	1	1	1	1	2	1	8
Distance	1 m	1.95 m	–	1.4 m	0.68,1.35 m	1 m	0.2–1.6 m
Stimulus type	GC	LS	GC	ML	LS	LS	SG
Mic position	BEC	BEC	BEC	BEC	BEC	BEC,BTE	BEC
Data format	.mat	.wav,.mat	.txt	.wav	.mat	.mat	.dat
No. of subjects	45	51	15	2	21	92	1
Raw data	no	yes	no	yes	no	yes	no
Onset/ITD data	yes	no	yes	no	no	no	no
Anthropometric data	yes	yes	yes	no	no	yes	no

[a] Algazi et al. (2001a)
[b] Gupta et al. (2010)
[c] Gardner and Martin (1995)
[d] Gómez Bolaños and Pulkki (2012)
[e] Qu et al. (2009)

Table 1.1: *Specifications of seven public HRIR databases. Legend: IP = interaural polar; VP = vertical polar, GC = Golay codes, LS = logarithmic sweep, ML = maximum length sequence, SG = spark gap impulse, BEC = blocked ear canal, BTE = behind the ear. The HRIR length specification refers to the final, post-processed HRIRs. The two values in the spatial grid specification refer to the number of azimuth angles in the horizontal plane and the number of elevations in the median plane respectively; the uneven angular spacing is omitted for brevity.*

tively to the assumed coordinate system: *interaural polar* or *vertical polar*. In the former case (e.g. CIPIC database), typically a loudspeaker (or multiple loudspeakers) sequentially plays the sound stimulus moving (or switching to the next one when multiple loudspeakers are used) along a semi-circle with the subject's head in its center until all the sampled positions are stored; then the stimuli are again played after having changed the elevation angle of the semi-circle and thus moved in a sheaf of planes with the interaural line as rotation axis. In the latter case (e.g. LISTEN database), the rotation axis of the semi-circle spanned by the loudspeaker(s) is the mid-dorsal line passing through the center of the head, and the azimuth angle is varied first. As a result, the spatial grids spanned by the two coordinate systems are different (see Fig. 1.2.1) and can be defined as follow:

- in the interaural polar system, elevation ϕ is the angle between the horizontal plane and the plane containing both the interaural axis and the sound source ($\phi \in [-180, +180)$); on this plane, azimuth θ is the angle between the line of intersection with the median plane and the line passing from the origin of the axes and the sound source ($\theta \in [-90, +90]$);

- in the vertical polar system, azimuth θ is the angle between the median plane and the plane containing both the vertical axis and the sound source ($\theta \in [-180, +180)$); on this plane, elevation ϕ is the angle between the line of intersection with the horizontal plane and the line passing from the origin of the axes and the sound source ($\phi \in [-90, +90]$).

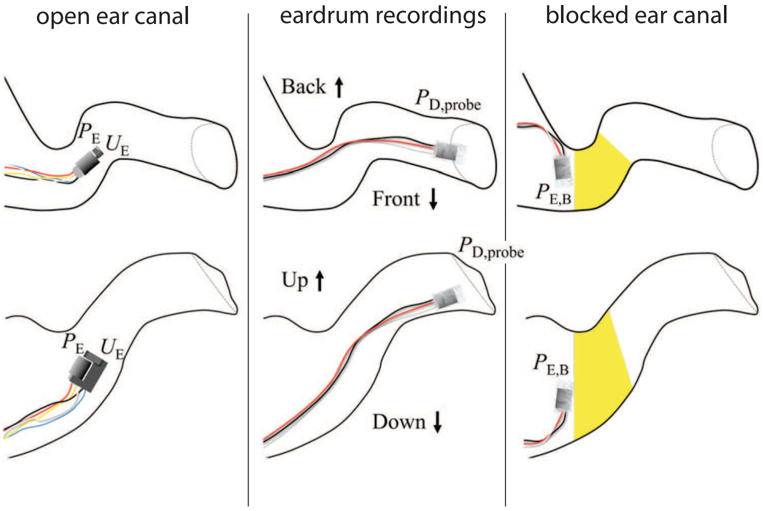


Figure 1.16: *Three HRTF measurement techniques used by Hiipakka et al. (2012). The upper sequence depicts the top view and the lower sequence the frontal one. Variables P and U represent pressure and velocity, respectively.*

At the transmitter side (the loudspeaker), the audio chain has to be calibrated with respect to the stimulus that maximizes the signal-to-noise ratio of the recordings. Some examples of stimuli used in the aforementioned databases are Golay codes and logarithmic sine sweeps. At the receiver side, the position of the microphone plays a predominant role in signal acquisition: distinction can be made between blocked-ear-canal, open-ear-canal and eardrum recordings, each of which corresponds to different acoustical information captured by the measurement setup (Fig. 1.16 and 1.9 sketch microphone positions). Furthermore, after recording all the collected raw data are processed and compensated in different ways (e.g. through inverse filtering of the stimulus plus free-field compensation) in order to extrapolate the impulse response. These and other differences among a number of public HRIR databases can be appreciated in Table 1.1.

Modeling

Computational models generate synthetic HRTFs from a physical (Lopez-Poveda and Meddis, 1996) or structural interpretation of the acoustic contribution of head, pinna, shoulders and torso. These models have different degrees of simplification, going from basic geometries (Teranishi and Shaw, 1968; Algazi et al., 2002a; Takemoto et al., 2010) to more accurate descriptions capable to reproduce the peaks and notches of the HRTF (Katz, 2001b; Fels and Vorlander, 2009). HRTF spectral details also emerge exploiting principal component analysis (PCA) (Kistler and Wightman, 1992) allowing to further tune the HRTF to a specific listener.

Several techniques for synthetic HRTF design have been proposed during the last two decades. According to Brown and Duda (1998), computational models of HRTFs can be classified in three groups:

1. *pole/zero models*: filter design, system identification, and neural network techniques are

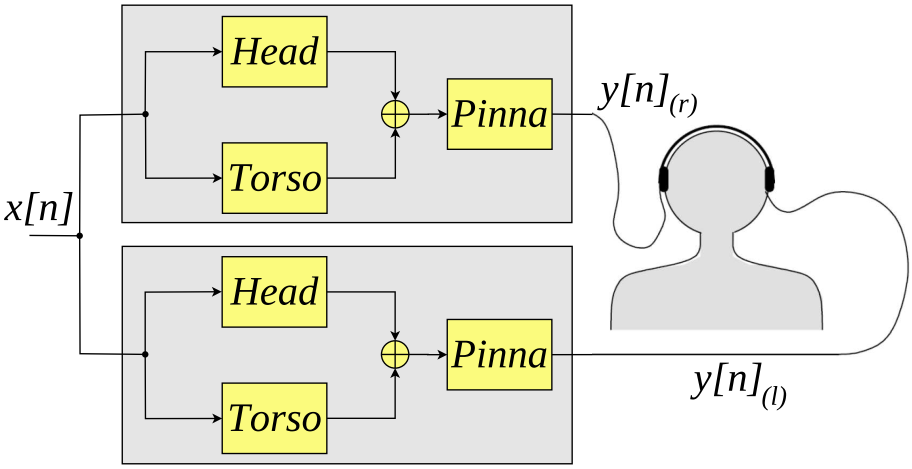


Figure 1.17: *A generic structural HRTF model.*

applied in order to fit multiparameter models to experimental data (e.g. (Durant and Wakefield, 2002)) leading to synthetic HRTFs approximated with low-order rational filters;

2. *series expansions*: modeled HRTF is represented as a weighted sum of simpler basis functions applied to collections of HRIRs or HRTFs, e.g. on principal component analysis (PCA) (Kistler and Wightman, 1992) or surface spherical harmonics (SSH) (Evans et al., 1998);

3. *structural models*: the contributions of the listener's head, pinnae, shoulders and torso to the HRTF are isolated and arranged in different filter structures each accounting for some well-defined physical phenomenon, as Fig. 1.17 roughly sketches. The linearity of these contributions allows reconstruction of the global HRTF from a proper combination of all the considered effects (Algazi et al., 2001c).

Although recent trends in HRTF customization mainly have focused on series expansions with self-tuning of weights (Hwang et al., 2008; Shin and Park, 2008) or simply non-individualized HRTF selection (Seeber and Fastl, 2003; So et al., 2010; Katz et al., 2012), structural HRTF modeling remains the most attractive alternative from both the viewpoints of computational efficiency and physical meaning: parameters of the rendering blocks sketched in Fig. 1.17 can be estimated from real data, fitted to low-order filter structures, and finally related to meaningful anthropometric measurements.

The *mixed structural modeling* proposed in this thesis and formalized in Ch. 2, follows the structural modeling approach, generalizing it so as to allow for the inclusion of acoustic measurements, numerical simulations, extrapolated data or one of the aforementioned modeling approaches in one or more of its components. For instance, one would desire to combine a numerically simulated model of the pinna (Takemoto et al. (2010) or the external ear model described in Sec. 3.3) with the measured HRTF of a generic pinnaless mannequin, or to feed a filter snowman model (Algazi et al., 2002b) with an individualized weighted combination of PCA basis functions extracted in the median HRTFs (Shin and Park, 2008) allowing vertical control of the virtual sound source.

1.2.2 Headphone impulse response data

The *headphone impulse response* (HpIR) describes both the headphone's transducing properties and the headphone to eardrum transfer function (Pralong and Carlile, 1996). In order to provide proper binaural signals and thus to reach high localization accuracy in VADs, headphones have to be carefully equalized.

In the typical sound transmission model, Møller (1992) assumes as prerequisite for 3D audio rendering

$$Z_{\text{headphone}} \approx Z_{\text{radiation}}, \tag{1.9}$$

where $Z_{\text{radiation}}$ denotes the equivalent impedance outside the ear canal in free-field listening conditions and $Z_{\text{headphone}}$ the equivalent impedance outside the ear canal with headphones. This equality holds for wavelengths greater than the ear canal's width, thus approximately under 10 kHz, and gives rise to the so-called *Pressure Division Ratio* (PDR):

$$\frac{P_{\text{open}}}{P_{\text{blocked}}} = \frac{P_{\text{open}}^{Hp}}{P_{\text{blocked}}^{Hp}}, \tag{1.10}$$

where P_{open} and P_{blocked} denote the free field sound pressure at the entrance of the open- and blocked-ear canal respectively, while P_{open}^{Hp} and P_{blocked}^{Hp} denote the same sound pressure observation points when the sound source is a headphone. Headphones with PDR ≈ 1 satisfy the *free-air equivalent coupling* (FEC) characteristic (Møller, 1992). In order to verify this basic behavior, several measurements with different subjects and recording conditions should be conducted and properly stored. However, headphones with no FEC, e.g. insert headphones or in-ear monitors, could be employed in spatial audio reproduction as well, provided that the HpIRs are measured with an unblocked ear canal (Hiipakka et al., 2012).

For low frequencies, inter-subject variability is limited up to ≈ 4 kHz because headphones act as an acoustic cavity only introducing a constant level variation. On the contrary, in the higher spectrum, headphone position and listener's anthropometry give rise to several peaks and this acoustic effect is mainly due to two reasons:

- standing waves start to grow inside headphone cups;
- outer ear's resonances yield to an individual characterization also for the HpIRs.

The availability of a large amount of individual measurements makes it possible to develop headphones correction procedures that are listener-specific (Møller et al., 1995; Pralong and Carlile, 1996) and robust to headphone placements (Kulkarni and Colburn, 2000). All these elements contribute to authentic and high-fidelity virtual auditory spaces. Furthermore, analysis of headphone to external ear transfer functions, ear canal entrance-to-eardrum transfer functions and impedance mismatch are crucial issues towards understanding and controlling sound artifacts introduced by headphones.

Measurements

No public HpIR databases nor standard HpIR repositories have been proposed to date. Typically, small sets of headphones are tested limited to the purpose of reproducing a *virtual auditory scene* for a specific experiment and subject (Pralong and Carlile, 1996). Recent auralization softwares, e.g. AM3D[14], cluster typical headphones by type (e.g. earphones, in-ear headphones, closed-back or open-back circumaural), each capturing common behaviours for a generic listener. However, given H headphone models and L listeners it is possible to measure $H \times L$ individual HpIRs.

The HpIR measuring setup is a special case of the HRIR's, discussed in Sec. 1.2.1. The peculiar difference lies in the focus on either the emitter or the receiver: emitter, i.e. headphones, inevitably perturbs acoustic waves before reaching the receiver, i.e. microphones, being physically connected to resonating body parts (external ear and ear canal) and thus shaping a new geometry. Generally speaking, the emitter encloses the acoustic meatus and becomes itself part of a unique resonating object which shares some acoustic properties of the listener and other from how and where the headphones is positioned. Furthermore, emitter's placement and measuring conditions (e.g. open vs. blocked-entrance ear canal measurements) depend strongly on headphone type, e.g. circumaural or supraaural headphones, earphones, insert earphones, bone conducted headset and assistive hearing devices (see Borwick (2001) for an exhaustive review on headphone design.).

The most common methods employed in the acquisition of HpIRs involve the use of an artificial ear, such as B&K 4153, or a head and torso simulator (HATS), such as a KEMAR mannequin, or human listeners. By using an artificial ear, one would exclude potential acoustic effects by the pinna, otherwise typical behaviors of the headphones coupled with a pinna mold are captured with HAT simulator. Finally, individual recordings measure listener specific characteristics of headphone coupling.

Obtaining in-situ robust and individual HpTFs with straightforward procedures in order to always apply listener specific corrections to headphones is a challenging research issue. Horiuchi et al. (2001) proposed a stereo earphone-plus-microphone to enhance externalization by calculating the *ear canal transfer function* (ECTF) in real-time and controlling an adaptive inverse filtering method. More recently, Kim and Choi (2005), Sankowsky-Rothe et al. (2011), Hiipakka (2013) and Kohler et al. (2013) are investigating innovative approaches to estimate the sound pressure in occluded ear canal, e.g. a listener wearing hearing aid or widespread inserted headphones.

It is worthwhile to notice that emitter characteristics and acquired signals need to be methodically stored in specific data structures sharing common features with individual HRIR databases, otherwise no systematic analysis can be shared within the international research community. In light of such observation, Ch. 5 suggests practical guidelines towards an organized collection of HpIR data.

[14] http://www.am3d.com/

Equalization

Headphones, when used for the binaural reproduction, have to be carefully equalized because high localization accuracy is needed. Unfortunately, the transfer function between headphone and eardrum heavily varies from person to person and with small displacements of the headphone itself. Such variation is particularly marked in the high-frequency range where important elevation cues generally lie. Thus, an inaccurate compensation likely leads to spectral colorations that affect both source elevation perception and sound externalization (Masiero and Fels, 2011). Although various techniques have been proposed in order to tackle such a delicate issue, modeling the correct equalization filter is still a hot open research theme.

It is worthwhile to notice that the fidelity of spatial audio reproduction increases with the amount of individualization in the headphone correction (Lindau and Brinkmann, 2012). In light of this, both measurement techniques (e.g. general and individual HpIRs) and equalization methods play a determinant role in the individualization process.

A intensity example in trying to capture the mean headphone behavior on any listener is described by Martens (2003) and Nishimura et al. (2010), where headphone response is estimated as the average on N measurements, i.e. available measurements for the same headphone model (the left, l, and right, r, channels are treated separately and they are omitted in the following notation):

$$\begin{gathered}\left|G^{hp}(w)\right| = 10^{-\overline{|H^{hp}(w)|}},\\ \overline{|H^{hp}(w)|} = \frac{1}{N}\sum_{i=1}^{N}\log_{10}\left|H_i^{hp}(w)\right|, \text{ with } N = L \times R,\end{gathered} \tag{1.11}$$

where L is the number of listeners, R is the number of headphone repositionements. Despite of the non satisfactory and solid results due to an efficient evaluation method, it takes advantage of all acquired information by the measurement phase which might be seen, to some extent, as a first step towards a deeper and comprehensive data analysis.

After having determined headphone response, several issues must be taken into account in considering the acoustic inversion problem of a HpIR, among which:

- acoustic perception might be affected by unstable inverse filter and phase distortion;
- any filter has 1-to-1 correspondence with its inverse filter, thus the inversion process is restricted to one HpIR, i.e. one listener (or dummy head) and one position; especially in the high frequency range, narrow notches might be over-compensated by strong peaks causing ringing artifacts after any device repositionings.

Regardless of what headphone contribution is considered, i.e. general/individual, measured or statistically computed, Schärer's taxonomy (Schärer and Lindau, 2009) (Fig. 1.18) collects most of the inverse filter design techniques for headphone equalization/correction.

Recent trends in headphone design involve the use of frontal emitter (supra-aural) headphones (Sunder et al., 2013; Farkas and Hajnal, 2013). These devices, and those which will be invented in the future, still require novel headphone equalization techniques, that would preserve embedded personal anthropometric cues, and possibly a rethinking of the equalization methods (like the one outlined in Hammershøi and Hoffmann (2011)).

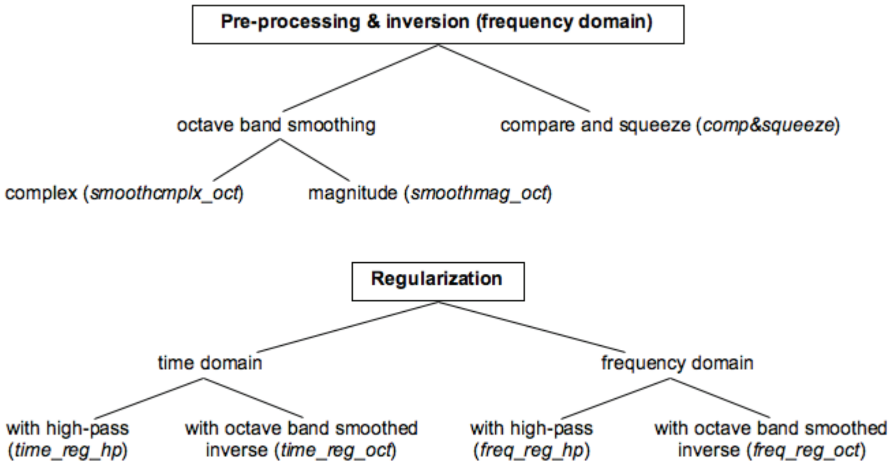

Pre-processing & inversion (frequency domain)
octave band smoothing
compare and squeeze (*comp&squeeze*)
complex (*smoothcmplx_oct*)
magnitude (*smoothmag_oct*)
Regularization
time domain
frequency domain
with high-pass (*time_reg_hp*)
with octave band smoothed inverse (*time_reg_oct*)
with high-pass (*freq_reg_hp*)
with octave band smoothed inverse (*freq_reg_oct*)

Figure 1.18: *Seven inverse filter design techniques examined in Schärer and Lindau (2009).*

1.3 Virtual and mixed reality

Mixed reality (MR) applications anchor rendering processes to the world's reference frame, rather than to the listener's reference frame, as is the case for pure virtual reality (VR). The degree of immersion and the definition of the spatial frame are qualities connected to the concept of *virtuality continuum* introduced in the literature by Milgram et al. (1994) for visual displays. These notions can be adapted to virtual auditory displays (VAD) and augmented audio reality (AAR), including sonic effects and overlaying computer-generated sounds on top of real-time acquired audio signals (Cohen et al., 1993). This section addresses the problem of characterizing immersive sound systems and applications over the different degrees of virtuality. This thesis focuses on headphone-based systems for binaural audio rendering, taking into account that possible disadvantages (e.g. invasiveness, non-flat frequency responses) are counterbalanced by a number of desirable features. Indeed, these systems might control and/or eliminate reverberation and other acoustic effects of the real listening space, reduce background noise, and provide adaptable audio displays, which are all relevant aspects especially in augmented contexts.

Nowadays most of MR systems are commonly able to control two dimensions of an auditory space, i.e. far field sound sources positioned in the horizontal plane referring to a head-centric coordinate system. Head-tracking technologies in conjunction with non-individualized HRTFs, customizable ITD models and virtual reverberation (Begault, 1994) allow for accurate discrimination between sound sources placed around the user and into the defined subspace. The third dimension, elevation or vertical control, requires a user specific characterization in order to simulate effectively position in the vertical plane mainly due to the shape of the external ear, and few spatial audio engines try to address this challenging problem.

1.3.1 Taxonomy of applications

A virtual 3D audio scene, synthesized by a binaural sound reproduction system, can be described in terms of each individual sound source signal and its associated meta-data[15], and then summing the left and the right signals to produce the final stereo signal sent to the earphones.

This architecture can allow for effective scalability depending on the available computational resources or bandwidth. Psychoacoustic criteria can define the rendering priority and attributes in scenario with multiple sources, such as audibility and grouping due to the *precedence effect*. Specifically, in relation to the amount of bandwidth available the least perceivable sources can be removed from the scene and this graceful degradation of the rendering scene would result in a satisfactory experience even in cases of limited quality of service.

In typical virtual audio applications the user's head is the central reference for audio objects rendering. Specifically, the location of the user's head establishes a virtual coordinate system and builds a map of the virtual auditory scene. In the case of a mixed environment, sound objects are placed in the physical world around the user and hence, conceptually, positioned consistently with a physical coordinate system. Thus, locating virtual audio objects into a mixed environment requires the superimposition of one coordinate system onto another.

Depending on the nature of the specific application, several settings can be used to characterize the coordinate system for virtual audio objects and their locations in the environment. A simple distinction is the choice between a global positioning system, or a local one. An ideal classification can help the definition of the possible applications that use spatial audio technologies. In some cases it is necessary to make the two coordinate systems match in a way that virtual sound sources appear in specific locations into the physical environment, while in other applications virtual sources are floating somewhere around the user because in accordance with the conceptual level of user interaction.

This section proposes a characterization that uses a simplified two-dimensional parameter space defined in terms of *degree of immersion* (DI) and *coordinate system deviation* (CSD) from the real world. This point of view is a simplification of the three-dimensional parameter space proposed by Milgram et al. (1994) consisting of

- *Extent of World Knowledge* (EWK): knowledge held by system about virtual and real object shapes and locations;

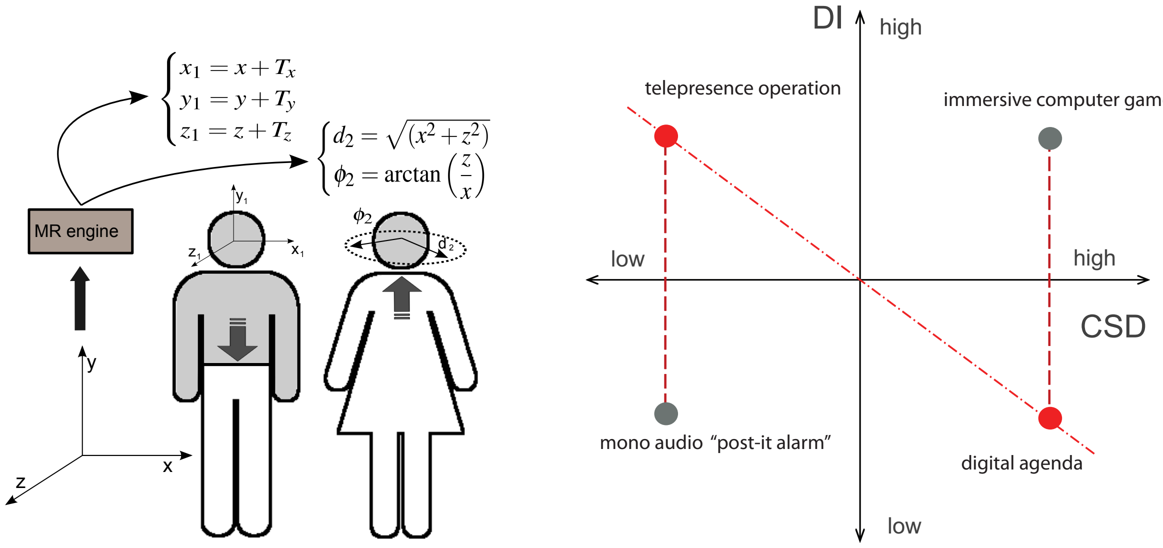
- *Reproduction Fidelity* (RF) - virtual object rendering: quality of the stimuli presented by the system, in terms of multimodal congruency with their real counterpart;

- *Extent of Presence Metaphor*(EPM) - subject sensations: this dimension takes into account observer's sense of presence;

Links between the aforementioned dimensions originate from the interaction of the three main entities involved: the real world, the MR engine and the listener.

The MR engine is an intermediate layer between reality and its representation perceived by the listener; in that sense CSD matches with EWK. A low CSD means a high EWK: the

[15] Information regarding sound source dynamic position and interaction with the environment.



(a) (b)

Figure 1.19: *(a) Coordinates distinction in two mixed reality scenarios. (b) Four applications placed into the taxonomy's parameter space.*

MR engine knows everything about the objects' position in reality and can render the synthetic acoustic scene in such a way that the listener perceives a coherent world.

On the other hand the issue of realism concerns the technologies involved in the MR engine, and the complexity of the taxonomy for such system increases considerably. EPM and RF are not entirely orthogonal and a choice is to define DI according to the following idea: when a listener is surrounded by a real sound, all his/her body interacts with the acoustic waves propagating in the environment, i.e. a technology with high realism rate is able to monitor the whole listener's body position and orientation (high DI).

In order to clarify such distinction, a visual scheme of two different applications is shown in Fig. 1.19. The subject on the left provides an example of high DI corresponding to a high percentage of the body being tracked into the virtual coordinate system (an almost fully gray-filled body), and it exhibits a low virtual coordinate deviation from the physical coordinate system, due to a simple translation. On the contrary the subject on the right exhibits a low DI and a high CSD, represented by a gray head and a listener-centered 2D virtual space.

Concrete examples of the two cases are the following. The male user is in a dangerous telepresence operation, and his head and torso are tracked to immerse his body in a distant real place. The female user is wearing a mobile device and is navigating throughout her many messages and appointments. The first scenario depicts the exact superposition of the virtual and real worlds, on the contrary the latter represents a totally virtual world and in many cases not fully three-dimensional.

1.3.2 Virtual coordinates: example scenarios

The most common case of a floating virtual coordinate system is the one where the only anchor point relative to which the event is localized is the user's head. Usually, virtual sound sources are rendered to different directions and are purely virtual (minimum DI and maximum CSD).

As an example, information services such as news, e-mails, calendar events or other types of messages can be positioned in the virtual acoustic space around the user's head (Brewster et al., 1993). Calendar events, in the form of speech messages, are rendered in different directions depending on the timetable of the user's agenda, so that noon appears in the front (Kajastila and Lokki, 2013). In the case of a hierarchical menu structure presentation, as commonly found on mobile devices, spatial user-interface designs such as the one presented in (Lorho et al., 2002) can be adopted.

Immersive virtual reality applications also use specific virtual coordinate systems, usually related to the geometry of a graphical virtual reality scene. For those computer games that use spatial audio techniques, the virtual coordinate system is defined according to the game scene and sometimes combined with information on the physical location of a user (e.g. head tracking via webcam).

Telepresence is another case of a floating virtual coordinate system and is similar to virtual auditory display systems once focused on the immersive experience of the user. An interesting mixed reality case is the bidirectional augmented telepresence application where a binaural telepresence signal is combined with a VAD (Härmä et al., 2004). The MR engine merges the local pseudoacoustic environment with a remote pseudoacoustic environment to acoustically produce the other person's environment. In this case the CSD related to the remote environment is very low.

In collaborative virtual environments, Benford *et al.* (Benford and Fahlén, 1993) have shown that spatial and visual cues can aid communication in a natural way. The well-known "cocktail-party" effect shows that people can easily monitor several spatialised audio streams at once, selectively focusing on those of interest. In multiparty teleconferencing, the positioning of each talker can be freely done in a virtual meeting room.

Walker and Brewster have explored the use of spatial audio on mobile devices, e.g. for addressing problems of visual clutter in mobile device interfaces (Walker and Brewster, 2000). Their work could help to disambiguate speakers in multi-party conferences, and affords further exploration for gesture-based spatial audio augmentation in mobile multi-party calling scenarios.

1.3.3 Physical coordinates: example scenarios

Once given locations have been established for virtual audio objects in the physical world around the listener,the virtual coordinate system must match with that of the physical environment. It would be ideally possible to put a virtual audio object near any physical object in the real world. Audio messages spatially rendered close to an artwork at the museum or used as an introduction to the exhibition are examples of audio guide systems (Bederson, 1995; Hatala et al., 2004): sound designer can stick an acoustic Post-it in the physical coordinate system. A pre-recorded message can be played or read by a synthetic voice to a visitor when he/she is in a specific location of the museum. The user location and orientation are updated and the acoustic environment monitored as well, resulting in an very high DI, thus an auralized dynamic soundscape and different spoken messages are played through his/her wireless headphones. The above remarks are particularly relevant for mobile applications and eyes-free mobile control.

Systems for aiding the navigation of the visually impaired people represent a socially relevant application of these technologies and virtual maps of the physical space can be global or local, i.e. the specific room (a more detailed review in Ch. 7), introducing a relative relation between CSD and the adopted real coordinate system.

In order to further clarify the concept of local physical coordinate systems, two final examples are mentioned: virtual auditory displays for flight simulators and collision alarm systems in the cockpit. In these applications, the associated physical coordinate system is moving with the airplane and a low CSD has to be considered, because the matching between virtual and physical coordinate systems is locally managed and effective in such domain. Accordingly, if a monophonic alarm signal is employed, a lower DI as compared to spatialized cues characterizes the VAD.

Chapter 2

Mixed structural models

Several techniques for synthetic HRTF design have been proposed during the last two decades. According to Sec. 1.2.1, these can be grouped into two main families: *pole-zero models* (Durant and Wakefield, 2002), in which the HRTF is approximated with low-order rational filters, and *series expansions* (Kistler and Wightman, 1992), in which the HRTF is represented as a weighted sum of simpler basis functions. On a different level of representation stand *structural HRTF models* (Brown and Duda, 1998). In this approach, by isolating the effects of different body components (head, pinnae, ear canals, shoulders/torso), and modeling separately each one of them with a corresponding filtering element, the global HRTF is approximated through a proper combination of all the considered effects (Algazi et al., 2001c). Moreover, by relating the temporal/spectral features (or equivalently, the filter parameters) of each component to corresponding anthropometric quantities, one can in principle obtain a HRTF representation that is both computationally economical and customizable.

More recent research has focused on the problem of HRTF customization for individual subjects. Although most approaches use series expansions with self-tuning of weights (Hwang et al., 2008; Shin and Park, 2008) or simply non-individualized HRTF selection (Seeber and Fastl, 2003; So et al., 2010; Katz et al., 2012), structural HRTF modeling remains the most attractive alternative in terms of both computational efficiency and physical interpretation: parameters of the rendering blocks can be estimated from real data, fitted to low-order filter structures, and related to anthropometric data (Raykar et al., 2005; Satarzadeh et al., 2007).

A novel approach to the modeling of head-related transfer functions (HRTFs) for binaural audio rendering is formalized and described in this chapter. Mixed structural modeling (MSM) can be seen as the generalization and extension of the structural modeling approach first defined by Brown and Duda back in 1998. Possible solutions for building partial HRTFs (pHRTFs) of the head, torso, and pinna of a specific listener are first described and then used in the construction of two possible mixed structural models of a KEMAR mannequin. Thanks to the flexibility of the MSM approach, an exponential number of solutions for building custom binaural audio displays can be considered and evaluated, the final aim of the process being the achievement of a HRTF model fully customizable by the listener.

This chapter is partially based on (Geronazzo et al., 2013c).

2.1 Mixed structural model formalism

In its commonly accepted meaning, the term “head-related” transfer function indicates in fact the full “body-related” transfer function, that also includes acoustic effects of body parts different from the head. Based on this remark, two additional definitions are introduced.

Def. 1 A *partial head-related transfer function* ($pHRTF$) contains acoustic information either recorded by isolating specific body parts (e.g. pinna-related transfer functions (Satarzadeh et al., 2007)), or estimated through DSP techniques from the decomposition of recorded HRTFs. We refer to its inverse Fourier transform as *partial head-related impulse response* ($pHRIR$).

Def. 2 A *synthesized partial head-related transfer function*, $p\widehat{HRTF}$, contains modeled acoustic information related to specific body parts, or computationally generated through acoustic simulations. We refer to its inverse Fourier transform as *synthesized partial head-related impulse response* ($p\widehat{HRIR}$).

The presented approach aims at building a completely customizable structural model through subsequent refinements, ranging from a selection of recorded $pHRTF$s to a totally synthetic filter model. Intermediate steps include mixtures of selected $pHRTF$s and synthetic components.

Let $HRTF_i$ be the individual HRTF set of a subject i. The *mixed structural modeling* (MSM) approach proposed here provides a possible approximation $\widetilde{HRTF}_i$:

$$HRTF_i \overset{MSM}{\leftrightarrow} \widetilde{HRTF}_i. \tag{2.1}$$

Such an approximation is constructed by connecting N components, i.e. N pHRTFs related to different body parts. Typically in structural models N is equal to 3 (head, torso, and pinna components), but it depends on whether some of these components are merged (e.g. in a complete HRTF, $N = 1$), further decomposed (e.g. concha and helix are modeled separately) or supported by additional components (e.g. the ear canal contribution or headphones responses, which are also strictly related to anthropometry).

Each component can be chosen within three different sets:

1. individual components ($pHRTF$s of subject i);

2. selected components ($pHRTF$s of different subjects);

3. modeled components (synthesized $p\widehat{HRTF}$s).

The approximation $\widetilde{HRTF}_i$ will include S selected components, I individual components, and M model components:

$$\widetilde{HRTF}_i = \mathop{\circledast}\limits_{k=1}^{S} pHRTF_{s_k^*} \circledast \mathop{\circledast}\limits_{k=1}^{I} pHRTF_{i_k} \circledast \mathop{\circledast}\limits_{k=1}^{M} p\widehat{HRTF}_{m_k}, \tag{2.2}$$

where

$$\begin{array}{ll} i, s \in \mathcal{S}, & m \in \mathcal{M} \\ I + S + M = N & \end{array}$$

The sets $\mathcal{S}$ and $\mathcal{M}$ represent the collections of subjects and models of which at least one $pHRTF$ or one $p\widehat{HRTF}$ is available; s_k and i_k denote the k^{th} partial component for a subject s and for the target subject i, respectively; m_k denotes the k^{th} modeled component. The $\circledast$ operator relates to the filter representation, and each of its instances denotes series or parallel filter connections.

Selected components in Eq. (2.2) are in general a subset of N components chosen based on the following optimization criterion:

$$\{s_k^*\} = \left\{ s \in \mathcal{S} - \{i\}, k = 1, ..., N |\ s_k \text{ minimizes } e_k^{\mathbf{S}} \right\}. \quad (2.3)$$

Here **S** represents a given selection technique,[1] and $e_k^{\mathbf{S}}$ is the associated selection error for the k^{th} component.

As a particular case, $S = M = 0$ and $I = N$ yields:

$$\widetilde{HRTF}_i = HRTF_i = \bigcircledast_{k=1}^{I} pHRTF_{i_k}. \quad (2.4)$$

Different combinations of S, I, M in the formalism include other relevant cases already proposed in previous literature:

- $S = N = 1$, $I = M = 0$ using a generic subject s: common use of non-individualized $HRTF$s (e.g., only mannequin HRTFs available).
- $S = N = 1$, $I = M = 0$ using one subject s^* that minimizes a given selection error: HRTF selection (Katz et al., 2012).
- $M = N = 1$, $I = S = 0$ using a model m^* that minimizes a given modeling error: direct HRTF modeling without structural decomposition (Durant and Wakefield, 2002).
- $M = N = 3$, $I = S = 0$ using customized models m_k for each component: structural HRTF modeling (Brown and Duda, 1998).

The goal of the MSM approach is twofold:

1. progressively remove all the individual partial components, i.e. $I = 0$, $S + M = N$;
2. provide reliable techniques to pHRTF modeling and pHRTF selection, and to evaluate their combinations (Geronazzo et al., 2013e) towards a complete structural model.

Ultimately, the optimal MSM solution corresponds to the case $M = N$, $I = S = 0$:

$$\widetilde{HRTF}_i = \bigcircledast_{k=1}^{M} p\widehat{HRTF}_{m_k^*}. \quad (2.5)$$

[1] One can also consider techniques based on series expansions with self-tuning of weights and perceptually-driven HRTF selections as candidate selection techniques, even if our focus lies on HRTF selection with respect to anthropometric features.

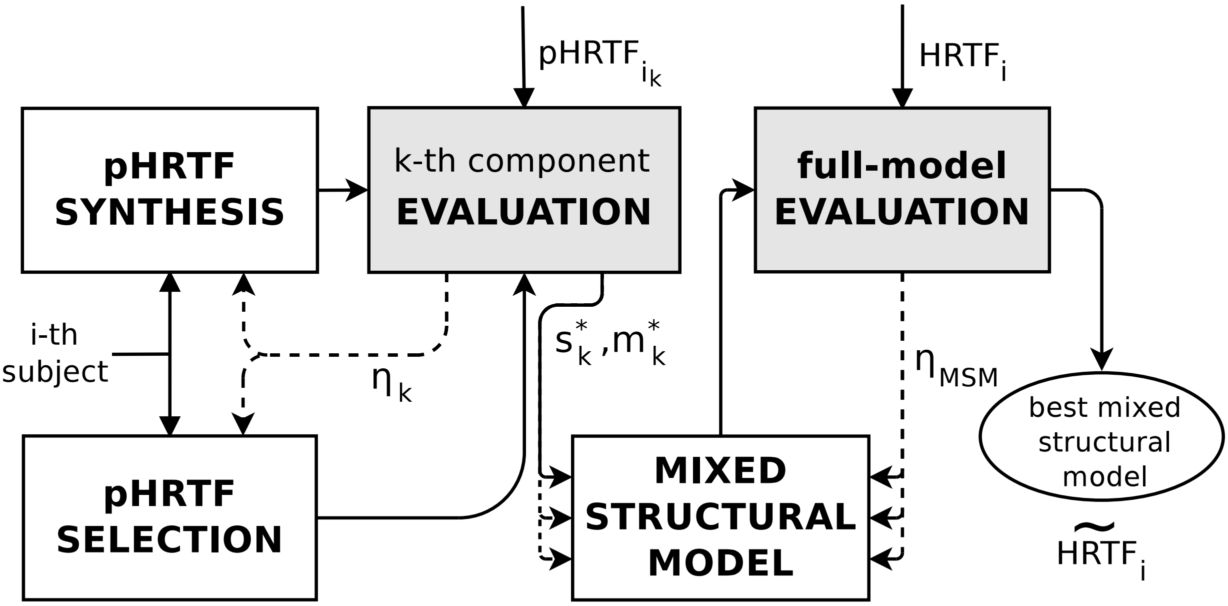


Figure 2.1: *Typical research workflow towards a mixed structural model.*

2.2 Evaluation

The process towards this case considers a wide group of candidate MSMs each described by a set of parameters. Fig. 2.1 depicts the workflow that leads to the construction of a specific MSM in the space of all possible model instances. Given the collections $\mathcal{S}, \mathcal{M}$, and given a test set of subjects with known HRTFs, the evaluation procedure in Fig. 2.1 provides the "best MSM", i.e. the best combination of modeled and selected components, including the relative balance between S and M.

A two-stage evaluation procedure, composed by a *single-component* and a *full-model* evaluation, guides the exclusion of certain instances and combinations of single components. The two fundamental evaluation parameters considered in the first stage are:

- *accuracy* $\alpha_k \in [0, 1]$, defined as the correlation between localization performances of the single $pHRTF_{s^*_k}$ or $p\widehat{HRTF}_{m_k}$ and $pHRTF_{i_k}$;
- *handiness* $\lambda_k \in [0, 1]$, which measures the ease in feeding the single model or selection procedure with individual parameters.[2]

For simplicity, accuracy may be measured on a dimensionally reduced localization space (e.g., for the pinna the error may be measured only on the median plane). These two parameters ultimately define the *efficiency* $\eta_k = \alpha_k \lambda_k$ of the considered m_k, that we aim to maximize:

$$\{m^*_k\} = \left\{ m \in \mathcal{M}, k = 1, ..., N \mid m_k \text{ maximizes } \eta_k \right\}. \tag{2.6}$$

[2] For instance, an acoustically measured individual HRTF implies $\lambda_k = 0$, while the use of a generic HRTF of a different subject has $\lambda_k = 1$ because no individualization is needed. All of the possible customization techniques ranging from the use of MRI scanning to the measurement of simple scalar anthropometric quantities have $\lambda_k = (0, 1)$ in decreasing order of customization burden.

The candidate m_k^* is then compared to the candidate s_k^*. If s_k^* provides an efficiency greater than η_k for m_k^*, it will be chosen as the k^{th} component, otherwise m_k^* will be chosen.

Subsequently, the full-model evaluation takes the best representative solutions of each k^{th} structural component in order to test the combined effects and the orthogonality of the models within full-space 3D virtual scenes. The same two evaluation criteria of the single-component evaluation procedure are used here, where α_{MSM} is the correlation between global localization performances of the resulting $\widetilde{HRTF}_i$ and $HRTF_i$, while

$$\lambda_{MSM} = \prod_{k=1}^{N} \lambda_k. \tag{2.7}$$

The minimization of $\eta_{MSM} = \alpha_{MSM}\lambda_{MSM}$ leads the mixing process over subsequent versions of the MSM.

2.3 Examples

This last section provides two basic examples of the mixed structural modeling approach. In the first one, frontal horizontal-plane HRTFs of a pinnaless KEMAR mannequin are approximated by the combination of a spherical head model parameterized on the mannequin's head dimensions and the selected torso response from the nearest subject in the CIPIC HRTF database with respect to the shoulder width parameter. In the second example, frontal median-plane HRTFs of a full KEMAR mannequin are derived from the application of a pinna model to the related recorded pinnaless responses.

2.3.1 Example #1

Right HRTF magnitudes of a pinnaless KEMAR mannequin in the horizontal plane up to 5 kHz are plotted in Fig. 2.2(d) for $\theta = [-80°, 80°]$, where $\theta > 0$ corresponds to the right hemisphere, hence the ipsilateral side. One can easily detect the different behaviour of the pinnaless mannequin in this zone, where shoulder reflections add up to the direct path of the sound wave, and in the contralateral side, where shadowing and diffraction by the head significantly attenuates any incoming sound.

In order to approximate such behaviour, the contributions of the head and shoulders to the pinnaless response are treated separately and then combined. Concerning the head, the spherical model with custom radius is the most straightforward choice. The optimal radius a^* for the KEMAR head is calculated as in Eq. (3.5), yielding $a^* = 8.86$ cm. A set of HRTFs from a spherical head are then derived from Eq. (3.1) by setting $\rho = 1$ m. These responses are reported in Fig. 2.2(b), where one can detect the substantial direct-path gain in the ipsilateral side and the effects of shadowing and diffraction in the contralateral side. The latter effect is however much shallower than in the pinnaless KEMAR responses and could be attributed to the intrinsic differences between an ideal sphere and a mannequin head, even though their gross behaviour is overall similar.

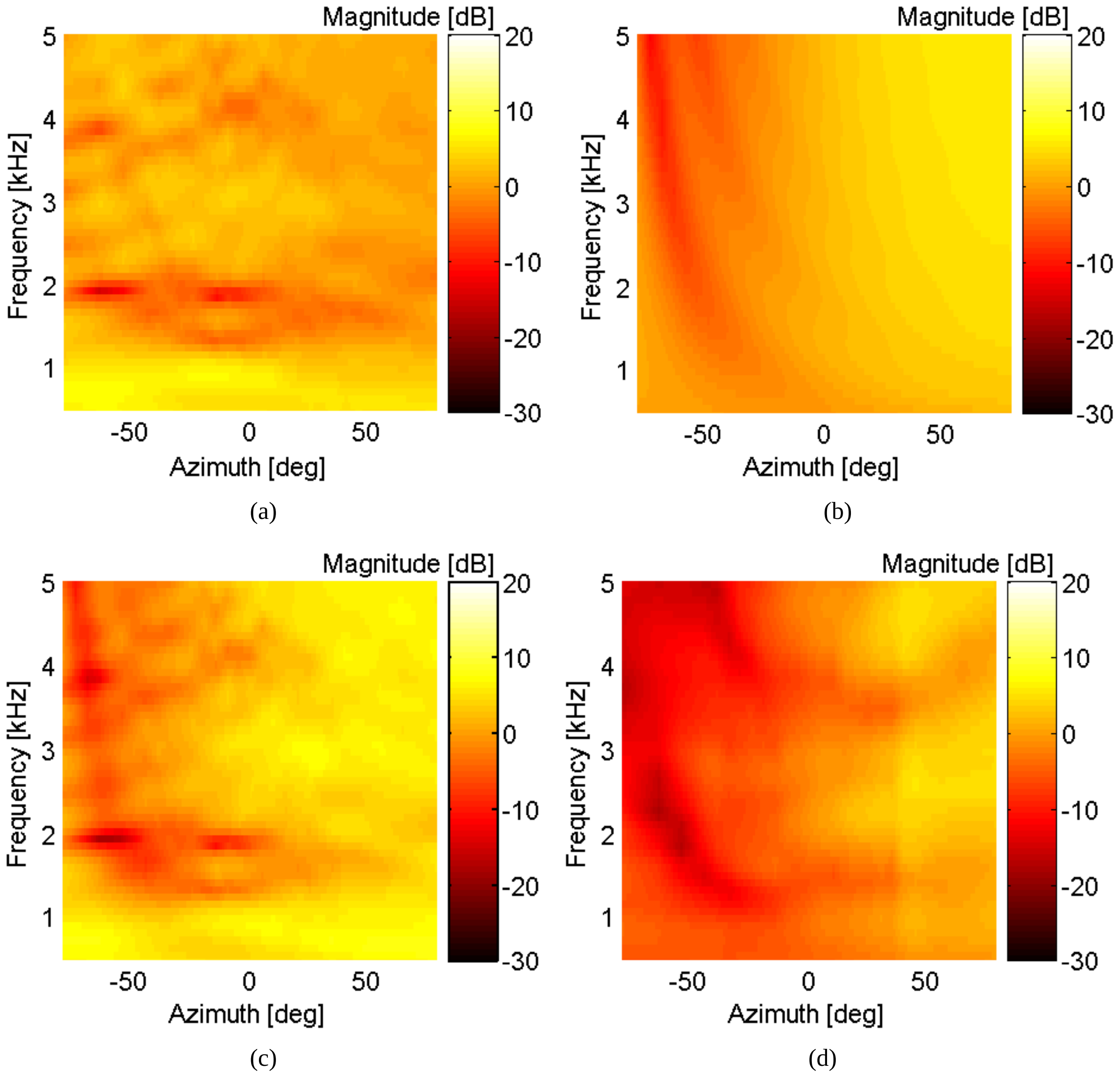


Figure 2.2: *Horizontal-plane right-ear HRTFs* ($\theta = [-80^\circ, 80^\circ]$)*. (a) Extrapolated torso of CIPIC Subject* 156*. (b) Rigid spherical head. (c) Combination of (a) and (b). (d) Pinnaless KEMAR mannequin.*

The shoulder's contribution is instead extrapolated from the HRTFs of the CIPIC database subject (KEMAR excluded) whose shoulder width is the closest to the KEMAR's, i.e. Subject 156. Even though the pinna modifies shoulder reflections, its contribution to the low-frequency range is negligible. For this reason, the torso response - i.e. the shoulder reflection - is isolated by simply subtracting a windowed version of the HRIR (1-ms Hann window) to the full HRIR. The magnitude plot in Fig. 2.2(a) shows a main reflection between 1 and 2 kHz followed by fainter comb-like harmonics in the contralateral side.

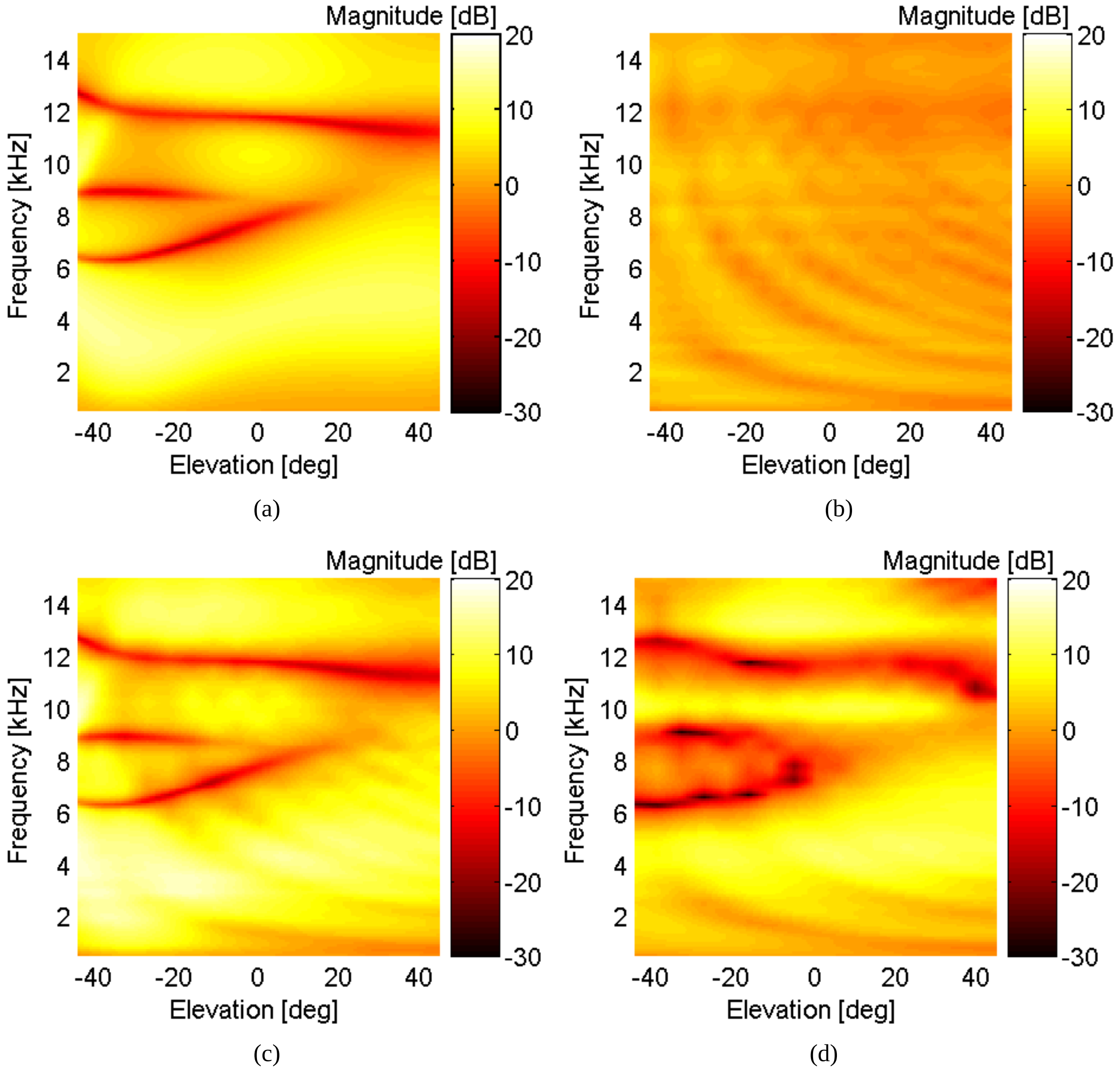
In this first MSM instance $N = 2$, and in particular $M = 1$, $S = 1$, and $I = 0$. The two separate contributions are simply combined by convolving the related HRIRs. The result, reported in Fig. 2.2(c), reveals that the head contribution in the contralateral side fails to overshadow the weak shoulder reflection as it happened in Fig. 2.2(d). The torso contribution is of course different; this is the price to pay when a non-individual response is used. However, the approximated response succeeds in replicating the lowest frequency notch and the gross behavior of the head. Of course, only psychoacoustic tests can evaluate the accuracy of the approximated pinnaless KEMAR responses, subject, however, to the high handiness of both contributions (only 4 anthropometric scalar quantities are needed overall).

2.3.2 Example #2

The pinnaless KEMAR responses used for comparison in the previous example are now used as a structural component of a more complete model including the pinna of the subject. In this case, the main aim is to recreate the full-body HRTFs of a KEMAR mannequin with small pinnae (i.e. Subject 165 of the CIPIC database) in the frontal side of the median plane, the region where the effect of the pinna and its subject intervariability is most prominent (Spagnol et al., 2011).

Median-plane HRTF magnitudes for $\phi = [-45°, 45°]$ of the pinnaless- and full-KEMAR mannequin are reported in Fig. 2.3(b) and Fig. 2.3(d) respectively. A quick comparison of these two plots reveals the massive effect of the pinna in the median plane, that literally overshadows the contributions of the head and torso with its three main notch ridges (beginning approximately at 6.5, 9, and 12.5 kHz) and its resonance patterns, the most prominent of which falls around 4.5 kHz at all elevations. The pinna contribution is provided by the filter model introduced in Sec. 3.3.1, with parameters of the characteristic peak and notch filters derived from an analysis of Subject 165's pinna-related transfer functions (hence not taken from its anthropometry). Transfer functions of this model, reported in Fig. 2.3(a), accurately reproduce the peak/notch patterns of the original response.

In this second MSM instance $N = 2$, and in particular $M = 1$, $S = 0$, and $I = 1$. The pinnaless KEMAR HRIRs are fed to the pinna model yielding the approximated HRTF plot in Fig. 2.3(c). Thanks to the use of individual contributions - either in their original or modeled form - differences between the approximated and original HRTFs are visually forgettable. Of course, despite the allegedly high α_{MSM}, the use of individual contributions pushes λ_{MSM} to 0.



(a)

(b)

(c)

(d)

Figure 2.3: *Median-plane right-ear HRTFs ($\phi = [-45^\circ, 45^\circ]$). (a) Pinna model of CIPIC Subject* 165 *(KEMAR with small pinna). (b) Pinnaless KEMAR mannequin. (c) Combination of (a) and (b). (d) Subject* 165*, full response.*

Chapter 3

Structural partial modeling

The key factor in the design of Mixed Structural Models can be summarized as follows: spatial cues for sound localization characterize each structural component according to a required localization accuracy and parameters' handiness.[1] As a matter of fact, each polar coordinate (azimuth θ, elevation ϕ, and distance r) has one or more dominant cues associated to a specific body component in a given frequency range depending on its physical dimensions. In particular,

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

This chapter will exhaustively describe how those three main components behave in each modeled pHRTF and how such behavior is approximated both in the literature and from the structural models proposed in this thesis.

In particular, an extremely low-order filter model for source distance rendering in binaural reproduction is proposed. The main purpose of such model is to cheaply simulate the effect that source-listener 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 is based on an analytical description of a spherical head response.

The last section proposes a model for real-time pinna pHRTF synthesis which allows to control separately the evolution of two different acoustic phenomena, i.e. ear resonances and reflections, through the design of distinct filter blocks. Parameters to be fed to the model can be derived from mean spectral features in a collection of measured HRTFs and anthropometric features of the specific subject (taken from a photograph of his/her outer ear, see the case study in Ch. 4), hence allowing model customization.

This chapter is partially based on (Geronazzo et al., 2010a; Spagnol et al., 2012a).

[1] Virtual reality applications determine such requirements; the parameter space of the simplified application taxonomy in Sec. 1.3, i.e. the pair := (*degree of immersion*, *coordinate system deviation* from the real world), defines which constrains delimit design and development of an audio engine.

Both models are objectively evaluated and, despite their simplicity, visual analysis of the synthesized HRTFs reveals a convincing correspondence between original and reconstructed spectral features in the chosen spatial range.

3.1 Head models

Azimuth cues can be reduced to two basic quantities[2] thanks to the active role of the head in the differentiation of incoming sound waves, i.e. ITD and ILD. 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. Conversely, frequency-dependent shadowing and diffraction effects introduced by the human head cause ILD to greatly depend on frequency.

Interaural cues are distance-independent when the source is in the so-called *far field* where sound waves reaching the listener can be assumed to be plane. For such ranges, distance dependence in an anechoic space can be approximated by a simple inverse square law for the sound intensity. On the other hand, when the source is in the *near field* interaural cues 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 (Brungart and Rabinowitz, 1999).

3.1.1 The spherical head model

The most recurring head model in the literature is the rigid sphere; therefore its transfer function will be referred to as *spherical transfer function*, or STF. The response related to a fixed observation point on the sphere's surface can be described by means of the following transfer function (Rabinowitz et al., 1993), based on Lord Rayleigh's diffraction formula (Strutt, 1904):[3]

$$H(\rho,\mu,\theta_{\text{inc}}) = -\frac{\rho}{\mu}e^{-i\mu\rho}\sum_{m=0}^{\infty}(2m+1)P_m(\cos\theta_{\text{inc}})\frac{h_m(\mu\rho)}{h'_m(\mu)}, \tag{3.1}$$

where a is the sphere radius, $\rho = r/a$ is normalized distance, 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 normalized frequency, defined as

$$\mu = f\frac{2\pi a}{c}, \tag{3.2}$$

where c is the speed of sound.

[2] For an extensive review of the main acoustic effects introduced by the physical presence of the head, see Sec. 1.1.1

[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.

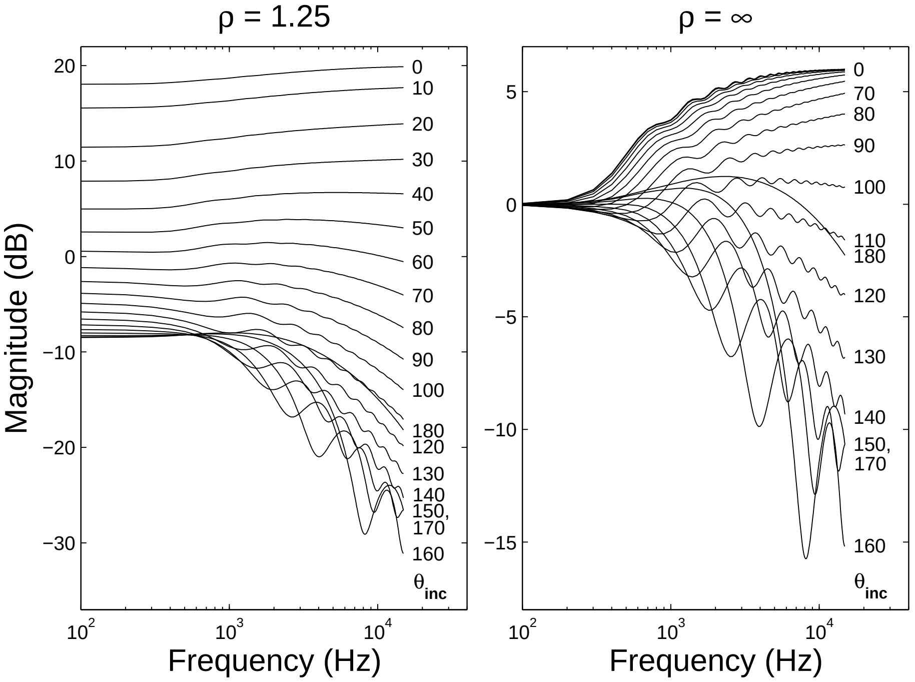


Figure 3.1: *Near-field and far-field Spherical Transfer Functions: $\rho = 1.25$ (left panel) and $\rho = \infty$ (right panel).*

The STF can be evaluated by means of Eq. (3.1) for each $\rho > 1$. A description of the evaluation algorithm, based on recursion relations, can be found in (Duda and Martens, 1998). Fig. 3.1 shows the magnitude of the so calculated transfer function for 19 different values of the incidence angle and two distances, corresponding to near- and far-field.

Previous works (Spagnol and Avanzini, 2009) explored the use of Principal Component Analysis (PCA) in order to study common trends and possible systematic variability in a set of STFs. The results indicated that angular dependence is much more prominent than distance dependence in the transfer function's frequency behaviour. Isolating distance information from the spherical response is thus the first goal towards the design of a cheap and effective model for the STF, as will be performed in the Sec. 3.1.2.

A first-order approximation of the transfer function produced by Eq. (3.1) for $r \to \infty$ was proposed by Brown and Duda (Brown and Duda, 1998) as a minimum-phase analog filter. In order to keep the overall structure as computationally efficient as possible, we choose to use the digital counterpart of the single-pole, single-zero minimum-phase analog filter that approximates head shadowing described in (Brown and Duda, 1998), obtained through the bilinear transform:

$$H_{\text{head}}(z) = \frac{\frac{\beta+\alpha f_s}{\beta+f_s} + \frac{\beta-\alpha f_s}{\beta+f_s} z^{-1}}{1 + \frac{\beta-f_s}{\beta+f_s} z^{-1}}, \tag{3.3}$$

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 (Brown and Duda, 1998),

$$\alpha(\theta_{\text{inc}}) = 1 + \frac{\alpha_{\text{min}}}{2} + \left(1 - \frac{\alpha_{\text{min}}}{2}\right) \cos\left(\frac{\theta_{\text{inc}}}{\theta_{\text{min}}}\pi\right). \tag{3.4}$$

θ_{inc} is the incidence angle that, assuming the interaural axis to coincide with the x axis for sake of brevity, relates to azimuth θ as $\theta_{\text{inc}} = 90° - \theta$ for the right ear and $\theta_{\text{inc}} = 90° + \theta$ for the left ear. For instance, a reasonably good approximation of real diffraction curves for frontal sound sources is heuristically found for parameters $\alpha_{\text{min}} = 0.1$ and $\theta_{\text{min}} = 180°$.

Typically, in spherical models the two observations points (i.e. the ear canals) are assumed to be diametrically opposed. As an alternative model, the spherical-head-with-offset-ears model described in (Algazi et al., 2001b) was obtained by displacing the ears backwards and downwards by a certain offset, introducing a nonlinear mapping between θ_{inc} 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 in measured contralateral HRIRs.

Note that Eq. (3.1) is a function of head radius, a. This is a critical parameter: as an example, a sphere having the same volume of the head approximates its behaviour much better than a sphere with diameter equal to the interaural distance (Katz, 2001b). Hence, in order to fit the spherical head filter model to a specific listener, parametrization of a on the subject's anthropometry shall be performed.

Furthermore, the head radius parameter a, whose value influences the cutoff frequency for the head shadowing, is defined by a weighted sum of the subject's head dimensions using the optimal weights obtained in (Algazi et al., 2001a) through a regression on the CIPIC subjects' anthropometric data. In (Algazi et al., 2001a) the ITD produced by spheres with different radii 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. A linear model for estimating the head radius given the three most relevant anthropometric parameters for the head (width w_h, height h_h, and depth d_h), is fitted to ITD-optimized radii of 45 different subjects through linear regression, yielding the optimal solution

$$a_{opt} = 0.26w_h + 0.01h_h + 0.09d_h + 3.2 \quad \text{cm}. \tag{3.5}$$

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

3.1.2 Near-field model

Relatively simple STF-like filter structures for sound source rendering in the far field have been proposed to date, e.g. Brown and Duda's first-order filter (Brown and Duda, 1998). These models, although replicating with some degree of approximation the mean magnitude characteristics of the far-field spherical response, do not simulate the rippled behaviour occurring for contralateral sources, and, more importantly, have no parametrization on source distance. Nowadays, to the author's knowledge, no real-time model including near-field effects is available in the literature. As a consequence, a proper approximation to distance effects on the spherical head model has to be introduced in order to grant an efficient and effective rendering.

Near-field distance dependence can be accounted for through the filter structure proposed in (Spagnol et al., 2012a), where spatial parameters ρ and θ_{inc} define all the inputs to its single components.

The near-field transfer function

In order to study the effect of source distance in the near field, a given STF can be normalized to the corresponding far field spherical response yielding a new transfer function, which is defined as Near-Field Transfer Function (NFTF):

$$H_{NF}(\rho,\mu,\theta_{\text{inc}}) = \frac{H(\rho,\mu,\theta_{\text{inc}})}{H(\infty,\mu,\theta_{\text{inc}})}. \tag{3.6}$$

Contrarily to STFs, NFTFs are almost non-rippled functions that slightly decay with frequency, in an approximately monotonic fashion. Furthermore, the magnitude boost for small distances is evident in ipsilateral NFTFs whereas it is less prominent in contralateral NFTFs. Such functions are now analyzed in detail in all of their magnitude features with the parallel aim of feeding the modeling process. In the following, parameter a is fixed to the commonly cited 8.75 cm average radius for an adult human head (Brown and Duda, 1998). Final results for a different head radius will just require a uniform rescaling of the frequency axis.

It could be questioned whether analytical NFTFs objectively reflect distance-dependent patterns in measured HRTFs. Unfortunately, most available HRTF recordings are performed in the far field or in its vicinity at one single given distance. Furthermore, a proper NFTF will become more and more sensitive to the geometric features of the complex scatterer (the head) as the sound source approches and, since the sphere can be considered as a simple scatterer, it could become an increasingly worse approximation of the real near field effects. Beside these lawful observations, various studies have verified the suitability of the spherical model for the head in the near field, at least at low frequencies (Zahorik et al., 2005).

DC gain

As a first step towards NFTF analysis, let us look more closely at how the DC gain varies as the source moves away along a given angular direction. For each of 19 incidence angles, $\theta_{\text{inc}} = [0°, 180°]$ at 10-degree steps, Eq. (3.1) is sampled at DC ($\mu = 0$) for a great number of different, exponentially increasing distances, specifically

$$\rho = 1.15^{1+\frac{k-1}{10}}, \quad k = 1,\ldots,250, \tag{3.7}$$

and its absolute value calculated, yielding dB gain $G_0(\theta_{\text{inc}},\rho)$. Fig. 3.2 plots DC gains as functions of distance and incidence angle.

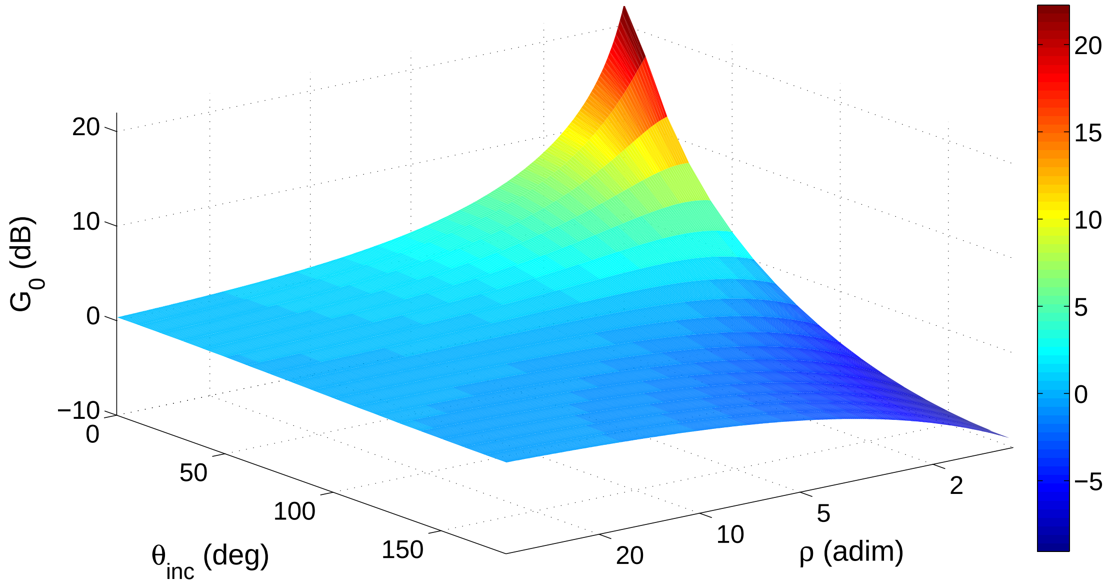


Figure 3.2: *NFTF gain at DC.*

In order to model distance dependence of NFTFs at DC, it is approximated as a second-order rational function for all the 19 different incidence angles. This function, that has the form

$$\tilde{G}_0(\theta_{\text{inc}}, \rho) = \frac{p_{11}(\theta_{\text{inc}})\rho + p_{21}(\theta_{\text{inc}})}{\rho^2 + q_{11}(\theta_{\text{inc}})\rho + q_{21}(\theta_{\text{inc}})}, \tag{3.8}$$

where $\theta_{\text{inc}} = 0°, 10°, \ldots, 180°$, is found with the help of the MatLab Curve Fitting Toolbox (`cftool`).

Coefficients $p_{11}(\theta_{\text{inc}})$, $p_{21}(\theta_{\text{inc}})$, $q_{11}(\theta_{\text{inc}})$, and $q_{21}(\theta_{\text{inc}})$ for each of the 19 incidence angles are reported in Table 3.1, 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 approximation of the resulting rational functions: in all cases, $RMS(G_0, \tilde{G}_0) < 0.01$ dB. In order to model DC gain for intermediate incidence angles, a simple linear interpolation between adjacent functions can be used. The accuracy of such an approximation on a dB scale will be objectively evaluated later in this Section, even for incidence angles different from those considered up to now.

Frequency behaviour

The behaviour of NFTFs at DC having been checked, it has to be studied how much NFTFs depend on frequency and how such dependence can be efficiently modeled. In order to do this, the DC gain G_0 can be used 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 15 kHz, sampled at 10-Hz steps:

$$\hat{H}_{NF}(\rho, \mu, \theta_{\text{inc}}) = \frac{H_{NF}(\rho, \mu, \theta_{\text{inc}})}{G_0(\theta_{\text{inc}}, \rho)}. \tag{3.9}$$

θ_{inc}	p_{11}	p_{21}	q_{11}	q_{21}	RMS [dB]
0°	12.97	−9.691	−1.136	0.219	0.0027
10°	13.19	234.2	18.48	−8.498	0.0223
20°	12.13	−11.17	−1.249	0.346	0.0055
30°	11.19	−9.035	−1.017	0.336	0.0034
40°	9.91	−7.866	−0.83	0.379	0.002
50°	8.328	−7.416	−0.666	0.421	0.0009
60°	6.493	−7.312	−0.503	0.423	0.0002
70°	4.455	−7.278	−0.321	0.382	0.0004
80°	2.274	−7.291	−0.11	0.314	0.0005
90°	0.018	−7.484	−0.13	0.24	0.0005
100°	−2.242	−8.04	0.395	0.177	0.0004
110°	−4.433	−9.231	0.699	0.132	0.0003
120°	−6.488	−11.61	1.084	0.113	0.0002
130°	−8.342	−17.38	1.757	0.142	0.0002
140°	−9.93	−48.42	4.764	0.462	0.0004
150°	−11.29	9.149	−0.644	−0.138	0.0006
160°	−12.22	1.905	0.109	−0.082	0.0003
170°	−12.81	−0.748	0.386	−0.058	0.0002
180°	−13	−1.32	0.45	−0.055	0.0002

Table 3.1: *Coefficients for Eq. (3.8) and approximation fitness.*

Fig. 3.3 shows the frequency behaviour of normalized NFTFs for a fixed small distance, $\rho = 1.25$, and the usual 19 incidence angles. All normalized NFTFs lie in the negative dB space, 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 ipsilateral and contralateral angles: as an example, 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 $\rho = 1.25$ gradually morphs from that of a shelving filter to that of a lowpass filter as the angle increases. The same behaviour is observed for all the other near-field distances, the faster switch rate being observed for small distance values.

In light of such result, one could think of approximating the magnitude response of the normalized NFTF through a shelving or lowpass filter, depending on incidence angle and distance. Unfortunately, the switch from a shelving to a lowpass filter at a given incidence angle needs to be smooth in order to avoid listening artifacts; furthermore, 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 the one proposed in (Zolzer, 2011):

$$H_{\text{sh}}(z) = 1 + \frac{H_0}{2}\left(1 - \frac{z^{-1} + a_c}{1 + a_c z^{-1}}\right), \tag{3.10}$$

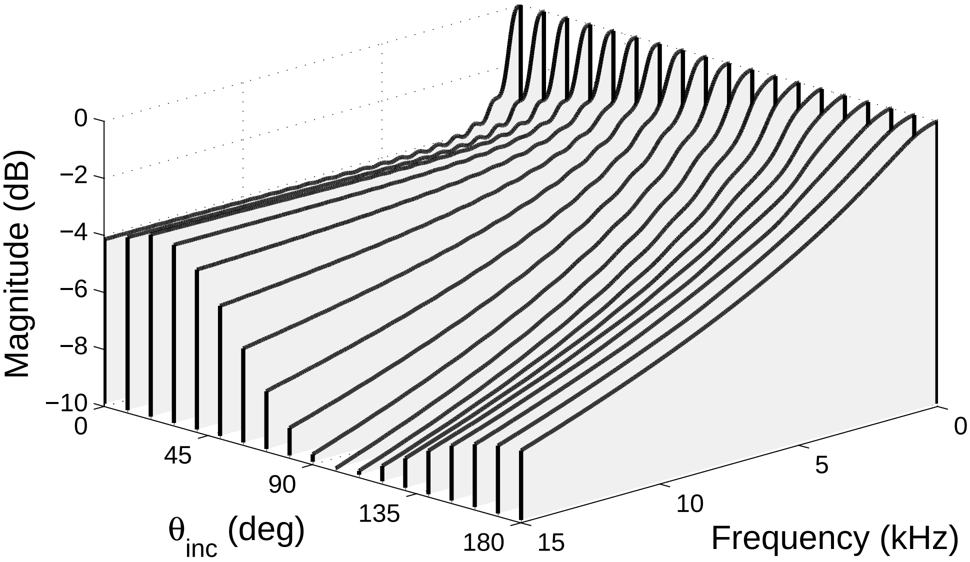


Figure 3.3: *Frequency trend of normalized NFTFs for* $\rho = 1.25$.

$$a_c = \frac{V_0 \tan\left(\pi \frac{f_c}{f_s}\right) - 1}{V_0 \tan\left(\pi \frac{f_c}{f_s}\right) + 1}, \qquad V_0 = 10^{\frac{G_\infty}{20}}, \tag{3.11}$$

where f_s is the 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 $\hat{H}_{NF}$ in order to yield a satisfactory approximation. First, the high-frequency gain is calculated as the (negative) dB gain of the NFTF at 15 kHz. The choice of a high frequency point is needed to best model the slope of near contralateral NFTFs in the whole audible range. Second, the cutoff frequency is calculated as the frequency point where $\hat{H}_{NF}$ has a negative dB gain which approximates two thirds of the high-frequency gain. This point is heuristically preferred 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 well depicted in Fig. 3.4 for three different distances at all incidence angles.

The variation of parameters G_∞ and f_c along distance and incidence angle was also studied. Similarly to what was done for DC gains, a second-order rational function was fitted as follows to the evolution of G_∞ and f_c along distance at given incidence angles:

$$\tilde{G}_\infty(\theta_{\text{inc}}, \rho) = \frac{p_{12}(\theta_{\text{inc}})\rho + p_{22}(\theta_{\text{inc}})}{\rho^2 + q_{12}(\theta_{\text{inc}})\rho + q_{22}(\theta_{\text{inc}})}, \tag{3.12}$$

$$\tilde{f}_c(\theta_{\text{inc}}, \rho) = \frac{p_{13}\rho^2 + p_{23}(\theta_{\text{inc}})\rho + p_{33}(\theta_{\text{inc}})}{\rho^2 + q_{13}(\theta_{\text{inc}})\rho + q_{23}(\theta_{\text{inc}})}. \tag{3.13}$$

Note the choice of a second-order numerator that allows greater flexibility in the approximation of the central frequency behaviour, which is more complex with respect to that of gains. Table 3.2

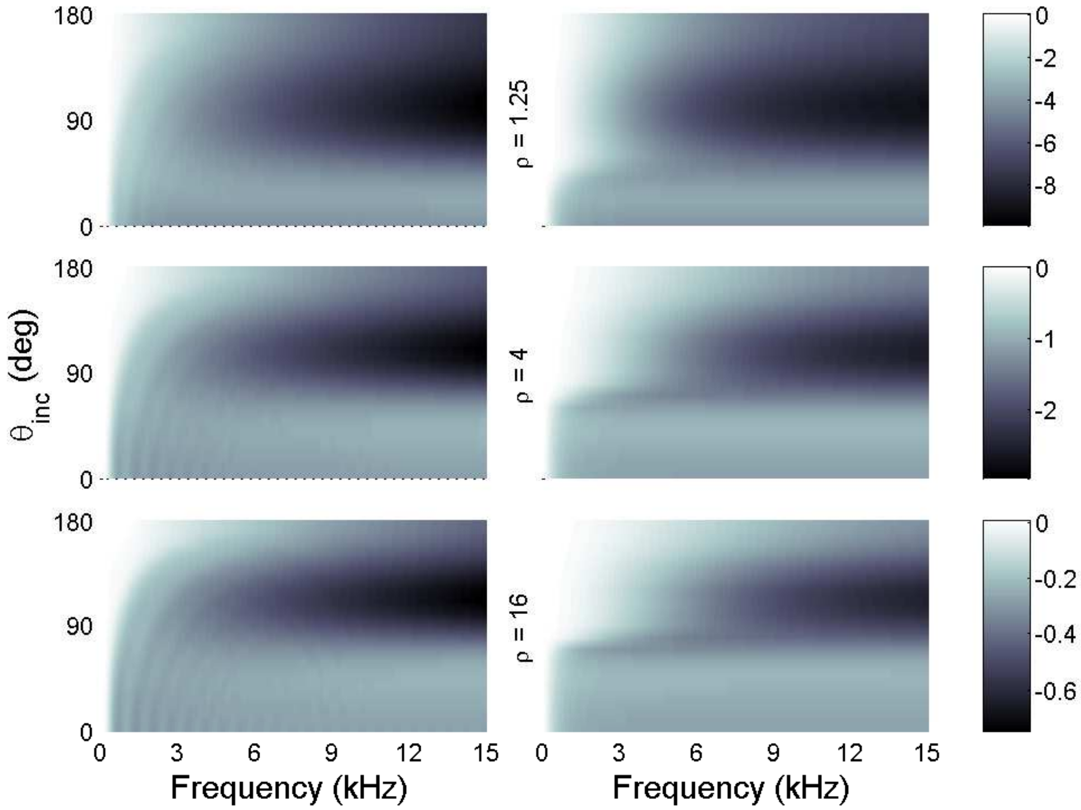


Figure 3.4: *Normalized NFTFs (left panels) versus shelving-filter approximation (right panels) for* $\rho = 1.25$*,* $\rho = 4$*, and* $\rho = 16$*.*

and Table 3.3 summarize fitness scores and parameter values for each of the two functional approximations.

The approximation of G_∞ is overall excellent, never exceeding a mean RMS error of 0.04 dB in the considered angular directions. Similarly, the approximation provided by $\tilde{f}_c$ yields a mean RMS error that is below the actual frequency resolution of 10 Hz for almost 70% of the considered incidence angles. Again, an interpolation of adjacent polynomials analogous to the one used for the DC gain is required to model parameters $\tilde{G}_\infty$ and $\tilde{f}_c$ for intermediate angular values.

Digital filter structure

The analysis performed in the previous section allows the straightforward construction of a filter model for distance rendering, that can be easily integrated with an infinite-distance spherical model of the head following one of the implementations available in the literature (Brown and Duda, 1998). In fact, if the latter is modeled through a filter $H_{\text{sphere}}^{\infty}$ that takes the incidence angle θ_{inc} as input, the information given by the NFTF can be provided by a cascade of a multiplicative gain G_0 and a shelving filter H_{sh}.

θ_{inc}	p_{12}	p_{22}	q_{12}	q_{22}	RMS [dB]
0°	−4.391	2.123	−0.55	−0.061	0.0007
10°	−4.314	−2.782	0.59	−0.173	0.0016
20°	−4.18	4.224	−1.006	−0.021	0.0057
30°	−4.012	3.039	−0.563	−0.316	0.0116
40°	−3.874	−0.566	0.665	−1.129	0.0199
50°	−4.099	−34.74	11.39	−8.301	0.039
60°	−3.868	3.271	−1.571	0.637	0.0151
70°	−5.021	0.023	−0.875	0.325	0.0097
80°	−6.724	−8.965	0.37	−0.083	0.0112
90°	−8.693	−58.38	5.446	−1.188	0.0179
100°	−11.17	11.47	−1.131	0.103	0.0217
110°	−12.08	8.716	−0.631	−0.12	0.0069
120°	−11.13	21.8	−2.009	0.098	0.0018
130°	−11.1	1.91	0.15	−0.401	0.0008
140°	−9.719	−0.043	0.243	−0.411	0.0014
150°	−8.417	−0.659	0.147	−0.344	0.0012
160°	−7.437	0.395	−0.178	−0.184	0.0006
170°	−6.783	2.662	−0.671	0.05	0.0006
180°	−6.584	3.387	−0.836	0.131	0.0008

Table 3.2: *Coefficients for Eq. (3.12) and approximation fitness.*

θ_{inc}	p_{13}	p_{23}	p_{33}	q_{13}	q_{23}	RMS [Hz]
0°	0.457	−0.668	0.174	−1.746	0.699	1.19
10°	0.455	0.142	−0.115	−0.01	−0.348	0.92
20°	−0.87	3404	−1699	7354	−5350	3.36
30°	0.465	−0.913	0.437	−2.181	1.188	7.01
40°	0.494	−0.669	0.658	−1.196	0.256	19.14
50°	0.549	−1.208	2.02	−1.59	0.816	30.67
60°	0.663	−1.756	6.815	−1.296	1.166	21.65
70°	0.691	4.655	0.614	−0.889	0.76	60.32
80°	3.507	55.09	589.3	29.23	59.51	29.59
90°	−27.41	10336	16818	1945	1707	36.16
100°	6.371	1.735	−9.389	−0.058	−1.118	4.54
110°	7.032	40.88	−44.09	5.635	−6.18	2.53
120°	7.092	23.86	−23.61	3.308	−3.392	2.72
130°	7.463	102.8	−92.27	13.88	−12.67	2.33
140°	7.453	−6.145	−1.809	−0.877	−0.19	2.9
150°	8.101	−18.1	10.54	−2.23	1.295	5.28
160°	8.702	−9.05	0.532	−0.96	−0.023	2.15
170°	8.925	−9.03	0.285	−0.905	−0.079	3.71
180°	9.317	−6.888	−2.082	−0.566	−0.398	3.87

Table 3.3: *Coefficients for Eq. (3.13) and approximation accuracy.*

The general filter structure is sketched in Fig. 3.5. Based on distance and incidence angle

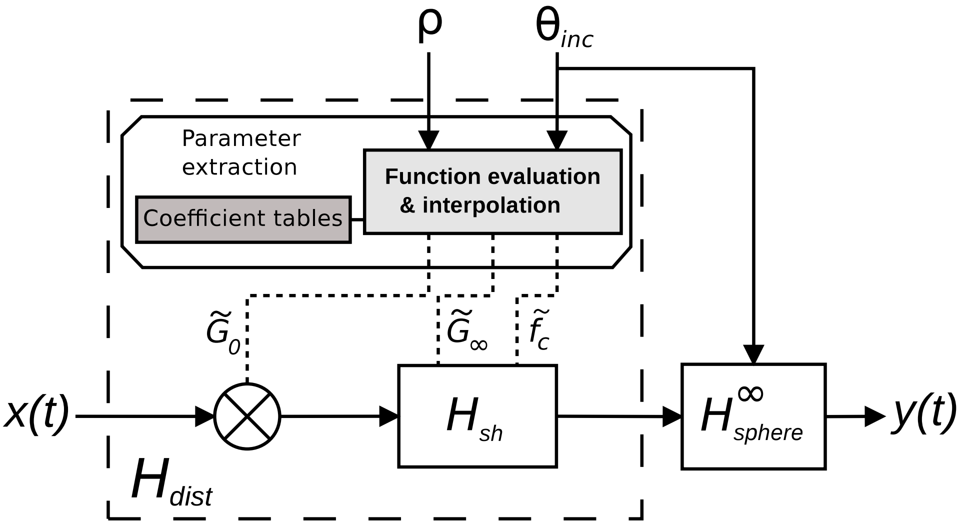


Figure 3.5: *A model for a spherical head including distance dependence.*

information, the "Parameter Extraction" computation block linearly interpolates functions $\tilde{G}_0$, $\tilde{G}_\infty$ and $\tilde{f}_c$ using Eq. (3.8), Eq. (3.12), and Eq. (3.13) respectively; afterwards, $\tilde{G}_0(\theta_{\text{inc}}, \rho)$ is used as multiplicative factor whereas $\tilde{G}_\infty(\theta_{\text{inc}}, \rho)$ and $\tilde{f}_c(\theta_{\text{inc}}, \rho)$ are feeded as parameters to the shelving filter.

A crucial question is the overall goodness of model H_{dist}, that is, an objective measure of how much all the introduced approximations distort the magnitude response of original NFTFs as computed through Eq. (3.1) and Eq. (3.6). The quality of the approximation offered by the model is assessed through a measure of spectral distortion widely used in recent literature (Otani et al., 2009):

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(20\log_{10}\frac{|H(f_i)|}{|\tilde{H}(f_i)|}\right)^2} \quad \text{dB}, \tag{3.14}$$

where H is the original response (here H_{NF}), $\tilde{H}$ is the reconstructed response (here H_{dist}), and N is the number of available frequencies in the considered range, that in this case is limited between 100 Hz and 15 kHz. The error measure was calculated either for spatial locations that were considered during the analysis process and new ones, thanks to the functional representation and interpolation of the key parameters. Specifically, the magnitude of H_{dist} was computed via the model 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. The distance-dependent spectral distortion plot for the 37 considered incidence angles is shown in Fig. 3.6.

Notice that the overall fitness of the approximation is excellent in almost the whole near field, being the SD lower than 1 dB in all of the considered source locations except for the very nearest ones around $\theta_{\text{inc}} = 90°$. This discrepancy is well explained by the frequency behaviour of the normalized NFTF at these positions which is halfway between that of a shelving filter and of a lowpass filter. Also note that there is no evident relative SD increase between reconstructed

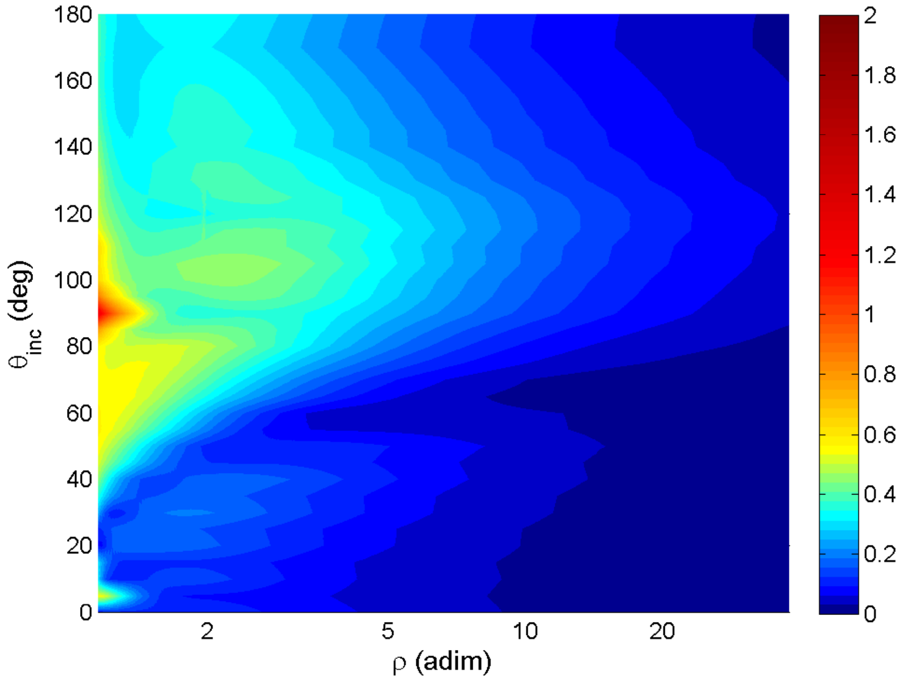


Figure 3.6: *Spectral distortion introduced by model H_{dist}.*

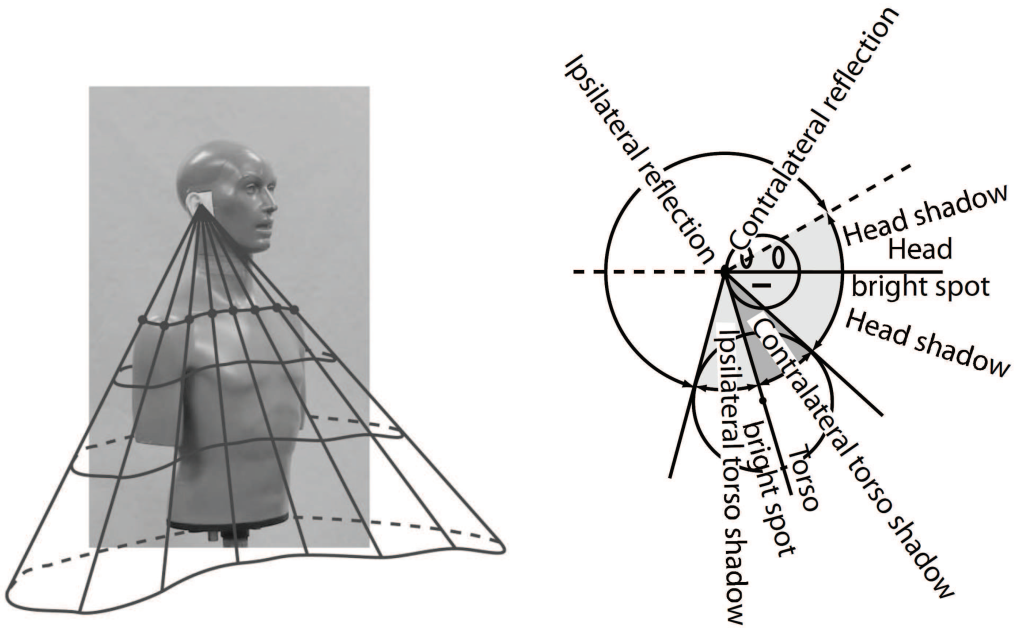
NFTFs for angles that were considered in the analysis process and for interpolated angular directions, except for a small dump at $\theta_{\text{inc}} = 5°$ and $\rho < 1.5$. As a consequence, linear interpolations of the key coefficients are already effective as they are, not needing to be improved through higher-order interpolations and/or a denser sampling along the angular axis during the analysis process.

Finally, the almost null SD for high ρ values indicates that near-field effects gradually dissolve with distance just like in the analytical NFTF, i.e. the contribution of H_{dist} tends to 1 as $\rho \to \infty$. This result confirms the validity of the above model for the whole spatial range, including the far field.

3.1.3 Alternative head models

The spherical model of the head provides an excellent approximation to the magnitude of a measured HRTF (Mokhtari et al., 2009). Still, it 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 (Duda et al., 1999). ITD estimation accuracy can be improved by considering an ellipsoidal head model that can account for the ITD variation and be adapted to individual listeners. A drawback of this formulation is that the analytical solution for the ITD is 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 (Novy, 1998; Jo et al., 2008a) as the sole alternative analytical model to a sphere. Although adding nothing new in



(a) (b)

Figure 3.7: *Mean height estimations (a) torso-shadow cone for the KEMAR mannequin (b) zone for the right-ear response of a snowman model*

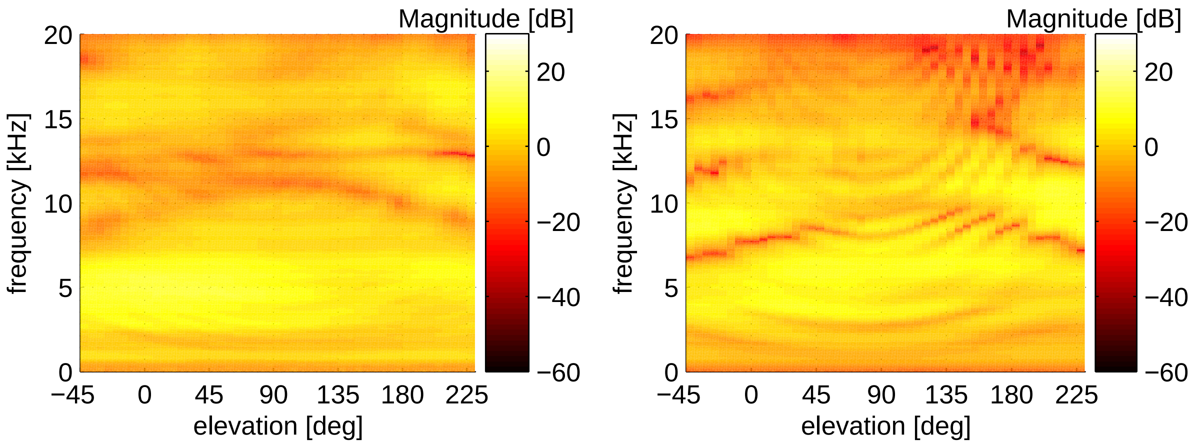
the ITD's point of view, comparison of spheroidal HRTFs against spherical HRTFs revealed a different behaviour in head-induced low-frequency ripples in the magnitude response at the contralateral ear, which is closer to responses of a KEMAR head (Burkhard and Sachs, 1975) for the spheroidal case (Jo et al., 2008b). Still, this model has been very little studied, and consistent advantages over the spherical model have not been made clear.

3.2 Torso models

At wavelengths where ray tracing is valid, torso generates specular reflections for sound sources above the shadow cone, i.e. the rays starting from the ear and tangent to the torso. Sound sources within this area can be divided into an ipsilateral and a contralateral zone shadowed by the so called *torso bright spot*. Fig. 3.7 schematically illustrates the aforementioned torso effects in a KEMAR mannequin (left panel) and in a spherical head and torso approximation (right panel).

The shoulder reflection translates into a series of comb-filter notches in the frequency domain (Algazi et al., 2002b). Nevertheless, the relative strength of this reflection with respect to the direct path of the sound wave seems to depend on both the subject's clothing (Kuhn, 1977) and his/her upper torso dimensions. For instance, as Fig. 3.8 demonstrates, lateral HRTFs of the two CIPIC database.[4] subjects having the smallest (a) and largest (b) shoulder width exhibit

[4] http://interface.cipic.ucdavis.edu/



(a) (b)

Figure 3.8: *Left HRTFs of CIPIC subjects* 018 *(a) and* 050 *(b) for* $\theta = -65°$.

different behaviours of the shoulder reflection (visible as peculiar arch-shaped patterns along elevation)

Furthermore, recent studies in the field of biometric identification have investigate head-to-shoulder signature for person recognition (Kirchner et al., 2012) and their promising results support customization of torso models in order to provide effective localization cues.

3.2.1 The snowman model

Similary to the head, in previous works the torso has been approximated by a sphere. 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 *snowman model* (Algazi et al., 2002b). The far-field behaviour of the snowman model was studied in the frontal plane both by direct measurements on two rigid spheres and by computation through multipole reexpansion (Algazi et al., 2002a). A structural head-and-torso model was also derived from the snowman model (Algazi et al., 2002b); its structure, illustrated in Fig. 3.9, distinguishes the two cases where the torso acts as a reflector or as a shadower, switching between the two filter sub-structures as soon as the source enters or leaves the torso shadow zone, respectively.

Additionally to the spherical model, an ellipsoidal model for the torso was studied in combination with the usual spherical head. This was done either by ray-tracing analysis (Algazi et al., 2001b) or through *boundary element method*(BEM) (Algazi et al., 2002a). Such model is able to account for different torso reflection patterns; listening tests confirmed that this approximation and the corresponding measured HRTF gave similar localization performances, showing larger correlations away from the median plane. Also, the ellipsoidal torso can be easily customized for a specific subject by directly defining control points for its three axes on the subject's torso (Algazi et al., 2002a).

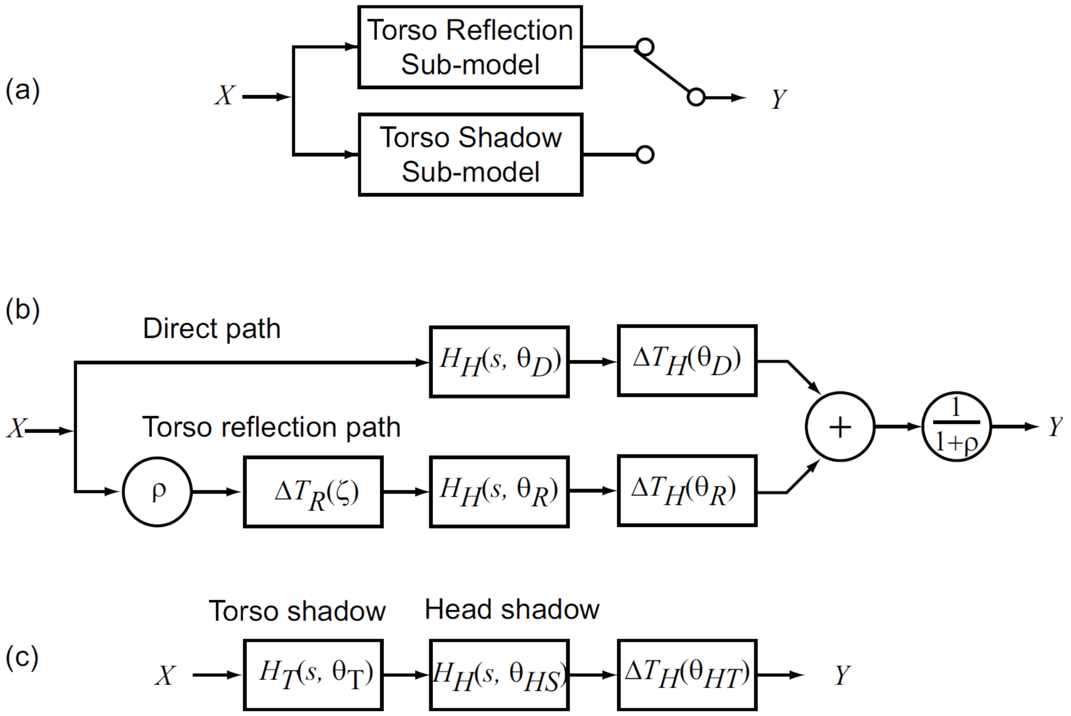


Figure 3.9: *The filter structure for the snowman model. (a) Major components. (b) The torso reflection sub-model. (c) The torso shadow sub-model. (figure reproduced from (Algazi et al., 2002b))*

3.3 External ear models

Following the structural modeling approach, this section investigates the contribution of the external ear to the HRTF, formally defined pinna pHRTF but also known as *Pinna-Related Transfer Function* (PRTF). 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 (Batteau, 1967) – and anthropometry is not fully understood. Recent related works (Katz, 2001b; Kahana and Nelson, 2007; Mokhtari et al., 2011) adopt a physical modeling approach in which PRTFs are simulated through computationally intensive techniques, such as finite-difference time-domain (FDTD) methods, or BEM. By altering the geometry of the pinna in the simulations, the relationship between PRTF features and anthropometry can be investigated experimentally.

Other works (Kistler and Wightman, 1992; Hwang et al., 2008; Evans et al., 1998) utilize series expansions, such as principal component analysis (PCA) or surface spherical harmonics (SSH) representations of HRTFs and PRTFs.

The history of structural models, one of which will be described in this section, begins with Batteau's reflection theory (Batteau, 1967). Following Batteau's observations, Watkins (Watkins, 1978) designed a very simple double-delay-and-add time-domain model of the pinna where the first reflection path is characterized by a fixed time delay of 15μs while the second path includes an elevation-dependent delay calculated from empirical data. Besides considering a very limited amount of reflections, no method for extracting parametric time delays and gain factors was proposed. Furthermore, 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 (Lopez-Poveda and Meddis, 1996). Nonetheless, the pioneering novelty of such model is undisputed.

A similar time-domain structural model, proposed by Faller II et al. (2010), is composed of multiple parallel reflection paths each including a different time delay, a gain factor, and a low-order resonance block. The model is fitted by decomposing a measured HRIR into a heuristic number of damped and delayed sinusoidals (DDS) using an adaptation of the Hankel Total Least Squares (HTLS) decomposition method, and associating the parameters of each DDS to the corresponding parameters of its relative model path. Still, no relation between model parameters and human anthropometry was explicitly found.

Moving from time domain to frequency domain, the approach followed by Satarzadeh et al. (2007) approximates PRTFs at elevations close to zero degrees through a structural model composed of two low-order bandpass filters and one comb filter which account for two resonance modes (Shaw's modes 1 and 4) and one main reflection, respectively. What's more relevant, a cylindrical approximation to the concha is exploited 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. The authors show that their model has sufficient adaptability to fit both PRTFs with rich and poor notch structures. One limitation 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 the only valuable anthropometry-based pinna model available to date.

According to Batteau's theory (Batteau, 1967), 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 direct wave and the reflected waves, sharp notches can be observed in the spectrum at high frequencies with a periodicity of $1/\tau_i$, where τ_i is the time delay of the i-th reflection. Such observation gave birth to a very simple multipath model of the pinna (see Fig.3.10), where the considered reflection paths are characterized by fixed reflection coefficients ρ_A and ρ_V, a fixed time delay τ_A and an elevation-dependent time delay τ_V. The fit with experimental data was found to be reasonably good; however, fixed reflection coefficients overestimate the effective number of notches in the spectrum.

A similar approach was adopted by Faller II et al. (2006), whose model includes four parallel paths that represent the multiple bounces of the incoming sound wave on the geometrical structures of the pinna, each modeled by a time delay τ_i and a magnitude factor ρ_i explaining energy loss. Furthermore, since sound waves are also influenced by the effect of the pinna cavities acting as resonators, the reflection structure is cascaded to a low-order resonator block. The model parameters are fitted by decomposing each specific measured head-related impulse response into

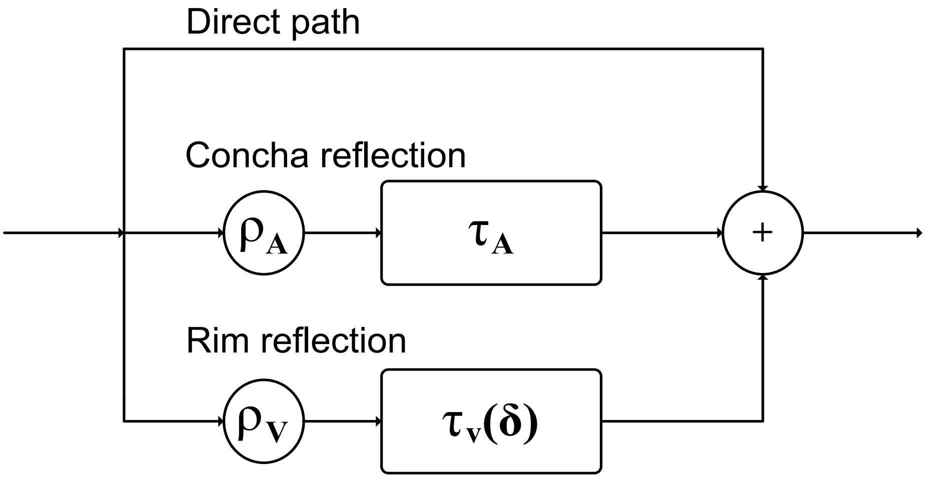


Figure 3.10: *Batteau's multipath pinna pHRTF model.*

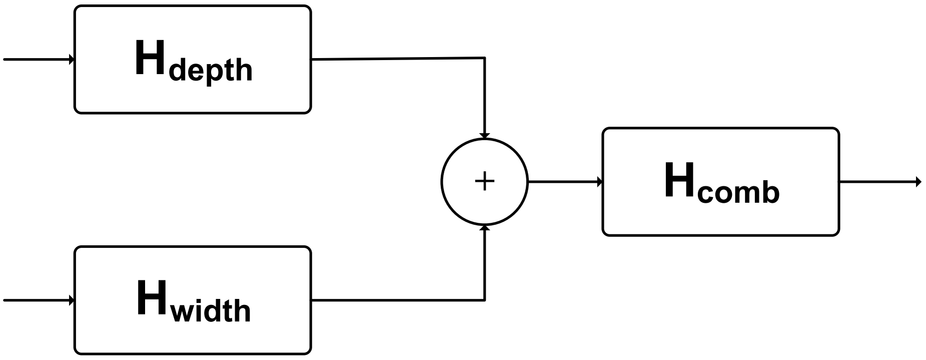


Figure 3.11: *Satarzadeh et al. (2007)'s pinna pHRTF model.*

four scaled and delayed damped sinusoidal 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. Multiple regression analysis was used in order to link the model parameters to eight measured anthropometric features (Gupta et al., 2004). Unfortunately, the considered measures are hard to obtain (a 3-D laser scanner is required); nevertheless, their work definitely emblematizes what this section assumes the typical pinna pHRTF model to be, that is, a "resonance-plus-delay" structure.

The approach taken by Raykar et al. (2005) for reflection modeling is different and operates both in the time and frequency domains. Moved by the observation that raw pole-zero models merely approximate the HRTF spectrum envelope without bringing out the specific features that one is looking for in the impulse response, the authors used robust digital signal processing techniques based on the residual of a linear prediction model for the HRIR to extract the frequencies of the spectral notches due to the pinna alone. Specifically, 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. The authors proved that the spectral notches estimated with this technique are related to the shape and

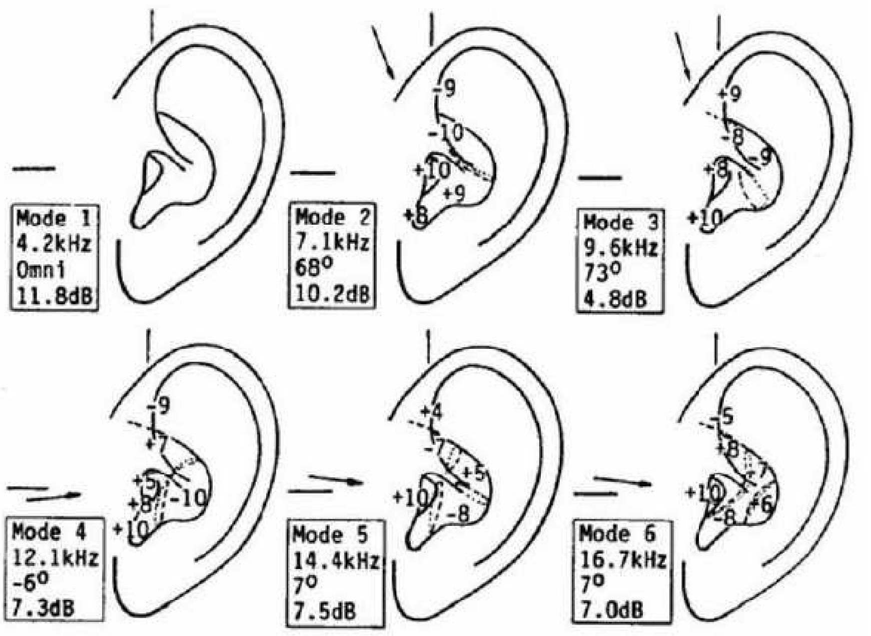


Figure 3.12: *The six pinna modes identified by Shaw (figure taken from (Shaw, 1997)).*

anthropometry of the pinna: for each of the extracted notches the corresponding distance was plotted on the image of the pinna, and by varying elevation such mapping appeared consistent with reflections on the back of the concha and on the crus helias. Spectral peaks were extracted in parallel by means of a linear prediction analysis, yielding results which match quite well the pinna resonant modes reported by Shaw (Shaw, 1997) (see Fig.3.12) and further justifying the "resonance-plus-delay" approach.

Yet another relevant contribution on low-cost modeling of pinna pHRTFs was provided by Satarzadeh et al. (2007). Here pinna pHRTFs for elevations close to zero degrees are synthesized through a model composed of bandpass and comb filters, which respectively approximate resonances and notches. Two major resonances related to the concha shape, specifically Shaw's resonant modes 1 (quarter wavelength concha depth mode) and 4 (concha width mode), and one main reflection are considered.

The two second-order bandpass filters and the comb filter are interconnected as in Fig. 3.11, the latter taking the form $[1 + \rho \exp(-s\tau)]$, where τ is the time delay of the considered reflection estimated from the spacing of notches in the pinna pHRTF spectrum and ρ a frequency-dependent reflection coefficient which strongly attenuates low-frequency notches, coming over Batteau's model aforementioned limitation. The model was proved to have sufficient adaptability to fit both pinna pHRTFs with rich and poor notch structures; furthermore, a cylindrical approximation to the concha was used with the purpose of directly parameterizing its coefficients. Specifically, depth and width of the cylinder uniquely define the depth resonance, while the width resonance is thought to be correlated to the time delay τ depending on whether the concha or the rim is the significant reflector. Although the anthropometric significance of the two parameters is

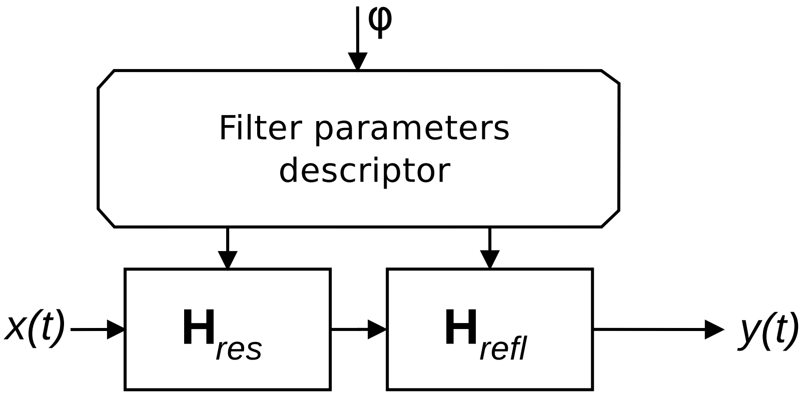


Figure 3.13: *The resonances-plus-reflections decomposition of PRTF model.*

not robust, 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 the experimental pinna pHRTF.

3.3.1 A new approach towards pinna pHRTF modeling

Satarzadeh et al. (2007) has shown that a very simple model can incorporate the gross magnitude characteristics of the pinna pHRTF, via straightforward parametrization on two physical measures. Unfortunately, besides considering solely the frontal direction of the sound wave, taking into account a single reflection seems to be a limiting factor: indeed, pinna pHRTFs with a poor notch structure do not exhibit a clear reflection in the impulse response. In addition, a cylinder is not an adequate physical model for the determination of the width mode of the pinna. Conversely, Raykar et al. (2005)'s work features a very accurate procedure for determining spectral notches in pinna pHRTFs which provides interesting insight on the understanding of pinna reflections. Nevertheless, no actual pinna pHRTF model is defined.

This section introduced a structural model based on a resonances-plus-reflections decomposition of pinna pHRTFs. Ch. 4 will prove that this approach results in an overall good accuracy for the approximation of measured data, thanks to an algorithm able to estimate the two distinct components; accordingly, a suitable anthropometric parametrization will be provided.

Different physical and structural models of the pinna have been proposed in the past, an exhaustive review of which can be found in the previous section. Restricting the attention to points near the median plane, we propose a pinna filter realization that acts as a synthetic PRTF (schematically reported in Fig. 3.13), consisting of two second-order peak filters (filter structure H_{res}) and three second-order notch filters (filter structure H_{refl}) synthesizing two resonance modes and three pinna reflections respectively. The associated parameters (peak/notch central frequency, bandwidth, and gain) are computed by evaluating a number of elevation-dependent polynomial functions constructed from single or average PRTF measurements or derived from the subject's anthropometry (see Sec. 4.4).

Spagnol et al. (2013a) exploited a simple ray-tracing law to show that in median-plane frontal

HRTFs the frequency of the spectral notches, each assumed to be caused by its own reflection path, is related to the shape of the concha, helix, and antihelix. This result allows direct parametrization of the reflective component of the pinna model onto the subject's anthropometry presented under the form of one or more side-view pictures of his/her head. Spectral distortion between real and synthesized PRTFs indicated that the approximation provided by the pinna model is objectively satisfactory.

The only independent parameter used in this model 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 ($F_n^j(\phi)$, $B_n^j(\phi)$, $G_n^j(\phi)$, $j = 1, 2, 3$). For a given subject, these parameters are straightforwardly estimated from the resonant or reflective component of median-plane PRTFs for all the available ϕ values.[5] An analytic description of peaks and notches, i.e. $\mathcal{P}_p^i$ 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 parametrization of the filters. Obviously, all the functions must be determined and computed offline previous to the rendering process.

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 (Zolzer, 2011)

$$H_{\text{res}}^{(1)}(z) = \frac{1+(1+k)\frac{H_0}{2}+l(1-k)z^{-1}+(-k-(1+k)\frac{H_0}{2})z^{-2}}{1+l(1-k)z^{-1}-kz^{-2}}, \tag{3.15}$$

where

$$k = \frac{\tan\left(\pi\frac{B_p^1(\phi)}{f_s}\right)-1}{\tan\left(\pi\frac{B_p^1(\phi)}{f_s}\right)+1}, \qquad l = -\cos\left(2\pi\frac{F_p^1(\phi)}{f_s}\right), \tag{3.16}$$

$$V_0 = 10^{\frac{G_p^1(\phi)}{20}}, \qquad H_0 = V_0 - 1, \tag{3.17}$$

and f_s is the sampling frequency. The second peak ($i = 2$) is implemented as in (Orfanidis, 1996),

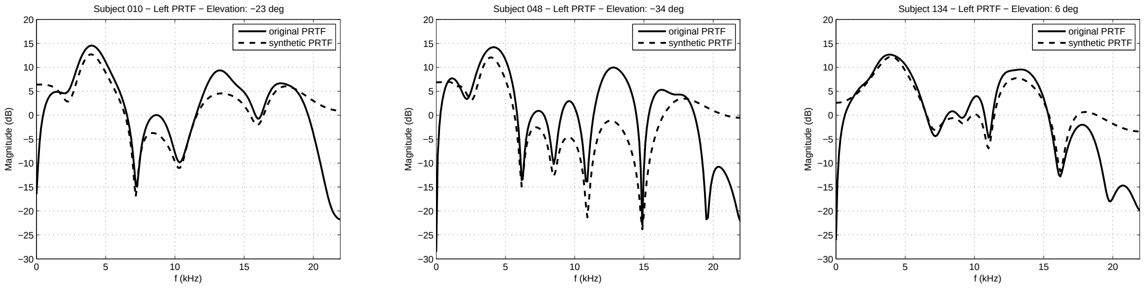
$$H_{\text{res}}^{(2)}(z) = \frac{V_0(1-h)(1-z^{-2})}{1+2lhz^{-1}+(2h-1)z^{-2}}, \tag{3.18}$$

$$h = \frac{1}{1+\tan\left(\pi\frac{B_p^2(\phi)}{f_s}\right)}, \tag{3.19}$$

while l and V_0 are defined as in Eqs. (3.16) and (3.17) with polynomial index $i = 2$. The reason for this distinction lies in 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, analytic functions $\mathcal{P}_p^1$

[5] 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) Subject 010, elevation -23 deg. (b) Subject 048, elevation -34 deg. (c) Subject 134, elevation 6 deg.

Figure 3.14: *Original vs Synthetic PRTF plots for three subjects at three different elevations.*

must be substituted by the corresponding notch counterparts $\mathcal{P}_n^j$, $j = 1, 2, 3$, and parameter k defined in Eq. (3.16) replaced by its "cut" version

$$k = \frac{\tan\left(\pi \frac{B_n^j(\phi)}{f_s}\right) - V_0}{\tan\left(\pi \frac{B_n^j(\phi)}{f_s}\right) + V_0}. \tag{3.20}$$

The cascade of the three notch filters yields a global sixth order multi-notch filter.

Fig. 3.14 shows an example of responses obtained with the proposed model. Specifically, the figure reports a comparison between real PRTF magnitudes estimated for three subjects of the CIPIC database[6] and synthesized PRTF magnitudes for three elevations:[7] the synthesis results are based on relevant parameter values, i.e. 3dB bandwidths, gains and central frequencies of all peaks and notches of a given PRTF obtained with the structural decomposition algorithm presented in Ch. 4. Note that these examples visually demonstrate the versatility of the model:

- the number of rilevant notches can be arbitrarily determined according to their psychoacoustic meaning; in this visual example, the number of notch filters is greater than three only to prove the reliability to signal approximation;

- since the effects of the PRTF are limited to the frequency range $3 - 18$ kHz, peak filter realization can be based both on Eq. (3.15) and Eq. (3.18), depending the filter structure of the whole MSM at low and high frequencies; in this example, one is free to cascade a high-pass filter which cuts out undesired frequencies.

The accuracy of the approximated PRTFs varies from case to case. For instance, while in Fig. 3.14(c) and, to a lesser extent, in Fig. 3.14(a) the re-synthesized PRTFs closely follow the original PRTFs' peaks and valleys, the approximation in Fig. 3.14(b) exhibits a systematic underestimation of resonance magnitudes. Such effect is due to the high number of deep frequency notches that appear at low elevations, for which the proposed multi-notch filter construction

[6] http://interface.cipic.ucdavis.edu/sound/hrtf.html

[7] No customization is realized in this section; Further considerations are presented in Ch. 4.

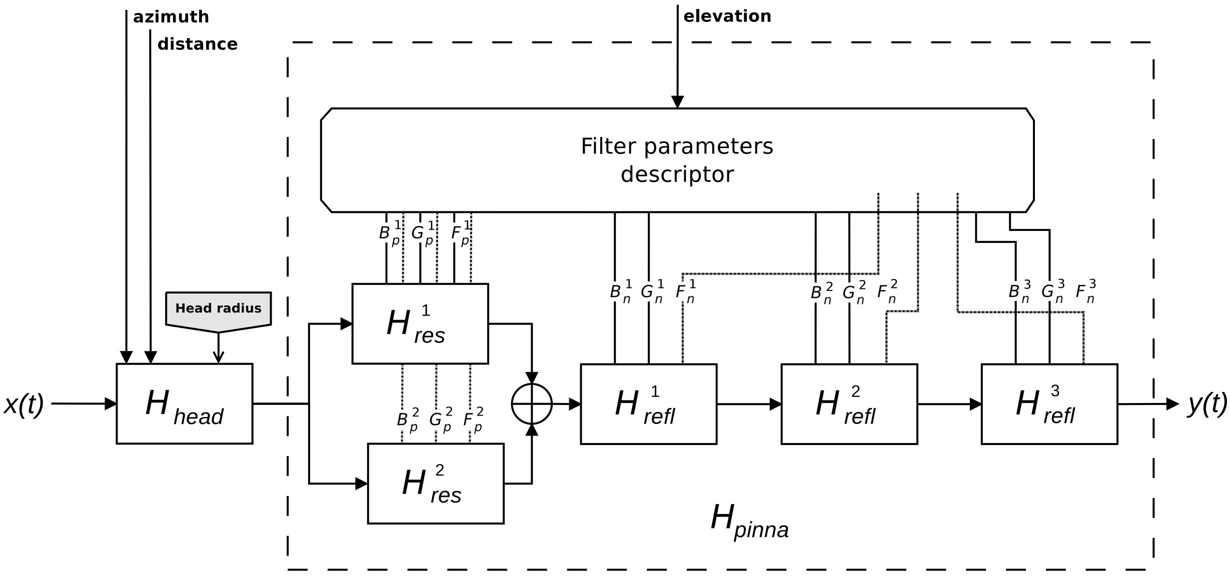


Figure 3.15: *The structural HRTF model.*

procedure turns out to be inadequate. As a matter of fact, interfering bandwidths results in underestimating resonances by several dBs. Using a different multi-notch filter design, which gives prominence to the bandwidth specification, during synthesis would grant a better rendering of resonances, to the detriment of notch depth accuracy. On the other hand, one can employ higher order filters for notches, to the detriment of the computational costs of such model. Nevertheless, only psychoacoustic evaluations are able to confirm the most suitable filter structure and realization.

Concerning source positions above the listener, the attenuation of frequency notches with increasing elevation observed in the literature (Raykar et al., 2005; Geronazzo et al., 2010a) and directly in HRTF sets suggests that notches could simply be gradually extinguished starting from $\phi = 45°$ up to $\phi = 90°$ while keeping their central frequency fixed. However, particular care should be reserved to the modeling of resonances in this elevation range, where the second peak generally disappears in favour of a broader first peak (Geronazzo et al., 2010a). Finally, the role of notches for posterior sources is not completely understood in current literature, although a regular presence of spectral notches has been observed in posterior HRTFs too (Kahana and Nelson, 2007). An assessment of the applicability of the ray tracing procedure to this elevation range is therefore left for future work.

3.3.2 Head and pinna structural model

In this Section, an extension of Satarzadeh et al. (2007)'s structural filter model for binaural rendering of frontal sound sources is proposed. The model includes acoustic contributions by head and pinna into two separate structures (see Fig. 3.15) thus exploiting the flexibility of the partial structural modeling approach.

Focusing on median-plane (azimuth angle $\theta = 0°$) HRIRs with elevation ϕ varying from

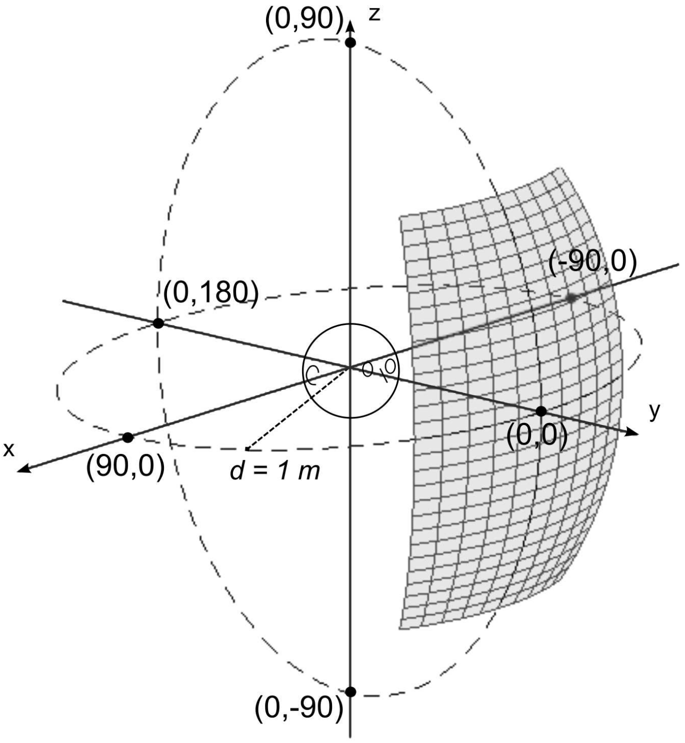


Figure 3.16: *Spatial range of validity of the proposed model.*

$\phi = -45°$ to $\phi = 45°$ at 5.625-degree steps (17 HRIRs per subject), relative azimuthal variations up to at least $\Delta\theta = 30°$ cause very small spectral changes in the PRTF Mokhtari et al. (2010); Lopez-Poveda and Meddis (1996); Raykar et al. (2005), hence PRTFs in this region can be considered elevation-dependent-only. The upper elevation limit ($\phi = 45°$) was chosen because of the high degree of uncertainty in elevation judgement for sources at $\phi > 45°$ Blauert (1983); Morimoto (2001) and the general lack of deep spectral notches in PRTFs in this region Kahana and Nelson (2007); Raykar et al. (2005), which may be in fact two faces of the same coin. Thus the angular range of validity of this model will be at least as broad as the shaded area depicted in Fig. 3.16.

In light of the above remarks, a fundamental assumption is introduced, i.e. elevation and azimuth cues are handled orthogonally throughout the considered frontal workspace. 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 their presence would generally add low-frequency secondary HRTF cues for elevation perception Algazi et al. (2001b). Furthermore, dependence on source distance is negligible in the pinna model but critical in the head counterpart in the near field, yielding both azimuth and distance control assigned to the head structural component. Two instances (one per ear) of such model, appropriately synchronized through interaural time delay (ITD) estimation methods, allow for real-time binaural rendering.

Examining its structure from left to right, the simple spherical model approximates head shadowing and diffraction also in the near field (see Sec. 3.1.2) and a "resonances-plus-reflections" block approximating the pinna effects described in Sec. 3.3.1 allows elevation control. It is worthwhile to notice that the above described pinna model can be easily combined with several ITD and ILD estimation methods available in the literature (see Minnaar et al. (2000) and Watanabe et al. (2007) for two examples), in alternative to the spherical head model previously discussed..

3.4 Future perspectives

In this chapter, two main original contributions have been presented: (i) a first-order filter model of the head for near-field effects rendering, thought for real-time binaural listening applications and (ii) a customized pinna model for real-time control of source elevation.

The fit to analytical responses provided by the head model was objectively proved to be overall appropriate, though experimental evaluations on its subjective effectiveness are a subsequent necessary step. Further work 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. Also, if one remains within the assumption that ITD does not change with distance, an all-pass section has to be included in the structure in order to compensate the effect of the shelving/lowpass filter's phase response on ITD. Last but not least, alternative choices of the far-field head filter (H_{dist}^{∞}) to be coupled with the distance rendering model can be explored in order to improve the STF approximation.

Ongoing and future work related to the proposed pinna structural model consist on automatic pinna contour extraction[8] and extension of the model to a wider spatial range, including the upper and posterior regions of the median plane. 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 is also required in order to develop a complete anthropometric parametrization of the pinna model.

Finally, it must be noted that, for $\phi < -45°$, the inclusion of the shoulders and torso's contribution becomes crucial, since they add relevant shadowing effects to the incoming waves (Algazi et al., 2002b). Thus, it should be verified whether a model of the torso can effectively compensate for the lack of a model for reflections due to the pinna at very low elevations, not forgetting that low-elevation HRTFs are usually heavily influenced by posture (Algazi et al., 2002b).

[8] The relation between anthropometry and filter parameters is discussed in Sec. 4.2, while a first prototype of automatic pinna contour extraction is introduced in Sec. 5.4

Chapter 4

Mixed structural modeling approach: the case of the pinna

This chapter considers the problem of modeling pinna-related transfer functions (PRTF) for 3D sound rendering. Thus, following a structural *modus operandi*, and presents an algorithm for the decomposition of PRTFs into ear resonances and frequency notches due to reflections over pinna cavities. Such an approach allows to control the evolution of each physical phenomenon separately through the design of two distinct filter blocks during PRTF synthesis. The resulting model is suitable for integration into a structural head-related transfer function model, and for parametrization over anthropometrical measurements of a HRTF selection mechanism.

Accordingly, the first step concerns feature extraction from the obtained PRTF. Reflections and resonances are treated as two separated phenomena and thus the PRTF is split into a “resonant” and a “reflective” component by means of a novel structural decomposition algorithm. 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 the final spectrum contains the resonant component, while the reflective component is given by direct combination of all the calculated multi-notch filters. An example of the algorithm output is reported in Fig. 4.1.

Then, the relationship between HRTFs and pinna reflection patterns in the frontal hemispace is investigated in detail. Up to three spectral notches are extracted from responses in the median plane of a pre-processed database of HRTFs. Ray-tracing analysis performed on the central frequency of the obtained notches is compared with a set of possible reflection surfaces directly recognizable from the corresponding pinna picture. Results of such analysis provide a 1-1 association between notches and pinna reflections, along with the sign of each reflection, which is found to be most likely negative.

Based on this finding, this Chapter formalizes and describes a novel approach to the selection of generic HRTFs for binaural audio rendering through headphones. The reflection model

This chapter is partially based on (Geronazzo et al., 2010a, 2011b; Spagnol et al., 2013a).

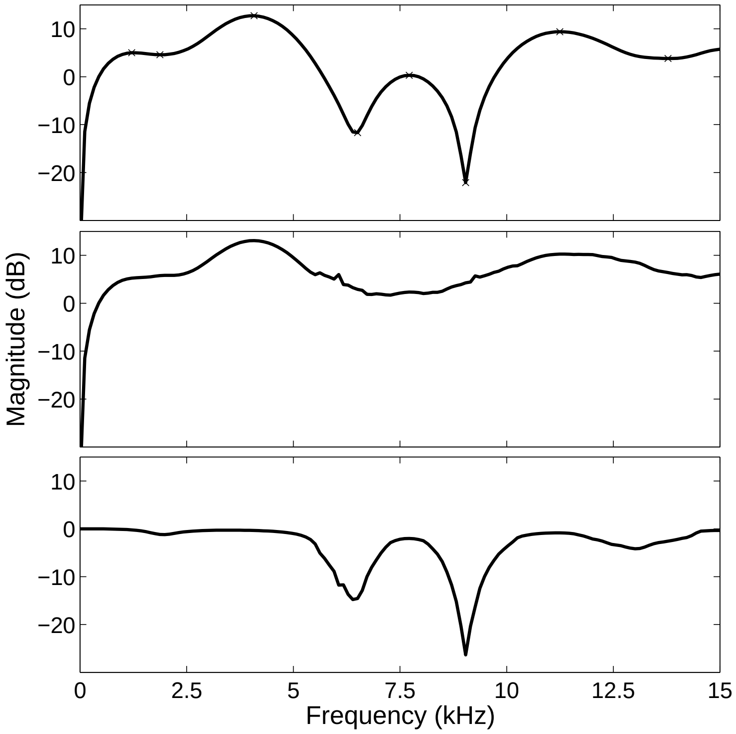
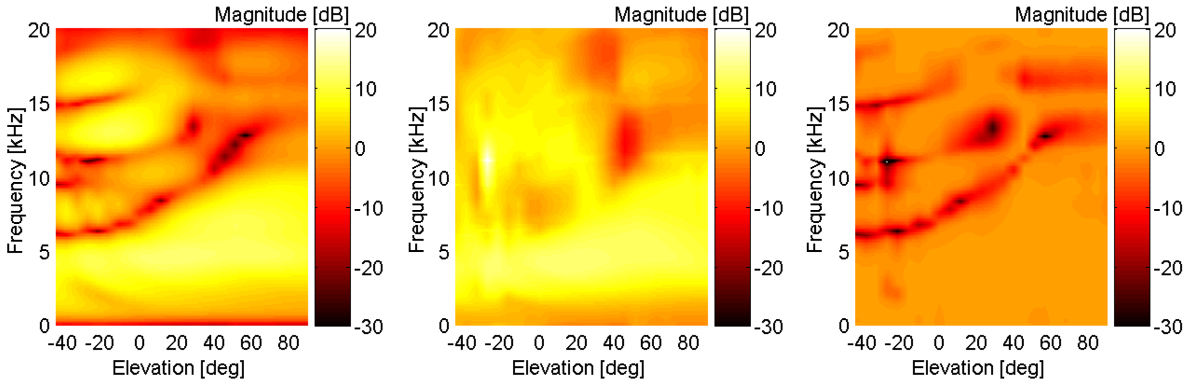


Figure 4.1: *Top panel: right PRTF of CIPIC subject* 003 *(*$\theta = 0°$*,* $\phi = -28.125°$*), magnitude spectrum. Middle panel: the PRTF resonant component extracted by the separation algorithm (Geronazzo et al., 2010a). Bottom panel: the PRTF reflective component extracted by the separation algorithm.*

applied to the user's ear picture allows to extract relevant anthropometric cues that are used for selecting from a database two HRTF sets fitting that user. Localization performance with the selected HRTFs are evaluated in a psychoacoustic experiment. The proposed selection increases the average elevation performances of 17% (with a peak of 34%) with respect to generic HRTFs from an anthropomorphic mannequin. It also significantly enhances externalization and reduces the number of up/down reversals.

The last section proposes a straightforward procedure to the individualization of the pinna model discussed in Sec. 3.3.1, in which parameters to be fed to the model are derived either from analysis or from specific anthropometric features of the subject in a handy process. Objective evaluations of reconstructed HRTFs in the chosen spatial range are performed through spectral distortion measurements.



(a) (b) (c)

Figure 4.2: *Separation of CIPIC Subject* 165*'s left PRTFs (a) into a resonant (b) and a reflective (c) component.*

4.1 Structural decomposition algorithm

As discussed in Ch. 1, both pinna peaks and notches seem to play an important function in vertical localization of a sound source. However, a previous work (Spagnol et al., 2010a) highlighted that while the resonant component of the pinna-related counterpart of the HRTF (known as PRTF) exhibits a similar behaviour among different subjects, the reflective component of the PRTF is critically subject-dependent. This result was found by separating the resonant and reflective components through an ad-hoc designed algorithm (Geronazzo et al., 2010a), an instance of which can be appreciated in Fig. 4.2.[1]

Such an algorithm is essential to study these two contributions separately. An analysis-by-synthesis approach drives the algorithm towards 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 the spectrum contains the resonant component alone, while the reflective component is given by direct combination of all the estimated multi-notch filters.

4.1.1 Pinna pHRTF analysis in the median plane

For the purpose of analysis, measured HRIRs from the CIPIC database are considered(see Table 1.1 for the specification). Since sensitivity of PRTFs to azimuth is weak, we focus on HRIRs sampled on the median plane, with elevation varying from -45 to 90 degrees. Given that the magnitude response of the pinnaless head with respect to a sound source in the median plane is essentially flat (it is ideally flat if the head is modeled as a rigid sphere), the only preprocessing

[1] In these and in all of the following plots, magnitude values are linearly interpolated across the available azimuth/elevation angles to yield a 1-degree resolution.

step that is needed in order to obtain an estimate of the PRTF is to window the corresponding HRIR using a 1.0 ms Hann window. In this way, effects due to reflections caused by shoulders and torso are removed from the PRTF estimate. A similar preprocessing procedure was adopted by Raykar et al. (2005) to estimate PRTFs from measured HRIRs.

Figure 4.3 reports the complete flow chart of the analysis algorithm. The idea beyond it is to iteratively compensate the PRTF magnitude spectrum with an approximate multi-notch filter until no significant notches are left. Once convergence is reached (say at iteration $\underline{i}$), the PRTF spectrum $H_{\text{res}}^{(\underline{i})}$ will contain the resonant component, while the combination $H_{\text{refl}}^{(\underline{i})}$ of the multi-notch filters will provide the reflective component. The initial conditions of the algorithm heavily influence the final result; 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 gain (depth) threshold for notches to be considered;

- ρ, the reduction factor for every notch filter bandwidth (its purpose will be discussed below).

Before discussing the core of the algorithm, let $H_{\text{res}}^{(1)}$ match the PRTF and set $H_{\text{refl}}^{(1)}$ to 1. These two frequency responses will be updated at each iteration, resulting in $H_{\text{res}}^{(i)}$ and $H_{\text{refl}}^{(i)}$ at the beginning of the i-th iteration. If $N_{\text{nch}}^{(i)}$ is the number of “valid” notches identified at the end of it, the algorithm will terminate at iteration $\underline{i}$ if $N_{\text{nch}}^{(\underline{i})} = 0$, while $N_{\text{res}}^{(\underline{i})}$ and $N_{\text{refl}}^{(\underline{i})}$ 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 $N_{\text{res}}^{(i)}$ is presented, assuming that $N_{\text{nch}}^{(i-1)} > 0$. For the sake of simplicity, in the following the apex (i) indicating iteration number is dropped from all notation.

4.1.2 Resonances and residue computation

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 N_{res} is calculated; then, by liftering the cepstrum with the first N_{ceps} cepstral coefficients and performing the FFT, an estimate of the spectral envelope of N_{res} is obtained, which we call C_{res}.

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. We experimentally found that the optimal number of coefficients that capture the resonant structure of the PRTF while leaving all the notches out of the spectral envelope is $N_{\text{ceps}} = 4$. This number also matches the maximum number of modes identified by Shaw which appear at one specific

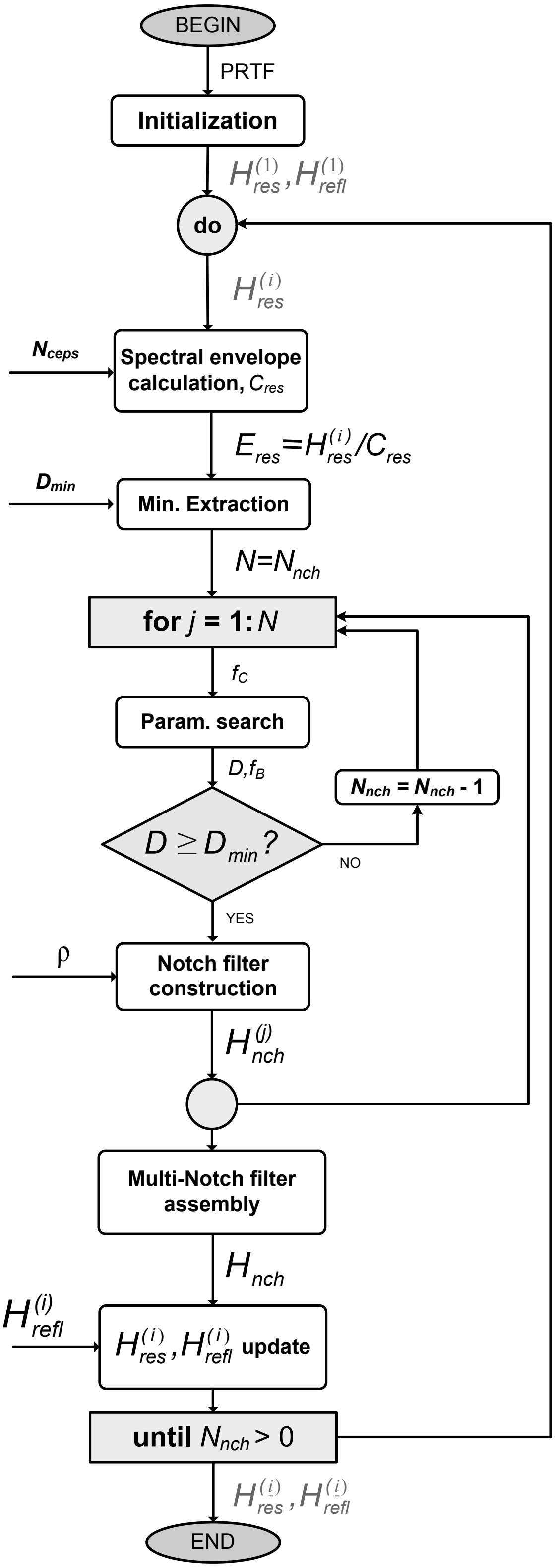


Figure 4.3: *Flow chart of the analysis algorithm.*

spatial location: for elevations close to zero, modes 1, 4, 5, and 6 are excited. Once C_{res} is computed, we subtract it from the dB magnitude of N_{res} and obtain the residue E_{res}.

4.1.3 Multi-notch filter parameter search

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. Our aim is to compensate each notch with a second-order notch filter, defined by three parameters: central frequency f_C, bandwidth f_B, and notch gain D.

Consider the j-th local minimum. The central frequency of the corresponding notch filter f_C is immediately determined, while notch gain is found as $D = |E_{\text{res}}(f_C)|$. Computation of f_B is less straightforward. Indeed, f_B is calculated as the standard 3DB 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 3DB 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, we vertically shift E_{res} until the 0-dB level meets the closest of the two. Then, f_B is calculated as before except if the new notch gain 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.

4.1.4 Multi-notch filter construction

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 (Zolzer, 2011) is used

$$H_{\text{nch}}^{(j)}(z) = \frac{1+(1+k)\frac{H_0}{2}+l(1-k)z^{-1}+(-k-(1+k)\frac{H_0}{2})z^{-2}}{1+l(1-k)z^{-1}-kz^{-2}}, \tag{4.1}$$

where

$$k = \frac{\tan(\pi\frac{f_B}{f_s}) - V_0}{\tan(\pi\frac{f_B}{f_s}) + V_0}, \tag{4.2}$$

$$l = -\cos(2\pi\frac{f_C}{f_s}), \tag{4.3}$$

$$V_0 = 10^{\frac{D}{20}}, \tag{4.4}$$

$$H_0 = V_0 - 1, \tag{4.5}$$

and f_s is the sampling frequency. Using such an implementation allows us to fit the required parameters directly to the filter model. 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 narrow, 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), an excessive notch bandwidth could lead to undesired artifacts in the final resonance spectrum. Here is where parameter ρ comes into play: if we divide f_B by $\rho > 1$, the new bandwidth specification will produce a filter whose notch amplitude will be further reduced, allowing us to reach a smaller bandwidth. Typically, in order to achieve a satisfactory trade-off between the size of ρ and the number of iterations, we set it to 2.

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

$$N_{\text{nch}}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}} = \prod_{j=1}^{N_{\text{nch}}} N_{\text{nch}}^{(j)}(z), \tag{4.6}$$

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

$$\boldsymbol{b} = [b_0, b_1, b_2] = \boldsymbol{b}^{(1)} * \boldsymbol{b}^{(2)} * \cdots * \boldsymbol{b}^{(N_{\text{nch}})} \tag{4.7}$$

$$\boldsymbol{a} = [a_0, a_1, a_2] = \boldsymbol{a}^{(1)} * \boldsymbol{a}^{(2)} * \cdots * \boldsymbol{a}^{(N_{\text{nch}})}. \tag{4.8}$$

Finally, before considering the next iteration, we must update the global multi-notch filter $N_{\text{refl}}^{(i+1)} = N_{\text{refl}}^{(i)} \cdot N_{\text{nch}}$ and compensate the PRTF by applying $N_{\text{res}}^{(i+1)} = N_{\text{res}}^{(i)} / N_{\text{nch}}$.

4.1.5 Examples of algorithm evolution

Figure 4.4 illustrates the algorithm evolution for a particular PRTF in this example. The specific choice of the initial parameters was $N_{\text{ceps}} = 4$, $D_{\text{min}} = 0.1$ dB, and $\rho = 2$. The top left panel illustrates Subject 010 PRTF for an elevation of -45 degrees. The bottom left panel that interfering spectral notches negatively influence the initial estimate.

Consider the range where acoustic effects of the pinna are relevant, i.e. the range from 3 to 18 kHz approximately. Figure 4.4 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 have 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 one considers the pinna contribution to the HRTF alone.

The behavior exemplified in Fig. 4.4 is common to several elevations and subjects.

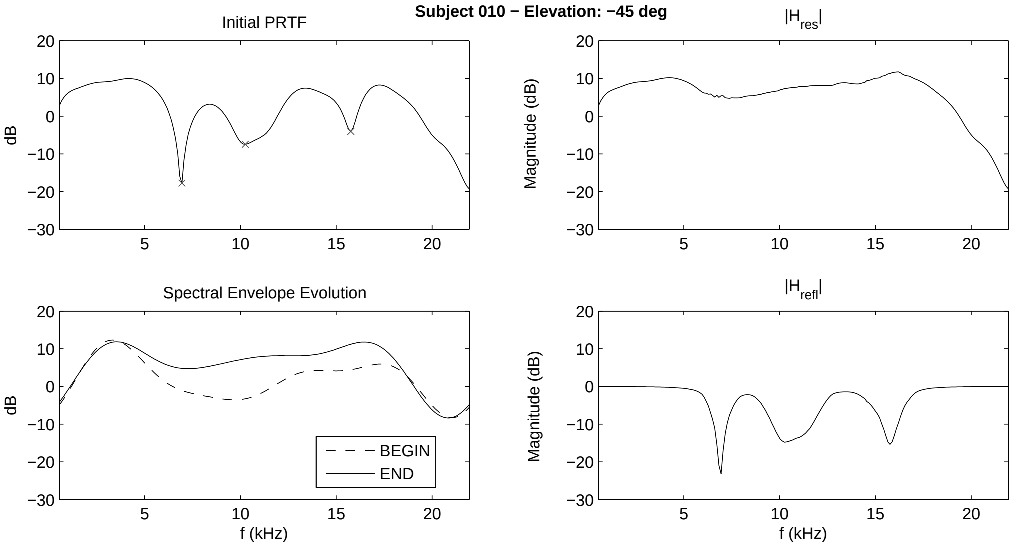


Figure 4.4: *An example of the algorithm evolution. The PRTF magnitude in the top left panel is decomposed into resonances (top right panel) and frequency notches (bottom right panel). The bottom left panel shows the evolution of the PRTF spectral envelope from the first iteration to convergence.*

4.1.6 Results

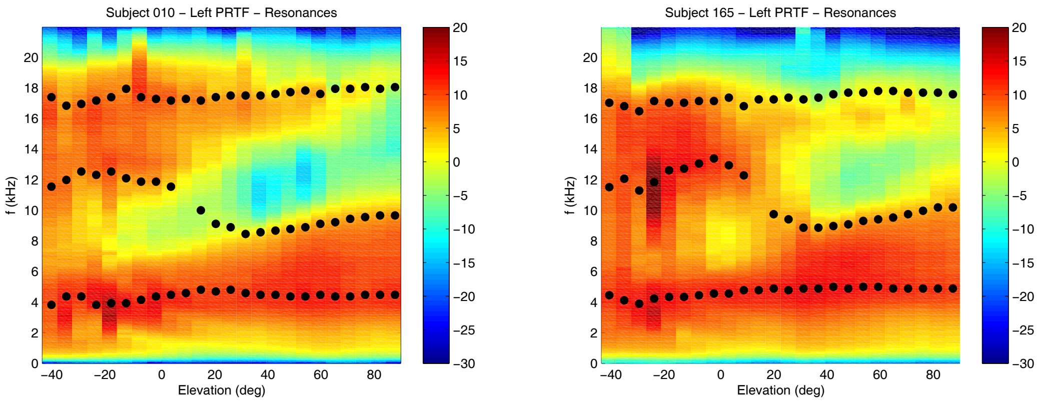
This section discusses the PRTF features identified through the decomposition carried out by the proposed algorithm. In order to facilitate comparison with previous works, this section reports results of the reflection analysis for the same CIPIC subjects that appear in (Raykar et al., 2005) and (Satarzadeh, 2006), specifically Subjects 010, 048, 134, and 165.[2]

Frequency peaks: resonances

The average magnitude spectrum for 45 left pinnae of CIPIC subjects is shown in Fig. 4.6 and it is computed by averaging $|H_{\text{res}}|$ over all subjects. Two brighter areas can be distinctly identified. The first one, centered at around 4 kHz, spans all elevations in both plots and appears to be very similar amongst all subjects in the CIPIC database. Given such observation, it can be immediately noticed that this area contains Shaw's omnidirectional mode 1. Its bandwidth apparently increases with elevation; however, inspection of the dotted tracks reveals that a second resonance is likely to interfere within this frequency range, in particular Shaw's resonant mode 2 which appears at around 7 kHz with a magnitude of 10 dB.

The second bright area in the average response covers a wider and less defined region. 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. The 3D plots in Fig. 4.5 represent the resonance contributions at all available elevations for Subjects 010 and 165, respectively. Every

[2] Subject 165 is a KEMAR head with small pinnae.



(a) Subject 010.

(b) Subject 165.

Figure 4.5: *Resonance plots of two subjects at all available elevations. Dotted tracks highlight resonance center frequencies.*

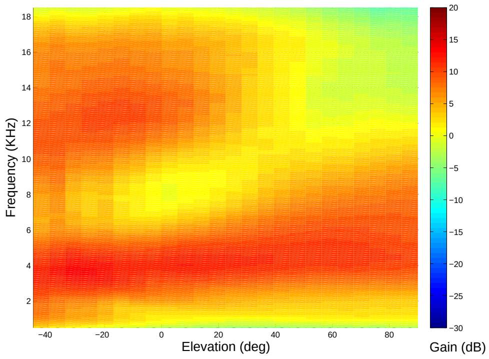
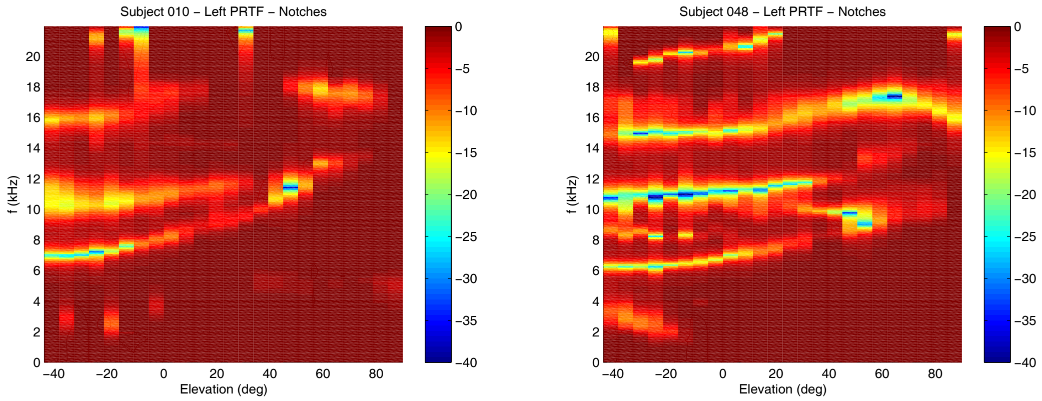


Figure 4.6: *Mean magnitude spectrum of the pinna resonant component, averaged on all* 45 *CIPIC subjects (left-ear responses).*

resonance's center frequency was extracted through an identification system based on a sixth-order ARMA model (Esquef et al., 2003) and spatially tracked along all elevations, resulting in the dotted tracks superposed on the plots. Note that the magnitude response around 12 kHz takes negative values at high elevations, especially for Subject 010: such an incongruity may be explained by phenomena other than reflections or resonances, e.g. diffraction around the pinna. One may make the same observation for very low and very high frequency zones; nevertheless, these anomalies are most likely due to computational limitations and lie in any case fairly outside the frequency range that interests the pinna.

Finally, note that the resonance at ~ 12 kHz and the one at ~ 7 kHz (associated to mode 2) are excited in mutually exclusive elevation ranges. This effect, which appears for all the analyzed subjects and is especially evident for Subject 165 in Fig. 4.5(b), gives the impression of a smooth transition from one resonance to the other. In light of this, one can consider and evaluate a two-



(a) Subject 010.

(b) Subject 048.

Figure 4.7: *Spectral notch plots of two subjects at all available elevations.*

or three-resonance filter design.

Frequency notches: reflections

As already mentioned, reflection patterns depend strongly on elevation and pinna shape. 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 gain of frequency notches grows to an extent that varies between subjects. These remarks can be immediately verified in Fig. 4.7, where the contribution of spectral notches for Subjects 010 and 048 are reported. In particular, Subject 048 exhibits at low elevations a clear reflection structure with three prominent notches.

Straightforward comparison of Fig. 4.7(a) to the findings by Raykar (Fig. 11(a) by Raykar et al. (2005)) shows an encouraging correspondence between notch structures.

Since common trends cannot be identified in the evolution of spectral notches, and following the common idea that notches are of major relevance for elevation detection in the frontal region (Iida et al., 2007; Moore et al., 1989; Wright et al., 1974; Hebrank and Wright, 1974b), a statistical analysis of the reflective component is performed for 45 CIPIC subjects based on the results provided by the structural decomposition algorithm.

Prominent notches are obtained through a simple notch picking algorithm (Spagnol et al., 2010a). In order to have a consistent labeling along subsequent PRTFs, extracted notches need to be grouped into tracks evolving through elevation. To this end, the McAulay-Quatieri partial tracking algorithm (McAulay and Quatieri, 1986) is exploited and fit it to these needs. The original formulation of the algorithm can be used to track the most prominent notch patterns along elevation, with elevation dependency conceptually replacing temporal evolution, and spectral notches taking the role of sinusoidal partials. With respect to its first formulation, it is sufficient to add that the notch detection (originally “peak detection”) step simply 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.

Two post-processing steps are performed on the obtained tracks. First, since it is convenient to restrict attention to the frequency range where pinna reflections are most likely seen (Hebrank and Wright, 1974b), we delete the tracks which start and terminate outside the range $4 - 16$ kHz. Second, we delete the tracks that do not present a notch deeper than 5 dB, since overall shallow notches are not likely to be associated with a major reflection. As a result, the majority[3] of the CIPIC subjects exhibits three notch tracks at a given elevation, which are defined in increasing frequency order as T_1, T_2 and T_3.

Starting from the the gain and bandwidth of the three prominent notches, a first-order statistical analysis among all 45 left pinnae of CIPIC subjects, subdivided by notch track and elevation, gives no systematic behavior nor evidence of correspondence with anthropometric quantities. Figure 4.8 shows results of such an analysis:

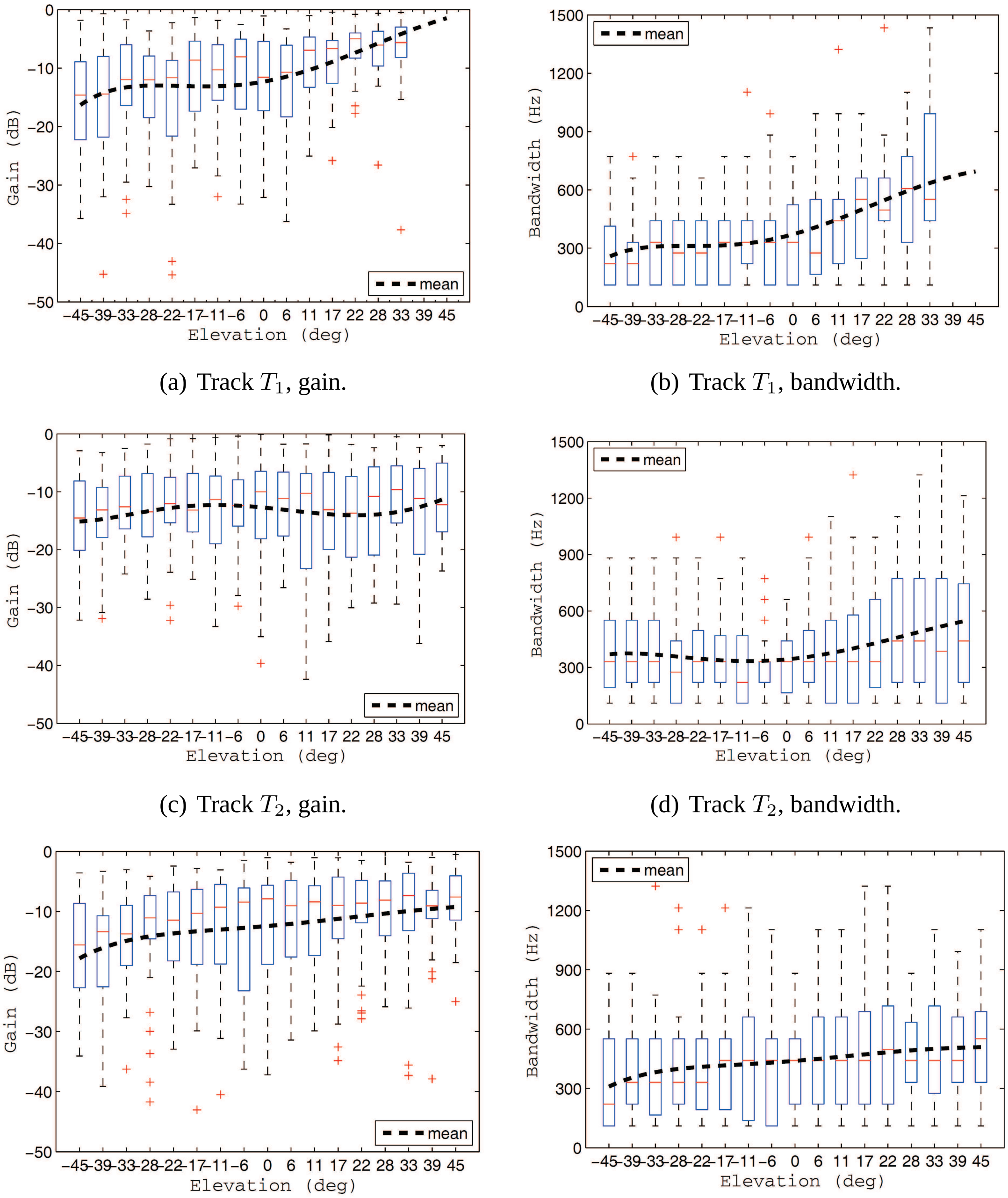
- all plots exhibit high variance among tracks and elevations ;
- mean values remain approximately constant with respect to elevation, except for a slight decrease in notch gain (Tracks T_1 and T_3) and increase in bandwidth (Track T_1 and T_3) as elevation increases up to $\phi = 45°$.
- in Track T_1, some box plots for higher elevations ($\phi = 39°, 45°$) are absent[4] indicating the tendency of these notches to disappear.

Nevertheless, similarly to (Raykar et al., 2005), each notch is treated as the result of a distinct reflection path. Also, similarly to previous works on reflection modeling (Raykar et al., 2005; Satarzadeh et al., 2007) 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 (Shaw and Teranishi, 1968; Hebrank and Wright, 1974b) to an extent that can definitely be detected by the human auditory system (Moore et al., 1989). Conversely, changes in notch bandwidth and amplitude along elevation are seen to be far less systematic, and their perceptual relevance is little understood in previous literature.

Average notch frequencies in the three tracks at each available elevation are reported in Fig. 4.9, along with their standard deviation: 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. Despite the significant variance in the central frequencies of the three tracks (T_3 in particular), these trends were seen to be consistent across subjects. Not reported in the figure 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 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.

[3] Only subjects 019 and 020 lack of one track, the lowest and the highest in frequency respectively.

[4] Less than the 40% of the CIPIC subjects has a notch in the PRTF for that track and for those elevations.



(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 4.8: *Box plot and mean of gain and bandwidth for the three prominent notch tracks among 45 CIPIC subjects. Dotted lines are fifth-order polynomials dependent on elevation fitted to each of the mean sequences in order to approximate average trends.*

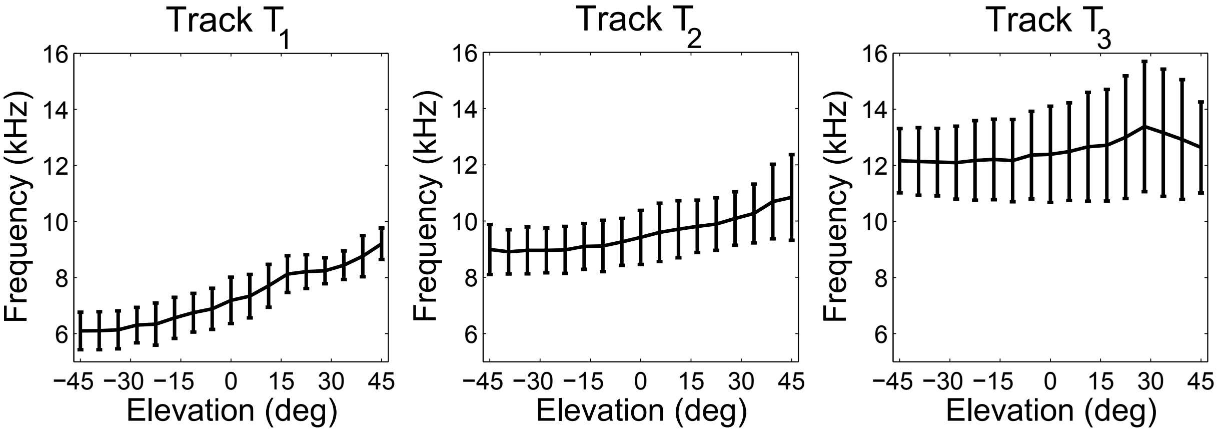


Figure 4.9: *Mean and standard deviation of notch frequencies per elevation and track across* 20 *subjects.*

4.2 Reflections and Anthropometry

Ray-tracing reflection models (Hebrank and Wright, 1974b) assume ray-like rather than wave-like behaviour of sound, providing a crude approximation of the wave equation. Despite this, the approach conveyed by such models is valid as long as the wavelength of the sound is small when compared to the dimensions of the involved reflection surfaces. This is definitely the case for higher frequencies of the audible spectrum, where spectral notches due to pinna reflections appear. In this context, one can intuitively observe that the elevation-dependent temporal delay $t_d(\phi)$ between the direct and the reflected wave corresponds to a the point of reflection at distance

$$d_c(\phi) = \frac{ct_d(\phi)}{2}, \tag{4.9}$$

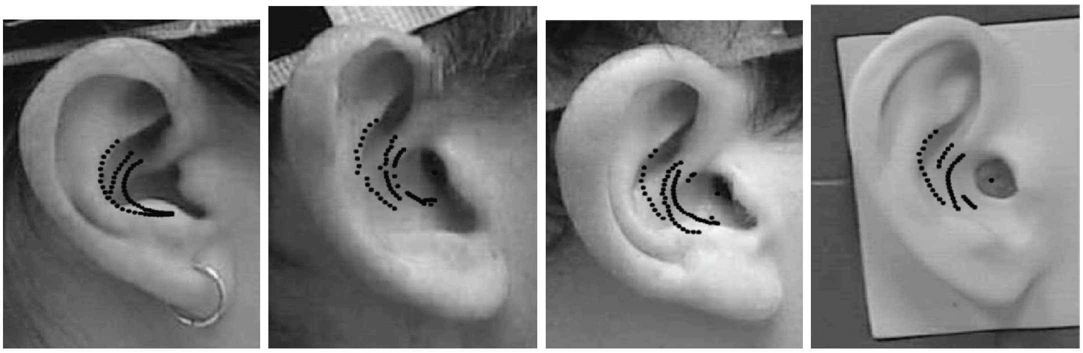
from the ear canal (where c is the speed of sound). Assuming the reflection coefficient to be positive, then we will have destructive interference (i.e., a notch) at all those frequencies where the reflection's phase shift equals π:

$$f_n(\phi) = \frac{2n+1}{2t_d(\phi)} = \frac{c(2n+1)}{4d_c(\phi)}, \qquad n = 0, 1, \ldots.. \tag{4.10}$$

Hence the first notch is located at frequency

$$f_0(\phi) = \frac{c}{4d_c(\phi)}. \tag{4.11}$$

The positive reflection assumption was also adopted by Raykar et al. (2005). Given a notch track $f_0(\phi)$, the corresponding reflection surface $d_c(\phi)$ is estimated using Eq. (4.11) and then superimposed to the pinna image. Figure 4.10 reproduces Raykar's results obtained using this approach. Note that since the extracted notch tracks are pairwise in non-harmonic relationship,



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

Figure 4.10: *The spectral notch frequencies for different elevations from* $45°$ *to* $90°$ *extracted for the right pinna for azimuth* $0°$ *of subject 10, 27, 134, and 165 in the CIPIC database, and marked on the pinna corresponding side images with positive reflection coefficient (figure reproduced from (Raykar et al., 2005)).*

both on average (see again Fig. 4.9) and for every single subject, a single reflection path cannot be assigned to any pair of tracks. Hence Raykar's assumption that each notch in the considered frequency range is the result of a distinct reflection path is well-grounded.

Nevertheless, Satarzadeh (2006) drew attention to the fact that almost 80% of CIPIC subjects exhibit a clear negative reflection in their HRIRs and proposed a physical explanation to this phenomenon. In case of negative reflection, destructive interference would not appear at half-wavelength delays anymore, but at full-wavelength delays. Eqs. (4.10) and (4.11) would then become

$$f_n(\phi) = \frac{n+1}{t_d(\phi)} = \frac{c(n+1)}{2d_c(\phi)}, \qquad n = 0, 1, \ldots \tag{4.12}$$

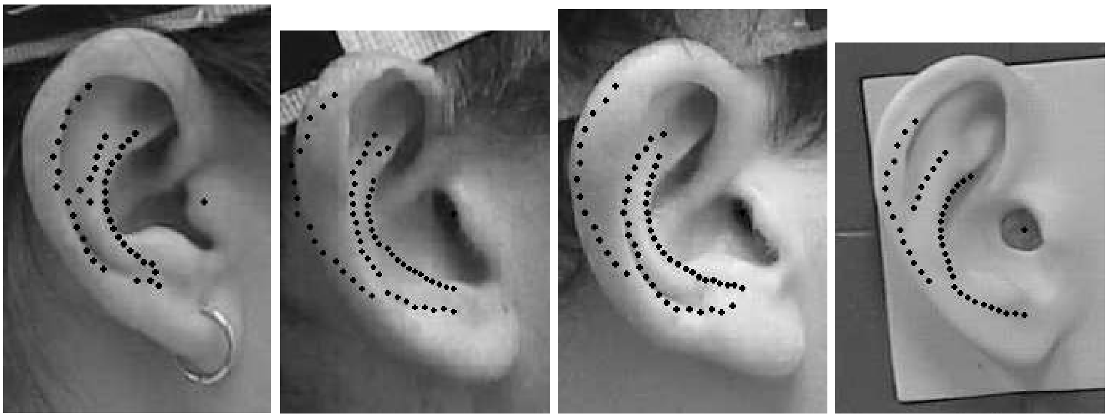
and

$$f_0(\phi) = \frac{c}{2d_c(\phi)}. \tag{4.13}$$

Results for subjects 010, 027, 134, and 165 of the CIPIC HRTF database are reported in Fig. 4.11 which shows the traced reflection surfaces computed using Eq. (4.13), one for each estimated notch track. For the sake of comparison, the same CIPIC subjects of Fig. 4.10 are visualized. For all these 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:

- 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.



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

Figure 4.11: *The spectral notch frequencies for different elevations from* $45°$ *to* $90°$ *extracted for the right pinna for azimuth* $0°$ *of subject 10, 27, 134, and 165 in the CIPIC database, and marked on the pinna corresponding side images with negative reflection coefficient (Spagnol et al., 2010a).*

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 we visually associate to a reflection on the rim's edge.

Finally, the furthest track follows the shape of the rim and 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. A strong evidence that supports connection of this track to the rim structure lies in the fact that the rim terminates in the vicinity of the point where the track disappears.

The following subsection investigates the correspondence between pinna anatomy and theoretical reflection points under different reflection sign conditions on a wide morphological variety of CIPIC subjects' pinnae.

4.2.1 Contour matching procedure

The basic assumption that drives the proposed analysis procedure is that each notch track is associated with 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 representation. We choose to investigate as possible reflection surfaces a set of three contours directly recognizeable from the pinna image, together with two hidden surfaces approximating the real inner back walls of the concha and helix. Specifically, as Fig. 4.12 depicts, the following contours are considered:

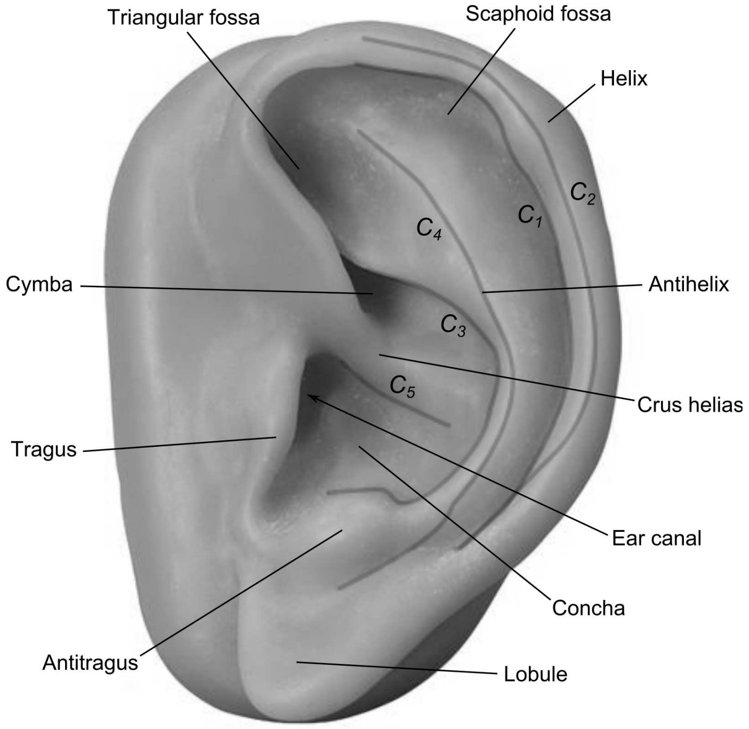


Figure 4.12: *Pinna anatomy and the five chosen contours for the matching procedure. C_1: helix border; C_2: helix wall; C_3: concha border; C_4: antihelix and concha wall; C_5: crus helias.*

1. helix border (C_1), visible on picture;
2. helix inner wall (C_2), estimated by 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), estimated by 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.

The extraction procedure was performed by manual tracing through a pen tablet (the topic of automatic contour extraction is covered within Ch. 5). Pinna images available from the CIPIC HRTF database for 20 subjects were accurately resized to match a $1:1$ scale based on the quantitative pinna height parameter (d_5 in (Algazi et al., 2001d)) also available from the database's anthropometric data, or based on the measuring tape visible in the pinna images (only 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 low resolution of the pictures; it is therefore necessary to point out that in some cases the lower part of this contour was almost blindly traced.

Before describing the contour matching procedure, let us formally state some useful definitions.

- 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.

Our goal is to discover which of the 8 configurations ($2 \times 2 \times 2$ possible combinations of the three reflection signs $s_j = \{0, 1\}$, $j = 1, 2, 3$) is the most likely 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 a wrong choice of the focus, its location is estimated through an optimization procedure over a rectangular search area A of the pinna image 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 first translated into Euclidean distances (in pixels) through a sign-dependent combination of Eqs. (4.11) and (4.13),

$$d_c^j(\phi) = \frac{c}{2(s_j + 1) f_0^j(\phi)}, \tag{4.14}$$

and subsequently projected onto the point

$$p_{\psi,\rho}^j(\phi) = (\psi_x + d_c^j(\phi)\cos(\phi + \rho), \psi_y + d_c^j(\phi)\sin(\phi + \rho)) \tag{4.15}$$

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:

$$\min_{\psi \in A, \rho \in I} \sum_{j=1}^{3} \min_i d_{\psi,\rho}(T_j, C_i)^2, \tag{4.16}$$

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.

Table 4.1: *Contour Matching Procedure Results*

	Fitness, F									
Subject	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$	$\boldsymbol{s}^{\text{opt}}$	Nearest contours
003	4.03	9.19	9.27	13.78	7.83	12.45	13.03	17.54	$[0,0,0]$	$1,4,3$
008	2.95	4.86	5.33	7.30	3.69	7.89	5.58	10.64	$[0,0,0]$	$1,4,3$
009	2.55	5.18	4.79	7.02	2.95	5.08	2.94	5.01	$[0,0,0]$	$2,4,4$
010	1.88	5.18	2.26	6.02	3.57	5.69	4.46	6.70	$[0,0,0]$	$1,4,3$
011	2.62	5.10	5.60	9.53	3.16	5.79	4.97	9.25	$[0,0,0]$	$1,4,3$
012	2.08	4.21	4.76	7.30	2.70	5.32	3.20	6.78	$[0,0,0]$	$2,4,3$
015	4.99	9.92	6.14	10.59	3.02	6.70	3.39	3.19	$[1,0,0]$	$3,1,4$
017	2.81	6.35	4.53	8.12	2.99	5.02	5.63	6.79	$[0,0,0]$	$1,4,3$
019	1.64	6.64	4.85	8.00	1.64	6.64	4.85	8.00	$[*,0,0]$	$-,4,3$
020	1.15	1.15	5.27	5.27	1.85	1.85	5.45	5.45	$[0,0,*]$	$2,4,-$
021	2.90	6.40	4.06	8.44	3.30	8.97	6.25	11.54	$[0,0,0]$	$2,4,3$
027	2.07	6.53	5.04	8.56	2.32	5.27	2.80	4.25	$[0,0,0]$	$2,4,3$
028	1.71	3.54	4.21	5.57	3.79	4.02	5.62	6.10	$[0,0,0]$	$2,4,3$
033	2.51	4.73	6.66	6.61	3.42	7.68	9.08	9.98	$[0,0,0]$	$1,4,3$
040	1.74	5.48	2.59	5.35	2.57	5.86	3.30	5.96	$[0,0,0]$	$1,4,3$
044	1.88	2.84	5.33	4.81	2.86	2.49	4.13	3.74	$[0,0,0]$	$2,4,3$
048	2.02	5.33	5.45	7.86	3.70	5.06	5.27	6.97	$[0,0,0]$	$1,4,3$
050	3.25	6.29	7.68	10.52	4.37	7.59	7.57	11.23	$[0,0,0]$	$2,4,3$
134	1.64	6.11	5.18	8.56	3.38	6.31	4.56	7.37	$[0,0,0]$	$2,4,3$
165	1.09	5.35	3.08	5.93	3.43	3.89	3.00	2.99	$[0,0,0]$	$2,4,3$

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

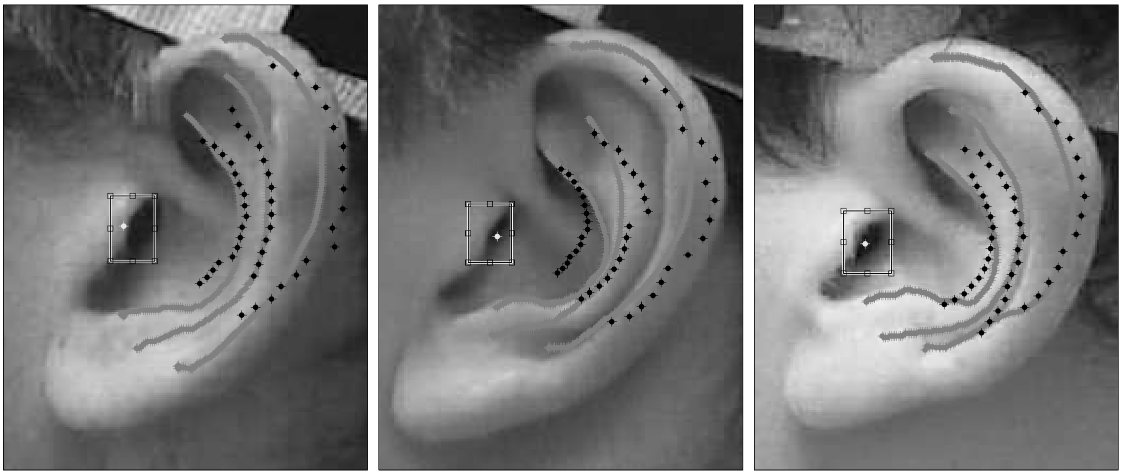
$$F(\boldsymbol{s}) = \frac{1}{3}\sum_{j=1}^{3}\min_{i}\frac{d_{\psi_{\mathbf{s}}^{\text{opt}},\rho_{\mathbf{s}}^{\text{opt}}}(T_j, C_i)}{2 - s_j}, \tag{4.17}$$

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. (4.17) is scaled by a factor of $1/2$ 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 better the fit, clearly.

4.2.2 Results

The above contour matching procedure was run for all the 20 CIPIC subjects for which a pinna image was available. Table 4.1 summarizes the final scores (fitness values) for all possible configurations, along with the resulting “best” configuration $\boldsymbol{s}^{\text{opt}}$ and the corresponding best matching contours. For subjects with two tracks only, the missing track’s reflection sign is conventionally labeled with “$*$”. As an example, Fig. 4.13 shows the optimal ray-traced tracks for three subjects: 027 (having a final score close to the median), 050 (second worst subject), and 134 (third best subject).

We 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, we noted 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



(a) Subject 027. (b) Subject 050. (c) Subject 134.

Figure 4.13: *Optimal ray-tracing for three subjects. The light grey point surrounded by the search area A is the optimal focus of the winning configuration $\boldsymbol{s}^{\text{opt}} = [0, 0, 0]$. Black points indicate the three projected tracks, and dark grey points the hand-traced nearest contours to the tracks.*

from being approximately aligned to the interaural axis, resulting in focus points 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 we regard Subject 009 and Subject 015 as outliers.

All the remaining subjects exhibit $\boldsymbol{s}^{\text{opt}} = [0, 0, 0]$ as the best configuration. Quantitative correspondence between tracks and contours varies from subject to subject, e.g. a much lower score is assigned 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 makes the hypothesis of an overall positive reflection sign very 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. 4.13 for graphical evidence), hence the hypothesis of two different signs for reflections onto the same surface seems unlikely. Second, as mentioned in Sec. 4.1.6, 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 reflection surface, and this is compatible with a reflection on the helix, which 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

Table 4.2: *Notch Frequency Mismatch between Tracks and Contours*

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%

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 $\phi = 0°$ being 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 the three cases depicted in Fig. 4.13, Subject 050 showing the greatest mismatch. This point is further discussed next.

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

$$m(T_j, C_i) = \frac{1}{n(T_j)} \sum_{\phi} \frac{|f_0^j(\phi) - F_n^j(\phi)|}{f_0^j(\phi)} \cdot 100\%, \tag{4.18}$$

where $n(T_j)$ is the number of available notch frequencies in track T_j, $f_0^j(\phi)$ is estimated from subject's PRTF using the structural decomposition algorithm described in Sec. 4.1, while $F_n^j(\phi)$

is estimated from the associated contour C_i using Eq (4.13).

Table 4.2 shows frequency mismatches computed for 18 CIPIC subjects. These results can be directly compared to the findings by Moore et al. (1989): 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 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 and conclude that only 5 tracks over 35 exhibit a mismatch greater than the threshold, suggesting that the frequency shift caused by contour extraction is not perceptually relevant on average.

4.2.3 Discussion

The above results provide a quantitative confirmation 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 (Satarzadeh, 2006). 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 this intuition, there is no evidence of the fact that waves are only reflected at borders and not onto inner pinna walls.

A recent study by Takemoto et al. (2009) on pressure distribution patterns in median-plane PRTFs reveals through FDTD simulations on four different subjects 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 all resonate in anti-phase with respect to the incoming wave, while when the source is placed in the antero-superior direction the same phenomenon appears at the back of the concha. 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. Thus, one can speculate about the following generation mechanism for notches in track T_1: a given frequency component of the incoming sound wave forms a negative pressure area in the vicinity of the helix wall or border, reflects back with inverted phase, and encounters 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 have not been studied in (Takemoto et al., 2009); still we can think of analogous generation mechanisms for these tracks too.

Shifting our 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. 4.13) 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 (Mokhtari et al., 2011). Concerning track T_1, these findings seem to conflict with the common idea that the first notch is due to a reflection on the concha wall (Hebrank and Wright, 1974b; Lopez-Poveda and Meddis, 1996; Raykar et al., 2005). In two works by Mokhtari *et al.* (Mokhtari et al., 2010, 2011), 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 the first notch, 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 the first notch crosses the whole pinna area, calling into question the above common belief and giving credit to the provided result instead.

Admittedly, as (Mokhtari et al., 2011) 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. Instead multiple reflection paths concur in determining the distinctive parameters of the notch. Nonetheless, 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 proposed approach is the adequateness of using a 2D 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 3D model of the pinna would be needed to investigate its horizontal section. Beside the unavailability of such kind of reconstruction for the considered subjects, the original aim was to keep the contour extraction procedure as low-cost and accessible as possible; furthermore, additional results in the following sections will confirm that the 2D approximation is, on a theoretical basis at least, already satisfactory.

It should be emphasized that the results of the ray-tracing analysis do not conclusively prove that negative reflections effectively occur in reality. In particular, it remains to be explained from an acoustical point of view why negative reflection coefficients are likely to be produced. Clearly, a negative reflection coefficient c_r will not have unitary magnitude in real conditions because of the soft reflective surfaces involved, hence it will always satisfy $-1 < c_r < 0$. This results in a partial cancellation of the frequency where the notch falls: the closer the reflection coefficient to -1 is, the deeper the corresponding frequency notch will be. In order to characterize the magnitude of the coefficient, it could be therefore worthy to study how notch gains change with elevation.[5]

To conclude this discussion, track T_3 shows much greater mismatches, mostly due to the anbove discussed "contour anticipation" effect. Beside possible improvements that may take into account such an effect while manully 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 Sec. 1.1.2, the third notch is of lesser importance than the first two in elevation perception (Iida et al., 2007), hence precise estimation of its center frequency is less critical from a perceptual viewpoint.

[5] Unfortunately, common HRTF recordings do not have a frequency resolution that allows detection of the exact local minimum characterizing a notch, i.e. notch gain is always underestimated.

4.3 Non-individual head-related transfer function selection

This section investigates a novel approach to the selection of non-individual HRTF sets from an existing database, according to two criteria extrapolated from the pinna reflection model of Sec. 4.2. The idea is that the two chosen HRTFs should render better spatial sounds than a generic one (KEMAR) thanks to the closer relation between pinna geometry and localization cues, especially in the vertical dimension.

4.3.1 Previous works

The last decade saw a notable increase of the number of psychoacoustic tests related to HRTF selection techniques. The most common approach, which is also adopted in this section, is to use a specific criterion in order to choose the best HRTF set for a particular user from a database. Seeber and Fastl (2003) proposed a procedure according to which one HRTF set was selected among 12 based on multiple criteria such as spatial perception, directional impression and externalization. Even though their selection minimized both localization error variance and inside-the-head localization, it was only tested on the frontal horizontal plane. Zotkin et al. (Zotkin et al., 2004) selected the HRTF set that best matched an anthropometric data vector of the pinnae (7 parameters), testing the $[-45°, +45°]$ elevation range in the frontal hemisphere in dynamic conditions. Results showed a general yet not universal decrease of the average elevation error.

Similarly, selection can be targeted at detecting a subset of HRTFs in a database that fit the majority of a pool of listeners. Such an approach was pursued e.g. by So et al. (So et al., 2010) through cluster analysis and by Katz and Parseihian (Katz et al., 2012) through subjective ratings. The choice of the personal best HRTF among this reduced set is, however, left to the listener.

A different selection approach was undertaken by Hwang et al. (Hwang et al., 2008) and Shin and Park (Shin and Park, 2008). They modeled HRIRs on the median plane as linear combinations of basis functions whose weights were then interactively self-tuned by the listeners themselves. Results of the respective tests on a few experimental subjects, although giving mixed results, showed how this method generally reduces the localization error with respect to generic HRTFs, as well as the number of front/back reversals.

4.3.2 Selection criteria

Thanks to the physical connection between the uniqueness of the listener's pinna shape and elevation cues in sound localization, this work exploits the use of the revised pinna reflection model of Sec. 4.2 on a 2D image as a selection mechanism for HRTFs. According to a ray-tracing method,[6] the three main frequency notches of a specific median-plane HRTF can be extracted with reasonable accuracy by calculating the distance between a point lying approximately at the ear canal entrance (which is referred to as the *focus* point) and each point lying on the three pinna contours thought to be responsible for pinna reflections:

[6] This is possible because in the frequency band where notches appear the wavelength is small enough compared to the dimensions of the pinna.

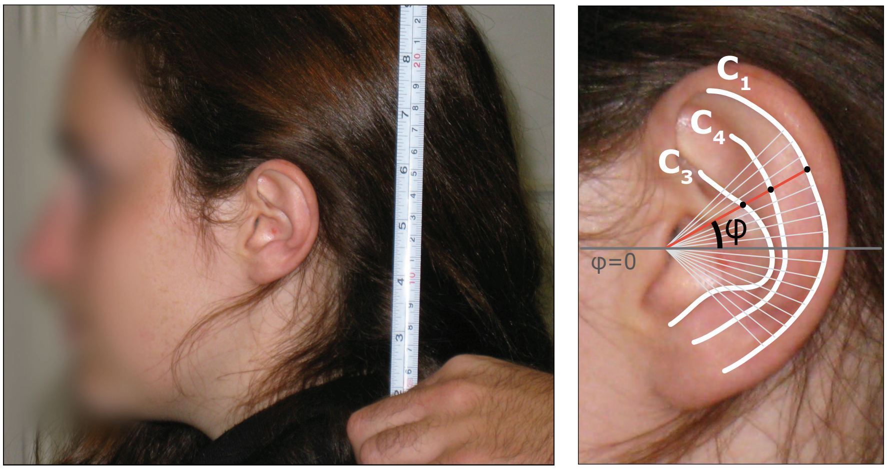


Figure 4.14: *Side-face picture and pinna contours of one subject.*

1. the helix border (C_1 in Fig. 4.14);
2. the antihelix and concha inner wall (C_4);
3. the concha outer border (C_3).

Specifically, given the i-th contour C_i, an elevation ϕ the distance between the pinna reflection point and the focus point is $d_i(\phi) = ct_i(\phi)$. Assuming each reflection to be negative and responsible for a single notch, the corresponding notch frequency, $f_0^i(\phi)$, is calculated using Eq. (4.13).

These frequencies were found to closely approximate notch frequencies appearing in the corresponding measured HRTFs of a number of CIPIC subjects already analyzed in the previous section. Given a subject whose personal HRTFs are not available, it is consequently possible for him to select the HRTF set in a database that has the minimum mismatch between the f_0^i frequencies extracted from his own pinna contours and the F_0^i notch frequencies of the available median-plane HRTF, extracted through a the structural decomposition algorithm (see Sec. 4.1.1). More formally, the above mismatch is defined as a weighted sum of the single contour mismatches given in Eq. (4.18):

$$m = \sum_i \frac{w_i}{n} \sum_\phi \frac{|f_0^i(\phi) - F_0^i(\phi)|}{F_0^i(\phi)}, \tag{4.19}$$

where n is the number of notches in the available HRTFs (typically 3), w_i (with $i = 1, 4, 3$) is a convex combination of weights and ϕ spans all the frontal elevation angles available in the HRTF database between $-45°$ and $45°$.

The relative importance of the pinna contours can be determined by tuning the w_i's. Once fixed, the HRTF set in the database whose mismatch is the lowest is selected.

Table 4.3: *Global mean results of the localization task.*

	S_1 (KEMAR)	S_2 ($w_1 = w_4 = w_3 = \frac{1}{3}$)	S_3 ($w_1 = 1, w_4 = w_3 = 0$)
Azimuth error (mean/SD)	$20.0°$±$3.0°$	$21.7°$±$5.3°$	$21.3°$±$4.5°$
Elevation error (mean/SD)	$31.6°$±$4.4°$	$29.9°$±$5.1°$	$26.2°$±$4.7°$
Linear fit slope (elevation)	0.20	0.30	0.40
r^2 goodness-of-fit (elevation)	0.10	0.17	0.31
Front/back reversal rate	36.6%	32.9%	34.3%
Up/down reversal rate	18.3%	14.7%	9.0%
Externalization rate	62.2%	64.7%	69.7%

4.3.3 Localization task

Eight subjects, 6 males and 2 females, whose age varied from 22 to 40 (mean 27.4, SD 6.1), took part to the localization task. All subjects reported normal hearing according to the adaptive maximum likelihood procedure proposed in (Green, 1993).

Apparatus

The listening tests were performed in a Sound Station Pro 45 silent booth. Sennheiser HDA 200[7] headphones were plugged to a Roland Edirol AudioCapture UA-101 external audio card working at 44.1 kHz sampling rate.

Subjects entered localization judgments in a GUI designed in MATLAB (see Fig. 5.7). In the GUI three different frames required judgments of elevation angle, azimuth angle, and externalization. Perceived elevation[8] was entered by manipulating a vertical slider spanning all elevations from $-90°$ to $90°$ which interactively controlled a blue marker moving onto an arc-shaped profile, very similarly to the input interface described in (Hwang et al., 2008). Perceived azimuth was selected by placing a point in a circular ring surrounding a top view of a stylized human head, inspired by the GUI described in (Begault et al., 2001). The externalization judgment simply required the subject to select one of two answers to the question “where did you hear the sound?”, i.e. “inside the head” or “outside the head”. More details on the software environment can be found in Sec. 5.2.

Stimuli

Stimuli used as sound source signal were a train of three 40-ms gaussian noise bursts with 30 ms of silence between each burst, repeated three times. This type of sound has already been proved to be more effective than a basic white noise burst (Katz et al., 2012). The average measured amplitude of the raw stimulus at the entrance of the ear canal was 60 dB(A).

[7] These dynamic closed circumaural headphones offer an effective passive ambient noise attenuation and high-definition reproduction of high frequencies.

[8] Azimuth and elevation are defined in this case according to the vertical polar coordinate system.

Experimental stimuli were then created by filtering the sound source signal through different HRTF sets and a headphone compensation filter obtained with the algorithm presented in (Lindau and Brinkmann, 2012) applied to measured responses of a KEMAR mannequin without pinnae. It has to be highlighted that compensation was not individual; however, such kind of processing offers an effective equalization of the headphone up to $8-10$ kHz on average and simulates a realistic application scenario where it is not feasible to design personal compensation filters. The HRTF sets were selected among the 45 subjects of the CIPIC database (Algazi et al., 2001d).

Procedure

Acquisition of pinna images was the first step performed in order to compute the mismatch defined in Sec. 4.3.2. An ad-hoc capture environment was created in order to acquire left side-face pictures of the experimental subjects (see Fig. 4.14). In a post-processing phase pictures were first rotated in order to horizontally align the tragus with the nose tip; then, the maximum protuberance of the tragus was chosen as the focus point. Contours C_1, C_4 and C_3 were manually traced and then used to calculate scaled distances from the focus point and consequently the f_0 frequencies as previously described.

For each subject, a fixed HRTF set corresponding to the KEMAR subject with large pinnae (CIPIC ID *21*) was included as control condition. Moreover, two different selection criteria were considered, corresponding to two different convex combinations of the weights in Eq. (4.19). In summary, for each subject three HRTF sets were selected based on the following criteria:

- criterion S_1: KEMAR subject;
- criterion S_2: minimum mismatch m, with $w_1 = w_4 = w_3 = \frac{1}{3}$;
- criterion S_3: minimum mismatch m, with $w_1 = 1$, $w_4 = w_3 = 0$.

It was verified that for each of the tested subjects S_2 and S_3 select different HRTF sets, denoting an adequate pool of subjects in the database and a reasonable differentiation between the two criteria. Subject *21* is also excluded from the candidate selected HRTF sets of S_2 and S_3.

Eighty-five stimuli per HRTF set, each repeated twice, were presented to each experimental subject, for a total of $85 \times 3 \times 2 = 510$ trials. These were generated considering all of the possible combinations of 10 azimuth values (from $-180°$ to $180°$ in $30°$-steps, excluding $\pm 90°$) and 8 elevation values (from $-45°$ to $60°$ in $15°$-steps), plus 5 presentations of the $90°$-elevation point in order to balance the number of stimuli per elevation. Subjects were instructed to enter the elevation, azimuth, and externalization judgments in this specific order for each trial. Each presentation of the 85 positions within a fixed HRTF set, proposed in random order, made up one block of trials, implying that each subject performed a total of 6 blocks. The sequence of presentation of the blocks followed a latin-square design. In order to reduce fatigue of the subject, a 3-minute pause was added between blocks.

Table 4.4: *Elevation results divided per subject.*

ID	Criterion	S_1	S_2	S_3
SA	Mean elev. error	$34.7°$	$37°$	$26.7°$
	Slope	0.094	0.016	0.281
	r^2	0.023	0.001	0.231
SB	Mean elev. error	$25.4°$	$20.4°$	$21°$
	Slope	0.444	0.670	0.606
	r^2	0.303	0.534	0.534
SC	Mean elev. error	$34.9°$	$31.8°$	$30.4°$
	Slope	0.162	0.231	0.252
	r^2	0.184	0.335	0.341
SD	Mean elev. error	$27.1°$	$29.6°$	$18°$
	Slope	0.286	0.231	0.677
	r^2	0.223	0.143	0.627
SE	Mean elev. error	$32.5°$	$30.5°$	$29.3°$
	Slope	0.077	0.115	0.159
	r^2	0.074	0.073	0.196
SF	Mean elev. error	$29.3°$	$27.6°$	$29°$
	Slope	0.309	0.355	0.317
	r^2	0.192	0.249	0.200
SG	Mean elev. error	$37.4°$	$32.4°$	$28.3°$
	Slope	0.026	0.477	0.500
	r^2	0.002	0.208	0.301

Results and discussion

Localization errors in azimuth and elevation were analyzed separately, ignoring front/back confusions on perceived azimuth (with the exception of a $30°$ cone of confusion around $\pm 90°$). Furthermore, linear fitting was performed on the front/back-corrected polar-angle evaluations. One subject who performed elevation judgments at chance performance, corresponding to guessing the direction of the sound (mean elevation error $\approx 45°$), for all three HRTF sets was treated as an outlier and discarded from the analysis.

The mean and standard deviation of localization errors for the three different selections, along with mean linear fit details, front/back and up/down confusion rates,[9] and perceived externalization, are shown in Table 4.3. Note that the adopted criteria have little effect on azimuth localization; this is reasonable as long as the selection is performed on pinna features only and not on the optimization of interaural differences. Similarly, the mean front/back reversal rate is not greatly affected by the HRTF choice, probably because of the number of dominant factors that contribute to its resolution such as dynamic localization cues. However, S_3 remarkably succeeds in significantly improving both the mean externalization and up/down reversal rates − up/down reversals

[9] The up/down confusion rate is calculated with a tolerance of 30° in elevation angle around the horizontal plane, and averaged over all target elevations except ϕ = 0°.

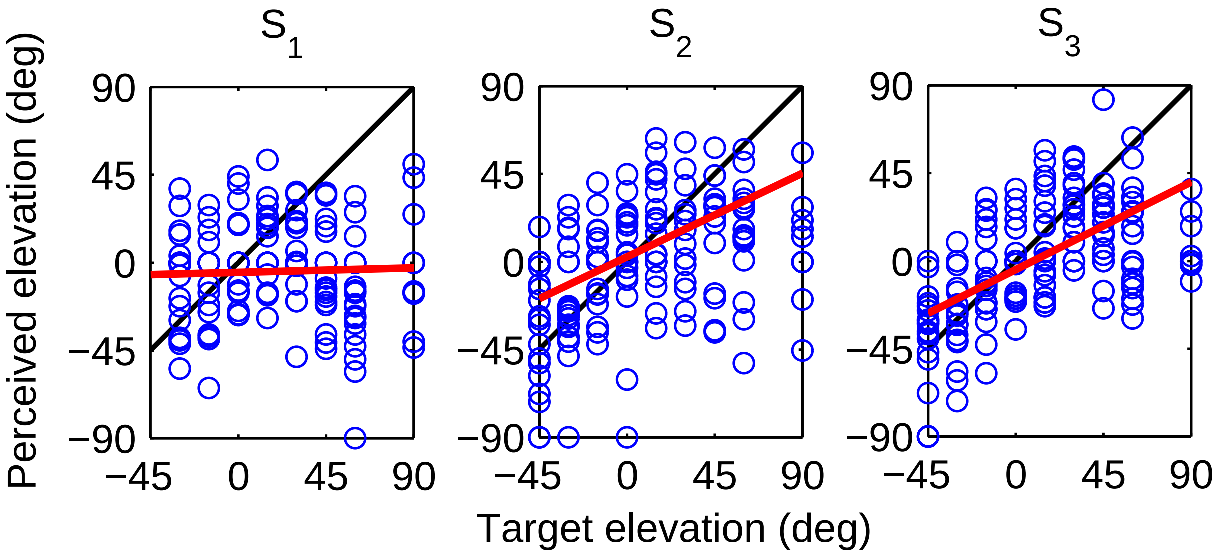


Figure 4.15: *Elevation scatterplots of subject SG.*

are more than halved with respect to S_1. We now concentrate on a more detailed analysis of the elevation results.

Table 4.4 illustrates the elevation-related scores of every subject, i.e. mean elevation error, slope of the linear fit, and r^2 goodness-of-fit. Note that S_1 has the average worst performance, while S_3 always scores better results. S_3 gives an average improvement of 17.4% in elevation error with a peak of 33.6% compared to S_1, suggesting that the most external contour, C_1, has high significance for elevation cues. Conversely, S_2 is found to be unreliable, as its performances are sometimes the best and sometimes the worst among the three criteria. This could be related to the non-individual headphone compensation that introduces spectral distortion starting from around $8 - 10$ kHz, where the spectral notches due to the two inner pinna contours generally lie. Consequently, weights assigned to the two inner contours should be differentiated with respect to that of C_1. More evidence of the benefits brought by S_3 can be appreciated in Fig. 4.15, which reports individual elevation scatterplots of subject SG. Notice the progressive improvement of the elevation judgments along with the three criteria, confirmed by the rise of both the linear fit slope (red line) and the goodness of fit.

As a separate note, a deeper analysis of the results highlighted that the best elevation performances of S_3 are achieved for sound sources coming from the back (with a mean improvement of the elevation error of 28% compared to S_1). This finding highlights that the HRTF selection criterion, even though developed in the front median plane, is robust and positively affects perception in posterior listening space too. Finally, since selection was based on a picture of the left pinna, the results for sources in the left and right hemispheres were compared. No significant differences were found, allowing to conclude that for the tested subjects the chosen ear did not influence elevation judgments.

The average improvement can be compared to the results found by Zotkin et al. (2004), where the increase of the elevation performance between a generic HRTF and a HRTF selected on anthropometric parameters was reported to be around 20-30% for 4 subjects out of 6. However, a more careful calculation of the average performance on all six subjects shows that the average elevation error decrease is about 6.5%. Still, these results are not directly comparable to theirs because of the different experimental conditions (e.g. presence of head tracking, use of a hand pointer for localization, different elevation range, small number of stimuli).

4.4 Anthropometric parametrization of the pinna model

In this section, three set of parameters for the pinna pHRTF structural model presented in Sec. 3.3.1 are objectively evaluated in terms of the spectral distortion introduced by their approximation of target PRTFs. In the following sections, we refer to:

- H^s, HRTFs given by the fully resynthesized model: all parameters are directly extracted from target responses using the structural decomposition algorithm;
- H^c, HRTFs resulting from the contour-parameterized model: notch central frequencies are estimated from contours of pinna images, while parameters for the resonant component and notch gains and bandwidth are estimated from the structural decomposition algorithm;
- H^a, HRTFs built through the fully synthetic model: notch central frequencies are estimated from contours of pinna images, while parameters for the resonance component, and notch gain and bandwidth are estimated as average values over subjects.

The aforementioned parameters lead to three models in decreasing order of customization. The last one might be probably the most suitable version for a commercial use in terms of handiness.

4.4.1 Fully resynthesized model

Section 4.1.6 has shown that a PRTF at one specific elevation includes three main resonances. Given a target PRTF, one can then deduce center frequency and bandwidth of each resonance directly from $|N_{\text{res}}|$ estimated by the structural decomposition algorithm and use these parameters to design three second-order bandpass filter which approximate them. An adequate filter structure uses the three bandpass filters in parallel. Note that the higher resonance may be perceptually irrelevant since it lies near the upper limit of the audible range. We could then consider two resonances only and consequently simplify this block with the filter realization of Sec. 3.3.1; nevertheless, psycho-acoustical criteria are needed to justify such simplification.

For what concerns the reflection block, the same analysis performed by the structural decomposition algorithm over $|N_{\text{refl}}|$ was followed and proceed as in Sec. 4.1.3 and 4.1.4 in order to produce a multi-notch filter. A stricter amplitude threshold for notches is considered, and no reduction factor is used ($\rho = 1$). Each notch in the spectrum of N_{refl} is then characterized by its central frequency, gain, and 3dB bandwidth, which are fed to the multi-notch filter construction procedure. Thus, the order of the resulting filter is twice the number of considered notches and the amplitude threshold specification is set to 3 dB.

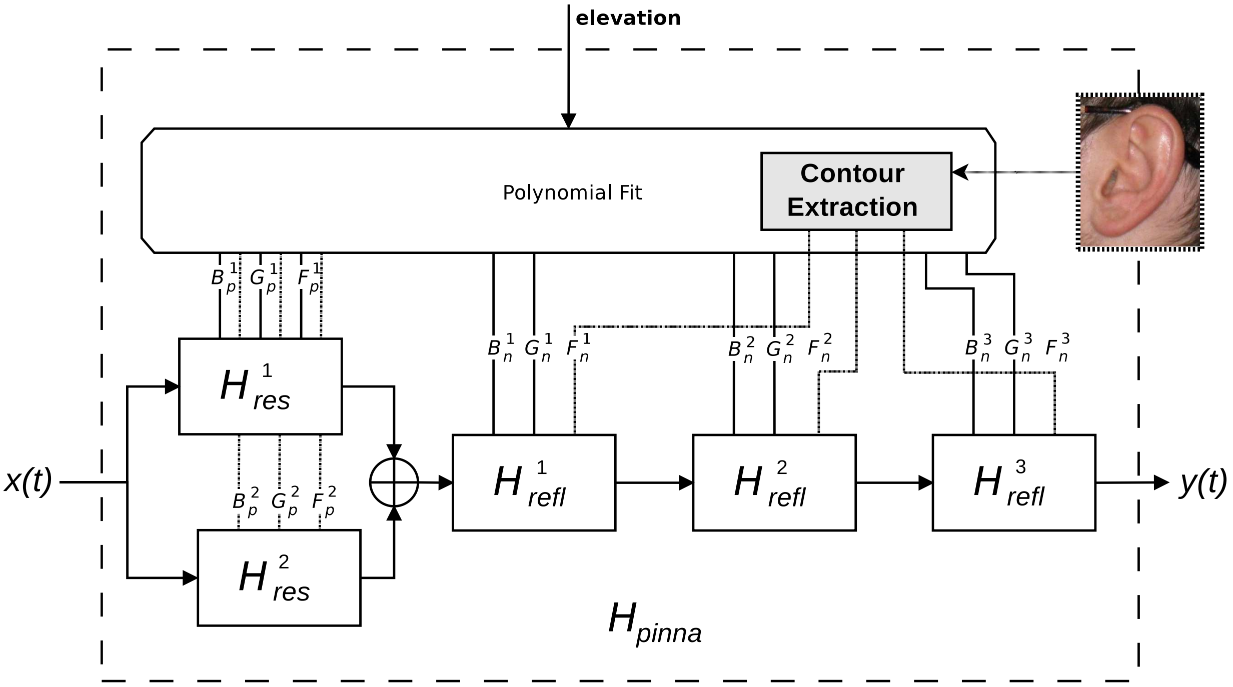


Figure 4.16: *The structural pinna pHRTF model.*

Finally, since the effects of the PRTF are limited to the frequency range $3-18$ kHz, one can cascade a bandpass filter to the whole structure in order to cut undesired frequencies.

4.4.2 Contour-parameterized model

With reference to Fig. 4.16, the only independent parameter used by the pinna block is the source elevation ϕ, which drives the evaluation of the analytic functions describing resonances' center frequencies $F_p^i(\phi)$, 3dB bandwidths $B_p^i(\phi)$, and gains $G_p^i(\phi)$, $i = 1, 2$, as well as the corresponding notch parameters ($F_n^j(\phi)$, $B_n^j(\phi)$, $G_n^j(\phi)$, $j = 1, 2, 3$). Only the center frequencies F_n^i are customized on the individual pinna shape, hence 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 parametrization of the filters. All the polynomials must be computed offline previous to the rendering process and can be extracted by means of the structural decomposition algorithm, except F_n^i, leading to the H^c model. These functions will be used in the model to continuously control the evolution of the reflection/resonant component when the sound source is moving along elevation.

Functions $F_n^j(\phi)$ can be extracted from the subject's anthropometry (in the form of a pinna picture): contours C_2 or C_1 (depending on whether the subject's ear is respectively protruding or not), C_4, and C_3 are converted into distances with respect to the ear canal entrance, and then translated into sequences of frequencies through Eq. (4.13), thus assuming overall negative reflection coefficients. Again, a fifth order polynomial is best fitted to these sequences, resulting in functions $F_n^j(\phi)$, $j = 1, 2, 3$.

The resonant part is modeled with a parallel of two different second-order peak filters, following Eqs. (3.15) and (3.18). The notch filter implementation has the same form as the first peak filter with the only difference in the k parameter description, see Eq. (3.20).

4.4.3 Fully synthetic model

It had been shown in Sec. (4.1.6) that there are no clear elevation-dependent patterns for gains and bandwidths of notches. Moreover, no anthropometric parametrization is available for these parameters yet. Therefore, the mean of both gains and bandwidths for all tracks and elevations ϕ among all subjects is computed, and again a fifth-order polynomial dependent on elevation is fitted to each of these sequences of points, yielding functions $G_n^j(\phi)$ and $B_n^j(\phi)$, $j = 1, 2, 3$. These statistically derived polynomials are fed to structural model H^a.

Similarly, the mean magnitude spectrum of the resonant component in the median plane among CIPIC subjects (see Fig. 4.6) is considered for the peaks parameter extraction of H^a. More in detail, a naïve procedure is applied: extracting for every available elevation angle the two maxima of the mean magnitude spectrum, which outputs the gain G_p^i and central frequency F_p^i of each resonance peak, $i = 1, 2$, and the corresponding 3DB bandwidth B_p^i. Then, a fifth-order polynomial (with the elevation ϕ as independent variable) was fitted to each of the former three parameters, yielding the functions $G_p^i(\phi)$, $F_p^i(\phi)$, and $B_p^i(\phi)$, $i = 1, 2$.

4.4.4 Evaluation

In order to objectively evaluate the first two models against the original measured HRTFs in the CIPIC database we consider an error measure based on spectral distortion (SD) between the original response, H, and the reconstructed response, $\tilde{H}$ (see Sec. 3.1.2 for more details). The frequency range is limit to [500,16000] kHz.

Fig. 4.17 reports SD values, averaged across the 18 non-outlier CIPIC subjects, 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^s_{res}, H^c_{res};
4. the all-round response of the fully resynthesized model, H^s_{tot};
5. the reflective component of the fully resynthesized model given by notch filters, H^s_{refl}.

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

As expected, H^c_{tot} is the response with the highest average SD. As a matter of fact, errors in the resynthesized resonant (H^s_{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 on average and is worse

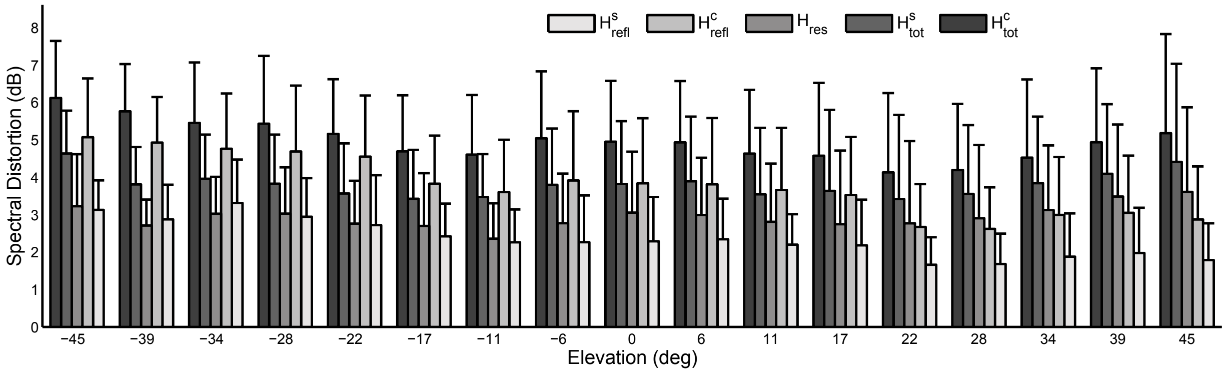


Figure 4.17: *Spectral distortion between reconstructed and measured median-plane HRTFs (mean and standard deviation over* 18 *CIPIC subjects).*

at negative elevations. This fact can be explained by the occurrence of very deep notches at these 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 mean SD of H^{c}_{tot} tends to decrease except for a new rise at the last elevation angles, which is conversely due to greater errors in the resonant component H_{res}. An informal inspection of resonant components at higher elevations revealed indeed that the second modeled high-frequency peak (horizontal mode) disappears, gradually letting non-modeled lower-frequency vertical modes in. The appearance of such modes also brings a significant rise of the SD variance in the all-round responses at the highest elevation angles.

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^{s}_{refl} distinctly decreases both in mean (3 dB or less) and variance, resulting in a noticeably lower average SD (about 4 dB) in the total response H^{s}_{tot}.

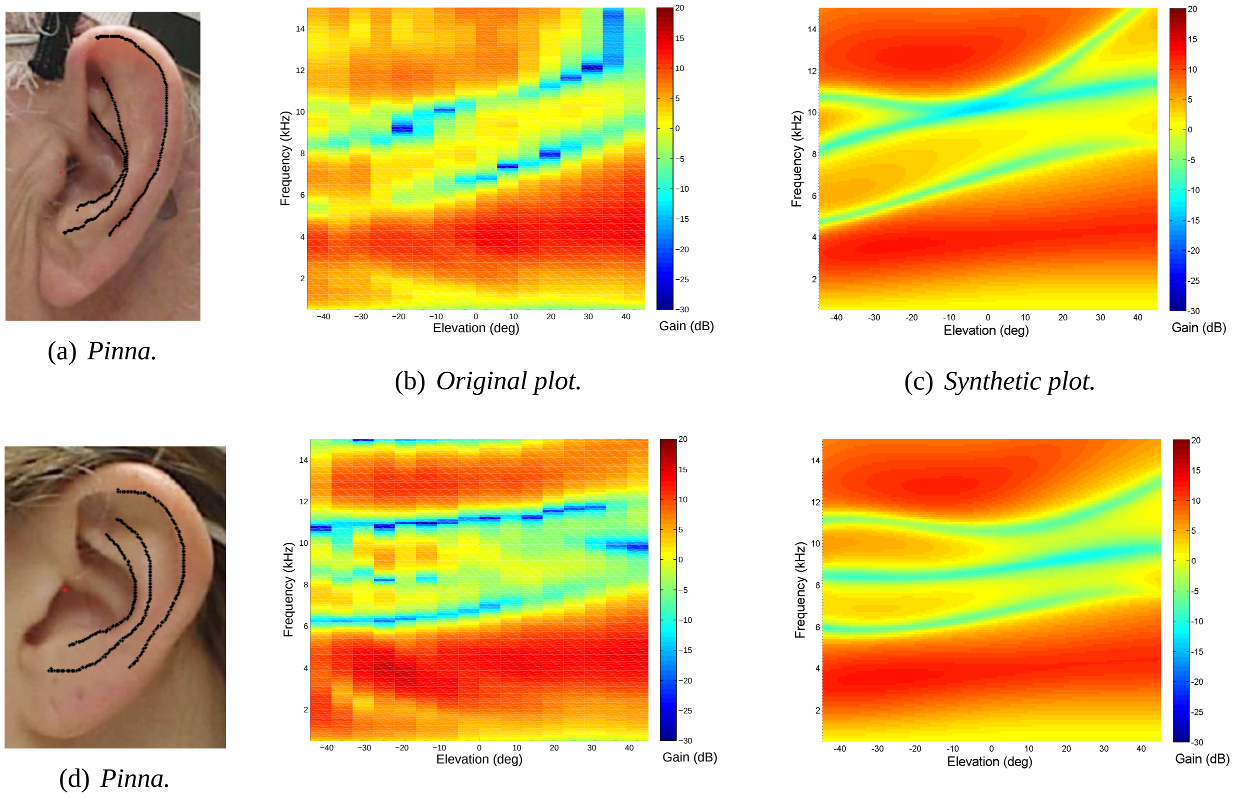
On the other hand, model H^{a} was tested on different CIPIC subjects; in the following, the results for two of them, Subject 020 and Subject 048, are presented. In both cases, the tracing procedure marks the rim border, concha wall and border, and antihelix. Since the concha back wall is not fully visible from the picture's lateral view of the head, a tentative contour for this surface was drawn (see Sec. 5.4 for details about the contour extraction procedure).

Having fed the model with all polynomial functions evaluated at half-degree elevation step, we are now ready to compare the original versus synthesized HRTF magnitude plots, shown in Fig. 4.18. We focus on the frequency range up to 15 kHz where all the relevant informations are included, spanning elevations between -45 and 45 degrees in the median plane.

Besides the 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 behavior in all subjects, is well approximated in both cases;

2. The extracted notch tracks, although much smoother than the original ones, closely follow



(a) *Pinna.*

(b) *Original plot.*

(c) *Synthetic plot.*

(d) *Pinna.*

(e) *Original plot.*

(f) *Synthetic plot.*

Figure 4.18: *(a) Contour extraction on Subject* 048*'s pinna. (b) Original and (c) synthetic HRTF magnitude,*$|H^a|$*, plots for Subject* 048*. (d) Contour extraction on Subject* 020*'s pinna. (e) Original and (f) synthetic HRTF magnitude,*$|H^a|$*, plots for Subject* 020*.*

the measured patterns, attesting fitness of the contour extraction and mapping procedure;

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

On a closer inspection, it can be noted that Subject 020 originally exhibits a wide dip around $\phi = 40°$ in the highest frequency range which is not correctly reproduced; this may be due to the superposition of two or more notches that cannot be detected when tracing the pinna contours. As for Subject 048, comparison of his pinna picture with the original HRTF plots suggests a relationship between the shorter antihelix and concha wall reflection surfaces and two distinct notch tracks, the first located around 8 kHz at negative elevation and the second around 10 kHz at positive elevation. Since three contours are modeled, these two notches are collapsed in one continuous track, see Fig. 4.18(f). A further notch appears around 15 kHz, yet it is likely associated with a mild pinna contour.

4.4.5 Discussion

As a conclusion to the presented results, if one assumes that the aforementioned mismatches are in most cases not perceptually relevant, one can then consider the mean SD of 4 dB in H^s_{tot} as a satisfactory result, being comparable to SD values found in similar works that deal with HRTF resynthesis by means of HRIR decomposition (Faller II et al., 2010) or anthropometric parametrization through multiple regression analysis on HRTF decomposition (Nishino et al., 2007). Furthermore, the proposed model is composed of first- and second-order filters only: given that many responses exhibit sharp notches whose shape cannot be reproduced 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.

In synthesized models of the Subject 020 and 048, the second resonance is clearly overestimated and its shape does not find a strong visual correspondence with its original counterpart. Such mismatch highlights a complex spectrum evolution due the presence of two or more resonances interacting in the higher frequency range for elevations in proximity of the horizontal plane Shaw (1997). 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. Furthermore, the CIPIC HRTF database used in this study does not include elevation data below $-45°$. Alternative HRTF data sets or BEM simulations should be considered in order to extend the ray tracing procedure to the range $-90° < \phi < -45°$.

Psychoacoustical evaluations in the context of virtual environments are needed to assess the effectiveness of this approach in improving user's sense of presence and immersion, together with perceptive relevance of using such homogeneous notch and peak shapes.

4.5 Future perspectives

In this chapter, a mixed structural approach for estimating, modeling and selecting the pinna pHRTF was presented. An algorithm that separates the resonant and reflective parts of the PRTF spectrum was firstly implemented and then such decomposition was used to resynthesize the original PRTF through a low-order filter model. Results showed an overall suitable approximation to the original PRTFs.

Ongoing and future work in order to extend the structural decomposition algorithm includes:

- improvements in the analysis algorithm: in particular through the use of a better multi-notch filter design and extending the analysis in sagittal planes;

- enhance the tracking of frequency notches through the McAulay-Quatieri partial tracking algorithm, in order to obtain a robust and continuous representation of frequency notches along elevation;
- performing regression of PRTF data over anthropometrical measurements towards functional representation of resonances and notches.

An analysis of real HRTF data in order to study the relation between HRTF features and anthropometry in the frontal median plane supports 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 an approach to HRTF customization, mainly based on structural modeling of the pinna contribution, was proposed. Spectral distortion and notch frequency mismatch measures indicate that this approximation is objectively satisfactory.

The pinna model as it was integrated in the structural model of Sec. 3.3.2 represents a notable extension of the one in (Satarzadeh et al., 2007) as it includes a large portion of the frontal hemispace, and could thus be suitable for real-time control of virtual sources in a number of applications involving frontal auditory displays, such as a sonified screen (Walker and Brewster, 2000). Further extensions of the model, such as to include source positions behind, above, and below the listener and also in sagittal planes, may be obtained in different ways.

Furthermore, the exploitation of the pinna reflection model for HRTF selection is promising and the reported experiment confirms these expectations. Compared to the use of a generic HRTF with average antropometric data, the pinna reflection approach increases the average elevation performances of 17%, significantly enhancing both the externalization and the up/down confusion rates. The selection criterion assigning the whole weight to contour C_1 gives the best results. Indeed, pinna contours may have different weights and could play different roles in the selection. As future work, one can exploit the three contours in a tuning process: while C_1 will be used to prune the candidate HRTF sets, the remaining contours will select the “best” HRTF set among the remaining.

Subjective evaluations that take into account both structural model and selection criteria will allow to understand the influence of notch gain and bandwidth in elevation perception as well as the relation between the resonant component of the PRTF and the shape of pinna cavities. All these information are essential requirements in order to have a complete anthropometric parametrization of the pinna model. It will be necessary to perform listening tests on subjects for which individual recorded HRTFs are available, in order to have a “ground-thruth” for the evaluation of structural models obtained with the MSM approach.

Finally, it is worthwhile to mention that the the listening setup comes closely to a feasible scenario for practical applications (e.g. no individual HRTFs for comparison, non-individual headphone compensation); in light of this, the next chapter presents a tool that automatically extracts pinna contours from a set of 2D images (Spagnol et al., 2013c). An extension of the reflection model to three dimensions, e.g. applied to 3D meshes of human pinna, would greatly improve the accuracy of the extraction, modeling and selection processes, provided that handiness of the system is not reduced too drastically.

Chapter 5

Personalized 3D audio environments

Headphone-based reproduction systems driven by head tracking devices – if properly designed – allow tailoring immersive and realistic auditory scenes to any user without the need of expensive and cumbersome loudspeaker-based systems.

This chapter gives an overview of a number of tools for the analysis and synthesis of HRTFs that we have developed in the past four years at the Department of Information Engineering, University of Padova, Italy. The main objective of our study in this context is the progressive development of a collection of algorithms for the construction of a totally synthetic customized HRTF set for a personalized 3D audio environment, replacing both time consuming and expensive individual HRTF measurements and the use of inaccurate non-individual HRTF sets. The research methodology is highlighted, along with the multiple possibilities of present and future research offered by such tools.

A system for customized binaural audio delivery based on the extraction of the relevant features from a 2D representation of the listener's pinna is then presented. Particular attention is given to the contribution of the external ear to the HRTF that heavily depends on the listener's unique anthropometry. An automatic procedure estimates the location of pinna edges starting from a set of pictures produced by a multi-flash imaging device. Image processing algorithms designed to obtain the principal edges and their distance from the ear canal entrance are described.

In particular, the shape of the most prominent contours of the pinna defines the frequency location of the HRTF spectral notches along the elevation of the sound source as described in Sec. 4.2. The resulting contours drive the parametrization of a mixed structural HRTF model that performs in real time the spatialization of a desired audio stream according to the listener's position with respect to the virtual sound source, tracked by sensor-equipped headphones. The low complexity of the model allows smooth implementation and delivery on any mobile device. The purpose of the developed system is to provide low-tech custom binaural audio to any user without the need of subjective measurements. Design guidelines for hardware and software requirements are discussed and the effectiveness of a first prototype is preliminarily evaluated on a small number of test subjects.

This chapter is partially based on (Geronazzo et al., 2013e; Spagnol et al., 2013c).

5.1 System architecture guidelines

Given an arbitrary sound scene and room by placing an array of microphones in a specific location, it is ideally possible to reproduce the recorded sound field on a fixed array configuration of loudspeakers; moreover, one can virtually render virtual sound sources on top of it. The virtual acoustic installation built in this way can be used by anyone, however the listener's movements are restricted to a small sweet-spot.

Among the diverse possibilities offered by spatial audio technologies, binaural headphone-based reproduction systems - if properly designed - allow tailoring immersive and realistic auditory scenes to any user everywhere without the need of loudspeaker-based systems. These technologies integrate well with mobile devices in particular (Fig. 5.1 schematizes such kind of system): no integrated loudspeaker can guarantee the same audio quality as a pair of common earbuds or headphones; additionally, headphones are typically used on the move without the need to hold anything in hand. However, the next generation of portable audio device has to overcome several limitations. Generally speaking, the mid-long term of these studies relies on the following goals:

- fully integrate 3D audios engine in mobile devices, ensuring an *"everywhere"* playback;
- make the devices be capable of analyzing and synthezing immersive sound fields *"every time"* it is needed, on demand;
- develop adaptive technologies that are able to fit the individual listening experience *"for everyone"*.

The research framework presented in the above sections answers both the requirements of structural modularity and systematic HRTF model evaluation. The modus operandi, defined with the aim of designing new synthetic filter models and HRIR/pHRIR selection processes is expected to progressively set the bar closer and closer to a complete individual structural HRTF model suitable for real-time auralization. Models, whose details can be found in Ch. 3, are designed so as to avoid expensive computational and temporal steps such as HRTF interpolation on different spatial locations, psychoacoustic selection of HRTFs, or the addition of further artificial localization cues, allowing implementation and evaluation in a real-time audio processing environment.

5.1.1 Mobile & portable devices

Listeners can experience being surrounded by realistic virtual acoustic scenes anywhere they desire. This idea is characterized by the use of portable devices: mobile computers (e.g. smart phones, tablets, etc.) as control and computational unit and headphones (e.g. ear-buds, bone-phones, etc.) as playback unit. The key features of such devices reside on the following requirements:

- being confortable for extended use;

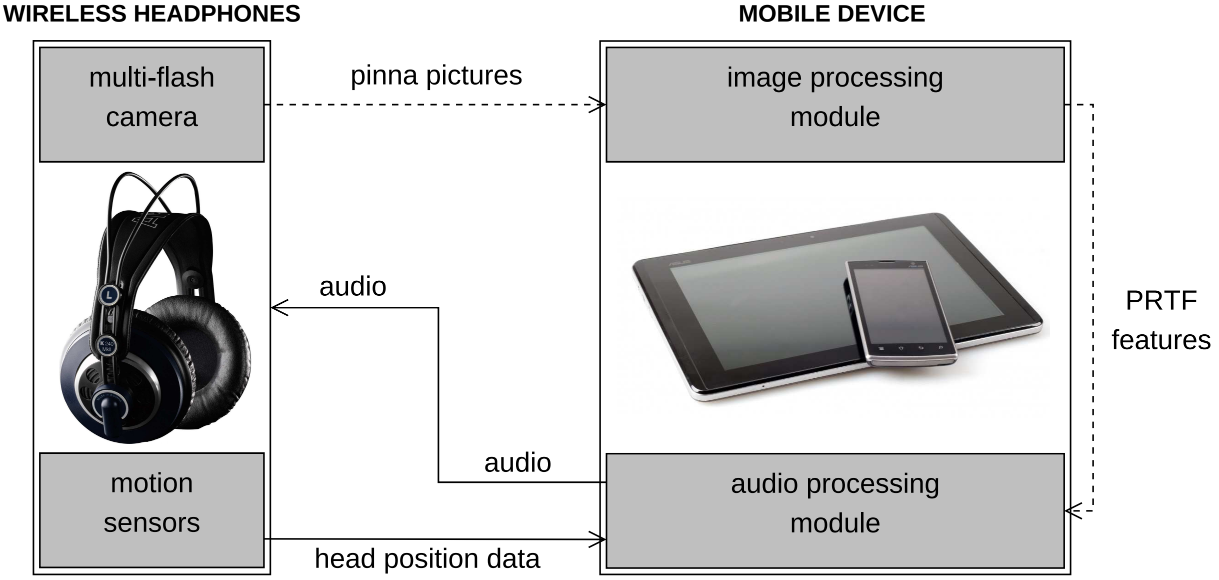


Figure 5.1: *A simplified scheme of the system's architecture and software. Broken line arrows refer to offline data exchange, solid line arrows to real-time data exchange.*

- providing different degrees of acoustic isolation from real environments.

The first aspect deals with the system invasiveness in activities of daily living introduced by its regular usage. Several distinctions can be addressed here depending on the application domain and resulting requirements; for instance, motor rehabilation (see Sec. 6.1 for a case study) or navigation aids for visually impaired people (see Sec. 7.2 for a case study) need special equipments and devices. Generally speaking, the system becomes invasive if no useful information counterbalances its physical presence and management.

On the other hand, the latter aspect controls the degree of superposition of virtual acoustic scenes on the surrounding audio reality. According to Sec. 1.2, individual headphones characterization able to adapt to listener anthropometry will strongly contribute to externalization and thus to the degree of immersion. Moreover, the devices might incorporate external acoustic knowledge through real-time recordings made by binaural microphones located on the headphone cups (Härmä et al., 2004; Christensen et al., 2013). Once properly compensated, these information acts as noise cancellation, otherwise it might contribute to the acoustic transparency of headphones allowing the superposition of virtual sound sources on top of the recorded soundscape.

5.1.2 Real listening scenarios and virtual acoustic scene

The "every time" requirements regards real-time constraints and, specifically, technologies that ensure different degree of complexity and immersion in the virtual scenarios.

In order to fully exploit the potential of mixed structural models in both static and dynamic listening scenarios, an appropriate audio device equipped with sensors able to detect the relevant parameters to fine tune the model both before and during listening is needed. Head pose estimation (head position and orientation) and body tracking with respect to the device local coordinate system and absolute positioning systems inside the real world are essential components in determining the degree of immersion and realism of the virtual/augmented acoustic scene.

Applications that require the device to be anchored to the *World Reference Frame* (WRF) can consider the use of GPS for outdoor environment and Wi-Fi signals, radio-frequency identification (RFID), Bluetooth or hybrid approaches for indoor environment; nowadays, position errors are still higher than 1 m (see Liu et al. (2007) for an extensive review) limiting the spatial precision and distance estimations.

Rendering of high realistic acoustic scenes is intrinsically multi-source, thus VADs require audio engines that are able to handle complex acoustic environments in terms of number of sources and their interaction with the virtual room. The most promising technologies to date are based on the increasingly ubiquitous *general purpose computing on graphics processing units* (GP-GPU) (Savioja et al., 2011; Belloch et al., 2013; Röber et al., 2007). HRTF convolution is parallelized on GP-GPU speeding up the computation when compared with CPUs. As a recent example, Belloch et al. (2013) are able to simultaneously render ≈ 240 virtual sound sources, i.e. parallel convolution with multiple measured HRTFs at different source positions.

Nothing prevents to apply these technologies to binaural rendering over headphones for sound field recordings acquired with microphones arrays (e.g. Eigenmike® [1]). In this research direction, the main challenges deal with:

- how to measure the sound field with an adequate spatial resolution (Zotkin et al., 2010);
- how to handle perceptual audio artifacts introduced by a non optimal HRTF spatial grid; this issue mainly regards the number of rendered virtual loudspeakers (Laitinen and Pulkki, 2009; Nishimura and Sonoda, 2013).

5.1.3 Why customize spatial audio?

One of the main limitations of binaural audio through headphones that cause its exclusion from commercial applications in virtual and augmented reality lies in the lack of individualization of the entire rendering process. Everyone should be able to benefit from a personal (in an acoustic point of view) portable VAD: it has to manage listeners anthropometry and psycho-acoustical sensitivities that exhibit high variance across subjects.

A common practice employs the trivial selection of an unique HRTF set for all listeners (i.e. recorded on a dummy head). However, as attested in the previous Chapters, anthropometric features of the human body have a key role in HRTF shaping. In this work, MSM approach is investigated in order to select non-individual HRTF sets from an existing database and model synthetic individual HRTF sets, according to anthropometric criteria extrapolated by easy to manage (handiness) geometrical information. The research methodology drives to the best MSM with similar individual HRTF localization performance thanks to the closer relation between listener's geometry and localization cues.

[1] http://www.mhacoustics.com/products#eigenmike1

It is worthwhile to notice that acquisition procedures for geometrical information have different handiness and precision, for instance an updated review of 3D scanning methods of pinnae can be found in Reichinger et al. (2013). Recent trends in HRTF individualization involve rapid numerical simulation (via boundary element methods) on individual 3D mesh models, far from finding an acceptable balance between handiness and accuracy. The continuing growth of available computing power (e.g. parallelism, cloud computing, quantum computing, etc.) suggests that this kind of approach will become feasible in the next years, although the time when it will be embedded in a portable device is unpredictable.

5.1.4 Multimodal integration

The modular approach described in this thesis merits a brief mention also in the multimodal domain where the integration of a 3D audio rendering engine with other sensory cues such as video and haptics requires new tools for the evaluation of integration, cross-augmentation and/or substitution of different modalities. Thus, all the previous guidelines have to be extended in a multimodal scenario with particular attention on the specific modality being studied.

A recent example is provided by Stamm and Altinsoy (2013), where the localization mechanism was investigated as a multimodal process (they consider hearing, haptics and proprioception). Localization prediction of binaural models were evaluated in a multimodal scenario in order to assess to which extent auditory cues contributes in human perception. Novel interfaces could surely benefit from such kind of studies, e.g synthesizing auditory localization cues that guide and assist proprioceptive space perception especially in large virtual workspaces. Moreover, the authors highlighted the relevance of employing individual HRTF in their future studies in order to determine the positive influence of spatial auditory cues. To this regard, the MSM approach can help to identify and quantify the contribution in each spatial dimension.

5.2 The research framework

This section gives a brief overview of a number of tools for the analysis and synthesis of HRTFs, highlighting in particular the mixed structural model research methodology proposed in Ch. 2.1 along with the diverse possibilities of present and future research offered by the mentioned tools. The main objective of these studies in this context is the progressive development of a collection of algorithms for the construction of a totally synthetic HRTF set suitable for real-time rendering of custom spatial audio, taking the listener's anthropometric parameters as the sole input to the audio chain. Such a modeling philosophy descends from the structural approach by Brown and Duda (Brown and Duda, 1998): the global contribution of the listener's body to the HRTF is split into smaller blocks or modules, and each module contains a measured, reconstructed or synthetic response, as will be made clearer throughout the following sections. This approach differentiates from recent trends in HRTF customization because no self-tuning of parameters or selection of responses from databases will be required in principle by the listener.

For the sake of clarity, the discussion follows this logical distinction among the fundamental components of the presented framework at file system level. The *database* folder acts as the main

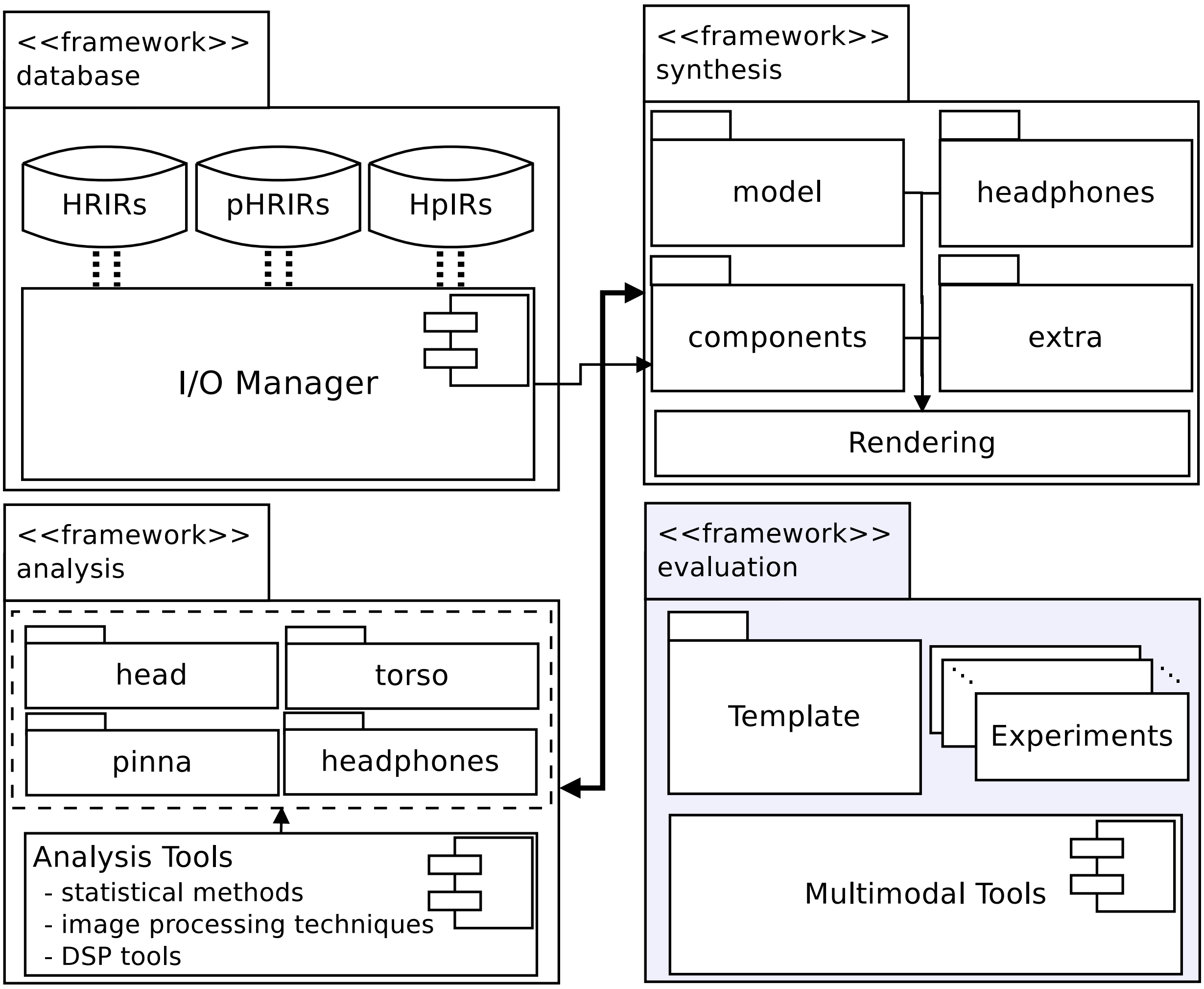


Figure 5.2: *UML-like representation of the research framework structure.*

data container, while all of the algorithms that extract relevant features in the available data are stored in the *analysis* folder. The *synthesis* folder contains the tools for spatial audio rendering designed and developed with an eye to real-time constraints. Finally, the experimental tools for subjective evaluation of models and the related data are organized in the *evaluation* folder. The UML-like diagram in Fig. 5.2 depicts all the components of the framework.

5.2.1 HRIR and HpIR databases

The included data is under the form of several sets of HRIRs recorded for a high number of subjects in different spatial locations, and sets of HpIRs for the characterization of different headphone models used for compensation in the reproduction process. Beside the full-body HRIR repository, a similar container includes the partial HRIRs, i.e. pHRIRs, recorded by isolating specific body parts (e.g. pinna-related impulse responses measured on the isolated pinna, or impulse responses measured on a pinnaless mannequin) or resulting from the decomposition carried on by the algorithms mentioned in Sec. 4.1.

HRIR and pHRIR databases

As already discussed, a number of publicly available HRIR databases exist, the most notables of which are the CIPIC HRTF database (Algazi et al., 2001d)[2] and the LISTEN HRIR database (Eckel, 2001)[3]. The main differences among these and other databases concern the type of stimulus used, the spatial grid of the measured HRIRs, and the microphone configuration (blocked- or open-ear-canal, distance from the eardrum, etc.).

An attempt to unify the aforementioned variability in HRIR databases gave birth to the MARL-NYU data format (Andreopoulou and Roginska, 2011); CIPIC, LISTEN, FIU (Gupta et al., 2010)[4] and KEMAR-MIT (Gardner and Martin, 1995)[5] databases were stored in this format, which organizes the information into data and specification sections. The described repository takes the MARL-NYU format as a starting point towards the introduction of some additional relevant information:

- the raw HRIR data in addition to the already available compensated version;
- the HRIR's onset sample;
- the coordinate system adopted for the measurements and the relative management of the measurement space.

Furthermore, four more databases were stored and fitted to the repository:

1. the Aalto HRIR database (Gómez Bolaños and Pulkki, 2012),[6] which includes actual source direction data;
2. the ARI HRTF database,[7] collecting both in-ear and behind-the-ear measurements;
3. the PKU&IOA HRTF database (Qu et al., 2009),[8] containing near-field recordings from a KEMAR dummy head;
4. the Aalto PRTF database (Spagnol et al., 2011),[9] which collects pinna-related impulse responses in the mid-sagittal plane as pHRIRs.

As Fig. 5.3 sketches, each single measured subject can be associated to N different HRIR sets. Each set corresponds to a different measurement session: for instance, open-ear canal

[2] http://interface.cipic.ucdavis.edu/

[3] http://recherche.ircam.fr/equipes/salles/listen/

[4] http://dsp.eng.fiu.edu/HRTFDB/main.htm

[5] http://sound.media.mit.edu/resources/KEMAR.html

[6] http://www.acoustics.hut.fi/go/aes133-hrtf/

[7] http://www.kfs.oeaw.ac.at

[8] http://www.cis.pku.edu.cn/auditory/Staff/Dr.Qu.files/Qu-HRTF-Database.html

[9] http://www.dei.unipd.it/~spagnols/PRTF_db.zip

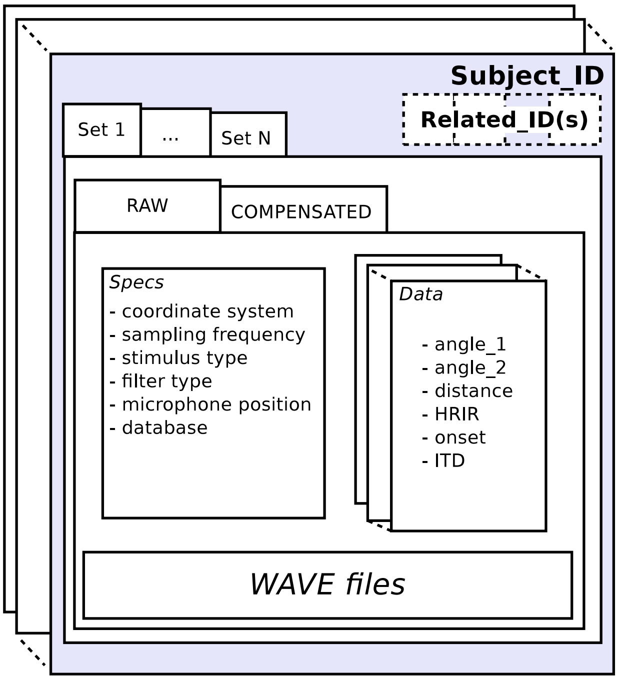


Figure 5.3: *HRIR/pHRIR repository structure.*

and closed-ear canal recordings of a given subject make up two different sets. Each HRIR set is in turn associated to a data structure composed of two parts storing raw and compensated information. Both raw and compensated datasets are stored as .mat and .wav files. The .mat structure is divided into *specs* and *data* sections.

The first section reports all the required information about the subject, to whom a unique ID number is assigned, and the adopted measurement technique. The following details are included:

- *coordinate system*: interaural polar or vertical polar;
- *sampling frequency* (in Hz) of the HRIRs;
- *stimulus type*: excitation signal used for measurements;
- *filter type* (in compensated version only): filter used in the compensation of the HRIR (e.g. minimum-phase or fixed filter);
- *microphone position*: e.g. at the entrance of the blocked/open ear canal, at the eardrum, etc.;
- *database* (optional): name of the database to which the responses belong.

A crucial issue with respect to the first point is how to interpret the polar ranges for azimuth $\theta = [-180°, 180°)$ and elevation $\phi = [-90°, 90°]$ of the MARL-NYU database container, because the inclusion of a database likely implies a conversion of the angular ranges. This is the main reason why the information about the coordinate system is included.

The data section defines the details of each different HRIR measurement. Three fields univocally determine the spatial position of the sound source: azimuth and elevation angles (in degrees), and *distance* (in meters) from the center of the subject's head.

Depending on the coordinate system, azimuth and elevation have different definition and range (see Sec. 1.2.1 for more details). Such a differentiation could cause confusion and ambiguity (as it happens in the MARL data format); this is managed in the repository by calling the two angles $angle_1$ and $angle_2$: $angle_1$ is the one with range $[-90, +90]$, while $angle_2$ has $[-180, +180)$ range. Points with equal $angle_1$ describe a circle which is parallel to the median plane (interaural polar case) or parallel to the horizontal plane (vertical polar case).

The remaining fields of the data section are dedicated to the result of the measurement:

- *HRIR*: pair of vectors relative to the left and right ears containing the HRIR samples;
- *onset*: pair of indices indicating the onset sample of each HRIR vector, calculated as the sample which precedes the last zero-crossing before the main impulse of the HRIR;
- *ITD*: difference between the left- and right-HRIR onsets; if the sound source is on the left (right) the ITD is negative (positive).

Finally, the subject's anthropometric data is saved in the *anthropometry* folder and associated to a single set of HRIRs. As a matter of fact, each HRIR set has to be seen as a still frame of the subject's anthropometry during those measurements, which is not guaranteed to remain unchanged in a future measurement session. The data format is not defined at the moment, yet it is desirable to uniform heterogeneous anthropometric information in a coherent data format in line with biometrical standards (Abaza et al., 2010).

The pHRIRs are organized in the same manner as the HRIRs; it is care of the one who includes the partial responses to keep track in a *comment* field of which structural component is associated to those signals. The HRIR and pHRIR repositories use different subject ID enumerations with the constraint that a subject whose partial responses are included in the pHRIR repository is linked to a corresponding subject in the HRIR repository through the *related_ID* field. Such a subject always exists, possibly with a corresponding HRIR data structure containing his/her anthropometric data only.

The resulting effort towards a standardization proposal has given birth to BTDEI format, schematized in Fig. 5.4. This data exchange format permits a transparent usage of HRTF/pHRTF data independently from original formats (passing through a conversion layer) and a modular approach aiming to easily add any required functionality, whether it is an analysis algorithm and/or a 3D audio application.

HpIR databases

The proposed standardization of HpIR databases follows an organization similar to the one introduced in the previous section for HRIR/pHRIR databases. The key observations (also identified in Sec. 1.2.2) that guide the design of such structure (see Fig. 5.5) are:

- HpIRs are highly sensitive to the positioning of headphones;

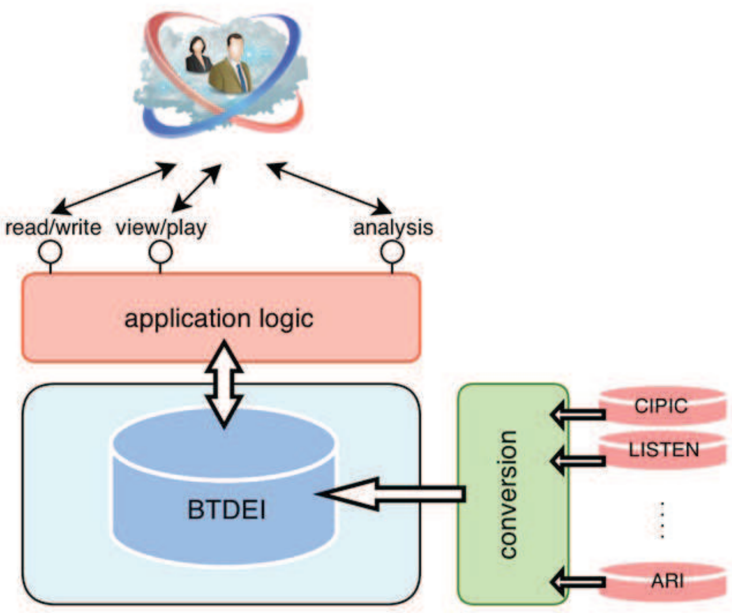


Figure 5.4: *BTDEI data exchange format.*

- both closed- and open-ear canal measurements are required for the evalution of binaural reproduction;

- how the headphones interact with the external ear is strictly subject-dependent.

Intersubjective variability is particularly marked in the high-frequency range, where important elevation cues generally lie. Thus, an inaccurate compensation likely leads to spectral colorations that affect both source elevation perception and sound externalization (Masiero and Fels, 2011).

A new file-system level is inserted on top of the *subject* folder: for each pair of headphones, a collection of individual HpIRs and equalization filters are stored. Indeed, one of the purposes of this archive is to compute the equalization filter that compensates the headphone starting from the raw HpIRs.[10] Every *subject* folder contains three subfolders - *raw*, *compensated* and *eq*. The *raw* subfolder contain raw data from the recordings in both .mat and .wav formats. The second subfolder contains the compensated impulse responses, while the latter stores the equalization filter (under the form of an impulse response) obtained from the inverse HpIR through one or more techniques, e.g. considering the mean HpIR measurement with respect to all the repositionments of that device (Nishimura et al., 2010).

The .mat structure is divided into *specs* and *data* sections. The former section includes, additionally to the *sampling frequency*, *stimulus type*, *filter type*, and *microphone position* fields (defined in the same way as in the HRIR repository), the following information:

- *headphone model*, producer included;

[10] Although various techniques have been proposed in order to face the equalization issue modeling the correct equalization filter is still a hot open research theme.

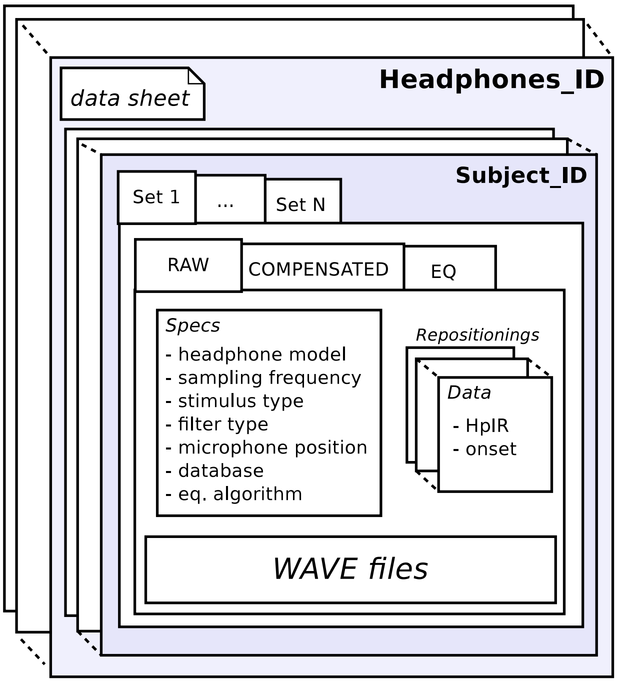


Figure 5.5: *HpIR repository structure.*

- *eq algorithm* (in *eq* version only): equalization algorithm used (e.g. preprocessed inversion transfer function, least-squares minimization of inversion error , etc.).

Furthermore, similarly to the HRIR / pHRIR repository, it is possible to include generic HpIRs measured in other laboratories and keep track of this information in the *database* field.

The *data* section is made of an array of structures of length R (= number of repositionings for that pair of headphones) with each cell containing the data from a single recording:

- *HpIR*: pair of vectors containing the HpIR samples, relative to the left and right ear/headphone respectively;

- *onset*: pair of indices indicating the left and right HpIR's onset sample.

Finally, specific characteristics of the headphones such as transducer type, acoustic coupling, and design are stored in their data sheet. This information resides in the root directory of that device and no re-organization has been made until now.

5.2.2 Signal analysis tools

The *analysis* folder (see again Fig. 5.2) contains all the Matlab scripts and data structures exploitable for HRTF analysis. The typical work flow follows an analysis-by-synthesis paradigm where the step-by-step modeling of salient features plays a significant role in the analysis of the acoustic signal. A notable instance of such paradigm is represented by the PRTF structural decomposition algorithm (see Sec. 4.1), which iteratively extrapolates the reflective component of a PRTF while keeping its resonant structure intact by direct subtraction of multi-notch filter structures. A similar algorithm, used in Sec. 3.1.2, separates the near- and far-field contributions of a rigid sphere approximating the head, allowing to model the two contributions independently through different filter structures.

Many other tools are included in order to support these analysis methods. For instance, an image processing algorithm that extracts the relevant anthropometric parameters from a picture of the pinna will be presented in Sec. 5.4. A script for PCA modeling of HRTF data that helps understanding the degree of variability of the transfer functions with respect to specific features is available (Spagnol and Avanzini, 2009). Last but not least, headphone equalization algorithms implementing various inverse filtering techniques are included.

5.2.3 Synthesized audio rendering

The audio engine, stored in the *synthesis* folder (see again Fig. 5.2), includes four modules organized in separate subfolders:

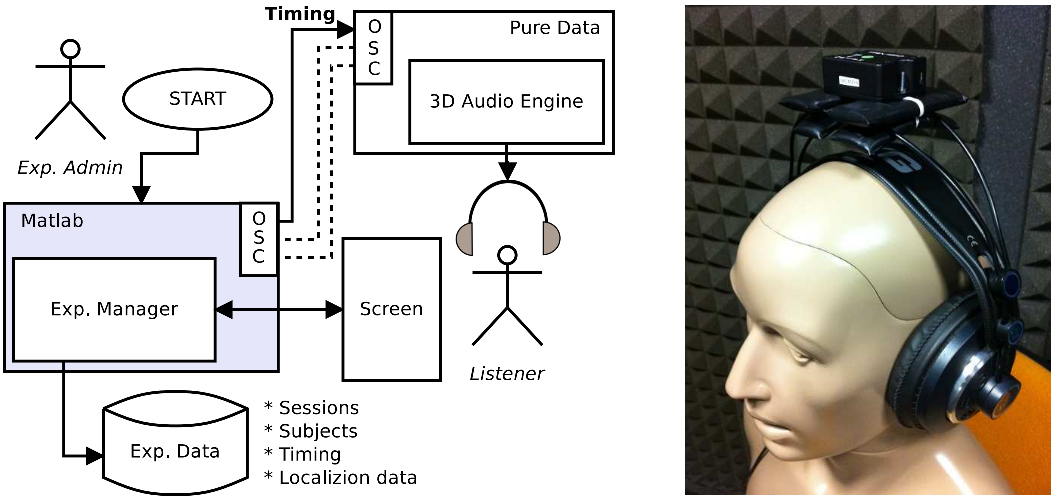
- *model*: real-time realizations of the synthetic structural components;
- *components*: collection of tools that perform real-time convolutions between audio files and HRIRs/pHRIRs;
- *headphones*: management tool for headphone compensation filters;
- *extra*: utility bundle for I/O operations, sensors and basic binaural processing tools.

Various combinations of one or more instances for each module are possible in order to realize a candidate version of the 3D audio engine. All instances are currently implemented in Pure Data,[11] a graphical programming environment for audio processing, in the form of C/C++ externals. All the tentative prototypes are catalogued in a further folder (the *rendering* folder), each accompanied by a descriptor file including the list of the modules used in that instance.

The intrinsic modularity of our approach leads to the implementation of one structural filter block for each relevant body part:

- a pinna filter realization that acts as a synthetic PRTF, consisting of a peak and notch filter structure (see Sec. 3.3.1) where each filter is fed by three parameters (peak/notch central frequency, bandwidth, and gain) each stored in a configuration file indexed by the current elevation angle;
- a spherical model of the head that takes into account far-field and near-field scattering effects around its rigid body. The parametrization is made onto the sphere radius selected as a weighed combination of the listener’s head dimensions (Algazi et al., 2001a) and the near-field contribution is reduced down to a first-order shelving filter (see Sec. 3.1.2);

[11] Puckette (1996), http://puredata.info/



(a) Experimental environment

(b) Head-tracker

Figure 5.6: *(a) A high-level technical description of a typical experimental scenario. (b) Headphones augmented with a head pose tracker.*

- a spherical torso approximation (as in the snowman model (Algazi et al., 2002b)) that models elevation cues at low frequencies.

The *extra* folder deserves a more detailed description. In this folder, a collection of third-party utility tools is kept updated and integrated in our prototyping environment mainly developed in Matlab, Pure Data and C/C++ libraries. A basic external for the rendering process is CW_binaural~ (Doukhan and Sédès, 2009) that implements real-time convolution of sound inputs with selected HRIRs (an antropometry-driven selection criteria is proposed and discussed in Sec. 4.3). It has the peculiar feature of being able to load an arbitrary discrete set of HRIRs in .wav format and realize different kinds of interpolations between adjacent spatial positions. This tool is at the basis of our dynamic 3D audio rendering system, where the successful transposition of dynamic sources into a virtual world not only depends on the accuracy of the interpolation scheme but is also heavily conditioned by the quality of the motion tracking system. In decreasing order of degree of immersion, a PhaseSpace Impulse MoCap system,[12] a head-pose estimation system via webcam with the faceAPI software library,[13] and a Trivisio Colibri wireless inertial motion tracker[14] mounted on top of a pair of headphones are already integrated in our environment.

5.2.4 Experimental environment

An environment for subjective localization tests is stored in the *evaluation* folder (see again Fig. 5.2). A collection of GUIs in Matlab offers the main environment for the playback and

[12] http://www.phasespace.com/impulse_motion_capture.html

[13] http://www.seeingmachines.com/product/faceapi/

[14] http://www.trivisio.com/products/inertial-motion-tracker/

evaluation of HRTF data and models (screenshot of Fig. 5.7 depicts an example of available GUI). The subject listens to the sound stimulus, interactively selects the perceived sound location and/or other properties, and switches to the next sound stimulus.

The basic features for the experimenter such as subject and session management are also available. Subjects' records and their personal information can be manipulated on demand. Each experimental session is stored in an independent file labeled by a session number id. The *task_info* struct contains the descriptive information of the task and the timestamps for each trial conducted within the session. The latter field operates as a primary key to read and understand the experimental data stored in a table specifically designed for the purpose of the experiment. Common statistical analysis software can directly import the already organized data.

From a technical point of view, the data exchange between Matlab environment and audio engine is granted by the OSC (Open Sound Control) protocol,[15] running on top of UDP.

Figure 5.6(a) reports a technical description of a typical experiment. A pair of common headphones augmented through motion sensors, as the one pictured in Fig. 5.6(b) (AKG K240 MKII), easily fits the most common applications. The Trivisio Colibri wireless motion tracker installed on top of the headphones incorporates indeed a number of sensors (a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis digital compass) able to track the 3D orientation of the user's head thanks to the 6-DoF motion processing they convey.

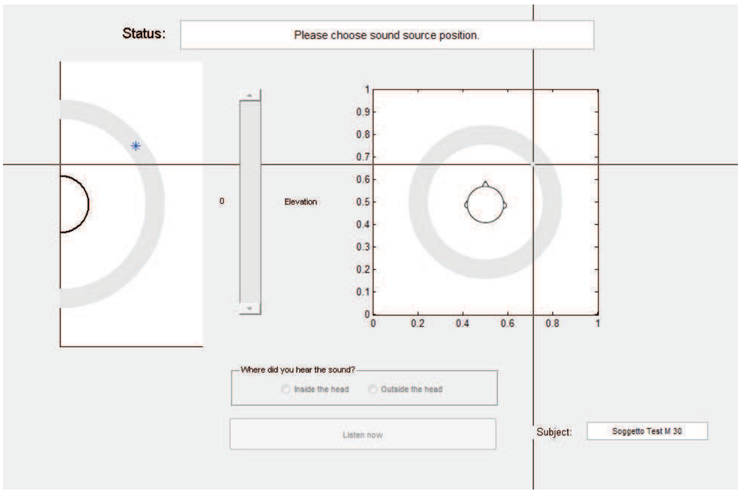
Data from the motion sensors (pitch, roll, and yaw rotations of the head) are sent in real time by radio transmission to the audio processing module and translated into a couple of polar coordinates (θ, ϕ) of a fixed or moving sound source. These coordinates finally represent the input to the structural HRTF model that performs the convolution between a desired sound file and the user's customized synthetic HRTFs. In this way, provided that the center of rotation of the head does not excessively translate during the rotation, the user will ideally perceive the position of the virtual sound source as being independent from his or her movement. For what concerns distance tracking between the user and the sound source, a 9-DoF head tracking device, e.g. deepth cameras, might be able to estimate the position of the listener within the local coordinate system of the device.

5.3 Modus operandi

The research process aims at building a completely customizable structural model through subsequent refinements, starting from a selection of recorded HRIRs to a totally synthetic filter model. The intermediate steps are balanced mixtures of selected pHRIRs and synthetic structural components (see Sec. 2.2 for the MSM formalism).

A candidate mixed structural model is described by a set of parameters and components; the evaluation step guides the exclusion of certain combinations of such components. The obtained 3D audio engine shall maximize both handiness and localization performances: this means that synthetic partial HRTF with a low handiness, e.g. high customization degree with cumbersome

[15] In the Fig. 5.6(a), a *pnet* library and the *dumpOSC* external are responsible for communication on the Matlab side and on the Pure Data side respectively.



Status: Please choose sound source position.

Where did you hear the sound? ○ Inside the head ○ Outside the head

Listen now

Subject: Soggetto Test M 30

Figure 5.7: *A GUI for localization experiments.*

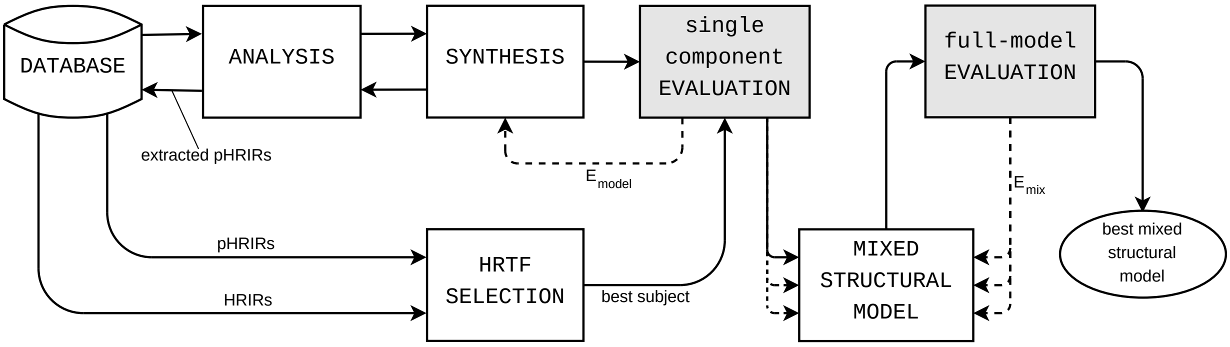


Figure 5.8: *Typical research workflow towards a complete structural model.*

procedure in order to acquire parameters, but poor localization performance compel us to conceive a different or modified submodel. In such a case, a previous version of that submodel, whether it be synthetic, directly recorded or extracted by analysis, remains the best option for the component's acoustic contribution up to that point. As soon as the partial selection process is optimized, the best solution for the complete structural model is provided. The typical research workflow is now described in more detail, referring to Fig. 5.8 throughout the whole section.

5.3.1 HRTF Selection

The simplest possible HRTF selection procedure is realized by choosing the same set of HRTFs for all of the listeners. Whether it is the best available localizer or a mannequin with mean anatomical features, no prediction can be made on the localization performance of a specific listener. An insight knowledge of the relation between localization cues and anthropometric features can guide the selection process as in (Middlebrooks, 1999). If measurements of pinna dimensions for each subject in the considered HRTF database are available, a simple "best match" with directly measured pinna dimensions of a common listener, can be exploited. Eq. (5.1) describes a similarity function, $s(l,j)$, on the left pinna applied to the CIPIC HRTF database:

$$s(l,j) = \min_{i=1\ldots N} \left(d_1^i + d_2^i + d_4^i - p^l \right), \tag{5.1}$$

where N is the number of subjects in the database, d_k is the vector associated to the k-th anthropometric feature of the pinna, i.e. cavum concha height (d_1), cymba concha height (d_3) and fossa height (d_4) respectively, and p^l is the distance from the superior internal helix border to the intertragic incisure of the l-th listener. At the end of this procedure the j-th database subject which has the minimum pinna mismatch with the l-th listener, is selected as the best-matching elevation localizer and this correspondence is saved in the experimental environment.

5.3.2 Structural HRTF selection & modeling

The keen reader shall criticize the restriction of the latter selection procedure to the sole contribution of the pinna. Of course, there exists no guarantee for the accuracy in azimuth localization. However, the former procedure can be applied to the Aalto PRTF database or alternatively to a collection of mid-sagittal PRTFs extracted in the analysis step. On the other side, the selection of a best matching contribution from the head by means of recorded impulse responses (such as the pinnaless KEMAR HRIRs) or of extracted ITD (interaural time difference) / ILD (interaural level difference) information may adopt the same principle. Such an alternative leads to a finer selection of each structural component and is at the base of the progression of our mixed structural modeling approach.

HRTF sets resulting from further refinements of selection criteria should now be compared to our candidate synthetic filter models, the parameters of which are strictly related to the anthropometric quantities used for HRTF selection. The simplest example is the head radius optimization for the spherical model in Sec. 3.1.1. Eq. (5.2) describes a trivial similarity measure between optimal head radii that is based on the application of Eq. (3.5) to a listener and all the database subjects:

$$s(l,j) = \min_{i=1\ldots N} a_{opt}^i - a_{opt}^l, \tag{5.2}$$

where a_{opt}^i is the optimal sphere radius for the i-th database subject and a_{opt}^l is the optimal sphere radius of the l-th listener. At the end of this procedure we obtain two alternatives to test and compare: (i) a parameterized spherical filter model with radius a_{opt}^l and (ii) a set of real ITDs extracted from the HRTFs of the selected j-th database subject.

Restricting our attention to the cited structural components (head and pinna), 3×3 instances of mixed structural models already arise from the combination of the following alternatives:

- pinna structural block: measured KEMAR PRTFs, Aalto PRTF database selection, selection of extracted PRTFs from HRTF databases;

- head structural block: measured pinnaless KEMAR HRTFs, selection of extracted ITDs, parameterized spherical filter model.

Further combinations can be also taken into account; for instance, if a pinna contour extraction procedure is performed instead of a simple pinna height measurement, the structural model discussed in Sec. 4.4 or the HRTF selection criterion in Sec. 4.3 can be counted with the previous alternatives.

5.3.3 Evaluation

The candidate models are subjected to three complementary evaluations:

- **objective evaluation**: signal-related error metrics such as spectral distortion and spectral cross-correlation;
- **auditory model evaluation**: auditory filterbanks and statistical prediction models;[16];
- **subjective evaluation**: listening tests of localization and realism.

The space of possible structural model instances is reduced by a two-stage evaluation procedure, made by a single-component and a full-model evaluation. The single-component evaluation focuses on the minimization of the error E_{model} in Fig. 5.8, defined as the mean localization error of the best synthetic version available. A dimensionally reduced localization space (e.g. mid-sagittal plane data only) supports this early stage.[17] The full-model evaluation takes the best representative solutions for each structural component in order to test the combined effects and the orthogonality of the models within full-space 3D virtual scenes. The minimization of E_{mix}, defined as the mean localization error of the mixed structural model, leads the mixing process.

Once the best MSM currently available is found in terms of E_{mix}, one can analyze its handiness and try to design a device that covers those hardware and software requirements. Otherwise, two other directions are contemplated in the MSM approach:

- find a more technologically handy procedure to acquire the same parameters;
- discover a new instance of MSM that exhibits the same or better localization performances and higher handiness;

5.4 Automatic extraction of pinna edges

As stated in previous Chapters, HRTFs are strictly personal and the pinna plays a fundamental role in the shaping of HRTFs and in the vertical localization of a sound source. Sec. 3.3 has presented the following analysis: while resonance peaks introduced by the pinna are similar among different subjects, frequency notch locations are critically subject-dependent. A ray-tracing law has strengthened the hypothesis that in frontal median-plane HRTFs the frequency of

[16] One valuable tools is the Auditory Modeling Toolbox, AMToolbox (http://amtoolbox.sourceforge.net/), a Matlab/Octave toolbox for developing and exploiting auditory perceptual models (such as Langendijk and Bronkhorst (2002)) which gives a particular attention on binaural models.

[17] A partial example of such a modus operandi at this stage can be appreciated in the external ear's case studi of Ch. 4

spectral notches, each assumed to be caused by a single reflection path, is related to the distance of the most prominent pinna edges to the ear canal (or meatus) entrance.

Such hypothesis allows for a very attractive approach to the parametrization or selection of the HRTF based on individual anthropometry, i.e. extrapolating the most relevant parameters that characterize the HRTF spectral shape from a representation of the principal pinna edges, which need to be in turn estimated from a picture. Having outlined this basic yet unexplored idea, the challenge addressed in this section concerns computational image processing side and may be summarized by the following questions:

- how to automatically derive a robust representation of the most prominent pinna edges from one or more side face pictures of a person?
- how to develop a handy procedure from both the design and computational points of view?

5.4.1 Previous works

It is commonly accepted that no two human beings have identical pinnae, and that the structure of the pinna does not change radically over time (Iannarelli, 1989). These two simple statements are at the basis of a recently growing interest in the field of biometrics in using the pinna as an alternative to face-, eye- or fingerprint-based subject recognition. A multitude of works addressing ear biometrics has surfaced in the last 15 years starting from Burge and Burger's rediscovery (Burge and Burger, 2000) of the work of Iannarelli (Iannarelli, 1989), the pioneer of ear recognition systems. These new works generally address the study and design of all the building blocks making up a complete recognition system, including ear localization in images or video, feature extraction and matching. A comprehensive review of the state-of-the-art in ear biometrics up to 2010 can be found in (Abaza et al., 2010).

Radically different approaches to the definition of a feature model that uniquely and robustly defines the pinna morphology from 2D or 3D images in ear recognition systems have been proposed (Chen and Bhanu, 2005). Some of these directly transport the input ear image to a different domain, e.g. using a 3D elliptic Fourier transform (Hetherington et al., 2003) that compactly represents the pinna shape or a force field transformation (Hurley et al., 2005) that treats pixels as an array of mutually attracting particles acting as the source of a Gaussian force field. Others extract an edge map from the original image containing the pinna; thanks to such map either the pinna is localized into the image or distinctive features are extracted. Since the focus of this research is in the extraction of pinna contours from 2D images, a brief review of these latter approaches are presented.

The most obvious way of extracting edges from a generic image implies the use of standard intensity edge detection techniques such as the Canny method (Canny, 1986). This method was exploited by Burge and Burger (Burge and Burger, 2000) to obtain a Voronoi diagram of the pinna edges, from which an adjacency graph was built for matching. Ansari and Gupta (Ansari and Gupta, 2007) also used the Canny method as the starting point towards extraction of the outer pinna edge for localization of the ear in side face pictures. However, in neither of the two works the effectiveness of the Canny method in the extraction of all pinna edges was analyzed in detail.

An analogous approach was adopted by Moreno et al. (1999). In order to obtain a profile of an ear picture, Sobel filters were applied both in the horizontal and vertical directions of a grayscale image. Then the most marked intensity discontinuities were derived from each resulting image by standard thresholding, and the sum of the thresholded images gave the profile. This was used either to automatically extract a biometric vector of feature points of the pinna or to compute a morphology vector capturing its shape; these vectors were the input for a perceptron performing classification. Thanks to such heuristic procedure, 90% of feature points were reported to be correctly found.

An alternative to the Canny and Sobel methods was proposed by Choraś (Choraś, 2005). Edge detection in a grayscale image of the pinna was performed through a pixelwise method which examined illumination changes within each 3×3 pixel window. Based on the minimum and maximum values of pixel intensities in such window and on an adaptive threshold value, the center pixel was either labeled as edge or background. Feature extraction from the edge map was then fulfilled by tracing concentric circles around the edge map's centroid and computing the intersections between circles and edges. Still, no quantitative results were given for the accuracy of both the edge map computation and the final classification.

Jeges and Máté (2007) proposed a very similar method, where the obtained edge map was used in conjunction with an orientation index image to detect the position of the ear in a video frame sequence and adapt a deformable template model (active contour method) to the ear itself. Similarly, in a very recent work González et al. (2012) used adaptation of an active contour model and ovoid fitting to localize the ear in side face pictures and estimate features under the form of distances between the outer and inner pinna edges and the inner edge centroid. No detail on how these edges were extracted is provided.

Jeges' edge extraction algorithm was also a critical component of the reconstruction method by Dellepiane et al. (2008) which interactively adapted a 3D head model to a specific user starting from a set of pictures of the head and pinna. Following a complementary approach to an anthropometry-based HRTF customization techniques, the resulting model was fed to a simplified boundary element method solver in order to simulate custom HRTFs for that user. Regrettably, few data supported the accuracy of this method for HRTF simulation.

5.5 Multi-flash camera-based approach to pinna edge extraction

Even though edge detection through intensity-based methods seems to be a valid choice in the extraction of pinna edges from 2D images, it is not the sole nor the most efficient option. One can initially try to process pictures with the Canny method, yet it turned out that it fails in low-contrast areas such as the pinna, and especially in cases where shadows are not projected below the considered edge. Fig. 5.9 shows an example of Canny edge extraction (with standard deviation of the Gaussian filter $\sigma = 2$, lower hysteresis threshold $t_l = 0.2$, and upper hysteresis threshold $t_h = 0.5$) on four pictures of the same pinna taken with different light sources. It can be clearly noticed that while in some cases the extraction is acceptable (rightmost image), in all

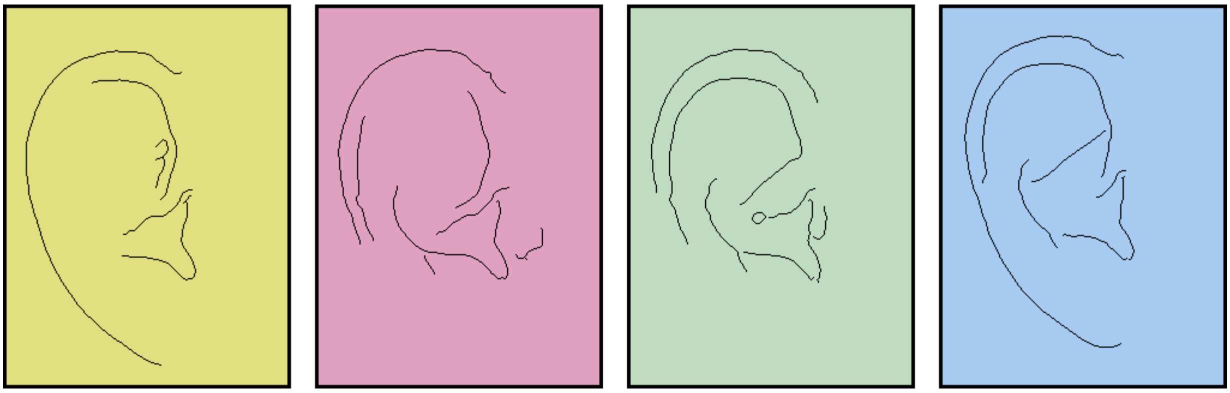


Figure 5.9: *Canny method ($\sigma = 2$, $t_l = 0.2$, $t_h = 0.5$) applied to four different pictures of the same pinna taken with different light sources.*

other cases either some important depth edges are lost or some minor depth edges or specular highlights are extracted.

A more robust depth edge extraction can instead be achieved through a technique known as *multi-flash imaging* (Raskar et al., 2004). The central concept to multi-flash imaging, which was born as a technical solution to non-photorealistic image rendering, is the exploitation of a camera with N flashes strategically positioned around the lens to cast shadows along depth discontinuities in the scene. For each of the N flashes a picture is taken; the location of the shadows abutting depth discontinuities, appearing only for a strict subset of the N pictures, represents a robust cue to create a depth edge map. Thus, thanks to this simple and computationally efficient method, one can robustly distinguish depth edges from texture edges due to reflectance changes or material discontinuities.

To the best of the author' knowledge, the method has never been systematically applied to pinna edge detection before. Since the pinna has a uniform texture, the main purpose of multi-flash imaging would reduce to the extraction of the *most marked* depth discontinuities, that usually correspond to the outer helix border, inner helix border, concha wall/antitragus border, and tragus border (see Sec. 5.5.3 for definitions of these anatomical components). We now describe the hardware and software components that implement the multi-flash camera-based pinna edge extractor.

5.5.1 Multi-flash camera prototype

A multi-flash camera prototype was built and the main electronic components of the device, pictured in Fig. 5.10, are:

- a battery-powered Arduino UNO microcontroller board;[18]

- an Arduino data logging shield;

[18] http://arduino.cc/en/Main/arduinoBoardUno

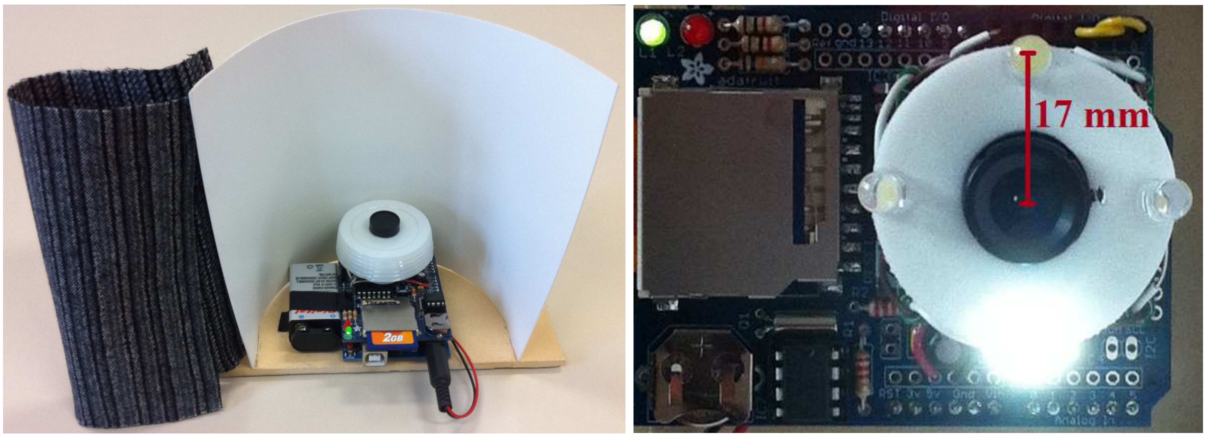


Figure 5.10: *The multi-flash device: full prototype and electronic parts.*

- a TTL serial JPEG camera;
- four Super Bright White LEDs;
- a common SD card.

The data logging shield manages data transmission from the camera to the SD card. The four LEDs, which represent the four flashes of our multi-flash camera, are symmetrically positioned around the camera lens along a 17mm-radius circumference and can be turned on independently by the microcontroller. As we will later see, their positions with respect to the pictured scene, i.e. in the up, down, left, and right directions, allow simplification of the post-processing phase. Since the light emitted by each LED has a high directional component that clearly appears in pictures, the application of a punched and reversed paper glass bottom right above the LEDs allows projection of a more diffuse light field.

The electronic components are secured to a rigid board by four long pins and enclosed in a hemi-cylindrical PVC foil, whose shape affords correct orientation as referred to the pinna. The height of the half-cylinder (15 cm) was chosen so as to entirely fit a big-sized pinna (8 cm height) in the pictured frame. Furthermore, the fixed distance between the lens and the pinna allows to maintain consistency among the dimensions of different pinnae. Lastly, because a dark environment is desirable to better project shadows, the open side of the hemi-cylinder is closed by a square of dark cloth with Velcro fastening strips before data acquisition.

Acquisition of the required data is managed as follows. By connecting the battery to the Arduino board, an Arduino sketch performing the following operations is run:

```
while no motion detected do
  wait; {wait for motion detection}
end while
delay 10 s;
for $k$ = 1 to 4 do
```

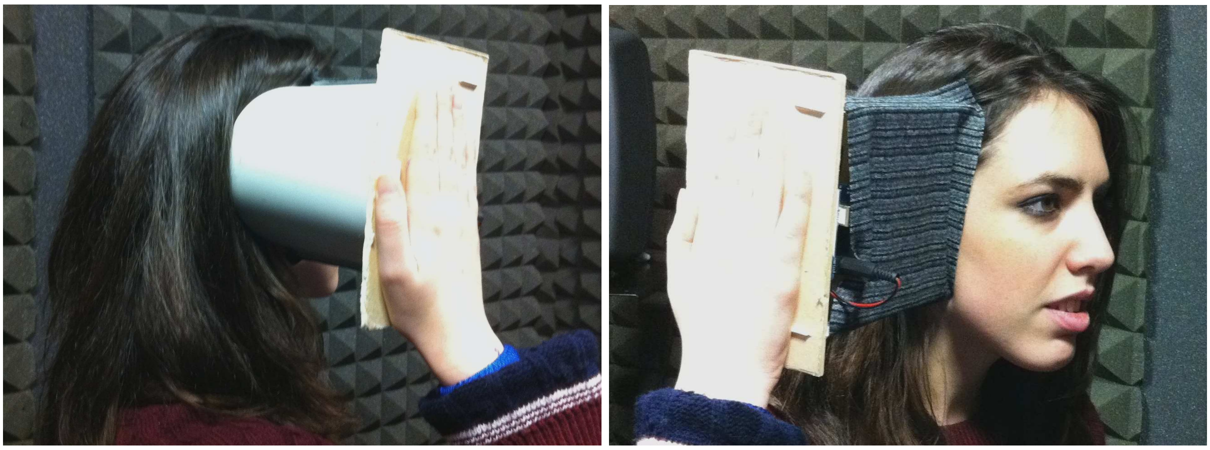


Figure 5.11: *Subject during acquisition of pinna images.*

```
    led_k ← turn on;
    take picture;
    led_k ← turn off;
    img_k.jpg ← picture; {save to SD card}
  end for
```

When the cap is removed from the lens, motion detection is triggered. During the following 10s pause, the subject presses the open top side of the device around his/her left or right ear trying to avoid hair occlusion and aligning the hemi-cylinder with the pinna (see Fig. 5.11). Afterwards, four pictures – each synchronized with a different flash light – are acquired. Because of the required storage time of the current prototype, this basic procedure takes approximately 30 seconds, during which the subject tries to hold the device as still as possible with respect to the pinna. The four pictures, stored in the SD card as 320×240 pixel JPEG files, are then passed to a PC for processing.

5.5.2 Depth edge map computation

After having associated each picture to the position of the corresponding flash light (i_1 = left, i_2 = right, i_3 = up, i_4 = down) depending on whether the left or right pinna has been acquired, the picture set is fed to a MATLAB processing script. The implemented procedure is divided into a pre-processing phase and an automatic depth map computation phase, whose core is the algorithm described in (Raskar et al., 2004).

The pre-processing phase consists of the following steps:

1. *grayscale conversion*: the four images are converted to grayscale;

2. *intensity normalization*: the four grayscale images are normalized with respect to their mean intensity;

3. *motion correction*: images are first rotated and then translated for the best possible relative alignment according to a standard 2D correlation function;

4. *low-pass filtering*: each motion-corrected image is median-filtered using a 7-by-7 neighbourhood.

Motion correction is critical in those cases where the subject moved during image acquisition. The best rotation is first calculated by rotating each image i_k, $k = 2, 3, 4$ in 1° increments, cropping the rotated image to fit the original image size, and finding the rotation of i_k that maximizes the correlation coefficient with respect to i_1. The best translation is instead calculated by considering each possible $(320 - w_c)\times(240 - w_c)$ pixel window of i_k, where w_c is an even positive user-defined parameter that one can typically set to $w_c = 20$, and by finding the window that maximizes the correlation coefficient with respect to the centrally $(320 - w_c)\times(240 - w_c)$ pixel cropped section of i_1. Finally, low-pass filtering was introduced *a posteriori* to remove the inherent noise introduced by hair in depth maps.

Shadows are now detected by taking a ratio r_k of each image with the pixelwise maximum of all images. Sharp transitions in r_k along the epipolar ray, i.e. the ray connecting the light epipole (defined as the position of the flash light with respect to the taken picture) to the shadowed area, are then marked as depth edges. In our case, since the four flash lights are in the plane parallel to the image plane that contains the camera lens, each light epipole is at infinity and the corresponding epipolar rays are parallel and aligned with the pixel grid. This reduces the problem to the detection of sharp transitions along the horizontal and vertical directions of the ratio images, that can be managed by standard Sobel filters.

More in detail, the depth edge map is calculated as follows:

- for all pixels x, create $i_{max}(x) = \max_k(i_k(x))$, $k = 1, \ldots, 4$;

- for each k, create ratio image $r_k(x) = i_k(x)/i_{max}(x)$;

- calculate e_k, $k = 1, \ldots, 4$ by applying a horizontal Sobel filter to r_1 and r_2 and a vertical Sobel filter to r_3 and r_4;

- keep only the negative transitions in e_1 and e_3 and the positive transitions in e_2 and e_4;

- extract the main depth edges from e_k, $k = 1, \ldots, 4$ through a Canny-like hysteresis thresholding, with upper threshold t_h defined by the user and lower threshold $t_l = 0.4t_h$;

- combine all the edges into a single depth edge map.

The final depth edge map is a $(320 - w_c)\times(240 - w_c)$ binary matrix whose black pixels represent the most prominent depth discontinuities of the pictured scene. As we will later see, the choice of t_h has a non-negligible impact on the extracted edges and on the final results. Fig. 5.12 reports an example of depth edge map extraction for two subjects with parameters $t_h = 0.35$ and $w_c = 20$.

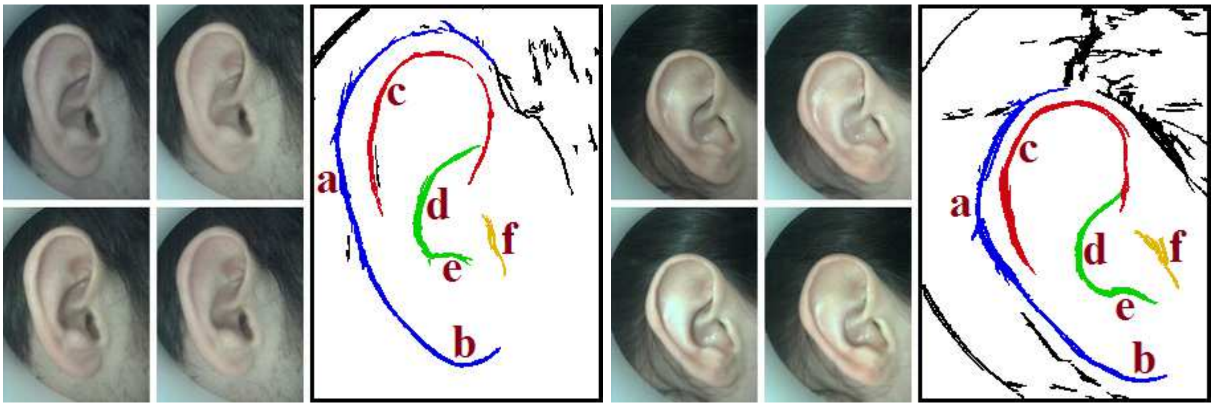


Figure 5.12: *Depth edge maps of two subjects. The outer helix/lobule (a/b), inner helix (c), concha wall/antitragus (d/e), and tragus (f) borders are highlighted in blue, red, green, and yellow respectively.*

5.5.3 Pinna edge extraction algorithm

The depth edge map of the subject's pinna allows extraction of the relevant features that characterize an individual acoustic response. The information contained in the depth edge map that reflects such characterization is included in the Euclidean distance from the points that form the outer helix, inner helix, and concha wall/antitragus borders to a point approximately situated at the meatus entrance, that one can conventionally assume to be located in the upper segment of the tragus border (definitions of all borders are given in Fig. 5.12).

In order to compute distance values, a second MATLAB script that sequentially executes the following steps is run:

1. *map refinement*: the connected components containing less than 100 pixels, i.e. the smallest blobs that usually correspond to spurious hair edges, are deleted;

2. *tragus detection*: the tragus edge is heuristically identified as the connected component lying in the central 200×150 pixel section of the depth edge map whose distance to the bottom left corner (left pinna) or bottom right corner (right pinna) of the map is the least;

3. *meatus point*: the tragus component is subtracted pixelwise to its convex hull and the northwestern/northeastern (left/right pinna) pixel is labeled as the meatus entrance point;

4. *radial sweep*: for each elevation angle $\phi \in [-90°, 90°]$ in $1°$ steps, all the transitions to a depth edge along the ray originating from the meatus point and heading towards the pinna edges with $-\phi$ inclination are stored as distances (in pixels);

5. *edge tracking*: a partial tracking algorithm (McAulay and Quatieri, 1986), originally used in audio signal processing to temporally group sinusoidal partials, is exploited to group

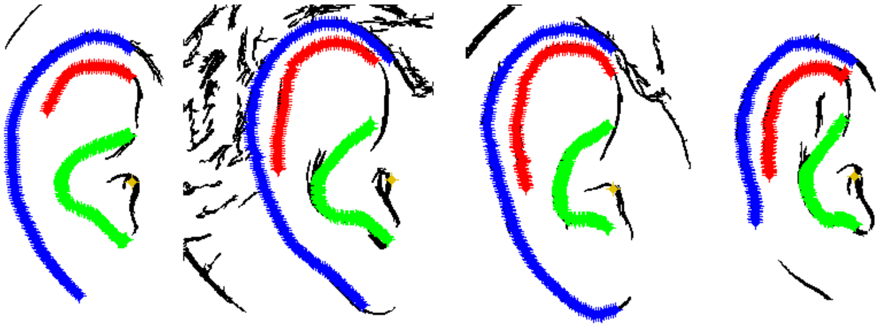


Figure 5.13: *Edge extraction of right pinna images of the multi-flash camera prototypes authors.*

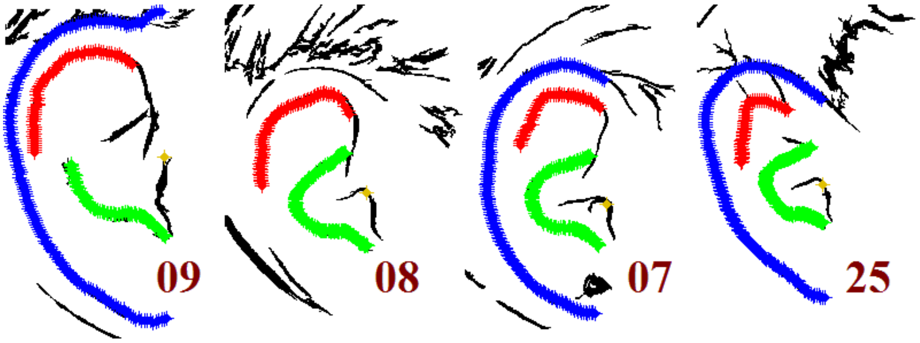


Figure 5.14: *Edge extraction of right pinna images of four test subjects.*

distances (i.e. edges) along consecutive frames into spatial tracks, where each frame corresponds to an elevation angle;[19]

6. *pinna edge detection*: the two longest tracks in increasing order of distance value as identified by the edge tracking algorithm, that we call d_1 and d_3, correspond to the concha wall and outer helix border respectively, and the longest track falling between these two tracks is called d_2 and corresponds to the inner helix border.

Fig. 5.13 depicts the results of the edge extraction algorithm as track points superimposed to the refined pinna depth edge maps of the multi-flash camera creators. This is achieved by simply projecting each point at distance $d_i(\phi)$, $i = 1, 2, 3$ from the yellow meatus point at $-\phi$ inclination.

5.5.4 Evaluation

The multi-flash-based approach to pinna edge extraction was tested on a small number of subjects. Right pinna images of 30 volunteers (aged 18 to 60, 12 female and 18 male, caucasian) were acquired with the multi-flash device and then processed. Parameter w_c was set to 20 for all subjects except for 5 of them who required a more substantial motion correction (in these cases,

[19] The maximum difference between two distances to allow grouping along consecutive frames is set to 5 pixels, while the maximum number of frames before a track being declared dead is set to 10.

Table 5.1: *Pinna Edge Extraction: Results.*

subject	t_h range	# tracks	bad tracks
01	0.29 – 0.33	3	
02	0.43 – 0.47	3	
03	0.43 – 0.56	3	
04	0.37 – 0.58	3	
05	0.23 – 0.34	3	
06	0.21 – 0.43	3	
07	0.27 – 0.60	3	
08	0.24 – 0.40	2	d_3 missing
09	0.27 – 0.49	2	d_1 interrupted
10	0.23 – 0.31	3	
11	0.27 – 0.38	3	
12	0.25 – 0.51	3	
13	0.25 – 0.32	3	
14	0.28 – 0.33	3	
15	0.40 – 0.60	3	
16	0.29 – 0.40	3	
17	0.28 – 0.39	2	d_2 missing
18	0.28 – 0.46	2	d_1 interrupted
19	0.37 – 0.43	3	
20	0.22 – 0.45	3	
21	0.24 – 0.50	3	
22	0.38 – 0.41	3	
23	0.33 – 0.40	2	d_3 missing
24	0.19 – 0.38	3	
25	0.36 – 0.44	3	
26	0.45 – 0.55	2	d_2 missing
27	0.20 – 0.57	3	
28	0.30 – 0.48	3	
29	0.31 – 0.40	3	
30	0.27 – 0.32	3	

w_c = 40). Parameter t_h was set from 0.1 to 0.7 in 0.01 steps in order to look for the range where edge extraction visually outputs the best results. Table 5.1 reports this information along with the number of correctly extracted edge tracks for each subject. This means that, in the reported t_h range,

- the meatus point is correctly placed in correspondence with the tragus edge and always falls in the same point;
- the three tracks follow the corresponding depth edge in its entirety.

If no t_h value satisfies the latter condition, the reported t_h range refers to two correctly extracted tracks out of three.

One can immediately notice that ranges for t_h are significantly different from subject to subject. The variability among pinna shapes is a first obvious cause of this finding: as an example, Subject 17 has a helix that folds into itself almost coming into contact with the antihelix, thus failing to project a consistent shadow. This results into a very shallow depth edge that is not recognized in the reported t_h range. Outside this range, either the number of edges is too high to discriminate the real depth edges from any artifact (low t_h) or some relevant depth edges are lost or broken (high t_h). Another factor that contributes to the determination of the lower t_h bound is the possible connection between the tragus and concha edges, that does not allow correct detection of the meatus point.

Two additional examples of how pinna morphology affects the final results are Subjects 09 (see Fig. 5.14) and 18, whose concha wall is not fully-extracted. This is due to the shape of the concha itself, resulting in two or more separate and non-intersecting edges (as in the pinna of Fig. 5.9). Since the grouping conditions of the edge tracking algorithm are not satisfied, no interpolation between these edges is performed and only partial extraction of the concha edge occurs.

Motion correction also plays an important role in the determination of the t_h range. As a matter of fact, often linear correction does not perfectly align the four pinna images. This causes the same edge to be considered twice or thrice in the final depth map in slightly different yet overlapping positions, resulting in thicker depth edges. At the same time, a non-perfect alignment allows extraction of the outer helix border when the back of the ear is surrounded by hair, as shadows on hair are only rarely detected by the multi-flash setup. The second pinna in Fig. 5.14 shows a case (Subject 08) for which a very good alignment is reached yet part of the outer helix border fails to be extracted.

However, if we consider a t_h value included in the reported range for each subject, the meatus point is correctly identified for all subjects, and 84 out of 90 edge tracks are correctly extracted (success rate: 93.3%). Statistically, the t_h value that guarantees a correct extraction of the edge tracks for the highest number of subjects is $t_h = 0.31$. These findings are conditioned by the fixed relation $t_l/t_h = 0.4$, hence further work is needed to check whether a different lower/upper threshold ratio improves the above success rate.

The described edge extraction procedure also seems to be robust in those cases where earrings, glasses or other objects appear in pictures (e.g. Subject 07 in Fig. 5.14). Even small amounts of hair occlusion causing the detection of depth edges due to hair (e.g. subject 25 in Fig. 5.14) do not corrupt the extracted tracks.

The above results refer to a preliminary study. An extensive analysis on a wide variety of subjects with different pinna sizes, shapes, and albedo values is required to robustly assess the effectiveness of the edge extraction procedure and to study how the t_h parameter can be automatically defined prior to the image post-processing routine. Nevertheless, a more robust motion correction is required before: possible solutions to this issue, whose feasibility still has to be investigated, include

- the exploitation of more complex feature-based image alignment (*image registration*) algorithms;
- *fast shooting* of pictures, in order to reduce the duration of the acquisition routine down to a few seconds and make motion correction become much less critical;
- *single-shot multi-flash* photography (Raskar et al., 2004; Vaquero et al., 2009), a little explored idea according to which four different flash colours can be used to take a single picture of the scene so that the Bayer filter of the camera should be able to decode the separate light wavelenghts and thus derive four different pictures each related to a single flash.

5.6 Future developments

The HRIR and HpIR repository described in this chapter represents a further step towards both a binaural impulse response database standardization and a basic tool for the analysis, synthesis and rendering of 3D audio. The mixed structural modeling formulation introduced in Ch. 2 requires a well-defined repository in order to support the analysis and design of novel synthetic filter models and HRIR/pHRIR selection processes. Thanks to the proposed solution, a flexible mixture of acoustic responses and synthetic models potentially increases possible pHRIR combinations together with their psychoacoustic accuracy.

The inclusion of computer-simulated HRIRs / pHRIRs calculated from mesh models of human heads (Katz, 2001b,a) and spatially discretized so as to be added to our repository would raise the number of available subjects in a limited amount of time. The unification of heterogeneous datasets also facilitates normalization and cluster analysis of impulse response data (e.g. through principal component analysis). Moreover, the storage of individual headphone compensation filters promotes the personal use of headphones. This is especially true in experimental environments, yet HpIR modeling and cluster analysis would allow the exploitation of this approach for commercial application.

The proposed technology was designed so as to be applied to automatic measurements of pinna anthropometry for binaural audio rendering, and represents a low-cost alternative to technologies involving 3D image acquisition (e.g. laser scanners). Still, the multi-flash device and head tracker are currently being used for anthropometry-based HRTF selection tests.

The presented real-time system is however currently lacking a solid implementation of the theorized components of the structural HRTF model, which is the topic of ongoing works. Once this fundamental component is integrated, extensive listening sessions will assess the degree of accuracy and realism of the presented 3D audio scenes with the two stage optimization procedure described in Sec. 5.3. Many improvements can be developed at the design level for such a system in order to increase its handiness and accuracy in parameter estimation:

- a completely wireless system may be developed through the use of wireless headphones;
- fast shooting is also desired to reduce the duration of the picture acquisition routine down to a few seconds and make motion correction become much less critical;
- a smaller and more compact version of the multi-flash camera device may be slotted inside one of the two headphones' cups, if space (both inside the cup and between the lens and the ear of the user wearing the headphones) permits; a reasonable compromise would be proposing the multi-flash as a separate yet wearable device;
- in-place sensing and processing on a mobile device having all the required motion and image sensors represents in perspective the optimal solution in terms of cost and flexibility.

The resulting hardware/software package will hopefully allow an easy and low-tech fruition of custom spatial audio to any user.

In particular, other improvements need to be introduced to the multi-flash edge extractor, especially at hardware level. For instance, the four flashes can be placed farther from the lens in order to project broader shadows and thus improve depth edge extraction. A similar result can be achieved by a configuration with more flash lights, e.g. 8. Other working ideas include the improvement of the outer shell of the device and the use of thermogram imagery to robustly detect the meatus location as well as to remove partially occluding hair (Burge and Burger, 2000). Furthermore, at software level, a combination of depth and intensity edge detection techniques will greatly improve extraction of the outer helix border.

Finally, ear biometrics represents a natural applicative area for the multi-flash edge extractor, as the feature vectors (distance tracks) it produces share analogies with those used in recent systems, especially (González et al., 2012). A deeper study of the applicability of this technology to a complete biometric system will disclose its real potential.

Chapter 6

Spatial sonification of movements

The traditional approach to perception investigates one sense at time (Fechner, 1889). This approach is useful to understand how single senses work, but it does not take into account that perception is intimately a multimodal process: sensations come all simultaneously, so that while touching an object we perceive its size with our eyes and hands.

The immersion in the environment is ecologically multimodal and in particular sound interactions are implicitly multimodal in their physical creation. The measurement of how relevant sound might be in a behavioral variable and how it can influence performances in complex and delicate tasks (e.g. walking, rehabilitation activities and learning of mobility and orientation cues), are crucial issues especially when activities are intrinsically multimodal and, thus, becomes more difficult to assess the accuracy and exploitability of the information carried by each modality (Gibson and Pick, 2000), and to determine how multiple modalities interact through integration and combination (Ernst and Bülthoff, 2004). As an example, vision estimates geometrical properties (e.g., size) better than audition. Therefore, audition has only a limited possibility to modulate size-information carried by eyes. Similarly, audition outperforms vision in temporal tasks, and vision has limited possibilities to modulate temporal information carried by our ears (Shams et al., 2002). These estimates guide people to adopt subjective combinations of rules that balance and calibrate the information arriving from reliable and unreliable cues. In the design process of human computer interfaces, measures of the effect of multimodal congruence can be exploited to enhance user perception and to merge different single modality presentations into a multisensory interactive display.

Positioned at the crossroads of ecological acoustics, computer science, experimental psychology, design, and music, sonic interaction design (SID) is the discipline which studies and exploits intentionally implemented auditory and vibrotactile feedback to convey abstract meanings, specific information about the state of a process or activity, or the inherent expressiveness of human-computer interaction (Franinovic and Serafin, 2013). Physics-based sound synthesis approaches potentially provide an intrinsically natural behavior, since the sound feedback is energetically consistent with the action performed (Delle Monache et al., 2010). The synthesized sound, described in terms of configurations, materials, geometries and interacting forces of vir-

This chapter is partially based on (Spagnol et al., 2012b).

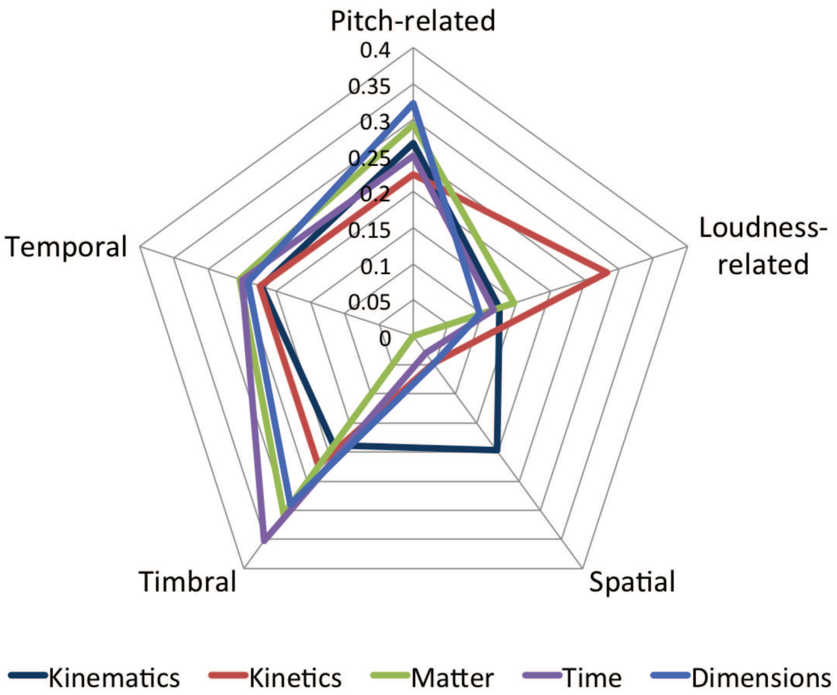


Figure 6.1: *Proportions of auditory mapping strategies normalized against main sonification categories (figure reproduced from (Dubus and Bresin, 2013)).*

tual objects, embodies audible affordances and as such provides information about the interaction with the virtual environment.

In a systematic review of the literature about the sonification of physical quantities, Dubus and Bresin (2013) summarized the main mapping strategies and their relation with physical world (see Fig. 6.1). In particular, they verified that the spatial auditory dimension is mostly used to sonify kinematic quantities. To this regard, this chapter investigates the behavioral importance of spatial attributes of sound and how continuous interaction with a virtual environment can benefit from using spatial audio rendering algorithms. It has to be mentioned that the ecological link between movement and perception represents a fundamental aspect in the understanding of which sonic features predominate, if any, in a specific task and how these features affect other auditory and multimodal components (Hendrix and Barfield, 1995).

Two two case studies related to the above concept are reported next: one involving lower limbs in the complex task of walking and a second one implying the use of upper limb in a reaching activity. Users movements within a virtual environment are continuously sonified in conjunction with other modalities (vision and touch).

The first application investigates the effect of spatial sonification of a moving target on the user's performance during the execution of a target following task with the upper limb. The starting hypothesis is that a properly designed multimodal continuous feedback could be used to represent temporal and spatial information that can in turn improve performance and motor learning of simple target following tasks. Sixteen subjects were asked to track the horizontal movement of a circular visual target by controlling an input device with their hand. Two different continuous task-related auditory feedback modalities were considered, both simulating the sound of a rolling ball, the only difference between them being the presence or absence of binaural spatialization of the target's position. Spatial auditory feedback significantly decreases the average tracking error with respect to visual feedback alone, contrarily to monophonic feedback. It was thus found how spatial information provided through sound in addition to visual feedback helps subjects improving their performance.

The second sonic interaction scenario described in this chapter is the localization of footstep sounds interactively generated during walking and provided through headphones. A pair of sandals enhanced with pressure sensors and a footstep synthesizer are capable to simulate two typologies of surface materials: solid (e.g. wood) and aggregate (e.g. gravel). Different sound delivery methods (mono, stereo, binaural) as well as several surface materials, in presence or absence of concurrent contextual auditory information provided as soundscapes, were evaluated. Three distinct experiments were conducted in which the realism and naturalness of the walking experience as well as the degree of disorientation of the walker were analyzed. Solid surfaces were localized significantly farther from the walker's feet than aggregate ones. This effect was independent of the used rendering technique, of the presence of soundscapes, and of merely temporal or spectral attributes of sound. The effect is hypothesized to be due to a semantic conflict between auditory and haptic information such that the higher the semantic incongruence the greater the distance of the perceived sound source from the feet.

The presented results contribute to the development of further knowledge towards a basis for the design of continuous multimodal feedback in virtual reality applications with high degree of immersion.

6.1 Auditory feedback in rehabilitation

Understanding how the human central neural system combines different kinds of simultaneous information such as proprioceptive, visual, or auditory feedback is today an open issue. The main goal of these researches in this first case study is to investigate the role of sound in motor learning and motor control as an additional or substitutive sensory information to the visual and proprioceptive modalities, with the final aim of incorporating optimized real-time auditory displays related to one or more variables (e.g., target velocity or position error) in augmented-feedback robotic rehabilitation systems.

Unfortunately, the consistent use of auditory feedback in robot-assisted rehabilitation has been largely overlooked in recent related literature. Despite the evidence that a proper sound may help individuals in learning a motor task (Rath and Rocchesso, 2005; Lemaitre et al., 2009), the precise ways in which mental engagement, repetition, kinematic error and sensory information in general translate into a pattern of recovery is not well defined for rehabilitation (Reinkensmeyer and Galvez, 2007; Rosati, 2010). Audio is used in many rehabilitation systems with the purpose of motivating patients in their performance; nevertheless, in the majority of these systems the audio component plays mostly a marginal role, for instance by offering a positive or negative feedback if the patient completes or fails a task, or by reinforcing the realism of a virtual reality environment (Rosati et al., 2013).

However, the use of auditory feedback could contribute to overcome some of the current main limitations of rehabilitation systems in terms of user engagement, acute phase rehabilitation, standardization of the rehabilitation process, and development of home rehabilitation devices (Franinovic and Serafin, 2013). In particular, sound is thought to be effective in the recovery of activities of daily living (ADLs). As a matter of fact, ADLs rely on an essentially continuous and multimodal interaction with the world, which involves visual, kinesthetic, haptic, and auditory cues. To this regard, in order to effectively represent the environment and/or the user's movements, continuous forms of auditory feedback ought to be used in conjunction with other sensory modalities. An incentive to the present case study is offered by the observation that audio, just like video, is more direct and requires less attention than proprioception as input modality (Seizova-Cajic and Azzi, 2010). Hence, auditory feedback can be potentially relevant not only as a stimulation to augment patient's engagement and motivation, but also as an additional or substitutive straightforward information towards the improvement of performance and learning.

A previous work on robot-assisted upper limb tracking movements revealed that providing subjects with auditory feedback of tracking error could effectively increase subjects' effort and reduce the effects of visual distraction (Secoli et al., 2011). Similarly, in a group of related experiments performed on healthy subjects and with no robotic assistance the authors argued that auditory feedback can also be effective in reducing tracking error (Rosati et al., 2012). In particular, continuous task-related information provided through sound in addition to visual feedback can improve not only performance but also the learning of a novel visuomotor perturbation. As a new work along this research thread, the experiment presented in the following sections can be seen as a further missing tile towards the definition of an effective auditory display for conveying informative content to the user during target following exercises. In particular, the aim of the experiment is to investigate whether the information given to the user by spatial task-related auditory feedback helps the subject improving his or her performance more than monophonic task-related feedback, or not.

6.1.1 Experimental setup

As pictured in Fig. 6.2, each participant was provided with a pair of common headphones that presented auditory feedback and a Wacom pen tablet[1] as controller. During the whole experiment, the subject was sitting in front of a Full HD screen and tablet suitably calibrated in order to match the screen size, positioned as depicted in the figure. The screen was backed by a blank wall.

A simplified scheme of the system's architecture is reported in Fig. 6.3. The graphical user interface for the experiment was implemented in MATLAB and consists of two color-filled, 25-pixel-radius dots displayed on the screen, one representing the controller's position (green dot) and one the target's position (red dot). The target performs a continuous horizontal movement (left to right and *vice versa*) with a minimum-jerk velocity profile. Each task has a specific target

[1] A stylus on a sensor board.

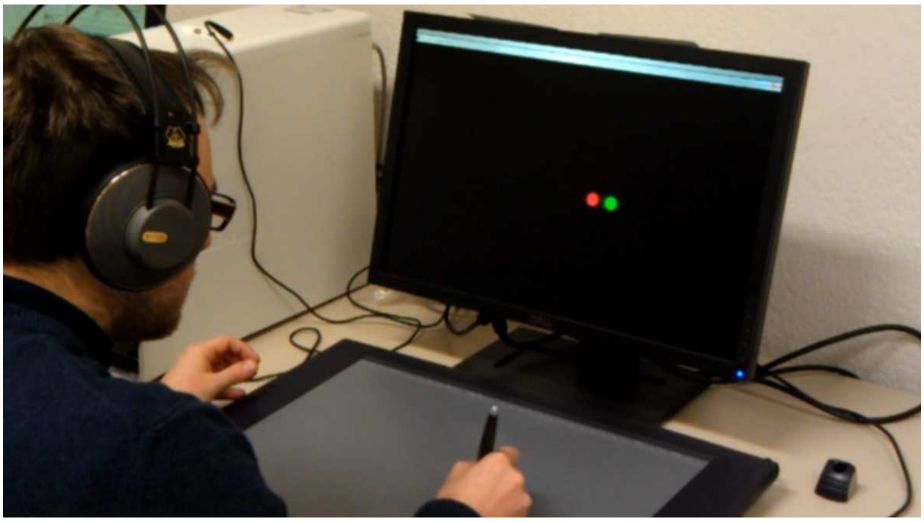


Figure 6.2: *Subject performing a trial of the experiment.*

motion profile, which can be either

- a *fixed-length* profile, where the length of every left-to-right or right-to-left segment is set to 60% of the screen width (corresponding to a range of motion for the subject's hand of approximately 30 cm) at each iteration; or

- a *random-length* profile, where the length of each segment pseudo-randomly varies from 20% to 90% of the screen width. At the end of the task, the total distance spanned by the target is equal to that travelled in the former case.

Auditory feedback was developed in Pure Data (PD) (Puckette, 1996). Target motion data (i.e., velocity in the x direction) is sent in real time to PD through the OSC (Open Sound Control) protocol. Auditory feedback was designed as a *task-related* sonification of the target roughly simulating the sound of a rolling ball. In order to efficiently obtain such feedback, the instantaneous velocity of the target was applied as a gain factor onto the output of a pink noise generator filtered through a bandpass filter with center frequency $f_c = 300$ Hz and Q-factor equal to $Q = 10$, as Fig. 6.4 sketches.

Two audio modalities were used. In the first modality, auditory feedback was provided monophonically through headphones, while in the second one HRTFs were exploited for sound spatialization. To this end, the previously described task-related audio signal was fed to a binaural spatialization filter (provided by the `earplug~` PD external) which renders the angular position of the sound source relative to the subject in the horizontal plane by convolving the incoming signal with left and right HRTFs of a KEMAR manikin. In order to minimize inconsistencies arising from the absence of a head tracking system during spatial audio listening, subjects were told to keep their head towards the center of the screen and to try not to move it during the exercise.

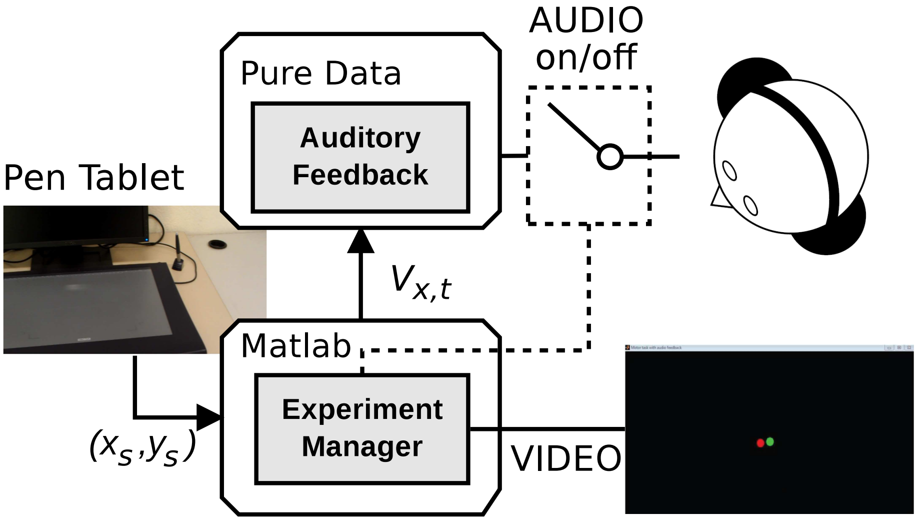


Figure 6.3: *A simplified scheme of the experimental setup.*

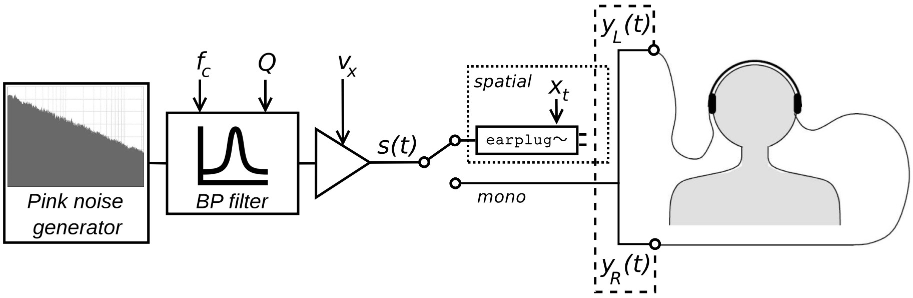


Figure 6.4: *Auditory feedback generation process. The switch, controlled by the experiment manager, selects one of the two audio modalities, i.e. spatial auditory feedback or monophonic feedback.*

6.1.2 Experimental protocol

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

Each participant was asked to complete six different tasks. During each task, the subject had to draw a trajectory onto the tablet with the pen in order to follow as closely as possible the target presented on the screen. The six tasks, presented in a random order, were:

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

Each task lasted 80 seconds and consisted of 14 repetitions of the left-right-left cycle. During each task, target position x_t, target velocity $v_{x,t}$, and 2D subject position (x_s, y_s) were sampled at a frequency $f_s = 50$ Hz. After a first warm-up task showing no target, during which the subject could get acquainted with the tablet, she or he executed the six tasks in a random order. During the three seconds preceding the beginning of each task, a countdown was simulated through a sequence of three tonal beeps.

6.1.3 Data analysis

All the task data mentioned above was recorded in real time and stored in a matrix. Hence, the full dataset comprised 96 matrices (6 tasks per 16 subjects). Prior to the analysis of such dataset, the subject's velocities in the two axes $(v_{x,s}, v_{y,s})$ were first calculated on the basis of the difference between the current $(x_s(k), y_s(k))$ and the previous $(x_s(k-1), y_s(k-1))$ indicator positions, and then smoothed with a fifth-order moving average filter.

For each task, the integral of relative velocity (i.e., the difference between subject's and target's velocities) and the weighted position error along the x-axis were measured. Each measure was calculated for every left-to-right and right-to-left segment, and then averaged over the whole task. A small number of segments (11 over all subjects and tasks, i.e. 11 over 2688) in which the participant clearly failed to follow the task, suddenly moving to the opposite direction (due to losing control of the pen or occasional distraction), were excluded from the analysis.

The *integral of relative velocity* for the k^{th} segment is formally defined as

$$R_v(k) = \frac{1}{L_k} \int_{t_k}^{t_{k+1}} |\vec{v}_r| dt, \tag{6.1}$$

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 start and end times of the segment, respectively. R_v was calculated using the classical rectangle method:

$$\sum_{h=1}^{N} \frac{\sqrt{\left(v_{x,s}\left(h\right) - v_{x,t}\left(h\right)\right)^2 + \left(v_{y,s}\left(h\right) - v_{y,t}\left(h\right)\right)^2} \cdot dt}{L_k}, \tag{6.2}$$

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

The position error along the x-axis was weighted with the target velocity sign and normalized to the target radius R. The *average weighted position error* for segment k is formally defined as

$$e_x(k) = \frac{1}{N} \sum_{h=1}^{N} \frac{\left(x_s\left(h\right) - x_t\left(h\right)\right) \cdot sign\left(v_{x,t}\left(h\right)\right)}{R}. \tag{6.3}$$

This equation 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. To this regard, *lead error* can be 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. (6.3) contribute to lead error calculation, while negative terms contribute to lag error calculation. A null value in the position error metric indicates that the subject had an average null delay with respect to target motion, even though the distance travelled around the target may have been not null.

A comparison between paired data (D'Agostino and Pearson omnibus normality test (D'Agostino, 1986)) was performed, resulting in a Gaussian distribution for tasks Br-C-Dr-E-Fr (integral of relative velocity), A-Br-Dr-E-Fr (weighted position error and lead error), 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. The significance level for the statistical analysis was set to $p = 0.05$.

6.1.4 Results and discussion

The only relevant result of the statistical analysis on the integral of relative velocity, reported in Figure 6.5, is that - as one may expect - the fixed-length task is always much better executed than the corresponding random-length task: subjects made significantly greater corrections in the latter, independently of the audio modality. This result confirms those found in (Rosati et al., 2011).

Conversely, no significant difference between fixed- and random-length tasks within the same audio modality is evidenced by the statistical analysis on the average weighted position error, as Fig. 6.6 points out. In this case, it is the auditory feedback modality that makes the difference.

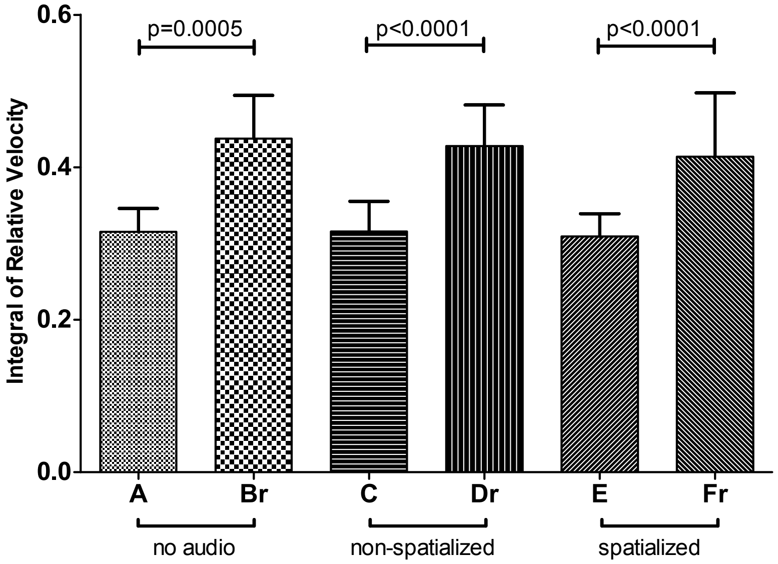


Figure 6.5: *Statistical analysis on the integral of relative velocity.*

Both the fixed-length audio tasks C and E present a smaller negative error with respect to task A, and the same applies to random-length audio tasks Dr and Fr with respect to task Br. However, only the spatial audio tasks report significant difference with respect to the no-audio tasks, while monophonic ones do not. In other words, only spatial task-related auditory feedback (tasks E and Fr) helps 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.

Monophonic feedback lies between the other two modalities in such terms, even though not reporting significant statistical differences with respect to both. It has however to be pointed out that, for fixed-length tasks, the number of outlier cycles (with respect to the weighted position error metric) in the executions of task C is much larger than that of the executions of tasks A and E: this could indicate that the sensory integration of video and audio was more difficult in the monophonic audio condition, especially during the very first cycles of the task.

When comparing these results to the related ones reported in (Rosati et al., 2012), the keen observer will note that Fig. 6.6 exhibits smaller differences in average tracking error values between tasks A-E and Br-Fr with respect to the equivalent feedback couples A-C and Br-Dr in (Rosati et al., 2012). This may be partly due to the slightly different settings of the rolling sound. However, the statistically significant upgrade given by the spatial task-related auditory feedback is preserved.

While the analysis of lag error does not add much with respect to the previous measure (means, standard deviations, and significance levels are similar to those reported in Fig. 6.6), lead error (reported in Fig. 6.7) is found to be statistically different both between fixed-length and equivalent random-length tasks and among fixed-length tasks themselves. In particular, lead

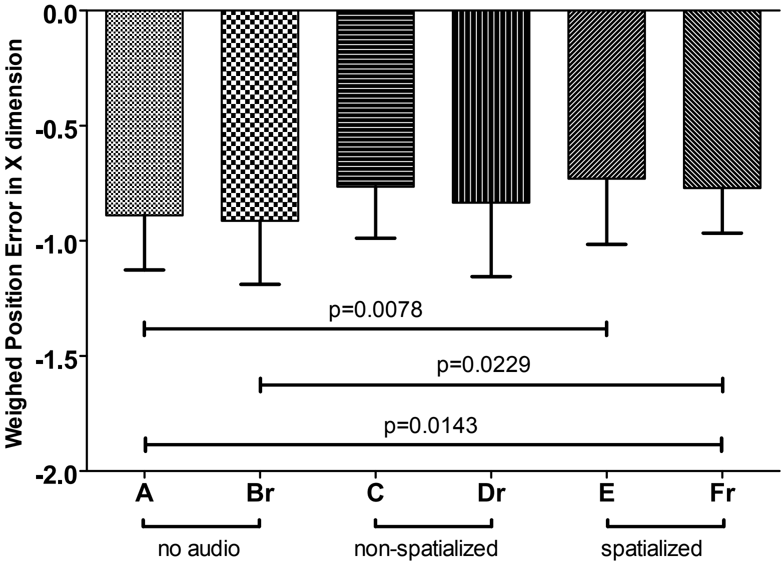


Figure 6.6: *Statistical analysis on weighted position error.*

error in task A is significantly lower than in tasks C and E. This result is harder to interpret than the previous ones; still, it could be suggested that the lead error component is greater in random-length tasks because of the sudden, unpredictable deceleration phase for short segments, whereas in fixed-length tasks lead error is lower but tends to increase in presence of consistent auditory feedback. Probably, the additional information allowed subjects to feel more confident while executing the task, tending sometimes to lead the target's movement. It is thus found that task-related auditory feedback involves actions that aim at increasing lead error.

To sum up, the effect of spatialization applied to task-related auditory feedback is found to be overall beneficial in the performance of tracking movements. However, the use of generalized HRTFs together with the absence of headphone compensation or reverberation could surely have limited the realism of the spatialization in a number of subjects, psychoacoustically resulting in a trivially panned, non-externalized version of the monophonic feedback (Begault et al., 2001). 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. Still, the other half peremptorily affirmed 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.

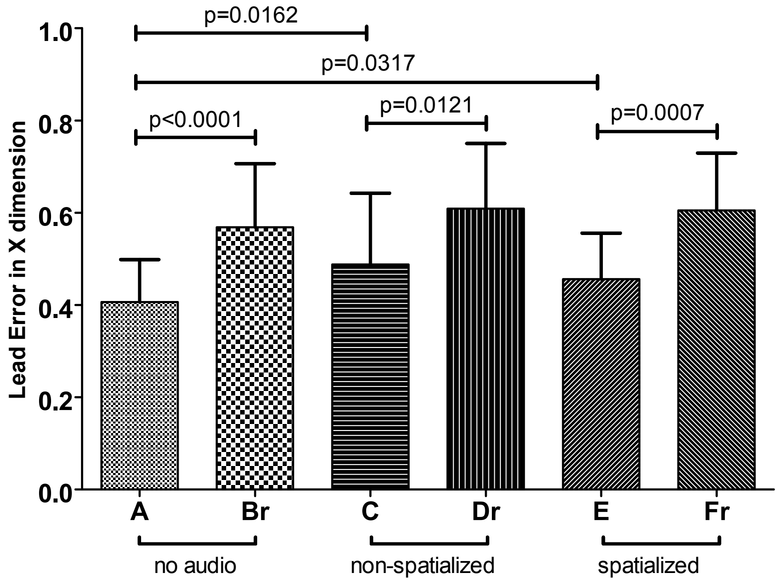


Figure 6.7: *Statistical analysis on lead error.*

6.2 Sonification of self generated footsteps

Recent research in the field of multimodal virtual environments has focused on the simulation of foot-floor interactions (Steinicke et al., 2013; Visell et al., 2009) by addressing the problem of enhancing their realism at auditory and haptic levels in order to achieve higher level of presence (Slater et al., 2009).

In particular, several results have indicated that the typology of the surface onto which we walk is processed very consistently in both the auditory and haptic modalities. The excellent somato-sensory capacities of the human feet have been demonstrated to be capable of discriminating with high accuracy different types of surfaces (Kobayashi et al., 2008; Giordano et al., 2012). Similarly, studies on the ability to identify ground materials simulated either with auditory or with haptic information (Serafin et al., 2010; Nordahl et al., 2010a) revealed that material typology is consistently recognized by using both modalities.

Turchet *et al.* have proposed a footstep sound synthesis engine, based on physical models, which allows the simulation of two typologies of ground materials: solid (i.e. homogeneous floors like wood or metal) and aggregate (i.e. grounds possessing a granular structure like gravel or snow) (Turchet et al., 2010b). The ecological validity of such simulations was assessed with experiments in which subjects were asked to recognize the synthesized materials (Nordahl et al., 2010b). Results showed that subjects were able to recognize most of the synthesized surfaces with high accuracy. Similar accuracy was noticed in the recognition of real recorded footstep sounds, which was an indication of the success of the proposed algorithms and their control.

A complicating factor is that various sound reproduction methods can be used to deliver the synthesized sounds to the walker: loudspeakers directly placed on top of the shoes (Papetti et al., 2010), on their soles (Papetti et al., 2011), or embedded in the walking surface (Visell et al., 2008). Also, the interactive delivery of footstep sounds can be achieved by means of a surround sound systems composed of loudspeakers, as illustrated in (Turchet and Serafin, 2011), and no extensive research has been conducted into headphone-based reproduction of interactive locomotion sounds.

The main goal in this second case study is to investigate the role of auditory information in modulating the localization of self generated footstep sounds, and to test whether differences in perceived localization of footstep sounds affect the realism and naturalness of the walking experience as well as the sense of disorientation associated to different layers of auditory information. To this end, different techniques for footstep sounds rendering by means of headphones are considered, allowing the delivery of stimuli with different degrees of spatiality, e.g. mono (= 0 dimensions), stereo (= 1 dimension), and binaural (= $2/3$ dimensions) reproduction by means of HRTFs. Furthermore, we assess the relative importance of auditory spatial cues with respect to semantic information such as walking surface and context as well as to signal-level features.

Three pychophysical experiments were conducted. Experiment 1 has the main goal of investigating whether different sound rendering techniques have an influence on the localization of solid and aggregate footstep sounds. The role of contextual information (soundscapes) is instead explored in experiment 2. The final experiment exploits a larger sonic palette to test whether signal-level features affect the results found in the previous two experiments.

6.2.1 Experiment #1: rendering techniques

This first experiment was designed so as to explore whether different audio rendering techniques over headphones (mono, stereo, binaural) affect localization judgments of synthetic self-generated footstep sounds on four different surface materials simulating two different surface typologies, i.e. aggregate and solid. Such a distinction is motivated by a previous work by Turchet and Serafin (2011) that highlighted significant differences (in terms of localization, realism, naturalness of the interaction, and sense of disorientation) between the perception of dynamically generated footsteps sounds on aggregate and solid surfaces provided via loudspeakers.

The starting hypothesis is that if the footstep sound has sufficient duration and high-frequency content (Vliegen and Van Opstal, 2004; Hebrank and Wright, 1974b) in order to enable vertical localization mechanisms, which is the case for aggregate surface sounds as opposed to solid surface sounds, then different rendering techniques should result in different localization ratings. In particular, binaural techniques should allow the walker to perceive synthesized aggregate footstep sounds as coming from below, despite the known difficulty in localizing virtual sources near the median plane, with an accuracy that shall depend on the degree of customization of the used HRTFs. Different localization ratings should in turn modulate the perception of the realism, naturalness and sense of disorientation of the walking experience.

Participants

Twelve participants, seven males and five females, aged between 19 and 31 (M = 22.41, SD = 4.23), took part in the experiment. All participants reported normal hearing and no impairment in locomotion.

Apparatus

The experiment was carried out in a quiet room where the setup was installed, and the walking area was 3.2×2.9 m wide. It consisted of a MacBook Pro laptop, running the sound synthesis engine described in (Turchet et al., 2010b); a pair of soft sole sandals enhanced with pressure sensors (placed in correspondence with the heel); an Arduino UNO board, managing the sensors' data acquisition; a Fireface 800 soundcard; a pair of Sennheiser HDA 200 headphones.

These headphones were chosen for three basic reasons. First, their closed form facilitates isolation from external noise. Second, as Fig. 6.8 depicts, the headphones' frequency response has a regular shape with no major peaks or notches between 100 Hz and 10 kHz that does not interfere with vertical localization cues (Masiero and Fels, 2011); this allows a non-compensated headphone rendering. Last but not least, the headphones' response is almost independent of positioning on the head (see again Fig. 6.8), which is a desirable feature while walking for maintaining a coherent audio rendering.

Footstep sound synthesis was interactively driven during locomotion of the subject wearing the shoes. The description of the control algorithms based on the analysis of the values of the pressure sensors, implemented in Max/MSP, can be found in (Turchet et al., 2010a). The generated audio stream was then sent in real time to a Pure Data patch responsible for the different audio rendering techniques. In particular, binaural synthesis was supported by the `cw_binaural~` external (Doukhan and Sédès, 2009), for real-time dynamic HRTF interpolation and convolution with any sound stimulus.

Stimuli

The used hardware allowed real-time control of the sound synthesis engine, which was set so as to synthesize footstep sounds on four surface materials, two solid (wood and metal) and two aggregate (snow and gravel).

Solid materials were simulated using an impact model (Avanzini and Rocchesso, 2001). In the simulation of impact with solids, the contact was modeled by a Hunt-Crossley-type interaction where the force f between two bodies combines hardening elasticity and a dissipation term (Hunt and Crossley, 1975):

$$f(x, \dot{x}) = -kx^{\alpha} - \lambda x^{\alpha}\dot{x} \quad \text{if } x > 0, \quad 0 \text{ otherwise.}$$

where x represents contact interpenetration (when $x > 0$ the two objects are in contact), $\dot{x}$ is compression velocity, k accounts for material stiffness, λ represents the force dissipation due to internal friction during the impact, and α is a coefficient which depends on the local geometry around the contact surface. The described model was discretized as proposed in (Avanzini and Rocchesso, 2001).

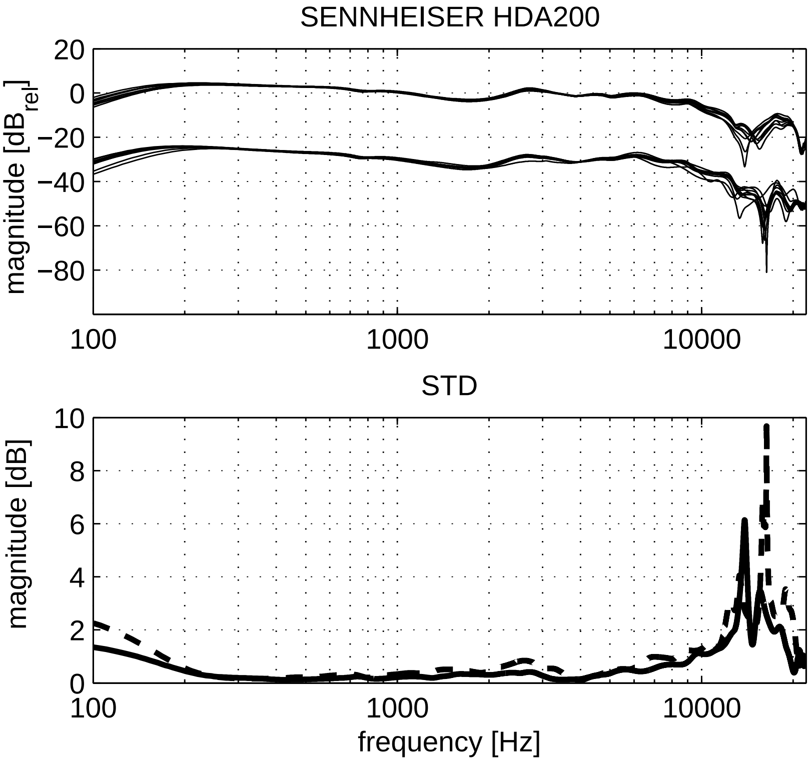
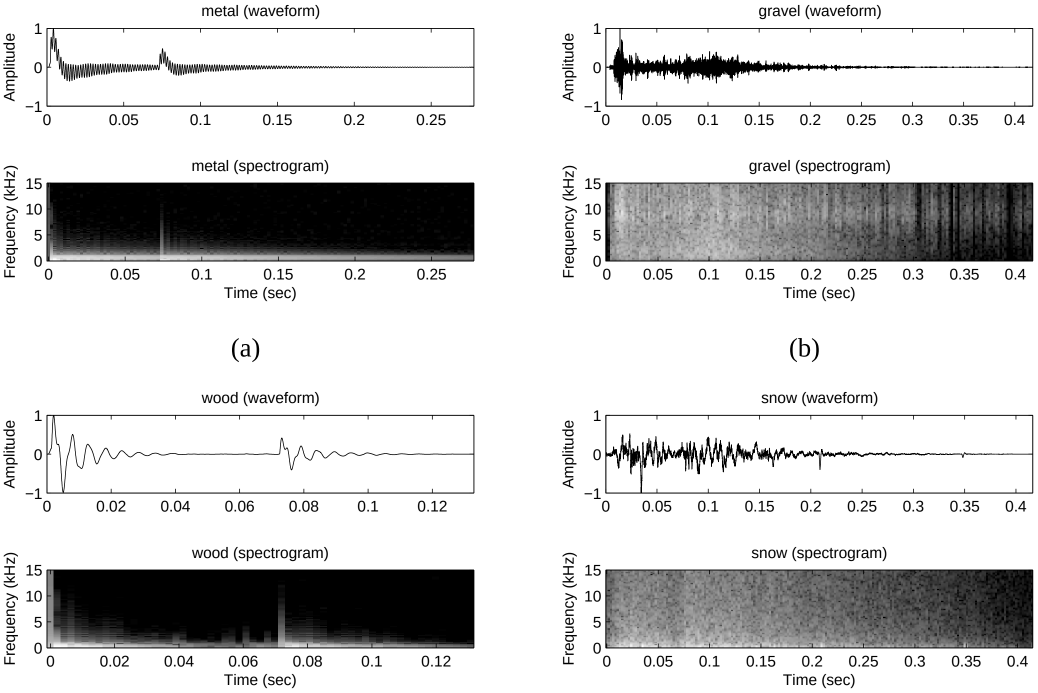


Figure 6.8: *Top: headphone impulse responses of Sennheiser HDA200 headphones measured on a KEMAR mannequin (Burkhard and Sachs, 1975) for ten different repositionings; the right channel is shifted down by* 30 *dB for convenience. Bottom: standard deviation among the ten repositionings' magnitude responses, left channel (solid line) and right channel (dashed line).*

To simulate aggregate surfaces, the physically informed sonic modeling algorithm was adopted (Cook, 1997). This algorithm simulates particle interactions by using a stochastic parametrization, thereby avoiding modeling each of the many particles explicitly. Instead, particles are assigned a probability to create an acoustic waveform. In the case of many particles, the interaction can be represented using a simple Poisson distribution, where the sound probability is constant at each time step. This gives rise to an exponential probability weighing time between events. The four signals had different features in terms of duration, amplitude, temporal evolution, and spectrum (see Fig. 6.9).

Since both males and females were involved in the experiment, footstep sounds were synthesized in order to avoid any specific cue about the gender of the walker, i.e. trying to simulate a sound which could generally be accepted as genderless. This was achieved by modeling the contribution of a type of shoe which fitted for both males and females, as ascertained in a gender recognition experiment reported in (Turchet and Serafin, 2013).

Three different sound reproduction techniques were considered: monophonic (mono, M), stereophonic (stereo panning, S) and binaural reproduction (B). In the mono condition, the peak level of the sounds was set to 55.4, 57.8, 54.2 and 61.5 dB(A) for snow, gravel, wood and metal



(a) (b)

(c) (d)

Figure 6.9: *Typical waveforms and spectrograms of the four simulated materials: (a) metal, (b) gravel, (c) wood, (d) snow.*

respectively;[2] these sound levels were taken as reference for the other reproduction conditions (S and B).

The stereo signals were obtained by adding half the mean Interaural Level Difference (ILD) of a KEMAR mannequin (Burkhard and Sachs, 1975) at ±5° azimuth to the ipsilateral channel and subtracting the same half-ILD from the contralateral channel.[3] The $5°$ value qualitatively corresponds to the displacement of each foot from the median vertical plane, allowing differentiation of left foot from right foot.

Binaural reproduction was provided by a generic or selected set of HRTFs from the CIPIC database (Algazi et al., 2001d) with the optional addition of a spherical torso approximation accounting for shadowing effects on sources coming from below (Algazi et al., 2002a). The combination of such choices gave rise to four more reproduction conditions:

1. non parametric binaural reproduction (B-NP): HRTFs of a KEMAR mannequin;

[2] Such values were chosen according to the results of a previous experiment whose goal was to find the appropriate level of amplitude for those synthesized sounds (Turchet and Serafin, 2013). Measurements were conducted by inserting the microphone of an SPL meter in a hole, having its same diameter, created in a piece of hardwood which was subsequently sealed against one of the two headphones. The amplitude peak value of the footstep sound was considered.

[3] The mean ILDs were extracted from the CIPIC HRTF database (Algazi et al., 2001d).

2. parametric binaural reproduction (B-P): HRTF selection of the best CIPIC subject according to an anthropometry-based distance metric (details follow);
3. non parametric binaural reproduction with torso (B-NPT): B-NP plus a spherical torso approximation;
4. parametric binaural reproduction with torso (B-PT): B-P plus a spherical torso approximation.

A previous study (Middlebrooks, 1999) highlighted the high correlation between the pinna cavity height, i.e. the distance from the superior internal helix border to the intertragic incisure, and an optimal frequency scaling factor aligning spectral HRTF features between subjects and thus minimizing inter-subject spectral differences. One can used such insight knowledge to guide the selection of the optimal HRTF set in the CIPIC database for a specific subject. Following the CIPIC database anthropometric parameters, the pinna cavity height p_h is given by the sum of d_1 (cavum concha height), d_3 (cymba concha height), and d_4 (fossa height). A simple "best match" of the mean measured p_h between the left and right pinnae detected the best subject for condition B-P.

Considering the impulsive nature of the footstep sound, one single spatial position for the left and right HRTFs is sufficient. Since no HRTF data for very low elevations is generally available in any public HRTF database, the lowest-elevation HRTFs were considered in all conditions. These correspond in the CIPIC database to the interaural polar coordinates $(\theta_l, \phi_l) = (-5°, -45°)$ and $(\theta_r, \phi_r) = (5°, -45°)$ for the left and right foot respectively, where θ denotes azimuth and ϕ denotes elevation.

It has to be recognized that since the used HRTFs were measured at knee height the elevation impression given to the listener might not be accurate. However, following the simplified geometry of the spherical torso approximation (Algazi et al., 2002a), we assumed that the sound wave coming from below travels around the sphere spanning an angle $\theta_{inc} = 135°$ before reaching the ear(s) at approximately $-45°$ elevation. This approximation was considered in the B-NPT and B-PT conditions, where the theoretical solution for diffraction around a rigid sphere (Strutt, 1904) with $\theta_{inc} = 135°$ was used to design a FIR filter reproducing its magnitude behavior. The only independent variable of the spherical model, i.e. the sphere radius, was adapted to the maximum circumference t_c of the subject's torso.

In order to maximize localization accuracy, even to the detriment of perceived realism, no reverberation was applied to the sound stimuli. The combination of the six rendering techniques and the four surface materials gave rise to 24 stimuli, each repeated twice for a total of 48 trials. Trials were randomized across participants.

Procedure

Participants were first subjected to a short anthropometric measurement session where parameters p_h and t_c were acquired. Then, each subject wore the pair of shoes and a belt which allowed the wires from shoes and headphones to be fixed to the user's back and to then be directed to the Arduino board. In addition, wires were attached to the subject's trousers with Velcro tape and secured to the waist. The wires were long enough (5 m) to allow free motion in the experimental

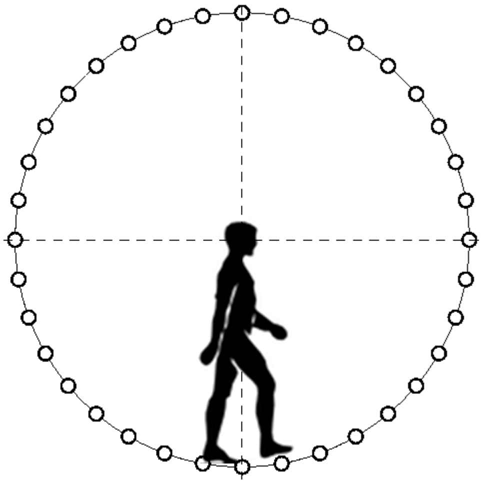


Figure 6.10: *Figure for questionnaire item Q1.*

space. The experiment was conducted in a laboratory whose floor was covered with carpet in order to mask the footstep sounds resulting from the interaction of sandals with the floor. Such a masking was further enhanced by the use of the closed headphone set, in addition to the softness of the sandals' sole.

Participants, who were never informed about which material was simulated at each trial, were instructed to walk freely inside the walking area and experience the trial as much as they wanted before concluding it. At the end of each trial, participants had to fill in the following questionnaire:

Q1: indicate in Fig. 3 the circlet corresponding to the point where the sound comes from;

Q2: evaluate the degree of realism of the sounds you have produced;

Q3: evaluate to what extent your way of walking seems natural to you;

Q4: evaluate to what extent you feel confused or disoriented while walking.

The circlets in Fig. 6.10 are $10°$ equally spaced because of the high localization uncertainty in the median vertical plane (Blauert, 1983). Questions Q2, Q3, and Q4 were instead evaluated on a visual analogue scale (VAS) [0 = not at all, 10 = very much]. Such questions were motivated by the necessity of having additional information concerning the subjective experience of interacting with the provided virtual world. Specifically, they were chosen because the realism of the provided sounds, the naturalness of the walking experience, and the sense of confusion or disorientation while walking are factors related to the sense of presence (Slater et al., 2009). Before performing the task, subjects were presented with six practice trials (one for each rendering technique) in order to become familiar with the system. To this purpose, the forest underbrush material was chosen (delivered at 53.5 dB(A)). This material was not among those involved in the experiment.

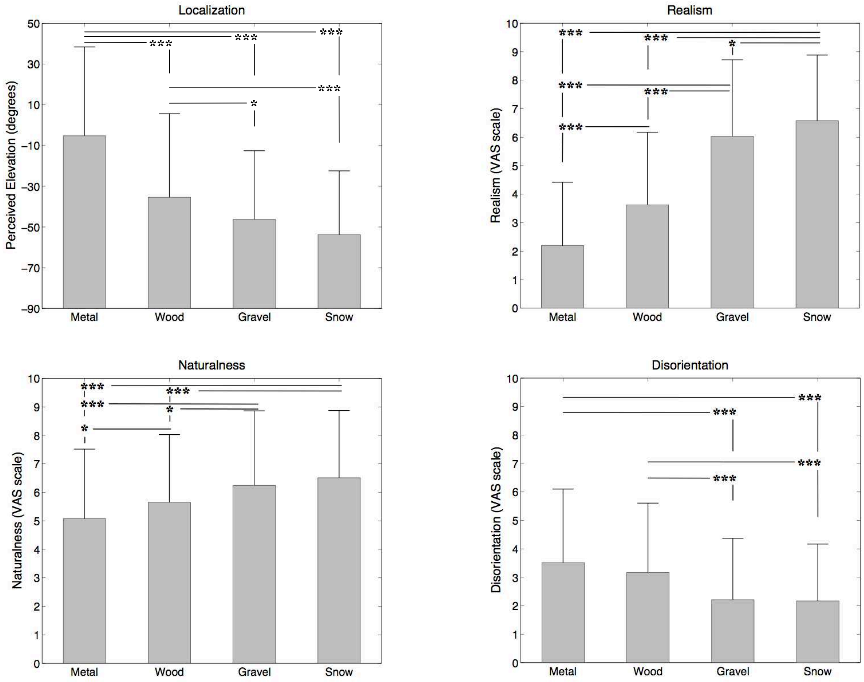


Figure 6.11: *Results of experiment 1: graphical representation of the mean and standard deviation for questionnaire items Q1 (top-left), Q2 (top-right), Q3 (bottom-left), and Q4 (bottom-right). Legend: * represents p ≤ 0.05 and *** p ≤ 0.001.*

Results and discussion

Data corresponding to questionnaire item Q1 were first analyzed with respect to scores corresponding to the circlets placed in the front and back half-circumferences (FHC and BHC) in Fig. 6.10 (i.e. the points in which the sound was perceived as coming from the front and from the back respectively). Such an analysis was performed in order to verify the presence of a preference for localization of the sound at the front or at the back. The number of scores in FHC and BHC was counted for each technique and each material separately, and subsequently analyzed by means of an exact binomial test. This statistical analysis revealed that in all cases the difference between the counts in FHC and BHC was not significant. Localization scores in the two half-circumferences were then subjected to a three-way repeated measures ANOVA having two levels of group (negative scores $[-18, 0]$ anticlockwise in the BHC and positive scores $[0, 18]$ anticlockwise in the FHC, where 0 is the lowest point in Fig. 6.10), six levels of rendering technique and four levels of material. A significant main effect was found only for material, $F(3, 33) = 125.3$, $p < 0.001$.

As a consequence, the localization scores corresponding to BHC were normalized in absolute value and added to those in FHC for further analyses. The resulting data were subjected to a two-way repeated measures ANOVA having six levels of rendering technique and four levels of material. The main effect of rendering technique was non-significant. The main effect of material was significant, $F(3,33) = 21.4$, $p < 0.001$. The interaction was non-significant.

As illustrated in the top-left panel of Fig. 6.11, the post-hoc analysis, performed by using Tukey's procedure, revealed that localization scores for the four materials were all significantly different except between the gravel and snow conditions. In particular, localization scores for the snow and gravel conditions were both significantly lower (i.e. towards the foot of the avatar in Fig. 6.10) than the metal and wood conditions.

Figure 6.11 also shows the evaluations expressed as VAS scores for questions Q2 (realism), Q3 (naturalness), and Q4 (disorientation) considering the data grouped by material. The three questionnaire items were subjected to a two-way repeated measures ANOVA having six levels of rendering technique and four levels of material. Concerning Q2, the main effect of rendering technique was non-significant, while the main effect of material was $F(3,33) = 19.1$, $p < 0.001$, and the interaction effect was also significant ($F(15,165) = 2.38$, $p < 0.01$). Tukey's post-hoc test indicated that realism scores were significantly different among all conditions and in ascending order for the metal, wood, gravel, and snow conditions. As regards Q3 and Q4, a significant main effect was again found only for material (Q3: $F(3,33) = 4.45$, $p < 0.01$, Q4: $F(3,33) = 5.09$, $p < 0.01$). For the sake of brevity, results of the respective post-hoc tests are reported in the figure.

In addition, linear mixed-effects model analyses were performed in order to search for correlations between each localization score (in absolute value) and each VAS evaluation expressed for Q2, Q3, and Q4. Such analyses revealed that the localization scores were not linearly related to either perceived realism, naturalness, or disorientation.

The four questionnaire items were subjected to a further one-way repeated measures ANOVA having two levels of surface typology (solid and aggregate). In all cases a significant main effect was found, showing that localization and disorientation scores were higher for the solid typology compared to the aggregate one ($F(1,11) = 25.73$, $p < 0.001$ and $F(1,11) = 10.81$, $p < 0.01$ respectively), and realism and naturalness scores were lower for the solid typology compared to the aggregate one ($F(1,11) = 29.67$, $p < 0.001$ and $F(1,11) = 6.69$, $p < 0.05$ respectively).

No significant differences among the six rendering techniques were found. This is in accordance with the initial hypothesis for solid surfaces, whose associated sounds do not have enough energy at high frequencies to enable vertical localization mechanisms (Hebrank and Wright, 1974b). As Fig. 6.9 shows, the frequency content of solid footstep sounds (wood and metal) only overshoots the 4-5 kHz threshold that enables vertical localization by the pinna in very short temporal windows. For footstep sounds in particular, the presence of high-frequency energy is needed to trigger not only pinna-related elevation cues (i.e., frequency notches), but also torso-related ones (i.e., shadowing effects).

However, binaural techniques were all unexpectedly found to be ineffective also for aggregate surfaces, independently of the degree of customization. Indeed, the interaction effect between material and rendering technique was not significant. Instead, results showed that materials belonging to the aggregate surface typology were always localized significantly lower than the solid ones. Therefore, taken together these results suggest that surface typology has an influence on the localization judgments, and that such an influence is strong enough to mask the differences between the involved rendering techniques.

Coherently, significant differences were also found between evaluations of aggregate and solid surfaces as far as the perceived realism of the simulations is concerned, as well as the naturalness of the walk and the degree of confusion or disorientation. Figure 6.11 shows that those judgments scale monotonically with the localization scores.

6.2.2 Experiment #2: contextual information

In order to test the strength of the surface typology effect in localization perception and to confirm the results of the first experiment concerning the absence of differences in localization judgments between the rendering techniques, a second experiment was designed. Specifically, the directionality of footstep sounds was studied in presence of sonically simulated virtual environments, i.e. adding a soundscape.

The role of contextual information, sonically provided as soundscape, on the perception of footstep sounds was studied in (Turchet et al., 2010c). Soundscapes sonically simulated either the environment typically associated with the surface material synthesized (i.e. coherently), or with a totally different one (i.e. incoherently). Results showed that adding a coherent soundscape significantly improved both recognition of surface materials and realism evaluations when compared to both footstep sounds alone and with footstep sounds with an accompanying incoherent soundscape.

In this second experiment, adding auditory information concurrent to the footstep sounds might decrease the accuracy of their localization, and such a decrement could be greater when incoherent soundscapes are provided compared to the case in which coherent ones are involved. However, if the effect is still present in such conditions this would mean that the effect is strong and that its causes might not only be due to the auditory channel per se but should be searched in the multimodal perceptual mechanisms involved in locomotion.

Participants

Twelve participants, six males and six females, aged between 19 and 26 (M = 22.66, SD = 2.49), not one of whom was involved in the previous experiment, took part in the experiment. All participants reported normal hearing and no impairment in locomotion.

Stimuli and procedure

The same apparatus was used as in the first experiment. In addition to footstep sounds, the soundscapes of the following four environments were used: a courtyard of a farm during summer; a ski slope; a house interior; a submarine. Such ad-hoc built soundscapes were the same adopted in (Turchet et al., 2010c) and were chosen in order to coherently fit with the synthesized footstep sounds (gravel, snow, wood, and metal respectively). When incoherently provided, they were coupled with metal, wood, snow, and gravel respectively. The used soundscapes were designed so as to provide a clear indication of the designed environments after the first few seconds.

The RMS amplitudes of the soundscapes were set to 54.1, 67.2, 62.7 and 63 dB(A) for the house, the submarine, the courtyard, and the ski slope respectively. Such values were again chosen according to the results of (Turchet and Serafin, 2013), whose goal was to find the appropriate sound level for those soundscapes in presence of synthesized footstep sounds set to the amplitudes indicated in Sec. 6.2.1 .

The experimental protocol was analogous to that of the first experiment. The familiarization phase consisted of presenting the footstep sounds of forest underbrush alone, with a coherent soundscape corresponding to a forest, and with an incoherent soundscape corresponding to a beach seaside in summer. Both the material and the two soundscapes were not among those involved in the experiment.

Footstep sounds were rendered using the M and B-PT techniques only. This choice was made in order to check whether the delivery method affects the quality of the results as far as the aggregate surfaces are concerned in presence of an accompanying soundscape. Results were expected to confirm those of the first experiment, i.e. no significant differences between M and B-PT. The combination of the two rendering techniques, the four surface materials, and the three soundscape conditions (coherent, incoherent, no soundscape) gave rise to 24 stimuli, each repeated twice for a total of 48 trials. Trials were randomized across subjects.

Results and discussion

Results of the second experiment are illustrated in Fig. 6.12. Localization scores were analyzed by means of a three-way repeated measures ANOVA having three levels of background context (footstep sounds alone, with coherent soundscape, with incoherent soundscape), two levels of rendering technique and four levels of material. Neither the main effect of rendering technique nor of stimulus type were significant. The main effect of material yielded an F ratio of $F(3,33) = 10.32$, $p < 0.001$. The interaction effects were non-significant. The post-hoc analysis revealed that localization scores were significantly lower for both the snow and gravel conditions when compared to both the metal and wood conditions.

The evaluations of Q2, Q3 and Q4 were subjected to the same three-way ANOVA. As regards Q2, the main effect of stimulus type yielded an F ratio of $F(2,22) = 12.6$, $p < 0.001$, the main effect of material yielded an F ratio of $F(3,33) = 17.29$, $p < 0.001$, and the interaction effect between stimulus type and material was significant, $F(6,66) = 2.57$, $p < 0.05$. The first post-hoc test indicated that realism scores were all significantly different except between the gravel and snow conditions; in particular, scores for the snow and gravel conditions were both significantly higher than the metal and wood conditions. The second post-hoc test indicated that realism scores were significantly higher for coherent soundscapes when compared to footstep sounds alone, which in turn were significantly higher than for incoherent soundscapes. Concerning Q3, the main effect of stimulus type yielded an F ratio of $F(2,22) = 8.96$, $p < 0.01$, the main effect of material yielded an F ratio of $F(3,33) = 6.53$, $p < 0.01$, and the interaction effect between stimulus type and material was significant, $F(6,66) = 2.37$, $p < 0.05$. The two post-hoc tests gave analogous results to those of Q2. Regarding Q4, the main effect of stimulus type yielded an F ratio of $F(2,22) = 11.82$, $p < 0.01$, the main effect of material yielded an F ratio of $F(3,33) = 6.72$, $p < 0.01$, while the interaction effects were non-significant. The first post-hoc test indicated that disorientation scores were significantly higher for the metal condition when compared to all of the other conditions. The second post-hoc test indicated that disorientation

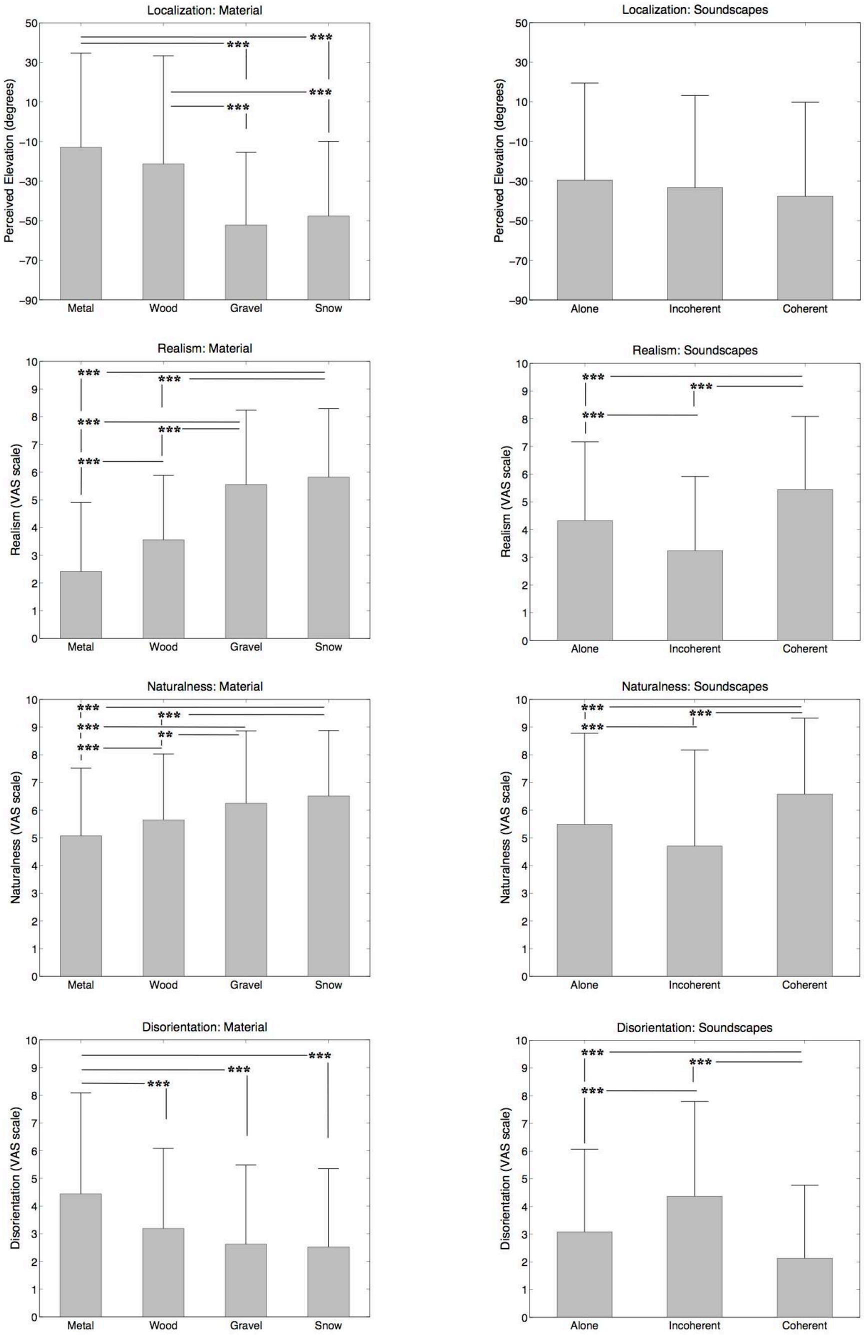


Figure 6.12: *Results of experiment 2: graphical representation of the mean and standard deviation for questionnaire items Q1, Q2, Q3, and Q4 analyzed by material (left) and by type of stimulus (right). Legend: ** represents p ≤ 0.01 and *** p ≤ 0.001.*

scores were significantly lower for coherent soundscapes when compared to footstep sounds alone, which in turn were significantly lower than for incoherent soundscapes.

Moreover, linear mixed-effects model analyses were performed in order to search for correlations between each localization score (in absolute value) and each VAS evaluation expressed for Q2, Q3, and Q4. Such analyses revealed that localization scores were not linearly related to either perceived realism, naturalness, or disorientation.

The four questionnaire items were subjected to a further one-way repeated measures ANOVA having two levels of surface typology (solid and aggregate). In all cases a significant main effect was found, showing that localization and disorientation scores were higher for the solid typology compared to the aggregate one ($F(1,11) = 14.38$, $p < 0.01$ and $F(1,11) = 7.64$, $p < 0.05$ respectively), and realism and naturalness scores were lower for the solid typology compared to the aggregate one ($F(1,11) = 22.81$, $p < 0.001$ and $F(1,11) = 7.73$, $p < 0.05$ respectively).

The results of this second experiment confirm, as expected, the prevalence of the information related to surface typology over the spatial rendering technique as far as perceived localization is concerned. Localization scores were not affected by the presence of the soundscapes provided, and analogously to the findings of the previous experiment, localization scores were not linearly related to judgments of realism, naturalness and disorientation. Moreover, concerning localization judgments, no significant differences were found between conditions in which footstep sounds were provided alone or with an accompanying soundscape. These results, therefore, would indicate that localization of footstep sounds is affected by the simulated surface typology and that this effect is independent of the presence of a soundscape, coherently or incoherently provided.

Concerning the perceived realism of footstep sounds, an influence of the presence of contextual information was noticed: footstep sounds accompanied by a coherent soundscape were judged significantly more realistic than when provided alone or with an incoherent soundscape. These findings confirm the results reported in (Turchet et al., 2010c).

The results of both the first and second experiment thus suggest that the influence of surface typology on localization judgments is a robust effect, since it is independent of the used rendering technique and of the presence of contextual information.

6.2.3 Experiment #3 signal-level features

The set of surface materials involved in the previous two experiments was relatively small. Only four synthesized materials were used, and no comparison against recordings of real footstep sounds was conducted. Another critical point arising from the first two experiments is that at signal level aggregate sounds are significantly longer in time and significantly richer in high-frequency content than solid sounds, hence the found effect could be merely dependent on temporal or spectral factors.

From all these considerations, a third experiment was designed with the goal of i) replicating the results of the first two experiments using a larger palette of surface materials; ii) test the effectiveness of synthesized footsteps sounds compared to recorded samples; and iii) assessing whether the found effect could be due to signal-level features of the involved sound stimulus.

Participants

Twelve participants, three males and nine females, aged between 19 and 39 (M = 25.75, SD = 6.09), all of whom were not involved in the previous experiments, took part in this experiment. All participants reported normal hearing and no impairment in locomotion.

Stimuli and procedure

The same apparatus was used as in the first two experiments. Both recordings of real and synthesized footstep sounds were used, for a total of 21 surface materials (9 solid, 10 aggregate, and 2 control conditions). In particular, the solid materials were wood, concrete and metal all provided as real and synthesized (54.2, 56.3 and 61.5 dB(A) respectively) sounds. Moreover, three sounds were created by coupling the synthesized materials with a reverberation tail corresponding to a room of size $9 \times 9 \times 2.5$ m ($T_{60} = 0.505$ s). Concerning the aggregate materials, the following surfaces were used (all provided as real and synthesized): snow, gravel, dry leaves, dirt pebbles, and forest underbrush (55.4, 57.8, 54.4, 53.5, 53.5 dB(A) respectively). The same amplitude for the corresponding real and synthesized materials was adopted and set according to the amplitude indicated in (Turchet and Serafin, 2013). The recordings of real surfaces were the same as those used in a recognition experiment described in (Nordahl et al., 2010b).

The recordings of real footstep sounds were used to increase the sonic palette and to search for possible differences with the synthesized sounds in the four questionnaire items. Analogously, the addition of reverberation to synthesized solid surfaces was used in order to verify possible differences in participants' evaluations compared to synthesized solid surfaces without reverberation: indeed, the duration of the reverberated stimuli lasted for a time long enough to cover the average duration of real footsteps, i.e. the whole temporal duration of the haptic stimulus, as opposed to the drier un-reverbed sounds.

Moreover, two control conditions were considered. They consisted of white noise bursts, lasting 80 and 420 milliseconds respectively, both provided at 56 dB(A). The two durations were set to the minimum and maximum duration of the involved solid and aggregate surface sounds respectively, while amplitudes were set to the average amplitude of all sounds. These control conditions were chosen to verify possible localization biases due to the stimulus' duration or frequency content. As a matter of fact, one of the salient differences between footstep sounds on aggregate and solid surfaces is the duration, which is longer for the first compared to the second. Furthermore, noise bursts have more high-frequency content than aggregate surface sounds; hence if frequency content were responsible for the localization bias then the noise bursts would be localized even lower.

Since the previous experiments revealed no significant differences between the techniques used for sound delivery, only one technique, M, was used. Each of the 21 stimuli was repeated twice for a total of 42 trials. Trials were randomized across subjects. The procedure was identical to that of the first two experiments, anthropometric measurements excluded. The familiarization phase consisted of presenting recordings of both real and synthesized footstep sounds on sand delivered at 51.9 dB(A). These stimuli were not among those involved in the experiment.

Results and discussion

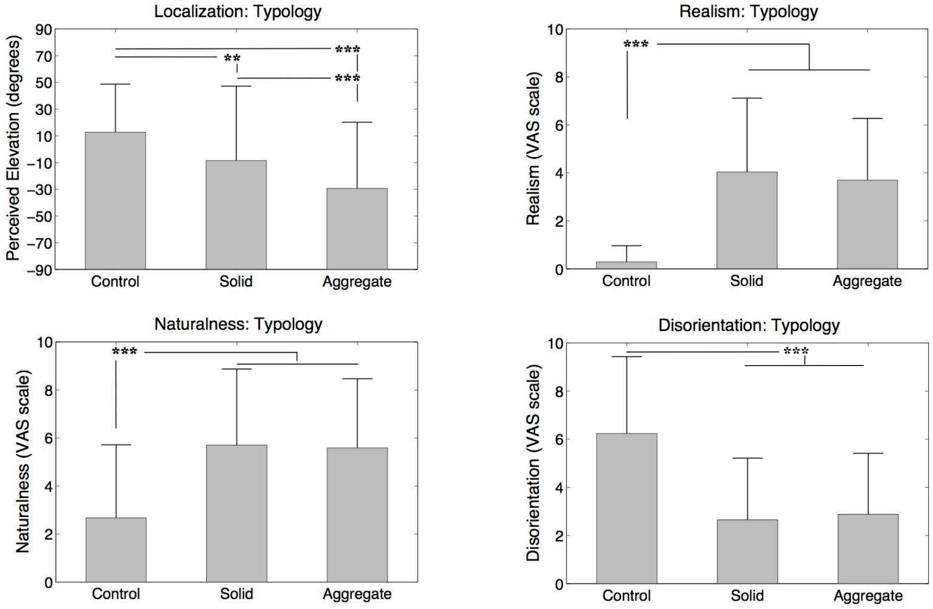


Figure 6.13: *Results of experiment 3: graphical representation of the mean and standard deviation for questionnaire items Q1, Q2, Q3, and Q4 analyzed by surface typology. Legend: ** represents p ≤ 0.01 and *** p ≤ 0.001.*

Figure 6.13 shows the results of the third experiment. Localization scores were analyzed by means of a one-way repeated measures ANOVA having three levels of surface typology (control, solid, aggregate). The ANOVA showed a significant main effect, $F(2,22) = 11.78$, $p < 0.001$. The post hoc comparisons indicated that the localization scores were significantly higher for the control condition when compared to solid and aggregate conditions and significantly higher for the solid condition when compared to the aggregate condition. A further one-way repeated measures ANOVA showed no significant differences between localization scores of the synthesized surfaces and the real ones. Similarly, a one-way repeated measures ANOVA showed no significant differences between localization scores of the synthesized solid surfaces with and without reverberation. Also, a one-way repeated measures ANOVA showed no significant differences between localization scores of the two noise bursts.

The evaluations of Q2, Q3 and Q4 were subjected to the same analyses. The main effect of surface typology yielded an F ratio of $F(2,22) = 28.68$, $p < 0.001$ for Q2, $F(2,22) = 15.77$, $p < 0.001$ for Q3, and $F(2,22) = 12.26$, $p < 0.001$ for Q4. The post-hoc test indicated that realism and naturalness (disorientation) scores were significantly lower (higher) for the control condition when compared to solid and aggregate conditions while no significant differences were found either between the synthesized surfaces and the real ones or the synthesized solid surfaces with and without reverberation.

As before, linear mixed-effects model analyses were performed in order to search for correlations between each localization score (in absolute value) and each VAS evaluation expressed for Q2, Q3, and Q4. Such analyses revealed that localization scores were not linearly related to either perceived realism, naturalness, or disorientation.

A further one-way repeated measures ANOVA was conducted on the four questionnaire items to compare the two control conditions. In none of the analyses statistical significance was noticed.

Taken together, results of the third experiment confirm that footstep sounds on aggregate surfaces are localized nearer to the feet than those on solid surfaces. Furthermore, both the noise bursts were localized in positions higher than those corresponding to the real and synthesized solid surfaces, and their localization scores did not differ significantly. Last but not least, no significant localization difference was found between solid surfaces with and without reverberation. Therefore these findings exclude any explanation of the cause of the found effect due to the duration or frequency content of the sound stimulus.

Contrarily to the previous two experiments, realism, naturalness, and disorientation scores were not significantly different for the solid and aggregate surface typologies, while as expected control conditions were judged as the least realistic. Furthermore, similar ratings were given for the real and synthesized sounds for all the questionnaire items; this suggests the success of the synthesis algorithms in mimicking real footsteps sounds. Analogously, in each of the four questionnaire items no significant difference was found for the synthesized solid surfaces with and without reverberation. This finding parallels the corresponding localization results.

6.3 General discussion and conclusions

The influence of auditory feedback was studied on healthy subjects first to characterize the normative response of the human motor system to auditory information, yet the experiment is ready to be adapted to a rehabilitative scenario in order to attest the absolute effectiveness of spatial sonification in target-following tasks. However, the results in Sec. 6.1 definitely provide a basis for a future comparison with impaired subjects. Whereas differences between spatial and monophonic task-related auditory feedback were not found to be particularly marked, spatial auditory feedback led to a statistically significant improvement in performance with respect to the no-audio condition. Such improvement was not observed for monophonic auditory feedback.

On the other hand, the main result common to the three experiments in Sec. 6.2 is that solid surfaces are localized significantly farther from the walker's feet than aggregate ones independently of rendering technique (including spatial audio feedback), presence or absence of contextual information, duration and frequency content of the sound stimulus. Such an effect could be explained by the presence of a semantic conflict between the haptic and auditory sensory channels, coupled with the hypothesis that the auditory system uses the information coming from the haptic channel to enhance sensitivity in the localization of sounds apparently coming from the walker's feet.

Semantic content of a multisensory stimulus plays a critical role in determining how it is processed by the nervous system (Laurienti et al., 2004). Visual system could make use of auditory cues during visual analysis of human action when there is a meaningful match between the auditory and visual cues (Thomas and Shiffrar, 2010). In these case studies the source of the auditory and visual (i.e., the movement of a rolling ball) was perceived from most of the subjects as unique and the two sensory channels rely on coherent information; otherwise, auditory and haptic (i.e., the foot-shoe contacts while walking) stimuli was not unique, and therefore the two sensory channels received conflicting information.

Still, this interpretation is supported by the evidence that audio-tactile interactions can happen independently of spatial coincidence in the region close to the head (see (Kitagawa and Spence, 2006) for a review). Several studies in the literature attest how information presented on one sensory modality can influence information processing in another sensory modality (e.g. the ventriloquism illusion (Howard and Templeton, 1966) and the "parchment-skin" illusion (Jousmäki and Hari, 1998)). In (Laurienti et al., 2004) the authors highlighted how the semantic content of a multisensory stimulus plays a critical role in determining how it is processed by the nervous system.

The information provided by spatial audio feedback using generalized HRTFs was found to reduce tracking errors only for a subgroup of subjects. In the first case study, along with improvements in the monophonic signal, a required step towards a better rendering of the used feedback is the exploitation of customized HRTFs (Møller et al., 1996). The use of customized HRTFs is expected increase the improvement in performance between the no-audio and spatial audio conditions, strengthening congruence between modalities in terms of spatial width and externalization.

In the second case study, technological limitations even forced the rendering of solid surfaces to be semantically incongruent: although the impact sound produced by hard sole shoes with a solid surface was realistically rendered, haptic stimuli induced by the actuators were not effective in masking the haptic sensation due to the softness of the sandals' sole and the presence of a carpeted floor. To this regard, the haptic sensation arising when walking with sandals over a floor covered with carpet is more semantically incongruent with the simultaneous presentation of an impact sound between a hard sole and a solid surface than with the simultaneous presentation of a footstep sound on an aggregate surface. The different localization ratings reported in the present study could be attributable to the different levels of semantic congruence between auditory and haptic information: the lower the semantic congruence, the greater the distance of the sound source from the feet. On a separate note, realism, naturalness, and disorientation scores were found to be unaffected by semantic congruence and by the induced localization bias, which consequently did not modulate the perception of the own way of walking.

In addition, it is worthwhile to notice that the present studies involved auditory stimuli that are both valid and not valid from the ecological point of view. In presence of non-ecological stimuli (e.g. noise in the second case study), behavioral/perceptual quantities (e.g. location of the sound source) were worst rated than the corresponding congruent and incongruent ecologically valid stimuli. This is a further indication that when the association between the information arriving to different modalities is not meaningful, interaction between sensory channels produces percepts which are not reliable.

The design process of audio-visual rehabilitation and audio-haptic locomotion interfaces for virtual reality contexts should take care of providing users with feedback fully valid from the ecological point of view, and capable to produce a meaningful association between the two sensory modalities. This aspect has received scarce attention from engineers in tools for rehabilitation protocols and designers of synthetic footstep sounds and vibrotactile feedbacks. Furthermore, the use of spatial sound reproduction techniques (through headphones) is an essential workbench where testing whether a performance improvement is detected or cue integration might effect human multimodal perception.

This research workflow contributes to the development of a theoretical framework applied to perceptual mechanisms involved in sonically augmented tracking/reaching tasks and simulated foot-floor interactions mediated by interactive interfaces. Ultimately, future researches will allow investigation of how audio-video-haptic interactions in humans contribute to the internal multisensory representation of the body.

Chapter 7

Auditory feedback for exploration tasks

In recent years spatial sound has become increasingly important in a plethora of application domains. Spatial rendering of sound is especially recognized to greatly enhance the effectiveness of auditory human-computer interfaces (Begault, 1994), particularly in those cases where the visual interface is limited in extension and/or resolution. Furthermore, it aids improving the sense of presence in augmented/virtual reality systems and adds engagement to computer games (see Sec. 1.3.1 for a VAD taxonomy). Among these ever-growing application domains, this Chapter focuses on the design and development of a real time, physically informed audio-tactile interactive system able to aid navigation in virtual 3D maps. Special emphasis on spatial audio models motivates the use of innovative binaural techniques for 3D audio rendering based on headphone reproduction, and their influence in spatial cognition.

In particular, design and prototypes of sensory substitution systems for visually impaired people aim to assess whether the acquisition of spatial information through non-visual modalities, i.e auditory and haptic, includes any properties of internal representations build upon visual cues (Cattaneo et al., 2008). The degree of dependency from visuospatial nature of images determines to which extent spatial knowledge should be considered as a more general property.

An effective Orientation & Mobility (O&M) educational program for visually impaired people based on real/virtual environments, should involve object and obstacle recognition, reliance on global landmarks, route and/or survey knowledge (Wiener et al., 2009). Active exploration of virtual environments might lead to the acquisition of such a spatial knowledge. Available idiothetic and allothetic.[1] information during exploration tasks interact with attention and working memory according to different weights (Afonso et al., 2010), affecting the spatial learning process.

The pinna reflection model, exhaustively described in Sec. 4.2, applied to user's ear image might be used in order to extract prominent anthropometric features that capture relevant elevation cues for binaural audio synthesis through headphones. The selection mechanism described

This chapter is partially based on (Geronazzo et al., 2013a).

[1] Idiothetic information refer to motor commands determining locomotion decisions, proprioception and vestibular (inertial) information for self-motion (Mittelstaedt and Mittelstaedt, 2001) Otherwise, allothetic information refer to external environment and include sensory cues.

in Sec. 4.3 used manually traced geometrical contours on a single pinna picture so as to match the extracted elevation cues within the best HRTF for the corresponding listener. The selected HRTF set is integrated in a virtual multimodal environment together with the haptic TActile MOuse (TAMO) device,[2] exploiting the contribution of audio feedback to spatial cognition and O&M education/training of visually impaired people. Reaching activities, object reconstruction and displacement/size estimation of virtual objects in space are three experimental scenario where these technologies were preliminarily tested in simple virtual maps incorporating a virtual object and a spatialized beacon sound. The promising results confirms the effectiveness of the proposed research methodology towards an effective sonification of movements and multimodal cross-augmentation in navigation tasks.

A closing preliminary experiment has the aim at investigating the influence of auditory feedback on haptic estimation of size. This work was motivated by the idea of developing an audio augmented minimalist tactile device, answering the following research questions: "Which is the more semantically congruent feedback for audio-haptic estimation of spatial height?" "Can audio modulate local perception of height?". Experimental subjects were instructed to explore a virtual 3D object (a stair-step) with a haptic device, and to return a verbal estimate of the step riser height. Haptic exploration was accompanied with a real-time generated sinusoid whose pitch varied as a function of the object's interaction point's height within two different ranges. Experimental results show that haptic estimation is robust and accurate regardless of the frequency range of the accompanying sound, leading to other kinds of auditory feedback and mapping strategies in future researches.

7.1 Motivation and open issues

3D audio has been increasingly applied to multimodal system in the last twenty years (Vorlnder, 2007; Hale and Stanney, 2002) as demostrated by the introduction of the word "auralization" back to the early 1990s. Along with already existing applications (see Sec. 1.3 for a detailed taxonomy), many unexplored scenarios exist, based either on incremental improvement of existing systems or on totally new approaches (Begault, 1994).

However, ensuring coherent sound generation, sound transmission and sound reproduction in virtual acoustic environments becomes crucial especially in those applications where the auditory modality is more than just a useful addition to vision in order to increase user immersion: in particular, auralized systems can aid blind people on the cognition and exploration of spatial maps. Identification of unknown obstacles and reliable cues for actual positions in space are required skills for people totally or partially without sight, particularly for late blinds (e.g. aged and diseased people).

The issue of sensory integration/combination related to spatial knowledge becomes more challenging considering dependencies of spatial learning from previous and current visual abilities (Cattaneo et al., 2008), i.e. late blind (or late-onset blindness) and congenitally blind (or

[2] http://www.iit.it/en/component/content/article/175-platforms/505-digeye.html, last access: December 10th 2013

early-onset blindness). To this regard, a growing body of evidence suggests that spatial information is partially independent from encoding modalities, especially vision, suggesting amodal representations in the learning process (Klatzky et al., 2003). Such an hypothesis on a common spatial substrate supports sensory substitution researches exploiting similarities between learning modalities.

Experimental protocols should emphasize interoperability and interdependency of modalities (Chrastil and Warren, 2011) keeping in mind the ecological distinction between near and far field. Furthermore, the exploration of large environments requires locomotion, on the contrary exploration of small objects (compared to human body dimensions) within a simple grasping action does not require a person to move and walk.

Consider the following application scenario: a detailed virtual map is available for download in order to help visually impaired people to learn the arrangement of a public space before actually going there. It has been suggested that recognition and reconstruction of cognitive spatial maps are more effective for those people who have benefited from practice in virtual environments than for those who need to explore a real space. Linear behavioral strategies are intrinsic in real space exploration while virtual environments can be freely handled, varying the starting point, scale factor and then moving in a three dimensional space along all the allowed directions, exploiting the vast combinations of multi-sensory feedback (Lahav and Mioduser, 2008).

More in detail, the main challenges to be faced in order to provide the user with effective ways of navigating and learning virtual spatial maps are: (i) resembling reliable real spatial knowledge (ii) complying with individual cognitive abilities and enhancing them, and (iii) providing an efficient system in terms of usability, scalability and technological resources.

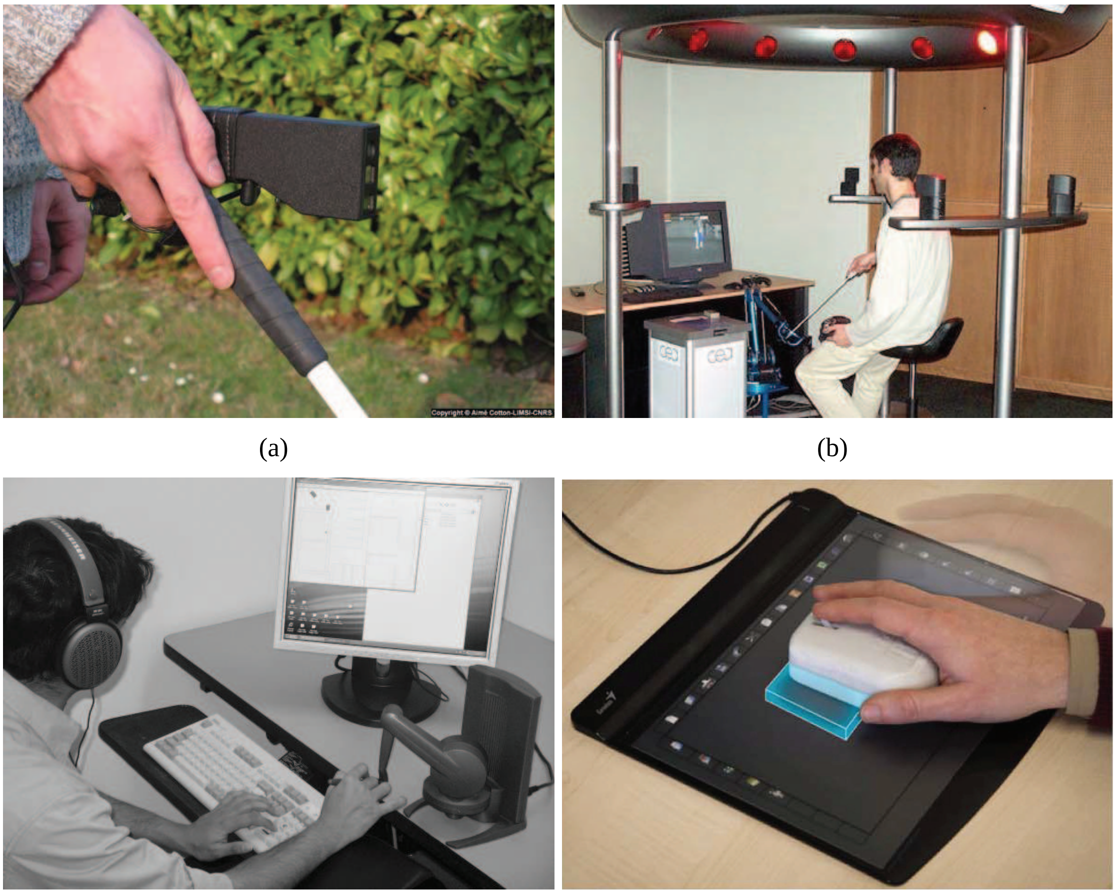
7.2 Orientation and mobility aids

Sighted people predominantly rely on vision to perform most of their everyday tasks, and most technological devices convey visual information as prominent and essential feature. Thus, easy and natural tasks for sighted people become challenging and nearly impossible for blind users due to design issues that are inadequate for their specific needs.

A simple classification of tactile displays for sensory substitution separates pictorial from encoded rendering approaches (Wall and Brewster, 2006). The first methodology consists in translating images/objects to high fidelity tactile ones, while the latter is based on a more abstract representation guided by affordances and limitations of skin as a communication medium, and by technological constraints imposed by design (e.g. Braille language).

In the domain of O&M activities, the most used navigation aid for visually impaired people is the white cane (see Fig. 7.1(a) for an augmented white cane). This tool is simple, lightweight and exhibits the following features:

- users have to be immersed in the exploration space (sensorimotor components);
- it provides mainly local information, collected by users through a continuous interaction with the environment (haptic modality);



(a) (b)

(c) (d)

Figure 7.1: *Examples of orientation & mobility aids: (a) Tom Pouce device - Farcy et al. 2006, (b) HOMERE system - Lecuyer et al. 2003, (c) BlindAid system - Schloerb et al. 2010, (d) DIGEYE system - Brayda et al. 2013.*

- global information emerge implicitly in the propagation of continuous impact sounds (auditory modality).

Tactile-vision substitution systems were conceptually the first to be developed, thanks to their low invasiveness and relatively low costs (Bujnowski et al., 2008). Several alternative technological solutions have been proposed over time, although affected by counterproductive side effects, e.g. excessive spatial dimensions, e.g. OPtical to TActile CONverter (Optacon) device (see Schiff and Foulke (1982) for an historical review), and suffering from interferences between sensory modalities, mostly at the expense of audition (Farcy et al., 2006). Nevertheless, multimodal virtual reality systems have increased in computational and expressive power during the last two decades; an overview of existing works related to O&M aids might help understanding which achievements have been reached and which challenges still need to be faced.

The *Haptic and audiO Multimodality to Explore and Recognize the Environment - HOMERE* system for virtual map exploration (Lecuyer et al., 2003) replicates a virtual white cane used during exploration tasks (Fig. 7.1(b)). This system is an example of passive aid because of its limitation to constrained and predefined paths. It also requires a large and expensive installation with an audio surround system.

Lahav and Mioduser (2004) performed some preliminary studies on map exploration by blind subjects in order to extract the main features required for the development of haptic virtual environments. Their results showed a reduction in the exploration time for participants who mastered the navigation in the unknown virtual space compared to the control group who explored a real space. A significant variety of exploration strategies, methods, and processes was reported by participants using this multi-sensory virtual environment, in comparison with those exploring real space. In a late study, the same authors presented a haptic-based multi-sensory virtual environment improving spatial knowledge of unknown spaces in blinds (Lahav and Mioduser, 2008).

A more complete and complex desktop virtual reality environment, depicted in Fig. 7.1(c), is *BlindAid* (Schloerb et al., 2010) based on a Phantom®[3] device for haptic rendering. The system has been designed to enhance exploration of unknown environments by visually impaired people; actions and commands are borrowed from standard human-computer interaction commands, such as zooming, scrolling, pausing, undoing, etc. *BlindAid* indeed intends to accelerate user's exploration and spatial learning process. However, there is no evidence nor evaluation about users management of such a palette of commands, simulated haptic textures and spatial audio rendered with generic HRTF (KEMAR). In this regard, the haptic system has complex behavioral requirements that eventually lead to a long training process.

The contribution of proprioception to active exploration of objects guided Lederman and Klatzky (1987) to perform two experiments to assess connections between hand movements and haptic exploration in blindfolded subjects. They argued that free haptic exploration of an object together with hand movements, are optimal or even necessary and are extremely efficient in 3D space. Nevertheless, haptic exploration needs to be further supported to achieve the perception of complex spatial layouts and structures, e.g. presentation of 2D environments with height information (Magee and Kennedy, 1980). In this direction Yu and Brewster (2002) noted that spatial perception and proprioception are helpful cues while exploring data non visually.

Walker and Lindsay (2006) examined the effects of nonspeech auditory beacons in navigation performances using sounds that change their timbre and position with different levels of subject's practice. The learning experience of the system improved both speed and accuracy in navigation performances. The system developed by Afonso et al. (2005, 2010) exploited a $3D$ audio virtual environment to investigate structural properties of spatial representations in visually impaired people. Auralization through headphones accompanied user's free locomotion in a real room and in a immersive virtual reality in which six landmarks were placed on the horizontal plane. Their virtual environment effectively drives the elaboration of spatial mental representations, with and without movements. These studies encourage the use of reliable binaural spatial audio as an essential element in the development of multimodal environments.

The technological solution described in this Chapter follows the idea of the "virtual tactile

[3] A 6-DOF position input/3-DOF force output device with a stylus grip. Phantom's website: http://www.sensable.com

tablet”, Vanderheiden (1989). In this approach, tactile displays and touch sensitive tablets allow users to acquire haptic information connected to interaction points on the working area of the tablet. A minimalist haptic feedback was chosen, the TActile MOuse(*TAMO*) device (Brayda et al., 2013) that provides a minimum tactile stimulus on a mouse-like interaction metaphor (see Fig. 7.1(d) and Sec. 7.3.1 for more details). Several motivations lie behind this choice:

- while most of the haptic devices (e.g, the Phantom)are expensive (Wall and Brewster, 2006), the TAMO provides a low-budget solution for visually impaired users;
- combination of haptics and active exploration appears adequate for th recognition and reconstruction of simple geometries;
- minimalist feedback guarantees a rapid learning curve and has relevant practical implications for end-users.

7.3 Audio-haptic exploration

In this Section, a multimodal virtual environment is evaluated in the exploration of simple virtual spatial maps. The system makes use of 2D/3D sounds and the TAMO device.

The haptic navigation metaphor is very similar to tactile sensing on a fingertip combined by kinesthetic exploration cues. Local haptic information is integrated with the spatial audio engine of Ch. 5, and two types of auditory feedback are defined in order to render spatial sounds with respect to listener’s position using HRTFs: spatialization on (i) a bi-dimensional space (i.e. azimuth and distance), and on (ii) a three-dimensional space (i.e azimuth, distance and elevation).

Three exploration experiments evaluate the proposed multimodal system with blindfolded subjects:

1. goal reaching;
2. object recognition;
3. spatial map reconstruction.

Experiments were conducted in order to determine mono-modal contributions and cross-modal augmentation in the exploration of simple virtual maps, e.g. a rectangular space containing one virtual object (Fig. 7.2 depicts a top view of an example map). In all the experiments, subjects were required to wear an eye-mask and were only provided haptic and auditory feedback, through the TAMO and headphones, respectively.

Participants

Eleven subjects (7 males and 4 females) whose age varied from 21 to 40 (mean 28.2, SD 5.5), took part to the three experiments. All subjects reported normal hearing according to the adaptive maximum likelihood procedure proposed by Green (1993). They had different levels of expertise in psychophysical experiments: only two subjects had previous experience with multimodal experiments, all the others were naive subjects.

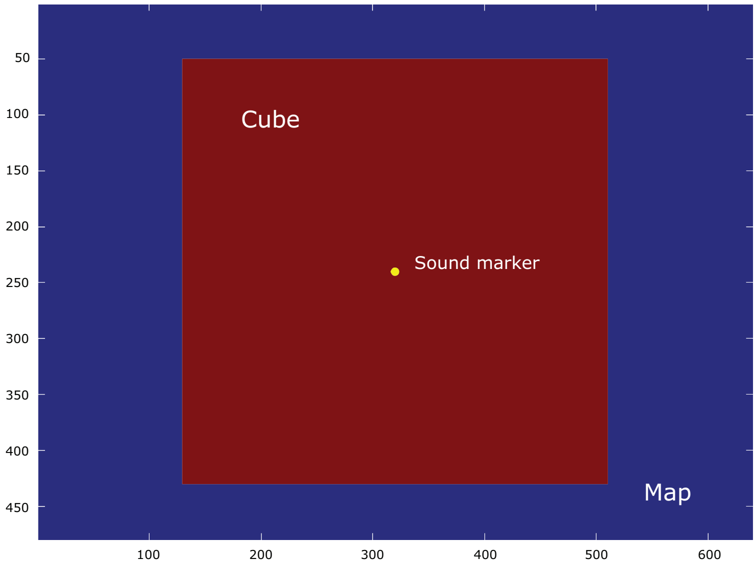


Figure 7.2: *Top view of a simple multimodal map.*

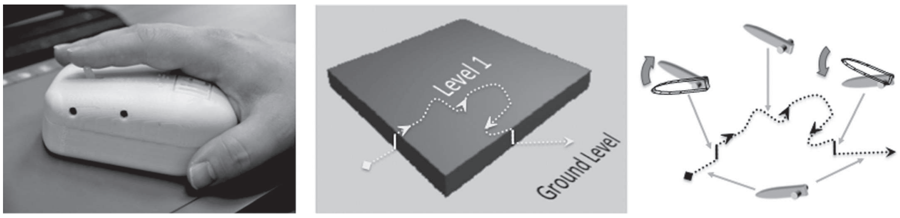


Figure 7.3: *The TAMO device (on the left), a single height virtual map (center), and movements of the TAMO's lever according to time-variant position on the map (right) (Campus et al., 2012).*

7.3.1 Apparatus

The three exploration experiments were performed in a Sound Station Pro 45 silent booth. Sennheiser HDA 200[4] headphones were plugged to a Roland Edirol AudioCapture UA-101 external audio card working at 44.1 kHz sampling rate. For each subject, HRTF selection was performed in the CIPIC database using the selection procedure described n Sec. 4.3, in order to optimize spatial impression especially for elevation The acquisition of one pinna image of each subject was required to compute the mismatch between manually traced contours and notch central frequencies of CIPIC HRTFs (with reference to Eq. (4.19), the weight values were $w_1 = 1$, $w_4 = w_3 = 0$).

In the experiments, the same GUI designed in MATLAB which includes menu and control buttons for direct management of the experimental conditions and timing. The *DIGEYE* System (Chellali et al., 2009) provided a complete test environment for haptic experiments and

[4] These dynamic closed circumaural headphones offer an effective passive ambient noise attenuation and high-definition reproduction of high frequencies.

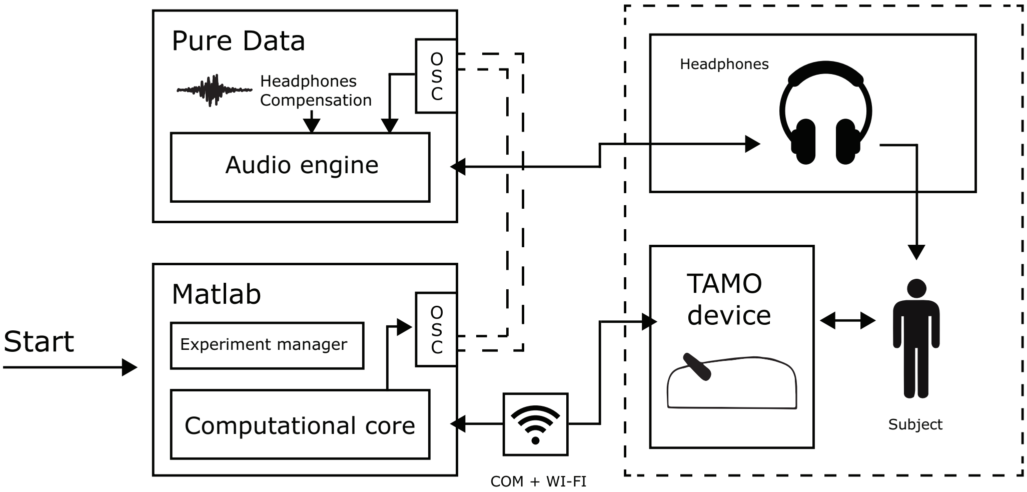


Figure 7.4: *A simplified scheme of the experiment setup (Bedin, 2013).*

some of its components (hardware and communication modules) were integrated in this new experimental interface. See Fig. 7.4 for a schematic view.

The TAMO device was developed for the *DIGEYE* System. It is capable to render height information related to any absolute position in its workspace, i.e. the supporting tablet where it moves (Brayda et al., 2011). The height information is provided by a lever, placed above the mouse-shaped device, at the location where the mouse-wheel is commonly placed. The user keeps a fingertip in contact with the lever and a stepper-motor raises it proportionally to the virtual height to be rendered (an example is depicted in Fig. 7.3).

The developed multimodal virtual environment is able to haptically render the presence of virtual objects by moving the TAMO's stepper (lever) and creating a tactile contact on subject's fingertip: when the pointer reaches a virtual object on the map, the stepper signals a virtual edge of a given height. TAMO generates a taxel for each pixel of the working area, i.e. a 210×297-mm sensing table, similarly to a tactile bas-relief representation. The lever moves from ground horizontal position $\phi_0 \approx 0°$, to a nearly vertical position, corresponding to a rotation of $\phi_{max} \approx 80°$ (more details are available in (Brayda et al., 2011)). This haptic minimalist feedback corresponding to a single movement in elevation is called *taxel*: information related to a single tactile unit are stored and rendered (Siegel, 2002) like pixels in vision. Previous studies showed that the just noticeable difference in height for the TAMO is $JND \in [3.5, 4.3]$ degrees (Brayda et al., 2013).

The TAMO already obtained several encouraging results: subjects properly identified simple geometric objects and navigated in virtual maps of small rooms, avoiding obstacles (Chellali et al., 2009). Furthermore, in recent investigations Brayda et al. (2013) discussed on which combination of factors among cognitive load, absence/presence of motivation and random/identifiable strategy, affects map construction quality.

7.3.2 Stimuli

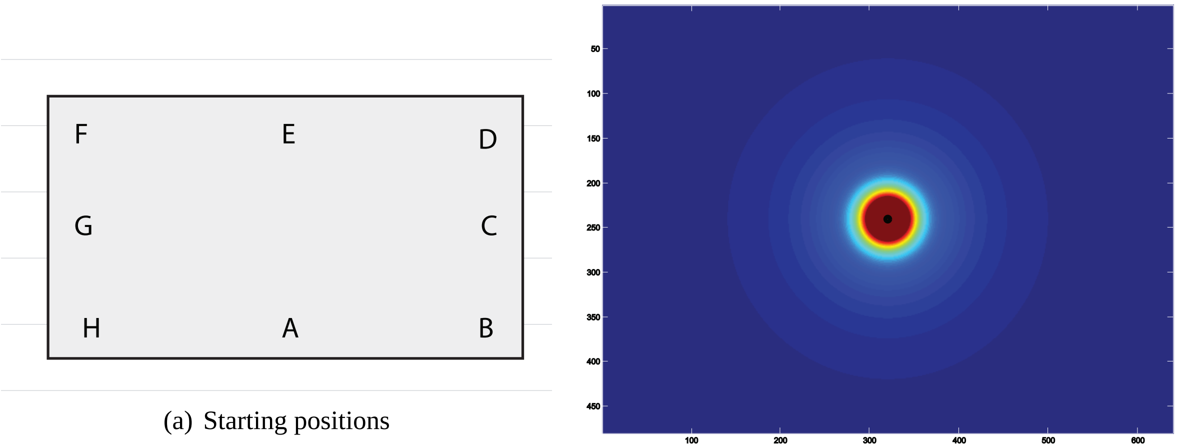
All maps were surrounded by virtual walls rendered with ϕ_{max} on the TAMO. When the device moved outside the working area the lever moved alternatively from ϕ_{max} to $\phi_{max} - 26°$ at refresh rate in order to signal that the subject crossed the boundaries delimited by virtual walls.

Depending on the feedback condition, a virtual sound source (beacon sound) was placed at the center of the map. The sound was spatially rendered by headphones according to the relative position between pointer and beacon sound's position. This rendering approach corresponds to an egocentric view of the virtual map in which the pointer corresponds to the listener's head. When the device exceeded wall boundaries the auditory feedback was stopped. Each auditory stimulus was a continuous train of repeated 40-ms Gaussian noise bursts with 30 ms of silence between each burst. A similar type of stimulus was employed in localization tasks and has already been proved to be more effective than a single white noise burst (Katz et al., 2012), although this latter stimulus is also used in navigation tasks (Walker and Lindsay, 2006). The maximum measured amplitude of the raw stimulus at the entrance of the ear canal was set to 60 dB(A). At the beginning of the experimental session, subjects could adjust the default sound volume in order to obtain a subjectively comfortable level.

Auditory stimuli were filtered through the selected HRTF set and a headphone compensation filter obtained with the algorithm presented by Lindau and Brinkmann (2012), applied to measured headphone responses on a KEMAR mannequin without pinnae. It has to be highlighted that compensation was not individual; however, such kind of processing guaranteed effective equalization of the headphones up to $8 - 10$ kHz on average and simulated a realistic application scenario where it is not always feasible to design personal compensation filters. Rendering through selected CIPIC HRTFs provided two dimensions (azimuth and elevation), of the auditory space. The third dimension (distance cue) was rendered through an inverse square law on sound attenuation level. The sound level decayed from a maximum sound pressure level when the pointer covered approximately the sound source position, to a minimum audible sound level set to the farthest reachable position along tablet borders. No near field calculation were performed, [5] and instead a 25 px neighborhood around the beacon sound was defined in which the auditory feedback remained constant (i.e., frontal direction with azimuth $\theta = 0$, and sound intensity at its maximum level).

Since the multimodal virtual environment exploited different and concurrent software/hardware for each modality, it was crucial to satisfy real-time constraints and to synchronize auditory and haptic stimuli within a coherent perceptual integration time window. In order to choose a unique refresh rate for the rendering process, the delay between audio and tactile stimuli was measured as follows: two condenser microphones connected to a Tascam 680 at 192 kHz sampling rate were placed at the headphones coupler and near the stepper, and latency was estimated as the time between the activation of the TAMO stepper (detected by means of the noise of the

[5] There exist several works proposing synthesized near-field HRTFs from far-field HRTFs (Kan et al., 2009) and it has to be proven their efficiency in such kind of exploration tasks compared to different sound decay laws (see (McMullen and Wakefield, 2013) for a preliminary study), adding new acoustic cues for distance perception (Parseihian et al., 2012) or mapping perceived and physical distance with physics-based models of an acoustic pipe (Devallez et al., 2008).



(a) Starting positions

(b) Virtual map

Figure 7.5: *Experiment #1. (a) Starting positions for the audio-haptic exploration. (b) The virtual maps: haptic elevations are visualized through a color map, starting from blue (ϕ_0) and reaching till dark red (ϕ_{max}) in proximity of the goal (central black dot).*

TAMO engine) and the audio output at the earphones. Based on these measurements, a refresh rate of 80 ms was chosen. This value is larger than the measured haptic-audio delay (68 ms) and therefore guarantees a consistent refresh. At the same time, it guarantees continuous integration of multimodal perception because neural information from different senses occur at approximately the same time, thus being associated with the same physical event (Holmes and Spence, 2005). Since signals coming from different sensory modalities have different time-of-arrivals and processing time in given brain area, a temporal window of about 200 ms ensures multisensory integration and enhancement (Meredith et al., 1987).

7.3.3 Experiment #1: goal reaching

This experiment has a twofold objective: (i) assessing horizontal auditory localization performances using the selected HRTF set in dynamic conditions and (ii) comparing performances with uni-modal, and bi-modal (audio-haptic) feedback.

Procedure

A brief tutorial section introduced the experiment. The subject was verbally informed that he/she had to explore a virtual map by means of a pair of headphones and a tactile mouse. At the same time, the exploration metaphor of the TAMO device was described; then, the subject was instructed to keep an egocentric view of the virtual world and (i) to reach a spatialized sounding object, (ii) the top of an haptic virtual object or (iii) a virtual object that exhibits both features as fast as he/she could. During the exploration task, the subject was blindfolded and had to keep a fixed head orientation. This indication helped the subject to keep a coherent auditory rendering, being his/her head the center of the 2D coordinate system for spatial audio. The experimenter guided the blindfolded subject to ehe experimental location and subsequently led him/her towards the starting position of each trial.

Figure 7.5(a) depicts the eight points at the workspace boundary from which each trial began. Three feedbak conditions were provided:

- TAMO: uni-modal haptic condition;
- 2D audio: uni-madal auditory condition;
- TAMO + 2D audio: bi-modal condition.

A spatialized sound source was placed at the center of the map, as depicted by the black dot in Fig. 7.5(b). The subject had to enter a 25 px neighborhood around it and to remain inside for 1.2 seconds in order to complete the trial. No height information was provided by the auditory feedback, i.e. only 2D information was mapped to azimuth and distance rendering.

On the other hand, gradually increasing haptic elevations were rendered the TAMO as subjects approached to the center. An inverse square law described the trend as depicted by the color scale in Fig. 7.5(b).

All the three feedback conditions were repeated 8 times (one for each starting position) leading to a total of 24 stimuli. Three sequences of presentation were built for the 24 stimuli and used in rotation. Since feedback conditions were randomized across trials, each sequence was composed by three 8-trial blocks where each starting position appeared once. Blocks were arranged in a 3×3 latin-square design across sequence and each block followed one of the first three row of a 8×8 balance latin-square.[6] This approach minimized motor memory effects by avoiding successive occurrences of the same starting point.

Results

Experimental results were preliminarily evaluated in terms of absolute reaching time, i.e. the time spent by the subject to reach the virtual object located in the center of the map. The average improvement between the unimodal haptic condition and the bimodal condition was 27% ± 14%, with a peak of 59% and a minimum improvement of 11%. The average improvement between unimodal auditory condition and the bimodal condition was 22%. The average standard deviation of the absolute reaching time in the bimodal condition was 1.57 s, whereas unimodal feedbacks resulted in 3.04 s and 1.57 s for audio and tactile conditions, respectively.

As shown in Fig. 7.6, the 2D audio condition usually performed better than the TAMO condition. However the adopted inverse square law for haptic height could have a determinant role. Different laws could provide better tactile performance and further studies are required in order to assess the supremacy of 2D audio.

These results confirm an effective multimodal integration between the TAMO and 2D spatial audio. The auditory feedback provided a constant global information about pointer's position on

[6] The 8×8 balances latin square was computed according to the lexicographic order defined in Fig. 7.5(a)

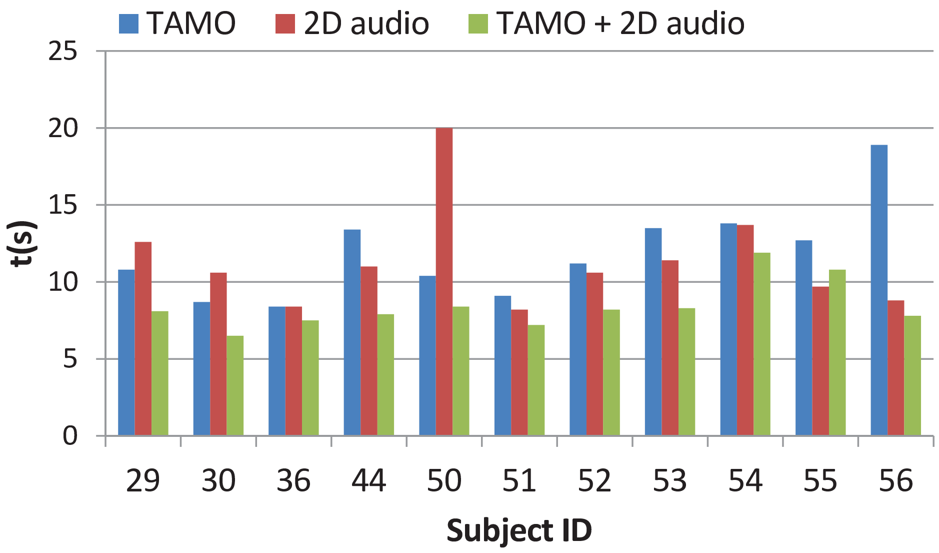


Figure 7.6: *Results for experiment #1, absolute reaching time for each subject.*

the map, while the haptic feedback provided just a local information. Their combination helped subjects to reach the goal. This observation has confirmed by the following informal observation: while in the unimodal haptic condition subjects were initially slow and could easily complete the task once close to the center, in the unimodal auditory condition they were initially faster but had difficulties in understanding that the pointer was above the marker.

Additional and more rigorous analysis has to be conducted on the collected data, using different metrics such as deviation from the optimal path and number of changes in direction.

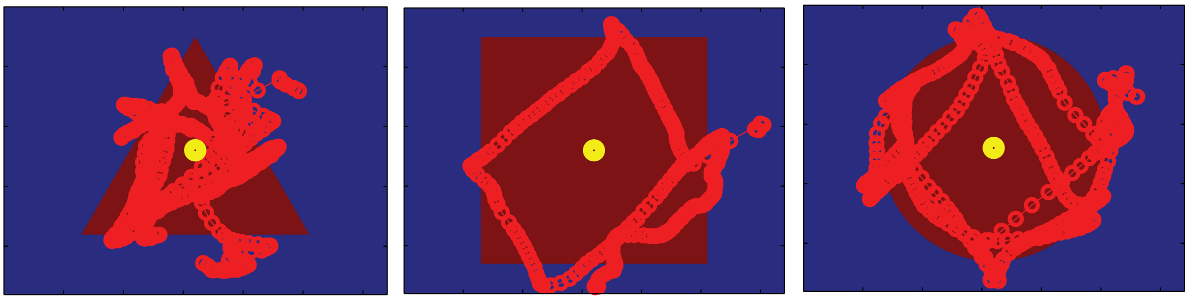
7.3.4 Experiment #2: object recognition

This experiment investigates exploration strategies in the recognition of virtual objects with basic geometric shapes, depending on the provided feedback. Similar tests were conducted by Chellali et al. (2009) in order to qualitatively describe the main strategies in reconstruction and recognition of virtual objects and rooms.

Subjects were asked to explore the virtual map using a TAMO device and 2D audio; they had to recognize simple geometries as quickly as possible. Objects were placed in the center of the map and a spatialized beacon sound was synthesized as global orientation cue.

Procedure

This experiment was performed by the same subjects after the previous one (Sec. 7.3.3) and shared most of the verbal description provided before. Each trial was completed when subjects verbally identified the object or after a maximum allowed amount of time (set to 150 s); during the exploration time, they could guess, but no feedback was provided by the experimenter



(a) Triangle.

(b) Square.

(c) Circle.

Figure 7.7: *Experiment #2. The three virtual objects to be recognized. Three trials of Subject 56 make use of bimodal feedback. Trajectories and exploration strategies adopted by Subject 56 are shown in red.*

until the right answer was given. If the subjects were not able to identify the object within the maximum allowed time, the trial was concluded with a negative outcome.

The basic virtual objects were:

- Parallelepiped with triangular base, Fig. 7.7(a);
- Parallelepiped with square base, Fig. 7.7(b);
- Cylinder,, Fig. 7.7(c).

Subjects always started from position “A” in Fig. 7.5(a). Object sizes were set according to previous study conducted by Brayda et al. (2010), yielding a TAMO movement range spanning from ≈ 10 m to 15 cm on the virtual object. With these constraints, object shapes had areas larger then 50% of the workspace (e.g. a cube of edge equals to 380 px was created). Finally, object heights were set to $\phi_0 + 18°$.

Each of the three objects was presented in two conditions:

- TAMO: unimodal condition;
- TAMO + 2D audio: bimodal condition.

As a result, six stimuli were constructed. Presentation sequences were arranged in latin square order with respect to object shape. Feedback condition were presented alternatively in order to minimize learning effects related to the use of 2D audio.

Results

A preliminary analysis of the results revealed an average recognition time which of 68.6 s in the TAMO condition and of 64.1 s in the bimodal condition, exhibiting not relevant differences and high standard deviation (36 s). Moreover, failed recognitions were sporadic.

Qualitative assessments of exploration strategies showed the same two main approaches found by Brayda et al. (2011):

- exploring the map on a grid and moving on vertical and horizontal lines;
- following the object contours trying to cross the edges orthogonally or diagonally.

These results suggest a weak role of the designed auditory feedback in the recognition task. However, it is still interesting to analyze micromotion (small precise movements) and macromotion (macroscopic quick changes on the map) of subject movements.[7] Further analysis could correlate those movement features to changes in feedback and strategies.

7.3.5 Experiment #3: spatial map reconstruction

The third preliminary experiment focuses on map reconstruction. In a recent work by Picinali et al. (2014), subjects explored a complex virtual environment, based on a real map, by means of an 3D audio engine which shares similar desirable features with the one designed in Ch. 5. Therefore, subject were asked to physically reconstruct the map using LEGO® bricks and annotated drawings.

A similar approach was adopted here on a simplified scenario: a single virtual cube was positioned in the map and subjects had to estimate its size and location. Particular attention was given to auditory feedback contribution in terms of dimensionality in spatial information, i.e. 2D and 3D localization cues.

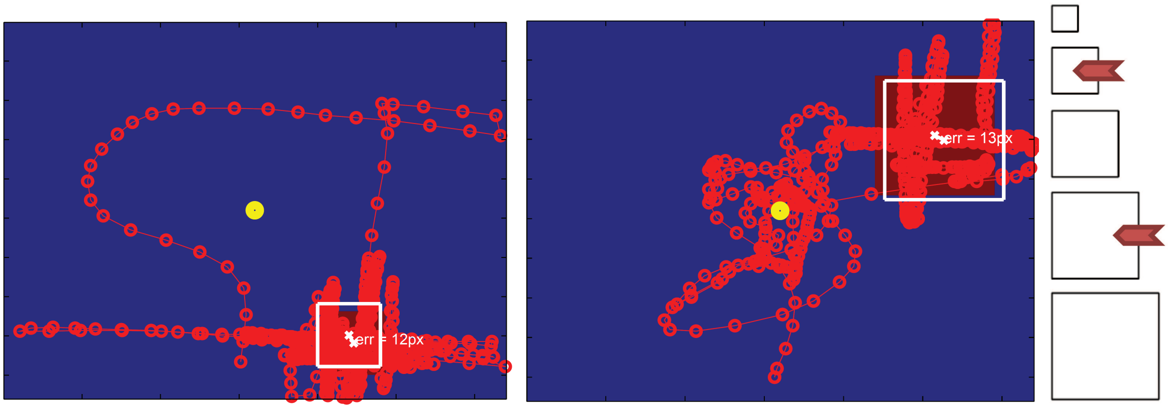
Procedure

The task was introduced to the subjects based upon their experiences in the previous experiments. Subjects were informed about the presence of a virtual cube on the map and that they had to memorize its size and location. They were also informed that the cube randomly changed and its size was different among trials. Subjects had one minute to explore the map. Afterward they had to reconstruct the explored map by picking one among a set of physical cubes, which were different in size, and to place it at the estimated location on a 210×297-mm paper sheet, which was next to their exploration position.

Feedback conditions are summarized as follow:

- TAMO: unimodal haptic condition as described in Sec. 7.3.4;
- TAMO + 2D audio: bimodal exploration with haptic feedback and beacon sound rendered according to 2D coordinates;
- TAMO + 3D audio: bimodal exploration with haptic feedback and beacon sound rendered according to 3D coordinates;

[7] This distinction was originally introduced on visuo-tactile object exploration by Zinchenko and Lomov (1960): micromotion is performed to maintain a constant level of stimulation on the tactile receptors and macromotion to manipulate and investigate object’s features.



(a) Cube - Size 1. (b) Cube - Size 2 (c) Sizes

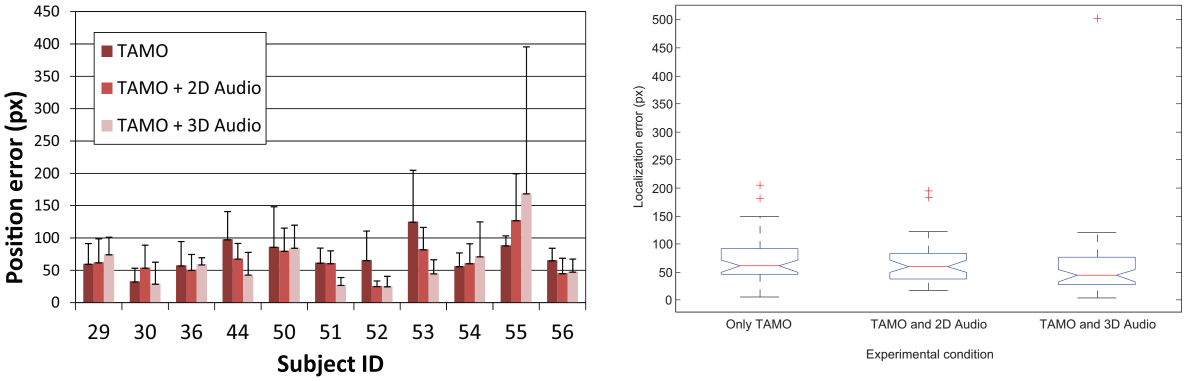
Figure 7.8: *Experiment #3. Two trials of Subject 30: (a) the cube with small (45 px) edge and (b) the cube with large (80 px) edge. (c) Available cube sizes for subject estimates. Subject trajectories are shown in red. The object location errors in subject estimates (12 px and 13 px, respectively) are also shown.*

Each feedback condition was associated to 2 cubes of different size, i.e. small edge (80 px) and large edge (150 px), with two repetitions, for a total of 12 stimuli arranged in latin square order with respect to feedback condition. Although only two cube sizes were used in the stimuli, when giving their estimates subjects had to choose between the following 5 sizes: 45 px, 80 px, 115 px, 150 px, 185 px (see Fig.. 7.8(c)).

Results

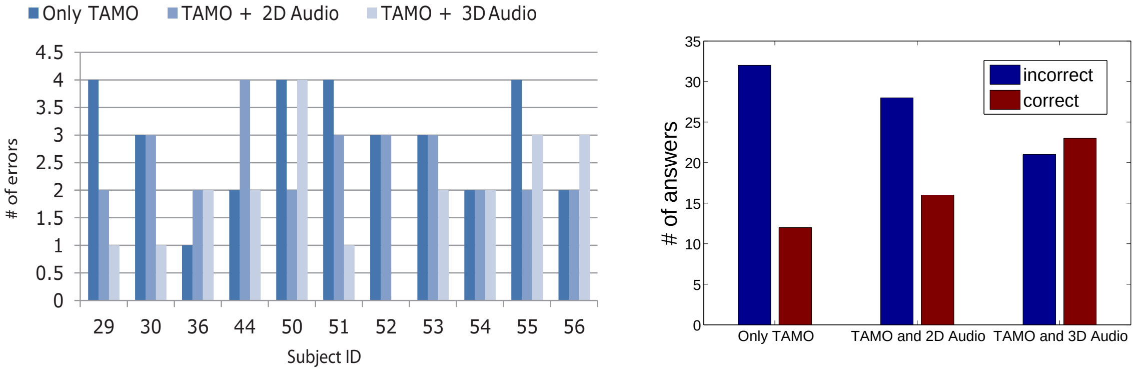
Figures 7.8(a) and (b) show the best trials of Subject 30: location errors are very small and estimated sizes are both correct. Figure 7.9(a) shows the average results divided by subject and grouped by experimental condition for object location errors. The average improvement between TAMO and TAMO + 2D audio was 5.9% and raised to 15.0% for the bimodal condition with 3D audio. Error data were subjected to a Kruskal Wallis nonparametric one-way ANOVA with three level of feedback condition, the test did not reach a complete significance level ($p = 0.0664 > 0.05$). A post-hoc Wilcoxon test would not be meaningful due to non significance of the ANOVA. However, a qualitatively analysis in Fig. 7.9(b) indicated that the position error progressively decreased between conditions.

Size estimation was investigated by performing a Pearson's Chi-Square test of independence of incorrect/correct answers on feedback conditions. The main effect of feedback condition was significant ($p = 0.05$). In particular, Fig. 7.10 visualizes the trends of incorrect/correct answers, and the number of size errors in the bimodal condition with 3D audio decreased of 35% with respect to unimodal condition. Furthermore, informal comments revealed that subjects were not aware of the presence of 2D or 3D audio feedback and did not consciously note the difference



(a) (b)

Figure 7.9: *Results for experiment #3, object location error: (a) mean and standard deviation of position error for each subject grouped by feedback condition; (b) global statistics on position error grouped by feedback condition.*



(a) (b)

Figure 7.10: *Results for experiment #3, size error: (a) number of errors in size estimation for each subject grouped by feedback condition; (b) global count of correct/incorrect size estimations grouped by feedback condition.*

between the two audio modalities. [8]

It should also be mentioned that subjects usually tended to underestimate the size of the cube: the smallest proposed size (45 px) was chosen 30 times, while the largest size (185 px) was chosen only 9 times on a total of 132 trials. This effect could be related to the presence of a scaling factor in subject spatial representation with the TAMO, i.e. the minimum physical movement of the TAMO device corresponds to a larger movement on the virtual map Brayda et al. (2010).

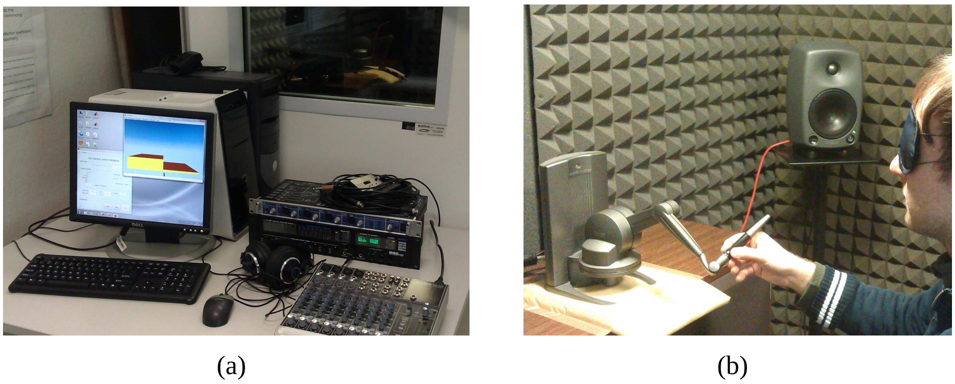
[8] They often reported virtual objects as 2D figures, i.e. “squares”, instead of 3D figures, i.e. “cubes”.

7.4 Audio-haptic perception of height

Chapter 6 has introduced the main concepts related to multisensory integration and combination. Research in multimodal perception provides the ground for the design of multimodal interfaces and virtual environments. It has been long recognised that properly designed and synchronized haptic and auditory displays can provide greater immersion in a virtual environment than a high-fidelity visual display alone (Hahn et al., 1998; Srinivasan and Basdogan, 1997).

Modulation across senses can also occur when the information carried by two senses is semanticaly coherent. As an example, musical pitch is often classified as “high” and “low”, i.e., with intimately spatial terms. Rusconi *et al.* (Rusconi et al., 2005, 2006) showed that this spatial connotation interacts with motor action so that, when we are asked to respond quickly whether a pitch is high or low in comparison to a reference, we are faster if the response is coherent with the spatial position of the response key (e.g., the response is “high” and the response key is in the upper part of the keyboard), rather than viceversa.

Empirical results on this type of interaction have led researchers to hypothesize that the representations of heterogeneous continua share a common nucleus. As an example, according to the ATOM’s theory (Walsh, 2003) the representation of space, time and number is processed by a common mechanism. Other authors, in contrast, suggest that some representations (i.e., time and numbers) are spatially mapped (Dehaene et al., 1993; Ishihara et al., 2008). At any rate, the work reported in (Rusconi et al., 2005, 2006) shows that musical pitch interacts with non-auditory continua such as motor space. Here, it was investigated whether a tone’s pitch (i.e., a stimulus subtly subtending a spatial representation) can influence a robust perception such as the haptic estimate of the height of an object in absence of vision. In the described experiment, blindfolded subjects explored a virtual 3D object by means of a haptic device and and had to return a verbal estimate of object’s height. Haptic exploration was accompanied with a continuous sound, a sinusoid whose pitch varied as a function of the interaction point’s height within two ranges. Experimental results, discussed in Sec. 7.4.4 show that the information carried by the frequency sweep (i.e., larger sweep, larger tonal space, therefore larger “space”) can modulate only to a limited extent a robust and inherently spatial perception such as the haptic perception.



(a) (b)

Figure 7.11: *Experiment manager station (a) and subject performing a trial (b).*

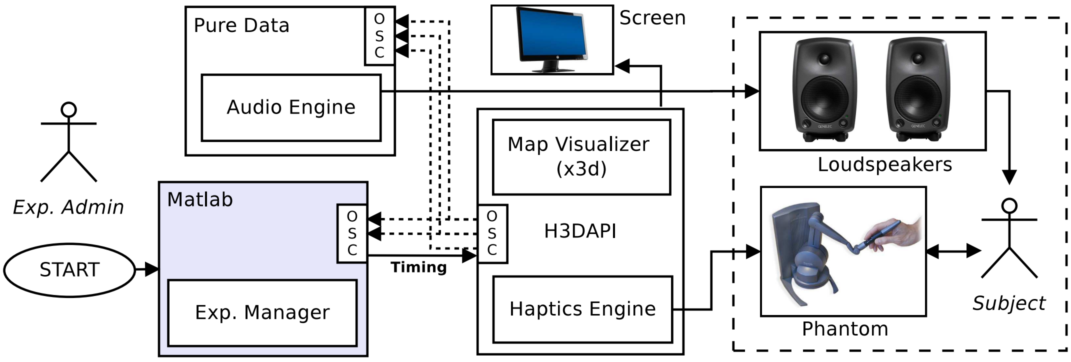


Figure 7.12: *A simplified scheme of the experimental setup.*

7.4.1 Apparatus

The experiment was carried out in a silent booth. The experimental setup is depicted in Fig. 7.11. A computer acted as the control unit and was connected to a Sensable PHANTOM Desktop haptic device and to a Motu 896mk3 sound card, which transmitted the auditory feedback to two loudspeakers (Genelec 8030A). The graphical user interface was implemented in MATLAB and is part of the framework described in Sec. 5.2. The developed software setup is outlined in Fig. 7.12. In order to build a extensible interactive multimodal virtual environment, the setup integrated H3DAPI[9] (an open source platform using OpenGL and X3D[10] with haptics in one unified scene graph) and Pure Data[11] (an open source real-time environment for audio processing). Communication was managed through Open Sound Control (OSC).

7.4.2 Stimuli

Test subjects were asked to haptically estimate the heights $h_i = 2 \cdot i$ $(i = 1 \ldots 5)$ of five virtual stair-steps. To guarantee a sufficient workspace, all the stair-steps spanned a $22 \times 22 = 484$ cm^2 horizontal square area (see Fig. 7.13) The step riser lay in the yz-plane of the virtual scene and in the midsagittal plane related to subject posture. The upper tread of the stair-step lay at the left or right side of the yz-plane, for a right-handed or a left-handed subject respectively. Normalized static and dynamic friction coefficients were set to 0.1 and 0.4 respectively. The normalized stiffness was set to 1 in order to render an impenetrable virtual stair-step without causing the device to become unstable.

Upon collision of the cursor with the stair-step, an auditory feedback was produced. The y coordinate of the *haptic interaction point* (HIP) was mapped to the frequency of a sine wave. Therefore, continuous interaction with the surface produced a dynamic sine sweep. The mapping was defined through three parameters: f_{min}, the lowest frequency associated to ground level; Δf, the frequency range spanned above f_{min} (thus the maximum frequency associated with the upper

[9] http://www.h3dapi.org/
[10] http://www.web3d.org/x3d/
[11] http://puredata.info/

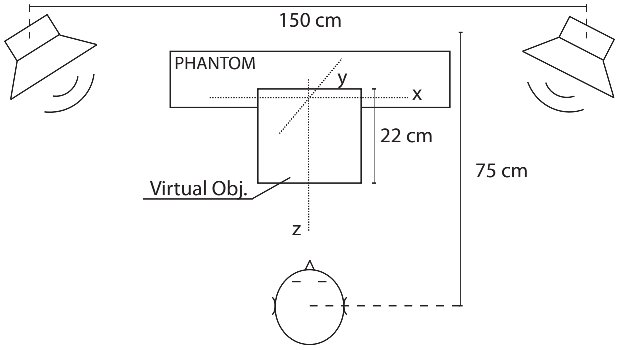


Figure 7.13: *A schematic representation of the experiment spatial arrangement.*

tread is $f_{max} = f_{min} + \Delta f$); the *mapping strategy*, i.e. the function that maps HIP to the frequency domain within the prescribed range.

In this work, $f_{min} = 200$ Hz and a linear *mapping strategy* were chosen. Thus, for the height h_i the fundamental frequency f of the sine sweep was:

$$f = f_{min} + \Delta f \frac{y}{h_i}. \tag{7.1}$$

The only varying parameter was Δf, which took the two values f_{min} and $3f_{min}$. These resulted in f_{max} values that were one and two octaves above f_{min}, respectively.

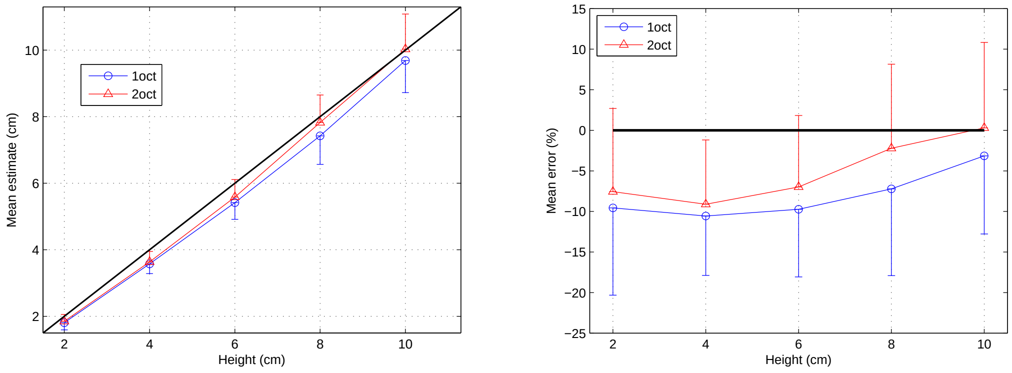
Along the x HIP coordinate, a simple linear panning approach spatially placed the auditory feedback between two loudspeakers in order to improve spatial orientation in the virtual scene (Stamm et al., 2011) and the localization of the step riser. The gains for the left and right channel were $G_{l,r} = \frac{1}{2}(1 \pm P)$, where $P \in [-1, +1]$ corresponds to the horizontal panning position, i.e. $P = \pm 1$ at the left/right loudspeaker positions, and $P = 0$ at the step riser. Levels were set so as to produce 70 dB SPL at the approximate listener head position.

The choice of sine waves was coherent with the initial research question aimed to quantify pitch interaction with haptic size estimation.

7.4.3 Procedure

Participants were informed about the use of the stylus for exploring objects, and no indication about their shape was provided. They were led to believe that they explored real, physical objects, and no mention of the haptic device was made. They were blindfolded before entering the silent booth and guided to the experiment-ready position (see Fig. 7.13).

The 10 stimuli (5 heights × 2 Δf values) were presented with 4 repetitions. The order of the 40 trials was randomized. Participants were instructed to be accurate to 1 mm in their verbal height estimations and to explore the whole objects surface (including step risers). They were allowed to interact with each stimulus for 10 seconds before answering. No feedback concerning the accuracy of their responses was given. The role of the auditory feedback was not explained or commented. At the end of the trials participants were guided out of the silent booth and asked to answer a questionnaire.



(a) (b)

Figure 7.14: *Mean height estimations (a) and mean relative error of height (b) ± standard deviation, as a function of the real stair-steps height, for the two sine-sweep frequency ranges across 20 subjects. Perfect estimations (black solid lines) lie on the bisector (a) and on the x axis (b).*

Subjects

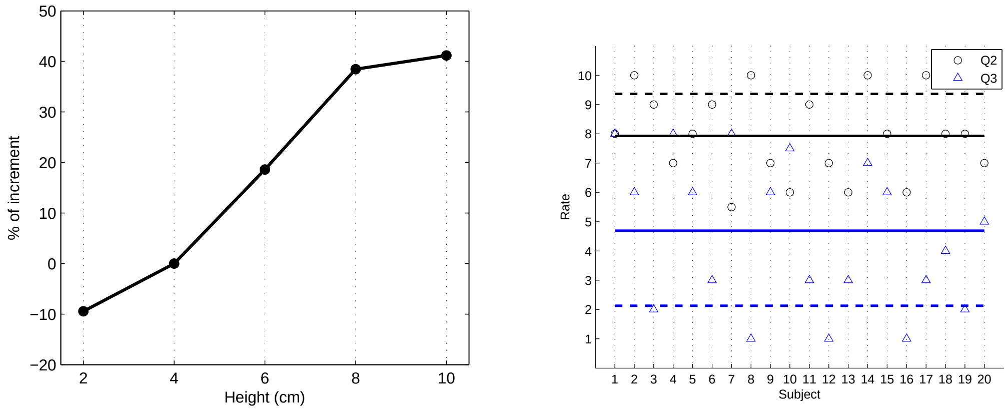
A total of twenty subjects (12 males and 8 females), aged between 20 and 30 (mean = 23, SD = 2.92), caucasian, 18 right-handed and 2 left-handed, took part to the experiment. All participants self-reported normal hearing and no impairment in limb movements. They took, on average, about 35 minutes to complete the experiment. They were students and apprentices of the University of Padova and had no knowledge nor experience of haptic force-feedback devices.

7.4.4 Results

Subject estimates were averaged separately for stair-step height and frequency range (see Fig. 7.14(a)). Two (frequency ranges) by five (stair-step heights) two ways analysis of variance (ANOVA) on the resulting values revealed that subjects' estimates increase as a function of the stair-step $F(4,74) = 66.00$, $p < .001$.

Figure 7.14(a) shows that absolute height estimates were accurate on average. A slight tendency to underestimate the stair-step height was observed for most subjects in all conditions. More importantly, subjects produced larger estimates when the stair-step was accompanied by a larger frequency range sweep. From Fig. 7.14(a) it can be seen that, for each height, averaged estimates were larger when the frequency range spans two octaves. However, the ANOVA showed that this result was not statistically significant: $F(1,19) = 1.79$, $p = .19$.

In order to asses how estimates changed as a function of the height of the stair-step, subjective estimates were transformed in percentages of under- or over-estimation of the stair-step (see Fig. 7.14(b)) and the two-ways ANOVA was recalculated. From Fig. 7.14(b), the effect of the auditory feedback can be appreciated more clearly. The subject's percent error did not change as a function of the stair-step size, $F(4,76) = 0.57$, $p > .05$. However, also in this case the ANOVA



(a) (b)

Figure 7.15: *(a) Increment in cardinality (%) of $\Delta f = 2 - oct$ sub-group related to $\Delta f = 1 - oct$ sub-group per stair-step height across 20 subjects. (b) Subjects' answers to Q2 (black solid line: mean; black dotted line: standard deviation), and Q3 (blue solid line: mean; blue dotted line: standard deviation.*

confirmed the non significant effect of sound: $F(1,19) = 1.58, p = .22$.

Although current results were not statistically significant, a further analysis is provided. For each subject, the mean value among equal-height repetitions defined the *personal perceived audio-haptic height*. The sum of trials that exhibited an estimated height greater than the corresponding reference formed the *personal over-estimation data pool* from which we could identify two sub-groups with respect to their frequency ranges. Figure 7.15(a) depicts the relative increment in cardinality (%) of $\Delta f = 2 - oct$ sub-groups related to $\Delta f = 1 - oct$ sub-groups in dependence of haptic stair-step heights. The resulting curve increased monotonically, especially in $4 - 8$ cm range encouraging further studies, particularly in haptic heights greater than 8 cm where a detectable audio effect occurred.

Finally, three questions from the post-experimental questionnaire are reported and the corresponding answers by subjects are discussed:

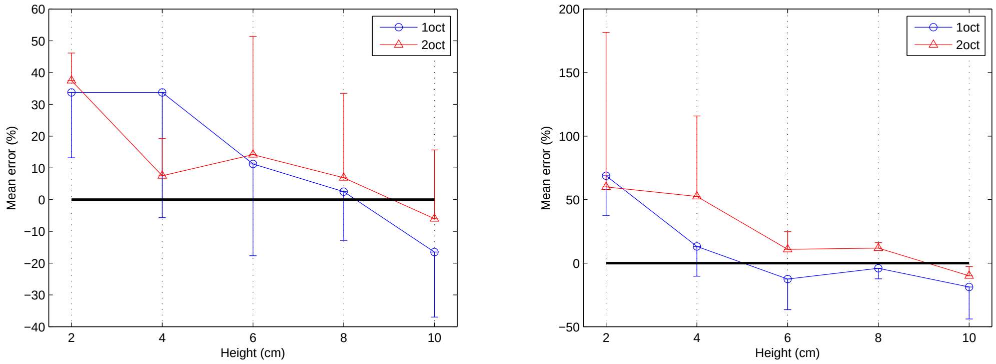
Q1: indicate if the object(s) was (were) real;

Q2: evaluate to which extent the haptic feedback helps your estimates;

Q3: evaluate to which extent the auditory feedback helps your estimates.

The first question was binary evaluated (yes or no). Questions Q2, Q3 were evaluated on a visual analogue scale (VAS) [0 = not at all, 10 = very much].

Interestingly, answers to Q1 revealed that the majority of subjects indicated the stair-steps as real (yes: 12; no: 6; do not know: 2), confirming the high degree of realism in the haptic simulation achievable by the PHANTOM device. Answers to for Q2 and Q3 are summarized in Fig. 7.15(b). It can be seen that most subjects considered the haptic feedback to be very



(a)

(b)

Figure 7.16: *Mean relative error ± standard deviation as a function of stair-steps height, for the two sine-sweep frequency range. (a) Subject 8; (b) Subject 7.*

helpful. On the contrary, subjects on average reported the auditory feedback to be substantially less helpful, although the answers to Q3 exhibited a larger variability. This again supports the experimental results, which show to some extent an effect of the auditory feedback.

In fact, by looking at individual subject performances, it can be seen that the frequency range of the sine-sweep had some effect also on subjects who reported no influence by the auditory feedback (e.g., Subject 8 in Fig. 7.16(a)). Furthermore, subjects who reported to equally take into account both modalities (e.g. Subject 7 in Fig. 7.16(b)) were clearly influenced by the frequency range.

According to current data, the effect was not statistically significant. This result may suggest that haptic information related to object size is extremely robust and that audition has only a limited possibility to modulate size information acquired haptically. This view is also supported by the overall accuracy of subjects in their absolute estimations of step heights.

However, this experiment was conceived as a pilot experiment from which more extensive tests can be designed. Therefore, firm conclusions can be drawn only upon these tests being completed. In particular, upcoming experiments will also measure subjects' performance without auditory feedback, in order to compare unimodal and bimodal conditions.

7.5 General discussion and future perspectives

Preliminary results of the first three experiments show different multimodal integration levels among different tasks in spatial cognition: results for goal reaching (Experiment #1) and map reconstruction (Experiment #3) are promising. On the other hand, the object recognition task (Experiment #2) has provided less clear results.

1. In the *goal reaching* experiment, a bimodal integration occurs as a combination of a global feedback (audio) and a local feedback (haptic). All subjects significantly improve their

performance with multimodal feedback (an average improvement of 27.5% in completion time) suggesting that this feedback is effective in this type of tasks.

2. In the *object recognition* experiment, the presence of 2D audio does not seem to haptic feedback. Further investigations must be then performed, considering different haptic devices and spatialized audio scenes.

3. In the *spatial map reconstruction* experiment, spatial audio feedback (both 2D and 3D) significantly helps navigation: it reduces the localization error and number of size errors (especially in the condition with 3D audio).

A pilot experiment aimed at studying the sonification of exploration tasks has also been presented: in this case the goal was to assess the influence of auditory feedback (and particularly pitch) on the haptic estimation of object size (and particularly height).

The collected data must be further analyzed in order to support future directions. However, the already available outcomes are a starting point for future improvements in the proposed multimodal system. several different directions can be explored in future studies:

- the system could render highly complex maps, e.g. with high numbers and different typologies of concurrent virtual objects and sound markers, and introducing obstacle avoidance. In general, high-level cognitive and problem-solving tasks can be investigated (Bowman et al., 2002), e.g. the creation of an optimal path between two points in a map. Once virtual maps of real environments will be easily accessible (or an ad-hoc real environment will be built upon a simplified virtual map), one can measure the navigation performances of blindfolded subjects walking in a real room after training sessions of the virtual counterpart.

- virtual reality has the potential to constrain the haptic exploration on certain directions or areas, reducing the amount of information provided by active exploration (Lederman and Klatzky, 1987);

- the multimodal environment could test different haptic devices (e.g. TAMO and Phantom), comparing performances in the same tasks.

It is worthwhile to stress the potential social impact of such a system, once it is tested with visually-impaired subjects. Since it is easy to create virtual maps of real places, e.g. from existing *CAD* planimetries, several applications could aid visually-impaired people in O&M training.

With particular attention to experiment of Sec. 7.4, other kinds of auditory feedback will be explored (e.g., using loudness, spectral centroid, etc.) in order to assess whether certain parameters are semantically more coherent than others with the concept of height. Although such a semantic link has been proven to exist for pitch (Rusconi et al., 2005, 2006), other parameters may provide an even stronger link.

Additionally, several design issues arise while using TAMO devices. The system could benefit by a larger workspace area and moreover, the orientation of the haptic device could be tracked and related to tablet orientation and also to subject head-pose. All these aspects become crucial if users are blindfolded or visually-impaired and thus they can not visually control and correct the pointer position.

However, the more important outcome of this conclusive chapter is the effectiveness of using VAD and spatial audio rendering for spatial navigation with the aim at investigating cognitive spatial mechanisms. To this regard, the experiments reported in Sec. 7.3.5 capture the essence of the MSM approach on designing dynamic spatial auditory feedback, showing preliminary and encouraging results to support the cognitive process of mental map creation.

Chapter 8

Conclusions and future work

The *Mixed Structural Modeling* approach presented in this doctoral thesis answers both the requirements of structural modularity and integration of heterogeneous contributions in virtual reality contexts. The MSM approach allows indeed an agile mixture of acoustic responses and synthetic models with the appealing merit of verifying such diversity. The well-defined characterization facilitates the design of novel synthetic filter models and HRTF selection processes.

Several studies have been reported towards the design and implementation of a research framework for the analysis, synthesis and evaluation of HRTFs in order to quantify the required degree of individualization of those spatial audio engines that will be integrated into novel multimodal interfaces.

Indeed, multimodal systems are expected to largely benefit from the integration of MSMs in the accurate description of the virtual acoustic scenes. Following the MSM approach, spatial audio rendering can provide accurate dynamic information about the relation between the sound source and the surrounding environment, including especially the listener's body which acts as a fingerprint for the individual perception of the world. Moreover, the concurrent presence of multiple senses make multimodal systems potentially flexible and adaptive, allowing users to integrate and/or switch between sensory and auditory cues as needed during his/her continuous interaction with the system.

8.1 Horizons

An exhaustive data analysis on the relation between individual anthropometric quantities and personal binaural signals behavior through headphones is still to come. Some possible future research directions are now grouped in three main points: (i) exploration of the MSM parameter space to find the best MSM given an individual HRTF set; (ii) individualization of the acoustic contribution of headphones for every listener; (iii) formalization of the evaluation methodologies for 3D audio technologies in multimodal virtual environments with a focus on orientation and mobility educational purposes.

8.1.1 Mixed structural modeling

In order to improve the localization accuracy provided by the MSM model in a full 3D space, the degree of orthogonality among structural components has to be tested. This implies an extension of the proposed pHRTF models outside their dominant spatial dimension. Among the possible options are an extension of the **pinna model** outside the median plane (a more complex analysis than that performed in Sec. 4.1 is required); the inclusion of elevation-dependent patterns in non-spherical head responses or ITD & ILD estimation from measured HRTFs; and a study of the behaviour of the torso in the near field.

Once a **complete structural decomposition** of measured HRTFs is achieved, a structural selection on extracted pHRTFs can be performed, e.g. ITD selection is a typical example of such an approach. To name but a few other options, one can conduct a PCA analysis on the collected pHRTFs to find a more compact representation of each structural component and thus compress HRTF data. Rliable psychoacoustic tests are needed in order to assess the optimality of such partial selection criteria and data reduction. Narrowband stimuli within different frequency ranges are often employed in spatial hearing tests and can help to disambiguate each structural acoustic contribution.

The extension of MSMs through the inclusion of **computer-simulated HRIRs/pHRIRs** (e.g., using boundary element methods) calculated from mesh models of human heads, spatially discretized so as to be included in a HRTF unified database, will speed up the entire research flow. Following the MSM approach, it will also be possible to manipulate meshes such as to hierarchically generate several mesh models that include a specific body part only (e.g. only head without pinna, pinna alone, smoothed head, etc.). In this regard, the definition of a standardized format for anthropometric features (possibly borrowed from biometric researches) so as to integrate this information in a HRTF/pHRTF database will be an essential step towards identification and manipulation of listener's geometry.

Moreover, the quality of the acquisition setup for the **anthropometric parameters** depends upon the precision of the scanning devices and sensors as well as level of handiness of the entire procedure. Thus, the objective of this activity will be twofold: obtaining as precise as possible acquisitions to maximize the precision of the acoustic simulations, and being less invasive as possible. Several methods will be investigated having different combination of precision and handiness, so as to include high definition scanning, 3D modeling using photographs and low-cost depth camera model reconstructions. A separate mention is needed for the ear canal geometry which is not externally detectable: fMRI (functional magnetic resonance imaging) scanning is required in order to obtain its 3D mesh model.

Particular attention will be given to the design of several MSM selection criteria exploiting the potentiality of using simulated HRIRs only as a robust dataset to select from.

Localization error minimization can be also achieved by increasing the number of structural components. As an example, the **ear canal** contributes to the approximation of the correct pressure at the eardrum both in free-field and headphone listening conditions. The aforementioned simulations of the sound pressure at the ear-drum of the listener, with and without headphones, provide insights on the personal directionally dependent response of the ear-canal whereby a structural model can be designed to proper compensate such behaviour;

On the other hand, the gradual increase of MSM instances requires reliable and complex **auditory models** so as to facilitate the systematic exclusion of weak HRIR/pHRTF models or selections in favour of the best MSM instances. With reference to the pinna case study discussed in Ch. 4, one can construct several instances of the pinna pHRTF model among all the possible combinations with different $F_{n/p}$, $G_{n/p}$ and $B_{n/p}$ values.

Design guidelines for innovative VR devices should incorporate more formally the MSM concepts of handiness and accuracy, specifically their relationship with auralization issues, individualization procedures and system ergonomics/usability. Thanks to the large amount of simulated pHRIRs, recorded HpIRs and the corresponding filter models, the MSM-guided criteria allow to select the best MSM instance for any listener who is not present in the data set. Finally, the candidate MSM, fed with the anthropometric parameters, will be evaluated by each subject against his individual HRTFs.

8.1.2 Individualized headphones

Once individual variations data on the repositioning of headphones are gathered independently for the left and right audio channel, a filter model can be designed for real-time control of the equalization parameters, in order to avoid spectral coloration and enhance externalization of the sound scene.

Individual headphones equalization filters will be first built with perceptually robust inverse filtering techniques; then a customizable filter model will be designed for robust real-time binaural audio delivery with arbitrary listeners. A real-time equalization filter could take into account each single headphones coupler, designing a feedback correction. It will be necessary to realize a collection of tools that permit such kind of headphone control (e.g. equipped headphones with pressure sensors, optical sensors, etc.). **Sensorized headphones** have the potential of introducing novel forms of customized communication and interaction with sound and music contents. **Sound quality** judgments will be collected in subjective listening tests. For instance, the judged realism and spatial impressions of the binaural experience will be related to the accuracy of sensorized headphones.

In a real listening scenario, the non constant section of the ear canal implies a directional dependency from the sound source position (Hudde and Schmidt, 2009). When headphones are used for virtual acoustic rendering, non negligible distortions of such localization cues occur especially due to the individual nature of ear-canal cross sections. The acoustic effect at the eardrum caused by headphones repositionings will be investigated in the headphones-to-eardrum transfer function simulations. Then, the introduction of an ear canal model to approximate the correct P_7 (see Fig. 1.14) at the eardrum will also support the formalization of a **structural HpIR model**. Moreover, the MSM selection process adopted for pinna shapes will be adapted so as to develop HpIR selection procedures and customize the designed HpIR equalization filters.

8.1.3 Orientation & mobility education

The use of of earcons (Brewster et al., 1993), auditory icons (Hoggan and Brewster, 2007) and continuous sound feedbacks enhanced by binaural MSM techniques will establish design guidelines for complex environments aimed at investigating higher-level cognitive functions , i.e. **spatial cognition**.

Physically-based sound synthesis (Rocchesso and Fontana, 2003) will be exploited for data and **gesture sonification**. Basic interactions with virtual environments (e.g. manipulation, pointing task, item recognition, etc.) will be spatially explored with increasingly dimensionality and shape complexity. Tactile textures and icons will also be controlled by physical model-generated audio signals.

Similar experiments to those discussed in Sec. 7.3.3 will be performed with the aim of testing audio and kinesthetic contribution to navigation in virtual environments. Data analysis of the exploration strategies adopted by subjects can reveal criticalities of HRTF representation (i.e. minimum localization requirements) and, more in general, of methodologies in sonic interaction design. Moreover, integration of virtual reverberation greatly contributes to improve distance perception. Especially, implementation of early reflections together with different intensity scaling factors will allow to investigate which combination of these effects is the most effective. Summarizing this research direction, the main general goal is to design spatialized auditory feedback able to convey semantic information. This will provide the basis for the development of a computationally efficient system for real-time simulations of acoustic scenes with increasing degrees of complexity with respect to those presented in Sec. 7.3.

Finally, all the designed combinations of devices, tactile feedback, and auditory feedback will be submitted to real-time performance benchmarks and usability tests to produce user centered applications (Bowman et al., 2002). The definition of novel efficiency metrics will guide the life-cycle of this system resulting in a low-cost solution which will be effectively validated and addressed to the **O&M education** community of blind and visually-impaired people.

8.2 Publications

The work presented in this thesis has produced the following publications.

8.2.1 International Journals

- Spagnol, S., Geronazzo, M., and Avanzini, F. (2013a). On the relation between pinna reflection patterns and head-related transfer function features. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(3):508–519

- Turchet, L., Spagnol, S., Geronazzo, M., and Avanzini, F. (2014). Localization of self generated synthetic footstep sounds. *ACM Trans. Applied Perception (submitted for publication)*, -(-):–

8.2.2 International Conferences

- 2014

  - Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F. (2014). Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014) (submitted for publication)*, pages –, Firenze, Italy

- 2013

  - Spagnol, S., Geronazzo, M., Rocchesso, D., and Avanzini, F. (2013b). Extraction of pinna features for customized binaural audio delivery on mobile devices. In *In Proc. 11th International Conference on Advances in Mobile Computing & Multimedia (MoMM'13)*, pages 514–517, Vienna, Austria
  - Geronazzo, M., Avanzini, F., and Grassi, M. (2013a). Influence of auditory pitch on haptic estimation of spatial height. In *Proc. 10th International Symposium on Computer Music Multidisciplinary Research (CMMR'13)*, Marseille
  - Spagnol, S., Rocchesso, D., Geronazzo, M., and Avanzini, F. (2013d). Automatic extraction of pinna edges for binaural audio customization. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP 2013)*, page 301–306, Pula, Italy
  - Geronazzo, M., Spagnol, S., and Avanzini, F. (2013d). Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, Santorini, Greece
  - Geronazzo, M., Granza, F., Spagnol, S., and Avanzini, F. (2013b). A standardized repository of head-related and headphone impulse response data. In *Audio Engineering Society Convention 134*, Rome, Italy
  - Geronazzo, M., Spagnol, S., and Avanzini, F. (2013f). A modular framework for the analysis and synthesis of head-related transfer functions. In *Proc. 134th Conv. Audio Eng. Soc.*, Rome, Italy

- 2012

  - Spagnol, S. (2012a). Are spectral elevation cues in head-related transfer functions distance-independent? In *Proc. XIX Colloquio di Informatica Musicale (XIX CIM)*, page 192–197, Trieste, Italy
  - Spagnol, S. (2012b). *Techniques for Customized Binaural Audio Rendering with Applications to Virtual Rehabilitation*. PhD thesis, Ingegneria dell'Informazione, Scienza e Tecnologia dell'Informazione

- 2011

  - Geronazzo, M., Spagnol, S., and Avanzini, F. (2011a). Customized 3D sound for innovative interaction design. In *Proc. SMC-HCI Work., CHItaly 2011 Conf.*, Alghero, Italy

  - Geronazzo, M., Spagnol, S., and Avanzini, F. (2011c). A head-related transfer function model for real-time customized 3-d sound rendering. In *Proc. INTERPRET Work., SITIS 2011 Conf.*, page 174–179, Dijon, France

- 2010

  - Spagnol, S., Geronazzo, M., and Avanzini, F. (2010b). Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP'10)*, page 194–199, Saint-Malo, France
  - Geronazzo, M., Spagnol, S., and Avanzini, F. (2010b). Estimation and modeling of pinna-related transfer functions. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, pages 431–438, Graz, Austria
  - Spagnol, S., Geronazzo, M., and Avanzini, F. (2010c). Structural modeling of pinna-related transfer functions. In *Proc. 7th Int. Conf. Sound and Music Computing (SMC 2010)*, page 422–428, Barcelona, Spain

8.2.3 National Conferences

- 2012

  - Geronazzo, M., Spagnol, S., Rocchesso, D., and Avanzini, F. (2012). Model-based customized binaural reproduction through headphones. In *Proc. XIX Colloquio di Informatica Musicale (XIX CIM)*, page 212–213, Trieste, Italy

- 2010

  - Spagnol, S., Geronazzo, M., and Avanzini, F. (2010d). Structural modeling of pinna-related transfer functions for 3-d sound rendering. In *Proc. XVIII Colloquio di Informatica Musicale (XVIII CIM)*, page 92–101, Torino, Italy

Bibliography

Abaza, A., Ross, A., Hebert, C., Harrison, M. A. F., and Nixon, M. S. (2010). A survey on ear biometrics. *ACM Trans. Embedded Computing Systems*, 9(4):39:1–39:33.

Afonso, A., Blum, A., Katz, B., Tarroux, P., Borst, G., and Denis, M. (2010). Structural properties of spatial representations in blind people: Scanning images constructed from haptic exploration or from locomotion in a 3-d audio virtual environment. *Memory & Cognition*, 38(5):591–604. 10.3758/MC.38.5.591.

Afonso, A., Katz, B. F., Blum, A., Jacquemin, C., and Denis, M. (2005). A study of spatial cognition in an immersive virtual audio environment: Comparing blind and blindfolded individuals. In *Proceedings of the 10th Meeting of the International Conference on Auditory Display (ICAD)*.

Algazi, V. R., Avendano, C., and Duda, R. (2001a). Estimation of a spherical-head model from anthropometry. *J. Audio Eng. Soc*, 49(6):472–479.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001b). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122.

Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002a). Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5):2053–2064.

Algazi, V. R., Duda, R. O., Morrison, R. P., and Thompson, D. M. (2001c). Structural composition and decomposition of HRTFs. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, page 103–106, New Paltz, New York, USA.

Algazi, V. R., Duda, R. O., and Thompson, D. M. (2002b). The use of head-and-torso models for improved spatial sound synthesis. In *Proc. 113th Conv. Audio Eng. Soc.*, page 1–18, Los Angeles, CA, USA.

Algazi, V. R., Duda, R. O., and Thompson, D. M. (2004). Motion-tracked binaural sound. *JAES*, 52(11):1142–1156.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001d). The CIPIC HRTF database. In *Proc. IEEE Work. Appl. Signal Process., Audio, Acoust.*, page 1–4, New Paltz, New York, USA.

Andreopoulou, A. and Roginska, A. (2011). Towards the creation of a standardized HRTF repository. In *Proc. 131st Conv. Audio Eng. Soc.*, New York, NY, USA.

Ansari, S. and Gupta, P. (2007). Localization of ear using outer helix curve of the ear. In *Proc. 17th IEEE Int. Conf. Comput. Theory Appl.*, page 688–692, Kolkata, India.

Asano, F., Suzuki, Y., and Sone, T. (1990). Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America*, 88(1):159–168.

Avanzini, F. and Rocchesso, D. (2001). Modeling collision sounds: Non-linear contact force. In *Proceedings of Digital Audio Effects Conference*, pages 61–66.

Barron, M. (1971). The subjective effects of first reflections in concert halls—The need for lateral reflections. *Journal of Sound and Vibration*, 15(4):475–494.

Batteau, D. W. (1967). The role of the pinna in human localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168(11):158–180.

Bederson, B. B. (1995). Audio augmented reality: a prototype automated tour guide. In *Conference companion on Human factors in computing systems*, CHI '95, page 210–211, Denver, Colorado, United States. ACM.

Bedin, A. (2013). *Head Related Transfer Function selection techniques applied to multimodal environments for spatial cognition*. Magistrali biennali, University of Padava.

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia*. Academic Press Professional, Inc., San Diego, CA, USA.

Begault, D. R., Lee, A. S., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. In *Audio Engineering Society Convention 108*.

Belloch, J. A., Ferrer, M., Gonzalez, A., Martinez-Zaldivar, F., and Vidal, A. M. (2013). Headphone-based virtual spatialization of sound with a GPU accelerator. *J. Audio Eng. Soc*, 61(7/8):546–561.

Benford, S. and Fahlén, L. (1993). A spatial model of interaction in large virtual environments. In *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW'93*, page 109–124.

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

Borwick, J. (2001). *Loudspeaker and Headphone Handbook*. Taylor & Francis.

Bowman, D. A., Gabbard, J. L., and Hix, D. (2002). A survey of usability evaluation in virtual environments: classification and comparison of methods. *Presence: Teleoper. Virtual Environ.*, 11(4):404–424.

Bradley, J. S. and Soulodre, G. A. (1995). Objective measures of listener envelopment. *The Journal of the Acoustical Society of America*, 98(5):2590–2597.

Brayda, L., Campus, C., Chellali, R., and Rodriguez, G. (2010). Objective evaluation of spatial information acquisition using a visuo-tactile sensory substitution device. In Ge, S., Li, H., Cabibihan, J.-J., and Tan, Y., editors, *Social Robotics*, volume 6414 of *Lecture Notes in Computer Science*, pages 315–324. Springer Berlin / Heidelberg. 10.1007/978-3-642-17248-9_33.

Brayda, L., Campus, C., Chellali, R., Rodriguez, G., and Martinoli, C. (2011). An investigation of search behaviour in a tactile exploration task for sighted and non-sighted adults. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, CHI EA '11, page 2317–2322, Vancouver, BC, Canada. ACM.

Brayda, L., Campus, C., and Gori, M. (2013). Predicting successful tactile mapping of virtual objects. *IEEE Transactions on Haptics*, pages 1–1.

Brewster, S. A., Wright, P. C., and Edwards, A. D. N. (1993). An evaluation of earcons for use in auditory human-computer interfaces. In *Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, CHI '93, page 222–227, Amsterdam, The Netherlands. ACM.

Brimijoin, W. O., Boyd, A. W., and Akeroyd, M. A. (2013). The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE*, 8(12):e83068.

Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *The Journal of the Acoustical Society of America*, 98(5):2542–2553.

Bronkhorst, A. W. and Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, 397(6719):517–520.

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

Brungart, D. S. (2002). Near-field virtual audio displays. *Presence*, 11(1):93–106.

Brungart, D. S., Durlach, N. I., and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. II. localization of a broadband source. *J. Acoust. Soc. Am.*, 106(4):1956–1968.

Brungart, D. S. and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479.

Bujnowski, A., Drozd, M., Kowalik, R., and Wtorek, J. (2008). A tactile system for informing the blind on direction of a walk. In *Human System Interactions, 2008 Conference on*, page 893–897.

Burge, M. and Burger, W. (2000). Ear biometrics in computer vision. In *Proc. 15th IEEE Int. Conf. Pattern Recog.*, page 2822–2826, Barcelona, Spain.

Burkhard, M. D. and Sachs, R. M. (1975). Anthropometric manikin for acoustic research. *The Journal of the Acoustical Society of America*, 58(1):214–222.

Busson, S., Nicol, R., and Katz, B. F. G. (2005). Subjective investigations of the interaural time difference in the horizontal plane. In *Audio Engineering Society Convention 118*.

Campus, C., Brayda, L., Carli, F. D., Chellali, R., Famà, F., Bruzzo, C., Lucagrossi, L., and Rodriguez, G. (2012). Tactile exploration of virtual objects for blind and sighted people: the role of beta 1 EEG band in sensory substitution and supramodal mental mapping. *J Neurophysiol*, 107(10):2713–2729. PMID: 22338024.

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, PAMI-8(6):679–698.

Cattaneo, Z., Vecchi, T., Cornoldi, C., Mammarella, I., Bonino, D., Ricciardi, E., and Pietrini, P. (2008). Imagery and spatial processes in blindness and visual impairment. *Neuroscience & Biobehavioral Reviews*, 32(8):1346–1360.

Chellali, R., Brayda, L., Martinoli, C., and Fontaine, E. (2009). How taxel-based displaying devices can help visually impaired people to navigate safely. In *Autonomous Robots and Agents, 2009. ICARA 2009. 4th International Conference on*, page 470–475.

Chen, H. and Bhanu, B. (2005). Contour matching for 3D ear recognition. In *Proc. 7th IEEE Work. Appl. Comp. Vision (WACV/MOTION'05)*, page 123–128, Breckenridge, CO, USA.

Cheng, C. I. and Wakefield, G. H. (2001). Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space. *J. Audio Eng. Soc.*, 49(4):231–249.

Choraś, M. (2005). Ear biometrics based on geometrical feature extraction. *Electron. Lett. Comput. Vision Image Anal.*, 5(3):84–95.

Chrastil, E. and Warren, W. (2011). Active and passive contributions to spatial learning. *Psychonomic Bulletin & Review*, -:1–23. 10.3758/s13423-011-0182-x.

Christensen, F., Hoffmann, P. F., and Hammershøi, D. (2013). Measuring directional characteristics of in-ear recording devices. In *Audio Engineering Society Convention 134*.

Cohen, M., Aoki, S., and Koizumi, N. (1993). Augmented audio reality: telepresence/VR hybrid acoustic environments. In *Robot and Human Communication, 1993. Proceedings., 2nd IEEE International Workshop on*, pages 361 –364.

Colburn, H. S. and Kulkarni, A. (2005). Models of sound localization. In Popper, A. N. and Fay, R. R., editors, *Sound Source Localization*, number 25 in Springer Handbook of Auditory Research, pages 272–316. Springer New York.

Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. *J. Acoust. Soc. Am.*, 34(3):345–346.

Cook, P. (1997). Physically informed sonic modeling (phism): Synthesis of percussive sounds. *Computer Music Journal*, 21(3):38–49.

D'Agostino, R. B. (1986). Tests for normal distribution. In D'Agostino, R. B. and Stephens, M. A., editors, *Goodness-of-fit Techniques*. Marcel Decker, New York.

Dehaene, S., Bossini, S., and Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122(3):371.

Delle Monache, S. D., Polotti, P., and Rocchesso, D. (2010). A toolkit for explorations in sonic interaction design. In *Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound*, AM '10, page 1:1–1:7, New York, NY, USA. ACM.

Dellepiane, M., Pietroni, N., Tsingos, N., Asselot, M., and Scopigno, R. (2008). Reconstructing head models from photographs for individualized 3D-Audio processing. *Comp. Graph. Forum*, 27(7):1719–1727.

Devallez, D., Fontana, F., and Rocchesso, D. (2008). Linearizing auditory distance estimates by means of virtual acoustics. *Acta Acustica united with Acustica*, 94(6):813–824.

Doukhan, D. and Sédès, A. (2009). CW binaural˜: A binaural synthesis external for pure data. In *PdCon09 - 3rd Puredata Int. Convention Proc.*

Dubus, G. and Bresin, R. (2013). A systematic review of mapping strategies for the sonification of physical quantities. *PLoS ONE*, 8(12):e82491.

Duda, R., Avendano, C., and Algazi, V. (1999). An adaptable ellipsoidal head model for the interaural time difference. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 2, pages 965 –968 vol.2.

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

Durant, E. C. and Wakefield, G. H. (2002). Efficient model fitting using a genetic algorithm: Pole-zero approximations of HRTFs. *IEEE Trans. Speech Audio Process.*, 10(1):18–27.

Eckel, G. (2001). Immersive audio-augmented environments - the LISTEN project. In *Proc. 5th IEEE Int. Conf. Info. Visualization (IV'01)*, page 571–573, Los Alamitos, CA, USA.

Ernst, M. O. and Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169.

Esquef, P. A. A., Karjalainen, M., and Välimäki, V. (2003). Frequency-zooming ARMA modeling for analysis of noisy string instrument tones. *EURASIP J. Appl. Signal Process.*, 2003(10):953–967.

Evans, M. J., Angus, J. A. S., and Tew, A. I. (1998). Analyzing head-related transfer function measurements using surface spherical harmonics. *J. Acoust. Soc. Am.*, 104(4):2400–2411.

Faller II, K. J., Barreto, A., and Adjouadi, M. (2010). Augmented hankel total least-squares decomposition of head-related transfer functions. *J. Audio Eng. Soc.*, 58(1/2):3–21.

Faller II, K. J., Barreto, A., Gupta, N., and Rishe, N. (2006). Accelerated method for the reduced-parameter modeling of head-related transfer functions for customizable spatial audio. In *Proc. 5th WSEAS Int. Conf. on Circuits, Systems, Electronics, Control & Signal Processing (CSECS'06)*, page 263–268, Dallas, TX, USA.

Farcy, R., Leroux, R., Jucha, A., Damaschini, R., Grégoire, C., and Zogaghi, A. (2006). Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view. In *Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Technology for Inclusion*, page 12.

Farkas, A. J. and Hajnal, A. (2013). Loudness threshold as a function of sound source location using circum-aural headphones in noisy and sound-proof acoustic environments. *Frontiers in Psychological and Behavioral Science*, 1(3):89–95.

Fechner, G. T. (1889). *Elemente der Psychophysik*. Breitkopf & Härtel, Leipzig.

Fels, J. and Vorlander, M. (2009). Anthropometric parameters influencing head-related transfer functions. *Acta Acustica united with Acustica*, 95(2):331–342.

Franinovic, K. and Serafin, S. (2013). *Sonic Interaction Design*. MIT Press.

Gardner, M. B. and Gardner, R. S. (1973). Problem of localization in the median plane: Effect of pinnae cavity occlusion. *J. Acoust. Soc. Am.*, 53(2):400–408.

Gardner, W. G. (1999). *3D Audio and Acoustic Environment Modeling*. Wave Arts, Inc. Wave Arts, Inc.

Gardner, W. G. and Martin, K. D. (1995). HRTF measurements of a KEMAR. *J. Acoust. Soc. Am.*, 97(6):3907–3908.

Geronazzo, M., Avanzini, F., and Grassi, M. (2013a). Influence of auditory pitch on haptic estimation of spatial height. In *Proc. 10th International Symposium on Computer Music Multidisciplinary Research (CMMR'13)*, Marseille.

Geronazzo, M., Granza, F., Spagnol, S., and Avanzini, F. (2013b). A standardized repository of head-related and headphone impulse response data. In *Audio Engineering Society Convention 134*, Rome, Italy.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2010a). Estimation and modeling of pinna-related transfer functions. In *Proc. 13th Int. Conf. Digital Audio Effects (DAFx-10)*, pages 431–438, Graz, Austria.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2010b). Estimation and modeling of pinna-related transfer functions. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, pages 431–438, Graz, Austria.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2011a). Customized 3D sound for innovative interaction design. In *Proc. SMC-HCI Work., CHItaly 2011 Conf.*, Alghero, Italy.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2011b). 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.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2011c). A head-related transfer function model for real-time customized 3-d sound rendering. In *Proc. INTERPRET Work., SITIS 2011 Conf.*, page 174–179, Dijon, France.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2013c). Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, Santorini, Greece.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2013d). Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*, Santorini, Greece.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2013e). A modular framework for the analysis and synthesis of head-related transfer functions. In *Proc. 134th Conv. Audio Eng. Soc.*, Rome, Italy.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2013f). A modular framework for the analysis and synthesis of head-related transfer functions. In *Proc. 134th Conv. Audio Eng. Soc.*, Rome, Italy.

Geronazzo, M., Spagnol, S., Bedin, A., and Avanzini, F. (2014). Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014) (submitted for publication)*, pages –, Firenze, Italy.

Geronazzo, M., Spagnol, S., Rocchesso, D., and Avanzini, F. (2012). Model-based customized binaural reproduction through headphones. In *Proc. XIX Colloquio di Informatica Musicale (XIX CIM)*, page 212–213, Trieste, Italy.

Gibson, E. J. and Pick, A. D. (2000). *An Ecological Approach to Perceptual Learning and Development*. Oxford University Press.

Giordano, B., Visell, Y., Yao, H.-Y., Hayward, V., Cooperstock, J., and McAdams, S. (2012). Identification of walked-upon materials in auditory, kinesthetic, haptic and audio-haptic conditions. *Journal of The Acoustical Society Of America*, 131:4002–4012.

Gómez Bolaños, J. and Pulkki, V. (2012). HRIR database with measured actual source direction data. In *Proc. 133rd Conv. Audio Eng. Soc.*, San Francisco, CA, USA.

González, E., Alvarez, L., and Mazorra, L. (2012). Normalization and feature extraction on ear images. In *Proc. IEEE 46th Int. Carnahan Conf. Security Tech.*, page 97–104, Boston, MA, USA.

Grantham, D. W. (1984). Discrimination of dynamic interaural intensity differences. *The Journal of the Acoustical Society of America*, 76(1):71–76.

Grantham, D. W. and Wightman, F. L. (1978). Detectability of varying interaural temporal differences. *The Journal of the Acoustical Society of America*, 63(2):511–523.

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *J. Acoust. Soc. Am.*, 93(4):2096–2105.

Gridi-Papp, M. and Narins, P. M. (2008). 3.04 - sensory ecology of hearing. In Masland, R. H., Albright, T. D., Albright, T. D., Masland, R. H., Dallos, P., Oertel, D., Firestein, S., Beauchamp, G. K., Bushnell, M. C., Basbaum, A. I., Kaas, J. H., and Gardner, E. P., editors, *The Senses: A Comprehensive Reference*, pages 61–74. Academic Press, New York.

Gupta, N., Barreto, A., and Choudhury, M. (2004). Modeling head-related transfer functions based on pinna anthropometry. In *Proc. 2nd Int. Lat. Am. Carib. Conf. Eng. Tech. (LACCEI'2004)*, Miami, FL, USA.

Gupta, N., Barreto, A., Joshi, M., and Agudelo, J. (2010). HRTF database at FIU DSP lab. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 169 –172.

Hafter, E. R. and Trahiotis, C. (1997). Functions of the binaural system. In Editor-in Chief, l. J. C., editor, *Encyclopedia of Acoustics*, page 1461–1479. John Wiley & Sons, Inc.

Hahn, J. K., Fouad, H., Gritz, L., and Lee, J. W. (1998). Integrating sounds in virtual environments. *Presence: Teleoperators and Virtual Environment*, 7(1):67–77.

Hale, K. S. and Stanney, K. M. (2002). *Handbook of Virtual Environments: Design, Implementation, and Applications*. Taylor & Francis.

Hammershøi, D. and Hoffmann, P. F. (2011). Control of earphone produced binaural signals. In Society, D. A., editor, *Proceedings of Forum Acusticum 2011*, pages 2235–2239. European Acoustics Association - EAA.

Härmä, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (2004). Augmented reality audio for mobile and wearable appliances. *J. Audio Eng. Soc*, 52(6):618–639.

Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *The Journal of the Acoustical Society of America*, 99(6):3678–3688.

Hatala, M., Kalantari, L., Wakkary, R., and Newby, K. (2004). Ontology and rule based retrieval of sound objects in augmented audio reality system for museum visitors. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, SAC '04, page 1045–1050, New York, NY, USA. ACM.

Hebrank, J. and Wright, D. (1974a). Are two ears necessary for localization of sound sources on the median plane? *J. Acoust. Soc. Am.*, 56(3):935–938.

Hebrank, J. and Wright, D. (1974b). Spectral cues used in the localization of sound sources on the median plane. *J. Acoust. Soc. Am.*, 56(6):1829–1834.

Heffner, R. S. (2004). Primate hearing from a mammalian perspective. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology*, 281A(1):1111–1122.

Hendrix, C. and Barfield, W. (1995). Presence in virtual environments as a function of visual and auditory cues. In *Virtual Reality Annual International Symposium, 1995. Proceedings.*, pages 74 –82.

Hess, W. (2012). Head-tracking techniques for virtual acoustics applications. In *In Proc. Audio Engineering Society Convention 133*. Audio Engineering Society.

Hetherington, C., Tew, A., and Tao, Y. (2003). Three-dimensional elliptic fourier methods for the parameterization of human pinna shape. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages V – 612–15 vol.5.

Hiipakka, M. (2008). *Measurement Apparatus and Modelling Techniques of Ear Canal Acoustics*. PhD thesis, HELSINKI UNIVERSITY OF TECHNOLOGY, Department of Signal Processing and Acoustics.

Hiipakka, M. (2013). Individual in-situ calibration of insert headphones. In *ICA 2013 Montreal*, volume 19, page 030091, Montreal, Canada. ASA.

Hiipakka, M., Kinnari, T., and Pulkki, V. (2012). Estimating head-related transfer functions of human subjects from pressure–velocity measurements. *The Journal of the Acoustical Society of America*, 131(5):4051–4061.

Hoggan, E. and Brewster, S. (2007). Designing audio and tactile cross-modal icons for mobile devices. In *Proceedings of the 9th international conference on Multimodal interfaces*, ICMI '07, page 162–169, Nagoya, Aichi, Japan. ACM.

Holmes, N. P. and Spence, C. (2005). Multisensory integration: Space, time, & superadditivity. *Curr Biol*, 15(18):R762–R764. PMID: 16169476 PMCID: PMC1578214.

Horiuchi, T., Hokari, H., and Shimada, S. (2001). Out-of-head sound localization using adaptive inverse filter. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 5, page 3333–3336.

Howard, I. P. and Templeton, W. B. (1966). *Human spatial orientation*. Wiley.

Hudde, H. and Schmidt, S. (2009). Sound fields in generally shaped curved ear canals. *The Journal of the Acoustical Society of America*, 125(5):3146–3157.

Hunt, K. H. and Crossley, F. R. E. (1975). Coefficient of restitution interpreted as damping in vibroimpact. *ASME Journal of Applied Mechanics*, 42(2):440–445.

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

Huttunen, T., Kärkkäinen, A., Kärkkäinen, L., Kirkeby, O., and Seppälä, E. T. (2007). Some effects of the torso on head-related transfer functions. In *Audio Engineering Society Convention 122*.

Hwang, S., Park, Y., and Park, Y.-s. (2008). 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.

Iannarelli, A. V. (1989). *Ear Identification*. Forensic Identification. Paramount Publishing Company, Fremont, CA, USA.

Iida, K., Itoh, M., Itagaki, A., and Morimoto, M. (2007). Median plane localization using a parametric model of the head-related transfer function based on spectral cues. *Applied Acoustics*, 68(8):835 – 850. <ce:title>Head- Related Transfer Function and its Applications</ce:title>.

Ishihara, M., Keller, P. E., Rossetti, Y., and Prinz, W. (2008). Horizontal spatial representations of time: Evidence for the stearc effect. *Cortex*, 44(4):454–461.

Jeges, E. and Máté, L. (2007). Model-based human ear localization and feature extraction. *Int. J. Intell. Comput. Med. Sci. Image Process.*, 1(2):101–112.

Jo, H., Park, Y., and Park, Y. (2008a). Approximation of head related transfer function using prolate spheroidal head model. In *Proc. 15th Int. Congr. Sound Vibr. (ICSV15)*, page 2963–2970, Daejeon, Korea.

Jo, H., Park, Y., and Park, Y. (2008b). Optimization of spherical and spheroidal head model for head related transfer function customization: Magnitude comparison. In *Proc. Int. Conf. Control, Automat., Syst. (ICCAS 2008)*, page 251–254, Seoul, Korea.

Jousmäki, V. and Hari, R. (1998). Parchment-skin illusion: sound-biased touch. *Current Biology*, 8(6):R190–R191.

Kahana, Y. and Nelson, P. A. (2007). Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of Sound and Vibration*, 300(3-5):552 – 579.

Kajastila, R. and Lokki, T. (2013). Eyes-free interaction with free-hand gestures and auditory menus. *International Journal of Human-Computer Studies*, 71(5):627–640.

Kan, A., Jin, C., and Schaik, A. v. (2009). A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function. *The Journal of the Acoustical Society of America*, 125(4):2233–2242.

Katz, B. F. G. (2001a). Boundary element method calculation of individual head-related transfer function. i. impedance effects and comparisons to real measurements. *J. Acoust. Soc. Am.*, 110(5):2449–2455.

Katz, B. F. G. (2001b). Boundary element method calculation of individual head-related transfer function. i. rigid model calculation. *J. Acoust. Soc. Am.*, 110(5):2440–2448.

Katz, B. F. G., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S., and Jouffrais, C. (2012). NAVIG: augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality*, 16(4):253–269.

Kim, C., Mason, R., and Brookes, T. (2013). Head movements made by listeners in experimental and real-life listening activities. *Journal of the Audio Engineering Society*, 61(6):425–438.

Kim, S.-M. and Choi, W. (2005). On the externalization of virtual sound images in headphone reproduction: A wiener filter approach. *The Journal of the Acoustical Society of America*, 117(6):3657–3665.

Kirchner, N., Alempijevic, A., and Virgona, A. (2012). Head-to-shoulder signature for person recognition. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1226 –1231.

Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Am.*, 91(3):1637–1647.

Kitagawa, N. and Spence, C. (2006). Audiotactile multisensory interactions in human information processing. *Japanese Psychological Research*, 48(3):158–173.

Klatzky, R. L., Lippa, Y., Loomis, J. M., and Golledge, R. G. (2003). Encoding, learning, and spatial updating of multiple object locations specified by 3-d sound, spatial language, and vision. *Experimental Brain Research*, 149(1):48–61.

Kleiner, M., Dalenbäck, B.-I., and Svensson, P. (1993). Auralization-an overview. *Journal of the Audio Engineering Society*, 41(11):861–875.

Kobayashi, Y., Osaka, R., Hara, T., and Fujimoto, H. (2008). How accurately people can discriminate the differences of floor materials with various elasticities. *IEEE Transactions on Neural and Rehabilitation Systems Engineering*, 16(1):99–105.

Kohler, S., Sankowsky-Rothe, T., Blau, M., and Stirnemann, A. (2013). A comparison of methods for estimating individual real-ear-to-coupler-differences (RECDs) in hearing aid fitting. In *Proceedings of Meetings on Acoustics*, volume 19, page 030096.

Kolarik, A., Cirstea, S., and Pardhan, S. (2013). Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues. *The Journal of the Acoustical Society of America*, 134(5):3395–3398.

Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *J. Acoust. Soc. Am.*, 62(1):157–167.

Kulkarni, A. and Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *The Journal of the Acoustical Society of America*, 107(2):1071–1074.

Kulkarni, A., Isabelle, S., and Colburn, H. (1995). On the minimum-phase approximation of head-related transfer functions. In , *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995*, pages 84–87.

Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (1999). Sensitivity of human subjects to head-related transfer-function phase spectra. *The Journal of the Acoustical Society of America*, 105(5):2821–2840.

Kuttruff, H. (2000). *Room Acoustics, Fourth Edition*. CRC Press.

Lahav, O. and Mioduser, D. (2004). Exploration of unknown spaces by people who are blind using a multi-sensory virtual environment. *Journal of Special Education Technology*, 19(3):15–23.

Lahav, O. and Mioduser, D. (2008). Construction of cognitive maps of unknown spaces using a multi-sensory virtual environment for people who are blind. *Computers in Human Behavior*, 24(3):1139 – 1155. <ce:title>Instructional Support for Enhancing Students' Information Problem Solving Ability</ce:title>.

Laitinen, M.-V. and Pulkki, V. (2009). *Binaural reproduction for Directional Audio Coding*. PhD thesis, HELSINKI UNIVERSITY OF TECHNOLOGY.

Langendijk, E. H. A. and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596.

Laurienti, P., Kraft, R., Maldjian, J., Burdette, J., and Wallace, M. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, 158(4):405–414.

Lecuyer, A., Mobuchon, P., Megard, C., Perret, J., Andriot, C., and Colinot, J.-P. (2003). HOMERE: a multimodal system for visually impaired people to explore virtual environments. In *Virtual Reality, 2003. Proceedings. IEEE*, pages 251 – 258.

Lederman, S. J. and Klatzky, R. L. (1987). Hand movements: A window into haptic object recognition. *Cognitive Psychology*, 19(3):342 – 368.

Lemaitre, G., Houix, O., Visell, Y., Franinovic, K., Misdariis, N., and Susini, P. (2009). Toward the design and evaluation of continuous sound in tangible interfaces: The Spinotron. *Int. J. Human-Comp. Studies, special issue on Sonic Interaction Design*, 67:976–993.

Lindau, A. and Brinkmann, F. (2010). Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings. In *3rd ISCADEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, page 137–142.

Lindau, A. and Brinkmann, F. (2012). Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. *J. Audio Eng. Soc*, 60(1/2):54–62.

Liu, H., Darabi, H., Banerjee, P., and Liu, J. (2007). Survey of wireless indoor positioning techniques and systems. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 37(6):1067–1080.

Liu, Y. and Xie, B.-s. (2013). Auditory discrimination on the distance dependence of near-field head-related transfer function magnitudes. *Proceedings of Meetings on Acoustics*, 19(1):050048.

Loomis, J., Klatzky, R., and Golledge, R. (1999). Auditory distance perception in real, virtual and mixed environments. In Ohta, Y. and Tamura, H., editors, *Mixed Reality: Merging Real and Virtual Worlds*. Springer.

Lopez-Poveda, E. A. and Meddis, R. (1996). A physical model of sound diffraction and reflections in the human concha. *J. Acoust. Soc. Am.*, 100(5):3248–3259.

Lorho, G., Hiipakka, J., and Marila, J. (2002). Structured menu presentation using spatial sound separation. In *Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction*, Mobile HCI '02, page 419–424, London, UK. Springer-Verlag.

Macpherson, E. (2011). Head motion, spectral cues, and wallach 's 'principle of least displacement' in sound localization. In *Principles and applications of spatial hearing*, pages 103–120. World Scientific, Singapore.

Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219.

Magee, L. E. and Kennedy, J. M. (1980). Exploring pictures tactually. *Nature*, 283(5744):287–288.

Majdak, P. and Laback, B. (2009). Effects of center frequency and rate on the sensitivity to interaural delay in high-frequency click trains. *J Acoust Soc Am*, 125(6):3903–3913. PMID: 19507973 PMCID: PMC3061452.

Martens, W. L. (2003). Individualized and generalized earphone correction filters for spatil sound reproduction. In Brazil, E. and Shinn-Cunningham, B., editors, -, pages 263–266, Boston, USA. Boston University Publications Production Department.

Masiero, B. and Fels, J. (2011). Perceptually robust headphone equalization for binaural reproduction. In *Audio Engineering Society Convention 130*.

McAulay, R. J. and Quatieri, T. F. (1986). Speech Analysis/Synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754.

McMullen, K. A. and Wakefield, G. H. (2013). The effects of attenuation modeling on spatial sound search. In *International Conference on Auditory Display 2013*, volume 1001, page 48109, Lodz, Poland.

Meredith, M. A., Nemitz, J. W., and Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. i. temporal factors. *J. Neurosci.*, 7(10):3215–3229. PMID: 3668625.

Mershon, D. H. and Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8(3):311–322.

Middlebrooks, J. C. (1999). Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3):1480–1492.

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1994). Augmented reality: A class of displays on the reality-virtuality continuum. In *IEICE TRANSACTIONS on Information and Systems*, page 282–292.

Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., and Møller, H. (2000). The interaural time difference in binaural synthesis. In *Audio Engineering Society Convention 108*.

Mittelstaedt, M.-L. and Mittelstaedt, H. (2001). Idiothetic navigation in humans: estimation of path length. *Experimental Brain Research*, 139(3):318–332.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2009). 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 2009)*, Zao, Miyagi, Japan.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2010). Acoustic sensitivity to micro-perturbations of KEMAR's pinna surface geometry. In *Proceedings of the 20th International Congress on Acoustics (ICA2010)*.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2011). Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2408 –2411.

Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4):171 – 218.

Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. (1995). Transfer characteristics of headphones measured on human ears. *JAES*, 43(4):203–217.

Møller, H., S\orensen, M., Friis, J., Clemen, B., and Hammershøi, D. (1996). Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc*, 44(6):451–469.

Moore, B. C. J., Oldfield, S. R., and Dooley, G. J. (1989). Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *The Journal of the Acoustical Society of America*, 85(2):820–836.

Moreno, B., Sánchez, A., and Vélez, J. F. (1999). On the use of outer ear images for personal identification in security applications. In *Proc. IEEE 33rd Int. Carnahan Conf. Security Tech.*, page 469–476, Madrid, Spain.

Morikawa, D., Toyoda, Y., and Hirahara, T. (2011). Impact of head movement on sound localization with band-limited noise. In *Proc. Inter-Noise*.

Morimoto, M. (2001). The contribution of two ears to the perception of vertical angle in sagittal planes. *J. Acoust. Soc. Am.*, 109(4):1596–1603.

Morimoto, M. (2002). The relation between spatial impression and the precedence effect. In *Proc. International Conf. on Auditory Display*, page 297–306.

Nam, J., Abel, J. S., Iii, S., and O, J. (2008a). A method for estimating interaural time difference for binaural synthesis. In *In Proc. Audio Engineering Society Convention 125*. Audio Engineering Society.

Nam, J., Kolar, M. A., and Abel, J. S. (2008b). On the minimum-phase nature of head-related transfer functions. In *In Proc. Audio Engineering Society Convention 125*. Audio Engineering Society.

Nishimura, R., Kato, H., Mokhtari, P., and Takemoto, H. (2010). Effects of headphone calibration functions on subjective impressions of spatial sound. In *Proceedings of 20th International Congress on Acoustics*, Sydney, Australia.

Nishimura, R. and Sonoda, K. (2013). B-format for binaural listening of higher order ambisonics. In *Proceedings of Meetings on Acoustics*, volume 19, page 055025.

Nishino, T., Inoue, N., Takeda, K., and Itakura, F. (2007). Estimation of HRTFs on the horizontal plane using physical features. *Applied Acoustics*, 68(8):897 – 908. <ce:title>Head- Related Transfer Function and its Applications</ce:title>.

Nojima, R., Morimoto, M., Sato, H., and Sato, H. (2013). Do spontaneous head movements occur during sound localization? *Acoustical Science and Technology*, 34(4):292–295.

Nordahl, R., Berrezag, A., Dimitrov, S., Turchet, L., Hayward, V., and Serafin, S. (2010a). Preliminary experiment combining virtual reality haptic shoes and audio synthesis. In *Haptics: Generating and Perceiving Tangible Sensations*, volume 6192 of *Lecture Notes in Computer Science*, pages 123–129. Springer Berlin Heidelberg.

Nordahl, R., Serafin, S., and Turchet, L. (2010b). Sound synthesis and evaluation of interactive footsteps for virtual reality applications. In *Proceedings of the IEEE Virtual Reality Conference*, pages 147–153. IEEE Press.

Novy, R. W. (1998). *Characterizing Elevation Effects of a Prolate Spheroidal HRTF Model*. PhD thesis, San Jose State University.

O'Donovan, A., Zotkin, D., and Duraiswami, R. (2008). Spherical microphone array based immersive audio scene rendering.

Orfanidis, S. J., editor (1996). *Introduction To Signal Processing*. Prentice Hall.

Otani, M., Hirahara, T., and Ise, S. (2009). Numerical study on source-distance dependency of head-related transfer functions. *The Journal of the Acoustical Society of America*, 125(5):3253–3261.

Papetti, S., Civolani, M., and Fontana, F. (2011). Rhythm'n'shoes: a wearable foot tapping interface with audio-tactile feedback. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 473–476.

Papetti, S., Fontana, F., Civolani, M., Berrezag, A., and Hayward, V. (2010). Audio-tactile display of ground properties using interactive shoes. In *Haptic and Audio Interaction Design*, volume 6306 of *Lecture Notes in Computer Science*, pages 117–128. Springer Berlin Heidelberg.

Parseihian, G., F, B., Katz, G., and Conan, S. (2012). Sound effect metaphors for near field distance sonification. In *Proceedings of the 18th International Conference on Auditory Display, Atlanta, USA, 18-21 June 2012*, pages 6–13.

Perrett, S. and Noble, W. (1997). The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332.

Perrott, D. R. and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *J. Acoust. Soc. Am.*, 87(4):1728–1731. PMID: 2341677.

Picinali, L., Afonso, A., Denis, M., and Katz, B. F. G. (2014). Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies*, 72(4):393 – 407.

Plenge, G. (1974). On the differences between localization and lateralization. *J. Acoust. Soc. Am.*, 56(3):944–951.

Popper, A. N. and Fay, R. R. (2005). *Sound source localization*, volume 25. Springer, New York.

Pralong, D. and Carlile, S. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America*, 100(6):3785–3793.

Puckette, M. (1996). Pure data: Another integrated computer music environment. In *Proc. 2nd Intercollege Computer Music Concerts*, page 37–41, Tachikawa, Japan.

Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., and Wu, X. (2009). Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(6):1124 –1132.

Rabinowitz, W. M., Maxwell, J., Shao, Y., and Wei, M. (1993). Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere. *Presence*, 2(2):125–129.

Rao, D. (2005). Head rotation and sound image localization in the median plane. *Chinese Science Bulletin*, 50(5):412.

Raskar, R., Tan, K.-H., Feris, R. S., Yu, J., and Turk, M. (2004). Non-photorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging. *ACM Trans. Graphics (Proc. SIGGRAPH)*, 23(3):679–688.

Rath, M. and Rocchesso, D. (2005). Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69.

Raykar, V. C., Duraiswami, R., and Yegnanarayana, B. (2005). Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *J. Acoust. Soc. Am.*, 118(1):364–374.

Reichinger, A., Majdak, P., Sablatnig, R., and Maierhofer, S. (2013). Evaluation of methods for optical 3-d scanning of human pinnas. In *3DV-Conference, 2013 International Conference on*, pages 390–397.

Reinkensmeyer, D. J. and Galvez, J. (2007). 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.

Röber, N., Kaminski, U., and Masuch, M. (2007). Ray acoustics using computer graphics technology. In *10th International Conference on Digital Audio Effects (DAFx-07), S*, page 117–124.

Rocchesso, D. and Fontana, F. (2003). *The Sounding Object*. Mondo Estremo.

Rosati, G. (2010). The place of robotics in post-stroke rehabilitation. *Exp. Rev. Med. Devices*, 7(6):753–758.

Rosati, G., Oscari, F., Reinkensmeyer, D. J., Secoli, R., Avanzini, F., Spagnol, S., and Masiero, S. (2011). Improving robotics for neurorehabilitation: Enhancing engagement, performance, and learning with auditory feedback. In *Proc. IEEE 12th Int. Conf. Rehab. Rob. (ICORR 2011)*, pages 341–346, Zurich, Switzerland.

Rosati, G., Oscari, F., Spagnol, S., Avanzini, F., and Masiero, S. (2012). Effect of task-related continuous auditory feedback during learning of tracking motion exercises. *J. Neuroeng. Rehab.*, 9:1–13.

Rosati, G., Rod&#xe0, Antonio, Avanzini, F., and Masiero, S. (2013). On the role of auditory feedback in robot-assisted movement training after stroke: Review of the literature. *Computational Intelligence and Neuroscience*, 2013.

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., and Butterworth, B. (2005). The mental space of pitch height. *Annals of the New York Academy of Sciences*, 1060(1):195–197.

Rusconi, E., Kwan, B., Giordano, B. L., Umiltà, C., and Butterworth, B. (2006). Spatial representation of pitch height: the smarc effect. *Cognition*, 99(2):113–129.

Sakamoto, N., Gotoh, T., and Kimura, Y. (1976). On -out-of-head localization- in headphone listening. *JAES*, 24(9):710–716.

Sankowsky-Rothe, T., Blau, M., Rasumow, E., Mojallal, H., Teschner, M., and Thiele, C. (2011). Prediction of the sound pressure at the ear drum in occluded human ears. *Acta Acustica united with Acustica*, 97(4):656–668.

Satarzadeh, P. (2006). *A Study of Physical and Circuit Models of the Human Pinnae*. PhD thesis, University of California Davis.

Satarzadeh, P., Algazi, R. V., and Duda, R. O. (2007). Physical and filter pinna models based on anthropometry. In *Proc. 122nd Conv. Audio Eng. Soc.*, page 718–737, Vienna, Austria.

Savioja, L., Välimäki, V., and Smith, J. O. (2011). Audio signal processing using graphics processing units. *J. Audio Eng. Soc*, 59(1/2):3–19.

Schärer, Z. and Lindau, A. (2009). Evaluation of equalization methods for binaural signals. In *Audio Engineering Society Convention 126*.

Schiff, W. and Foulke, E. (1982). *Tactual Perception: A Sourcebook*. Cambridge University Press.

Schloerb, D., Lahav, O., Desloge, J., and Srinivasan, M. (2010). BlindAid: virtual environment system for self-reliant trip planning and orientation and mobility training. In *Haptics Symposium, 2010 IEEE*, pages 363 –370.

Secoli, R., Milot, M.-H., Rosati, G., and Reinkensmeyer, D. J. (2011). Effect of visual distraction and auditory feedback on patient effort during robot-assisted movement training after stroke. *J. NeuroEng. Rehab.*, 8(21).

Seeber, B. U. and Fastl, H. (2003). Subjective selection of nonindividual head-related transfer functions. In *Proc. 2003 Int. Conf. Auditory Display (ICAD03)*, page 259–262, Boston, MA, USA.

Seizova-Cajic, T. and Azzi, R. (2010). A visual distracter task during adaptation reduces the proprioceptive movement aftereffect. *Exp. Brain Res.*, 203:213–219.

Serafin, S., Turchet, L., Nordahl, R., Dimitrov, S., Berrezag, A., and Hayward, V. (2010). Identification of virtual grounds using virtual reality haptic shoes and sound synthesis. In *Proceedings of Eurohaptics symposium on Haptic and Audio-Visual Stimuli: Enhancing Experiences and Interaction*, pages 61–70.

Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, 14(1):147–152.

Shaw, E. A. G. (1997). Acoustical features of human ear. In *Binaural and Spatial Hearing in Real and Virtual Environments*, page 25–47. R. H. Gilkey and T. R. Anderson, Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Shaw, E. A. G. and Teranishi, R. (1968). Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *The Journal of the Acoustical Society of America*, 44(1):240–249.

Shin, K. H. and Park, Y. (2008). Enhanced vertical perception through head-related impulse response customization based on pinna response tuning in the median plane. *IEICE Trans. Fundamentals*, E91-A(1):345–356.

Siegel, M. (2002). Tactile display development: the driving-force for tactile sensor development. In *Haptic Virtual Environments and Their Applications, IEEE International Workshop 2002 HAVE*, pages 115–118.

Slater, M., Lotto, B., Arnold, M. M., and Sanchez-Vives, M. V. (2009). How we experience immersive virtual environments: the concept of presence and its measurement. *Anuario de Psicologia*, 40(2):193–210.

So, R. H., Ngan, B., Horner, A., Braasch, J., Blauert, J., and Leung, K. L. (2010). Toward orthogonal non-individualised head-related transfer functions for forward and backward directional sound: cluster analysis and an experimental study. *Ergonomics*, 53(6):767–781. PMID: 20496243.

Song, M.-S., Zhang, C., Florencio, D., and Kang, H.-G. (2010). Personal 3D audio system with loudspeakers. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1600 – 1605.

Spagnol, S. (2012a). Are spectral elevation cues in head-related transfer functions distance-independent? In *Proc. XIX Colloquio di Informatica Musicale (XIX CIM)*, page 192–197, Trieste, Italy.

Spagnol, S. (2012b). *Techniques for Customized Binaural Audio Rendering with Applications to Virtual Rehabilitation*. PhD thesis, Ingegneria dell'Informazione, Scienza e Tecnologia dell'Informazione.

Spagnol, S. and Avanzini, F. (2009). Real-time binaural audio rendering in the near field. In *Proc. 6th Int. Conf. Sound and Music Computing (SMC09)*, pages 201–206, Porto, Portugal.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2010a). 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.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2010b). Fitting pinna-related transfer functions to anthropometry for binaural sound rendering. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP'10)*, page 194–199, Saint-Malo, France.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2010c). Structural modeling of pinna-related transfer functions. In *Proc. 7th Int. Conf. Sound and Music Computing (SMC 2010)*, page 422–428, Barcelona, Spain.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2010d). Structural modeling of pinna-related transfer functions for 3-d sound rendering. In *Proc. XVIII Colloquio di Informatica Musicale (XVIII CIM)*, page 92–101, Torino, Italy.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2012a). Hearing distance: A low-cost model for near-field binaural effects. In *Proc. EUSIPCO 2012 Conf.*, pages 2005–2009, Bucharest, Romania.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2013a). On the relation between pinna reflection patterns and head-related transfer function features. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(3):508–519.

Spagnol, S., Geronazzo, M., Avanzini, F., Oscari, F., and Rosati, G. (2012b). Employing spatial sonification of target motion in tracking exercises. In *Proc. 9th Int. Conf. Sound and Music Computing (SMC 2012)*, pages 85–89, Copenhagen, Denmark.

Spagnol, S., Geronazzo, M., Rocchesso, D., and Avanzini, F. (2013b). Extraction of pinna features for customized binaural audio delivery on mobile devices. In *In Proc. 11th International Conference on Advances in Mobile Computing & Multimedia (MoMM'13)*, pages 514–517, Vienna, Austria.

Spagnol, S., Hiipakka, M., and Pulkki, V. (2011). A single-azimuth pinna-related transfer function database. In *Proc. 14th Int. Conf. Digital Audio Effects (DAFx-11)*, pages 209–212, Paris, France.

Spagnol, S., Rocchesso, D., Geronazzo, M., and Avanzini, F. (2013c). Automatic extraction of pinna edges for binaural audio customization. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP 2013)*, pages 301–306, Pula, Italy.

Spagnol, S., Rocchesso, D., Geronazzo, M., and Avanzini, F. (2013d). Automatic extraction of pinna edges for binaural audio customization. In *Proc. IEEE Int. Work. Multi. Signal Process. (MMSP 2013)*, page 301–306, Pula, Italy.

Speigle, J. M. and Loomis, J. M. (1993). Auditory distance perception by translating observers. In *Proc. IEEE Symp. on Research Frontiers in Virtual Reality*, page 92–99, San Jose, CA, USA.

Srinivasan, M. A. and Basdogan, C. (1997). Haptics in virtual environments: taxonomy, research status, and challenges. *Comput. & Graphics*, 21(4):393–404.

Stamm, M., Altinsoy, M., and Merchel, S. (2011). Influence of the auditory localization direction on the haptic estimation of virtual length. In Cooper, E., Kryssanov, V., Ogawa, H., and Brewster, S., editors, *Haptic and Audio Interaction Design*, volume 6851 of *Lecture Notes in Computer Science*, pages 101–109. Springer Berlin / Heidelberg. 10.1007/978-3-642-22950-3_11.

Stamm, M. and Altinsoy, M. E. (2013). Assessment of Binaural–Proprioceptive interaction in human-machine interfaces. In Blauert, J., editor, *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pages 449–475. Springer Berlin Heidelberg.

Stanley, R. M. (2009). *Measurement and validation of bone-conduction adjustment functions in virtual 3D audio displays*. PhD thesis, Georgia Institute of Technology.

Steinicke, F., Visell, Y., Campos, J., and Lècuyer, A., editors (2013). *Human Walking in Virtual Environments: Perception, Technology, and Applications*. Springer.

Strumillo, P., editor (2011). *Advances in Sound Localization*. InTech.

Strutt, J. W. (1904). On the acoustic shadow of a sphere. *Phil. Trans.*, 203:87–110.

Sunder, K., Tan, E.-L., and Gan, W.-S. (2013). Individualization of binaural synthesis using frontal projection headphones. *Journal of the Audio Engineering Society*, 61(12):989–1000.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., and Iida, K. (2009). 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 2009)*, Zao, Miyagi, Japan.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., and Iida, K. (2010). A simple pinna model for generating head-related transfer functions in the median plane. In *Proc. of 20th Int. Congress on Acoustics (ICA 2010)*, Sydney, Australia.

Teranishi, R. and Shaw, E. A. G. (1968). External-ear acoustic models with simple geometry. *The Journal of the Acoustical Society of America*, 44(1):257–263.

Thomas, J. P. and Shiffrar, M. (2010). I can see you better if i can hear you coming: Action-consistent sounds facilitate the visual detection of human gait. *Journal of vision*, 10(12).

Thurlow, W. R., Mangels, J. W., and Runge, P. S. (1967). Head movements during sound localization. *The Journal of the Acoustical Society of America*, 42(2):489–493.

Thurlow, W. R. and Runge, P. S. (1967). Effect of induced head movements on localization of direction of sounds. *The Journal of the Acoustical Society of America*, 42(2):480–488.

Turchet, L., Nordahl, R., Berrezag, A., Dimitrov, S., Hayward, V., and Serafin, S. (2010a). Audio-haptic physically based simulation of walking on different grounds. In *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, pages 269–273. IEEE Press.

Turchet, L. and Serafin, S. (2011). A preliminary study on sound delivery methods for footstep sounds. In *Proceedings of Digital Audio Effects Conference*, pages 53–58.

Turchet, L. and Serafin, S. (2013). Investigating the amplitude of interactive footstep sounds and soundscape reproduction. *Applied Acoustics*, 74(4):566–574.

Turchet, L., Serafin, S., Dimitrov, S., and Nordahl, R. (2010b). Physically based sound synthesis and control of footsteps sounds. In *Proceedings of Digital Audio Effects Conference*, pages 161–168.

Turchet, L., Serafin, S., and Nordahl, R. (2010c). Examining the role of context in the recognition of walking sounds. In *Proceedings of Sound and Music Computing Conference*.

Turchet, L., Spagnol, S., Geronazzo, M., and Avanzini, F. (2014). Localization of self generated synthetic footstep sounds. *ACM Trans. Applied Perception (submitted for publication)*, -(-):–.

Valimaki, V., Parker, J. D., Savioja, L., Smith, J. O., and Abel, J. S. (2012). Fifty years of artificial reverberation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(5):1421 –1448.

Vanderheiden, G. C. (1989). Nonvisual alternative display techniques for output from graphics-based computers. *Journal of Visual Impairment and Blindness*, 83(8):383–90.

Vaquero, D. A., Raskar, R., Feris, R. S., and Turk, M. (2009). A projector-camera setup for geometry-invariant frequency demultiplexing. In *Proc. IEEE Conf. Comput. Vision Pattern Recog. (CVPR 2009)*, page 2082–2089, Miami, FL, USA.

Visell, Y., Cooperstock, J., Giordano, B., Franinovic, K., Law, A., McAdams, S., Jathal, K., and Fontana, F. (2008). A vibrotactile device for display of virtual ground materials in walking. *Lecture Notes in Computer Science*, 5024:420–426.

Visell, Y., Fontana, F., Giordano, B., Nordahl, R., Serafin, S., and Bresin, R. (2009). Sound design and perception in walking interactions. *International Journal of Human-Computer Studies*, 67(11):947–959.

Vliegen, J. and Van Opstal, A. J. (2004). The influence of duration and level on human sound localization. *J. Acoust. Soc. Am.*, 115(4):1705–1713.

Vorlnder, M. (2007). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Springer Publishing Company, Incorporated, 1st edition.

Walker, A. and Brewster, S. (2000). Spatial audio in small screen device displays. *Pers. Technol.*, 4(2):144–154.

Walker, B. N. and Lindsay, J. (2005). Navigation performance in a virtual environment with bonephones. In *Proceedings of the International Conference on Auditory Display (ICAD2005)*, volume 3, page 1–26.

Walker, B. N. and Lindsay, J. (2006). Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(2):265–278.

Wall, S. and Brewster, S. (2006). Feeling what you hear: tactile feedback for navigation of audio graphs. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, page 1123–1132, Montr&#233;al, Qu&#233;bec, Canada. ACM.

Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in cognitive sciences*, 7(11):483–488.

Watanabe, K., Ozawa, K., Iwaya, Y., Suzuki, Y., and Aso, K. (2007). Estimation of interaural level difference based on anthropometry and its effect on sound localization. *J. Acoust. Soc. Am.*, 122(5):2832–2841.

Watkins, A. J. (1978). Psychoacoustical aspects of synthesized vertical locale cues. *J. Acoust. Soc. Am.*, 63(4):1152–1165.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123.

Wiener, J. M., Büchner, S. J., and Hölscher, C. (2009). Taxonomy of human wayfinding tasks: A knowledge-based approach. *Spatial Cognition & Computation*, 9(2):152–165.

Wightman, F. L. and Kistler, D. J. (1989a). Headphone simulation of free-field listening. i: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867.

Wightman, F. L. and Kistler, D. J. (1989b). Headphone simulation of free-field listening. II: psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878.

Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661.

Wightman, F. L. and Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853.

Wilska, A. (2010). *Studies on Directional Hearing*. English translation, Aalto University School of Science and Technology, Department of Signal Processing and Acoustics. PhD thesis originally published in German as "Untersuchungen über das Richtungshören", University of Helsinki, 1938.

Woodworth, R. S. and Schlosberg, H. (1954). *Experimental Psychology*. Holt, Rinehard and Winston, NY, USA.

Wright, D., Hebrank, J. H., and Wilson, B. (1974). Pinna reflections as cues for localization. *J. Acoust. Soc. Am.*, 56(3):957–962.

Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display*. J ROSS PUB Incorporated.

Yost, W. A. and Hafter, E. R. (1987). Lateralization. In Yost, W. A. and Gourevitch, G., editors, *Directional Hearing*, Proceedings in Life Sciences, pages 49–84. Springer US.

Yu, G.-z., Xie, B.-s., and Chen, X.-x. (2012). Analysis on minimum-phase characteristics of measured head-related transfer functions affected by sound source responses. *Computers & Electrical Engineering*, 38(1):45–51.

Yu, W. and Brewster, S. (2002). Comparing two haptic interfaces for multimodal graph rendering. In *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2002. HAPTICS 2002. Proceedings. 10th Symposium on*, page 3–9.

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, 111(4):1832–1846.

Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420.

Zinchenko, V. and Lomov, B. (1960). The functions of hand and eye movements in the process of perception. *Problems of Psychology*, 1:12–26.

Zolzer, U. (2011). *DAFX: digital audio effects, 2nd Edition*. Wiley, Chichester, West Sussex, U.K., 2 edition. edited by Udo Zölzer.; Includes bibliographical references and index.

Zotkin, D., Duraiswami, R., and Davis, L. (2004). Rendering localized spatial audio in a virtual auditory space. *Multimedia, IEEE Transactions on*, 6(4):553 – 564.

Zotkin, D., Duraiswami, R., and Gumerov, N. (2010). Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):2–16.