UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Felipe Leonel Grijalva Arevalo

Manifold Learning for Spatial Audio Rendering

Aprendizado de Variedades para a Síntese de Áudio Espacial

Campinas
2018

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Felipe Leonel Grijalva Arevalo

Manifold Learning for Spatial Audio Rendering

Aprendizado de Variedades para a Síntese de Áudio Espacial

Thesis presented to the School of Electrical and Computer Engineering of the University of Campinas in partial fulfillment of the requirements for obtaining Doctor degree in Electrical Engineering, in the area of Computer Engineering.

Tese apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica, na área de Engenharia de Computação.

Supervisor/*Orientador*: Prof. Dr. Luiz César Martini
Co-Supervisor/*Co-orientador*: Prof. Dr. Bruno Sanches Masiero

Este exemplar corresponde à versão final da tese defendida pelo aluno Felipe Leonel Grijalva Arevalo, e orientada pelo Prof. Dr. Luiz César Martini

_____

Campinas
2018

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2014/14630-9; CAPES
ORCID: https://orcid.org/0000-0001-6162-3429

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

G878m

Grijalva Arevalo, Felipe Leonel, 1984-
Manifold learning for spatial audio rendering / Felipe Leonel Grijalva Arevalo. – Campinas, SP : [s.n.], 2018.

Orientador: Luiz César Martini.
Coorientador: Bruno Sanches Masiero.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Percepção auditiva. 2. Som. I. Martini, Luiz César, 1952-. II. Masiero, Bruno Sanches, 1981-. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Aprendizado de variedades para a síntese de áudio espacial
Palavras-chave em inglês:
Auditory perception
Sound
Área de concentração: Engenharia de Computação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora:
Luiz César Martini [Orientador]
Tiago Fernandes Tavares
Levy Boccato
Luiz Wagner Pereira Biscainho
Laurindo de Sousa Britto Neto
Data de defesa: 05-07-2018
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidato: Felipe Leonel Grijalva Arevalo **RA:** 134117
Data da Defesa: 05 de julho de 2018

Título da Tese: Manifold Learning for Spatial Audio Rendering (*Aprendizado de Variedades para a Síntese de Áudio Espacial*)

Prof. Dr. Luiz César Martini (Presidente, FEEC/UNICAMP)
Prof. Dr. Tiago Fernandes Tavares (FEEC/UNICAMP)
Prof. Dr. Levy Boccato (FEEC/UNICAMP)
Prof. Dr. Luiz Wagner Pereira Biscainho (UFRJ)
Prof. Dr. Laurindo de Sousa Britto Neto (UFPI)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no processo de vida acadêmica do aluno.

Acknowledgments

I would like to thank São Paulo Research Foundation (FAPESP) under Grant #2014/14630-9 and CAPES for the financial support[1].

[1] Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of FAPESP or CAPES.

Abstract

The objective of binaurally rendered spatial audio is to simulate a sound source in arbitrary spatial locations through the *Head-Related Transfer Functions* (HRTFs). HRTFs model the direction-dependent influence of ears, head, and torso on the incident sound field. When an audio source is filtered through a pair of HRTFs (one for each ear), a listener is capable of perceiving a sound as though it were reproduced at a specific location in space.

Inspired by our successful results building a practical face recognition application aimed at visually impaired people that uses a spatial audio user interface, in this work we have deepened our research to address several scientific aspects of spatial audio. In this context, this thesis explores the incorporation of spatial audio prior knowledge using a novel nonlinear HRTF representation based on manifold learning, which tackles three major challenges of broad interest among the spatial audio community: HRTF personalization, HRTF interpolation, and human sound localization improvement. Exploring manifold learning for spatial audio is based on the assumption that the data (i.e. the HRTFs) lies on a low-dimensional manifold. This assumption has also been of interest among researchers in computational neuroscience, who argue that manifolds are crucial for understanding the underlying nonlinear relationships of perception in the brain.

For all of our contributions using manifold learning, the construction of a single manifold across subjects through an Inter-subject Graph (ISG) has proven to lead to a powerful HRTF representation capable of incorporating prior knowledge of HRTFs and capturing the underlying factors of spatial hearing. Moreover, the use of our ISG to construct a single manifold offers the advantage of employing information from other individuals to improve the overall performance of the techniques herein proposed. The results show that our ISG-based techniques outperform other linear and nonlinear methods in tackling the spatial audio challenges addressed by this thesis.

Key-words: spatial audio; manifold learning; Head-Related Transfer Functions; binaural technique.

Resumo

O objetivo do áudio espacial gerado com a técnica binaural é simular uma fonte sonora em localizações espaciais arbitrárias através das Funções de Transferência Relativas à Cabeça (HRTFs) ou também chamadas de Funções de Transferência Anatômicas. As HRTFs modelam a interação entre uma fonte sonora e a antropometria de uma pessoa (e.g., cabeça, torso e orelhas). Se filtrarmos uma fonte de áudio através de um par de HRTFs (uma para cada orelha), o som virtual resultante parece originar-se de uma localização espacial específica.

Inspirados em nossos resultados bem sucedidos construindo uma aplicação prática de reconhecimento facial voltada para pessoas com deficiência visual que usa uma interface de usuário baseada em áudio espacial, neste trabalho aprofondamos nossa pesquisa para abordar vários aspectos científicos do áudio espacial. Neste contexto, esta tese analisa como incorporar conhecimentos prévios do áudio espacial usando uma nova representação não-linear das HRTFs baseada no aprendizado de variedades para enfrentar vários desafios de amplo interesse na comunidade do áudio espacial, como a personalização de HRTFs, a interpolação de HRTFs e a melhoria da localização de fontes sonoras. O uso do aprendizado de variedades para áudio espacial baseia-se no pressuposto de que os dados (i.e., as HRTFs) situam-se em uma variedade de baixa dimensão. Esta suposição também tem sido de grande interesse entre pesquisadores em neurociência computacional, que argumentam que as variedades são cruciais para entender as relações não lineares subjacentes à percepção no cérebro.

Para todas as nossas contribuições usando o aprendizado de variedades, a construção de uma única variedade entre os sujeitos através de um grafo Inter-sujeito (Inter-subject graph, ISG) revelou-se como uma poderosa representação das HRTFs capaz de incorporar conhecimento prévio destas e capturar seus fatores subjacentes. Além disso, a vantagem de construir uma única variedade usando o nosso ISG é o uso de informações de outros indivíduos para melhorar o desempenho geral das técnicas aqui propostas. Os resultados mostram que nossas técnicas baseadas no ISG superam outros métodos lineares e não-lineares nos desafios de áudio espacial abordados por esta tese.

Palavras-chave: Áudio Espacial, Aprendizado de Variedades, Funções de Transferência Anatômica, Técnica Binaural.

List of Figures

List of Tables

List of Publications

1. F. Grijalva, L. Martini, D. Florencio and S. Goldenstein, "A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 559-570, March 2016. doi: 10.1109/TASLP.2016.2517565

2. F. Grijalva, L. C. Martini, D. Florencio and S. Goldenstein, "Interpolation of Head-Related Transfer Functions Using Manifold Learning," in *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 221-225, Feb. 2017. doi: 10.1109/LSP.2017.2648794

3. F. Grijalva, L. C. Martini, B. Masiero and S. Goldenstein, "A Recommender System for Improving Median Plane Sound Localization Performance Based on a Nonlinear Representation of HRTFs," in IEEE Access, vol. 6, pp. 24829-24836, 2018. doi: 10.1109/ACCESS.2018.2832645

4. L. Neto, F. Grijalva, V. Maike, L. Martini, D. Florencio, M. Baranauskas, A. Rocha and S. Goldenstein, "A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users," in *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 52-64, Feb. 2017. doi: 10.1109/THMS.2016.2604367

Contents

Chapter 1

Introduction

Motivation

The objective of binaurally rendered spatial audio is to simulate a sound source in arbitrary spatial locations through the *Head-Related Transfer Functions* (HRTFs). HRTFs model the spectral filtering of a sound source caused by the head, pinna (i.e. the outer part of the ear) and torso before it reaches the eardrum. By filtering a sound source with these filters, a listener is capable of perceiving a sound as though it were reproduced at a specific location in space [1].

Spatial audio has a wide range of applications from hearing aids and entertainment (e.g. home theaters, video games) to virtual reality [2] (e.g. Oculus Rift™, Google Glass™, air traffic controllers [3]). In fact, as virtual reality applications become more important, there is increasing research effort in the spatial audio research. In this sense, several works has proposed the use of spatial audio as natural user interface for sensory substitution and augmented reality prototypes aimed at visually impaired people [4, 5, 6].

It is precisely in this type of application that the project "Vision for the blind: translating 3D Visual Concepts into 3D Auditory Clues" focused on [1]. The goal of this project was to construct and validate a complete proof-of-concept assistive device for the blind. This device uses computer vision algorithms to extract high-level 3D information from a Microsoft Kinect Sensor and communicates this information to the visually impaired user using 3D audio to exploit the inherent spatial sense of the auditory system.

Based on the successful results of this project, in this thesis we decided to deepen our knowledge of spatial audio beyond the aforementioned application. In this context, we have shifted our attention to three important challenges of broad interest among the spatial audio community that we describe in the next section.

[1]Project approved through the Microsoft Research/Fapesp cooperation agreement under Grant 2012/50468-6

Challenges

A significant challenge in the implementation of 3D sound systems is the wide difference of spectral features of HRTFs among individuals, which vary depending on their anatomical structure (e.g. head dimensions, pinna shape and size). As a consequence and given that non-individualized HRTFs hinder the listener's sound localization capabilities [7], HRTFs should ideally be measured for each person. However, since HRTF measurement is a complex and non-scalable task that requires specialized and costly equipment (e.g. an anechoic or semi-anechoic chamber, and a loudspeaker array), it is necessary to personalize HRTFs to guarantee high quality 3D sound perception while avoiding such measurements.

Although HRTFs are continuous functions of sound source location, in practice, they are measured only at discrete positions in space [8]. Under these circumstances, HRTF interpolation techniques might be applied to achieve high spatial resolution with as few measurements as possible. A small set of measurements might considerably reduce the overall HRTF measurement time and equipment costs, and minimize undesired audio artifacts (e.g. clicks) [9].

On the other hand, when direct HRTF measurement is not an option, it is possible to use a set of HRTFs from another individual at the expense of losing some localization accuracy. Under this scenario, to improve sound localization performance, a listener might select, through a series of listening tests, the best HRTFs from a database composed by other subjects' HRTFs. However, this procedure might be time-consuming and tiring for the participants, especially when the database is composed by HRTFS from a large number of subjects. In order to speed up this process, it is desirable to find a way to reduce the number of listening tests while still minimizing the localization error.

Moreover, since HRTFs vary in a complex and non-intuitive way, especially at high frequencies, it is more suitable to use a compact representation of HRTFs obtained through, e.g, dimensionality reduction methods such as *Principal Component Analysis* (PCA) [10]. In contrast to PCA and similar linear methods, in *nonlinear dimensionality reduction* techniques, also known as *manifold learning*, it is assumed that the data (i.e. herein the HRTFs) lie on a low-dimensional manifold. This assumption has also been of interest among researchers in computational neuroscience, who argue that manifolds are crucial for discovering and understanding the underlying nonlinear relationships of perception in the brain [11].

To gain some insight into manifold learning in the context of spatial audio, Figure 1.1 presents an intuitive example where an HRTF might be considered as a point in a nonlinear high-dimensional space. As the elevation increases smoothly, it intrinsically describes a one-dimensional manifold (see the red line) but embedded in a higher dimensional space (in this example a 3D space). Intuitively, a one-dimensional manifold is an appropriate representation in this case since the perceived sound depends solely on the elevation.

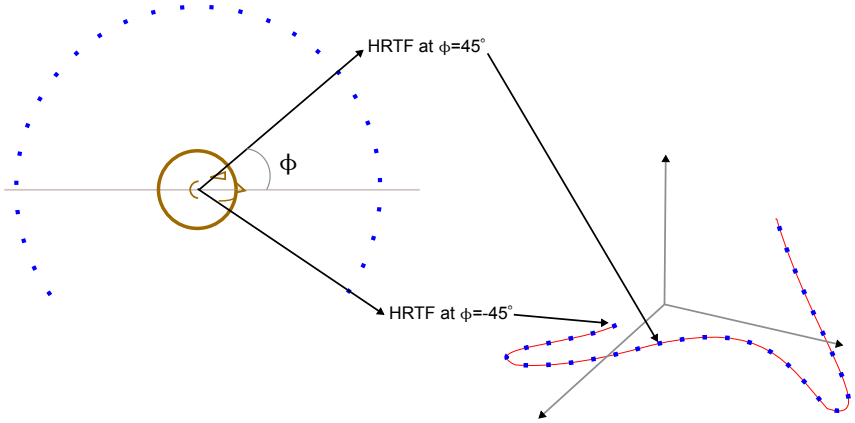


Figure 1.1: An intuitive example of a one-dimensional manifold embedded in a higher dimensional space. Figure inspired by [12].

Now, observe in the same figure the two points corresponding to HRTFs at $-45°$ and $45°$ respectively. Linear dimensionality reduction methods such as PCA attempt to preserve the euclidean distance between these two data points, which are very close to each other despite the fact that their two corresponding elevations are perceptually far apart (at i.e. at $-45°$ and $45°$ respectively). In contrast, manifold learning techniques aim at preserving the distance over the manifold (i.e. the distance measured over the red line, also known as the geodesic distance). Using the geodesic distance, the two data points remain far away which is consistent with the fact that they are also perceptually far apart.

Aims of this thesis

To address the aforementioned challenges in implementing 3D sound systems, in this thesis [2] we propose a novel nonlinear representation of HRTFs to:

- personalize HRTFs from anthropometric features.
- interpolate HRTFs.
- improve human sound localization performance by minimizing the number of perceptual tests.

Summary of Contributions and Thesis Outline

This thesis is presented as a compilation of four papers published in technical journals, which are appended in Chapters 3 through 6. The contents presented here are a faithful representation of the published materials. The corresponding permissions granted to reproduce the published papers are attached in Appendix A. The remainder of the thesis is structured as follows:

[2]This thesis was supported by FAPESP under Grant 2014/14630-9 and CAPES.

Chapter 2 provides an overview of human spatial hearing, HRTFs, and manifold learning.

Chapter 3 contains the first contribution of this work, entitled "A Wearable Face Recognition System Built With the Kinect to Aid the Visually Impaired User", which was published in the *IEEE Transactions on Human-Machine Systems*. This work was one of the main results of the "Vision for the blind: translating 3D Visual Concepts into 3D Auditory Clues" project that inspired us to deepen our knowledge in spatial audio. This wearable system to aid the blind and people with low vision was named Kinect Vision Blind (KVB). The KVB's development was a multi-year and multi-disciplinary project[3] involving several researchers, Master's degree students and Ph.D candidates from different research areas. While my main contribution to this project focused on the development of the spatial audio user interface, I also actively participated in some phases of the experiments' implementation related to the face recognition algorithms.

Specifically, we present a real-time wearable face recognition system aimed at individuals with visual disabilities. The system uses a Microsoft Kinect sensor as a wearable device, performs face detection and uses temporal coherence along with a simple biometric procedure to generate sounds associated with the identified person, virtualized at his/her estimated 3D location. The system also exploits the inherent spatial sense of the auditory system by using spatial audio to convey the location of a face in the environment. Conveying the directional location of a face in the environment using 3-D Audio proved to be an efficient feedback that does not overwhelm the person's auditory sense. Our main contribution herein is to show that it is possible to have a high-accuracy face recognition system in this limited-scope application with a simple, yet computationally efficient, approach that allows real-time performance on wearable devices by taking advantage of depth information and temporal coherence.

Chapter 4 contains the second contribution of this work, entitled "A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features", which was published in the *IEEE/ACM Transactions On Audio, Speech, And Language Processing*. In this paper, we introduce a new HRTF personalization method using a domain-specific manifold learning technique, denominated Isomap, in both azimuth and elevation angles with a single regression model that does not break the inherent multifactor nature of HRTFs. Our main contribution is the incorporation of important prior knowledge of spatial audio in Isomap's manifold that aims at exploiting the correlations existing among HRTFs, using a recently developed spectral decomposition that isolates the subject-dependent cues of HTRFs. Our findings show that introducing this prior knowledge is a powerful way to capture the underlying factors of spatial hearing.

[3]This project was supported by the Microsoft-FAPESP agreement under grant 2012/50468-6.

Chapter 5 presents the third contribution of this work, entitled "Interpolation of Head-Related Transfer Functions using Manifold Learning", which was published in the *IEEE Signal Processing Letters*. This work introduces a new HRTF interpolation method using Isomap, a domain-specific manifold learning technique. Our main contribution is the novel use of Isomap to address HRTF interpolation, where we apply important prior knowledge of spatial audio to construct a single Isomap's manifold that can represent simultaneously all subjects in all directions, taking advantage of the existing correlations among HRTFs. Unlike our contribution in Chapter 4, here we deal with HRTF interpolation by introducing a different and novel way of constructing the Isomap's graph. We also introduce a distortion metric based on variance, which is a more suited distance criterion to evaluate this domain's results. Our findings show that a single manifold representation has proven to be a powerful way to allow measured HRTFs from different subjects to contribute to reconstructing the HRTFs for new directions. Finally, our results suggest that a small number of spatial measurements capture most HRTFs acoustical properties.

Chapter 6 describes the fourth contribution of this work, entitled "A Recommender System For Improving Median Plane Sound Localization Performance Based on a Nonlinear Representation of HRTFs", which was published in the technical journal *IEEE Access*. This work addresses the problem of improving sound localization performance in the median plane and, simultaneously, minimizing the number of listening tests performed by a subject. This allows for shorter psychoacoustics experiments and avoids the difficult endeavor of acquiring HRTF data for new users – all of which are of broad interest among the spatial audio community. With this aim, we introduce an HRTF recommender system to improve median plane sound localization performance using manifold learning. Our first contribution is the novel use of manifold learning to address this problem, where we apply important prior knowledge of spatial audio to construct a single manifold that can represent simultaneously all subjects in all directions, taking advantage of the existing correlations among HRTFs. The results show that our method is capable of recommending a set of HRTFs that improves the median plane localization performance with respect to the mean localization performance using non-individualized HRTFs. Moreover, our recommender system finds the best HRTF set by mixing HRTFs from multiple subjects without having to perform listening tests with all individuals' HRTFs from a database.

Chapter 7 provides a discussion of all the contributions presented in Chapters 3 to 6 as a whole.

Chapter 8 presents the set of conclusions of this thesis and recommends possible directions for future research.

Chapter 2

Theoretical Background

In this chapter, we present the basic concepts spatial of the two key components of this thesis: audio and manifold learning. In Section 2.1, we formally introduce spatial audio and the Head-Related Transfer Functions. Finally, in Section 2.2, we intuitively describe manifold learning.

Spatial Audio and Head-Related Transfer Functions

The term spatial audio or 3D audio refers to the set of techniques that model the direction-dependent influence of ears, head, and torso on the incident sound field using digital filters to generate virtual sounds. If we filter an audio source through them, the resulting virtual sound seems to originate from a specific spatial location [1]. The main objective of 3D audio systems is to manipulate a listener's spatial audio perception by taking into account engineering specifications as well as psychoacoustic considerations [1]. Among the different spatial audio generation (e.g Panning, Binaural [13], and Sound Field Synthesis [14, 15]) and reproduction techniques (i.e. loudspeakers and headphones), this thesis focuses on binaurally rendered spatial audio through headphones for a single sound source [13], in which the key components are the *Head-Related Transfer Functions* (HRTFs).

In this section, we will describe the basic concepts of spatial audio perception and HRTFs. In Section 2.1.1, we present the common coordinate systems used in spatial audio. Next, in Section 2.1.2, we introduce the main cues that determine our directional perception of sound. It is worth mentioning that distance perception is beyond the scope of this study. In Section 2.1.3, we analyze several experimental results regarding the human sound source localization performance. Finally, we formally define the HRTFs and their properties in Section 2.1.4.

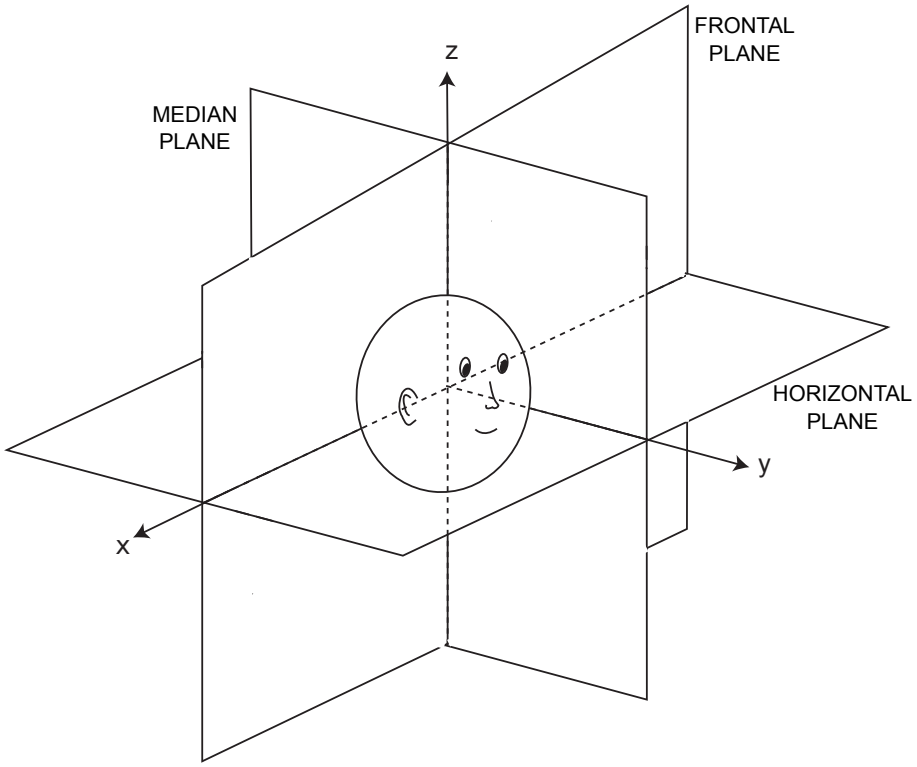


Figure 2.1: References planes

Coordinate systems

In spatial audio, we specify the sound source position with respect to the center of the listener's head. Figure 2.1 shows three important reference planes we define to locate a source:

- Median plane: y-z plane
- Frontal plane: x-z plane
- Horizontal or transversal plane: x-y plane

In the same figure, *the interaural axis* is the line segment connecting the two ears, and the *ipsilateral and contralateral* ears are the closest and the most distant ear with respect to the sound source, respectively.

There are two widely used coordinate systems in spatial audio [16]: the spherical coordinate system and the interaural polar coordinate system.

Figure 2.2 shows the spherical coordinate system where the sound source position is given by (r, Θ, Φ). The distance $0 \leq r \leq \infty$ is measured with respect to the origin. The angle $-90° \leq \Phi \leq +90°$ represents the elevation where $-90°$ and $+90°$ correspond to positions above and below the listener respectively, and $0°$ corresponds to positions in the horizontal plane. The angle $0° \leq \Theta < 360°$ represents the azimuth where $0°$, $90°$, $180°$ and $270°$ correspond to positions at the front, right, behind and left in the horizontal

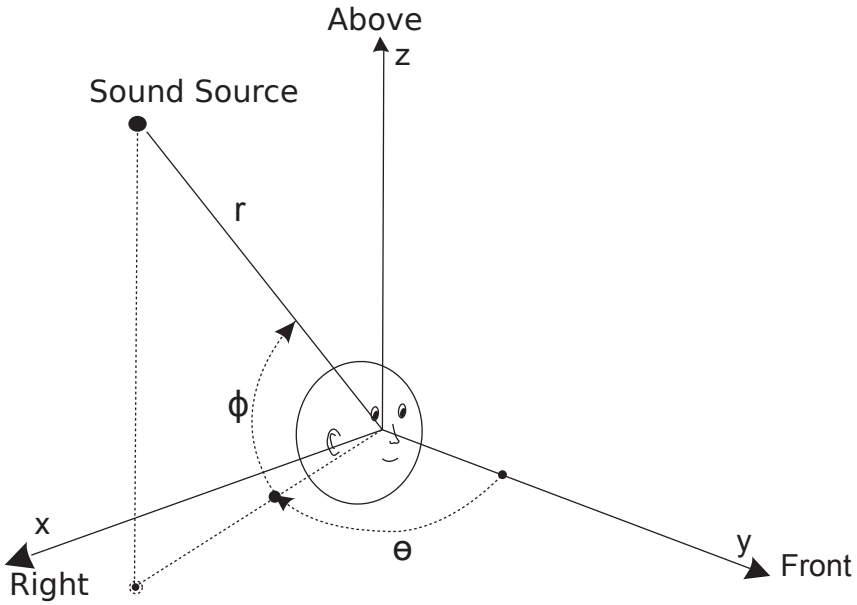


Figure 2.2: Spherical coordinate system. Adapted from [16].

plane, respectively. The azimuth angle might also lie in the interval $-180° < \Theta \leq +180°$ where 0°, 90°, 180° and −90° correspond to positions at the front, right, behind and left in the horizontal plane, respectively.

Figure 2.3 shows the interaural polar coordinate system. A sound source is defined by (r, θ, ϕ) where the distance with respect to the origin is $0 \leq r \leq \infty$. The azimuth and elevation angles lie in the intervals $-90° < \theta \leq +90°$ and $-90° < \phi \leq +270°$, respectively. In this system, the coordinates

$$(\theta, \phi) = (0°, 0°), (0°, 90°), (0°, 180°), (0°, 270°), (90°, 0°), (-90°, 0°)$$

correspond to sources at front, above, behind, below, right, and left respectively.

Directional Localization Cues

Our directional perception of sound integrates several cues that might fall into four categories: binaural, spectral, dynamic and cognitive cues.

Binaural Cues

The binaural cues are the most important cues to localize a sound source in the horizontal plane. They involve the relative differences between the waves arriving at both ears. These differences were described by Lord Rayleigh's duplex theory of sound localization [17] as the *Interaural Time Difference* (ITD) and the *Interaural Level Difference* (ILD).

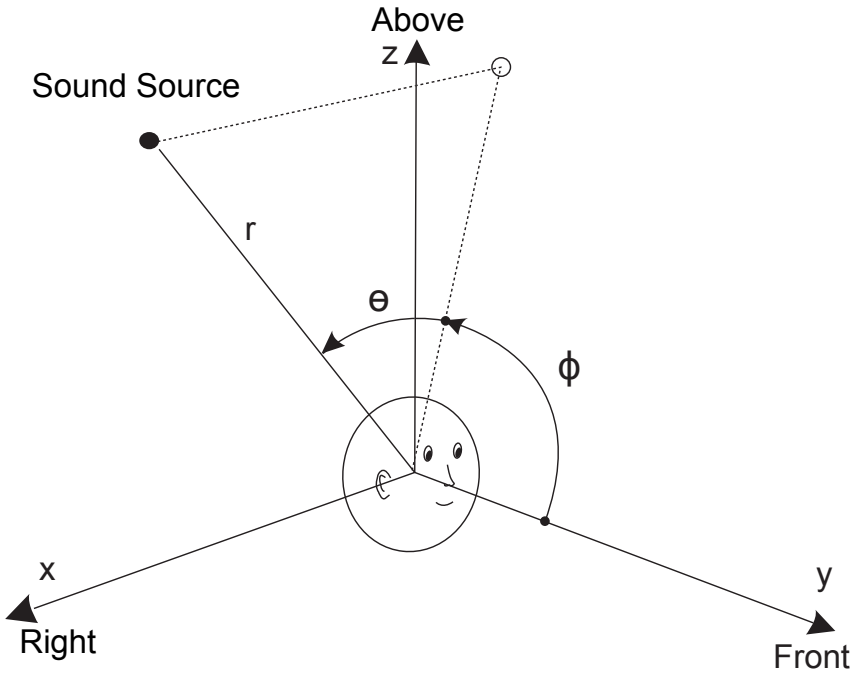


Figure 2.3: Interaural polar coordinate system. Adapted from [16].

Interaural Time Difference (ITD): the ITD refers to the difference in arrival time of a sound between two ears. It depends on both the frequency and the direction [18]. The ITD is the most important localization cue for frequencies below 700 Hz since in this band the head dimensions are smaller than the sound source wavelength. This allows the auditory system to detect the phase delay differences without confusion [16]. Depending on the type of stimulus, the human ear can distinguish ITDs from 0.005 to 1.5 ms [1]. On the other hand, at high frequencies (i.e. from 1.5 kHz), the ITD becomes ambiguous because the perception of the lateral position is no longer proportional to the perceived phase difference. Even so, the ITD at high-frequencies is considered a secondary cue since the auditory system is able to extract the interaural delay differences from the envelopes of the sound waves [19].

Interaural Level Difference (ILD): When an audio source moves away from the median plane, the sound pressure in the farthest ear (contralateral with respect to the audio source) is attenuated due to the head shadow. This phenomenon produces a sound pressure difference between both ears, called the ILD, which is especially noticeable at frequencies larger than 1.5 kHz when the wavelength becomes smaller than the diameter of the head. The ILD reaches values between 10 and 35 dB for frequencies from 3 to 10 kHz respectively, allowing to detect the position of the audio source at frequencies in which the ITD is ambiguous [20, 1].

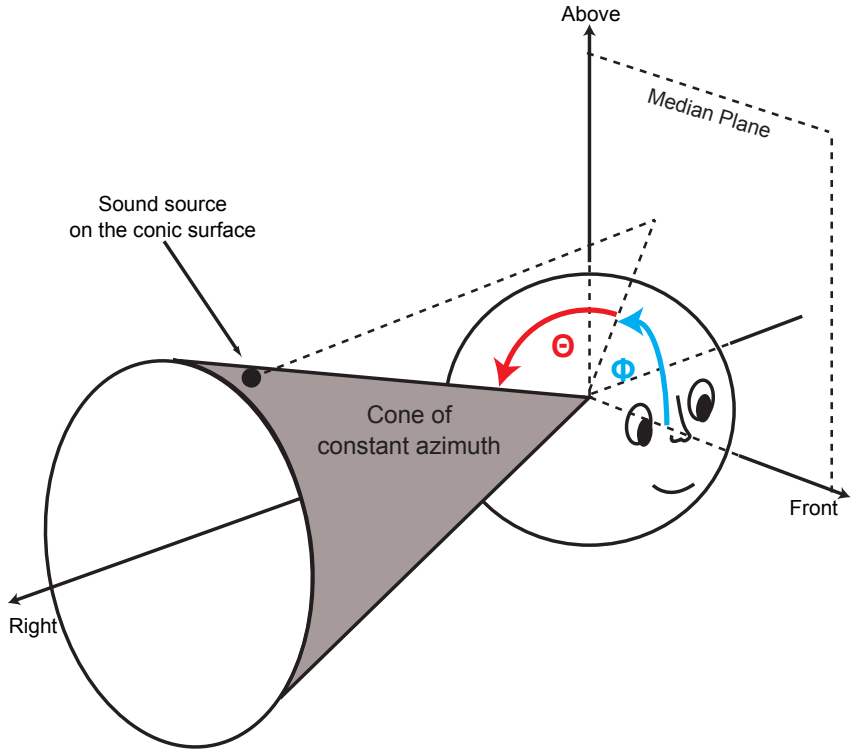


Figure 2.4: By keeping the azimuth constant in the interaural polar coordinate system, a cone is formed on whose surface the ILD and ITD values are identical. This cone is known as the cone of confusion.

Dynamic Cues

The binaural cues described so far to locate sound sources (i.e. the ILD and ITD) might become ambiguous since, in theory, it is possible to create identical ILDs and ITDs for different positions of an audio source. In fact, equal ITD and ILD values may exist for an audio source on a conical surface. In Figure 2.4, note that by keeping the azimuth constant in the interaural polar coordinate system, a cone is formed on whose surface the ILD and ITD values are theoretically identical. Even though the ITDs and ILDs are seldom identical for a real person, when they are very close for two different locations, ambiguities might arise [1].

The aforementioned cone is called the *cone of confusion* because it produces front-back reversals or up-down reversals. Front-back reversals refer to the perception that a sound projected in a position ahead or behind the subject is behind or ahead of the same, respectively. The same concept applies for up-down reversals. One way to minimize these ambiguities is through dynamic cues like head movements. Several studies have demonstrated the effectiveness of head movements to resolve front-back and up-down confusions [21, 22].

Spectral Cues

The spectral or monoaural cues are caused by anatomical characteristics such as the outer ear (i.e. the pinna), the head and the torso. Among them, the geometry of the pinna is the most important especially at frequencies above 3 kHz when its size is comparable to the wavelength of the source [23]. These spectral differences are used by the auditory system to resolve the cone of confusion.

Both binaural and monoaural factors are unique characteristics of each person's anatomy. These anatomical differences are reflected in the *Head-Related Transfer Function* (HRTF). A pair of these functions, one for each ear, uniquely defines the position of an audio source in space. In general, HRTFs are non-transferable among individuals since nonindividualized HRTFs increase the localization error [7]. In Section 2.1.4, we will talk about HRTFs in more detail.

Cognitive Cues

Beyond the cues described so far, there are cognitive cues such as familiarity with the sound source and visual cues that contribute to the directional perception process.

Familiarity refers to prior knowledge of the type of the sound source [1]. If a source is associated with a particular position after repeated experiments (e.g. speech), the sound is more simple to simulate. For instance, it is easier to simulate the sound of an airplane above us than to simulate the same sound from below.

On the other hand, there are visual cues such as the ventriloquism effect [24] whereby we perceive the apparent position of a sound as though it originated from a correlated visual object [1]. For example, when we watch a film in a movie theater, the actors' voices seem to originate from their mouths although the actual sound originates from the loudspeakers.

Human Sound Source Localization Performance and Localization Blur

Human sound source localization accuracy depends on the frequency band and the position of the stimuli [25]. In this section, we will briefly describe the results of several experiments performed on human listeners to determine the human performance for localizing sound sources. With this aim, Blauert [25, p. 38] proposed the concept of localization blur as "the amount of displacement of the position of the source that is recognized by 50% of experimental subjects as a change in the position of the auditory event"

Figure 2.5 shows the localization blur for four directions (0°, 90°, 180° and 270°) in the horizontal plane calculated by Blauert [25] from the results of the experiments under anechoic conditions performed on two experiments with 600 and 900 individuals using

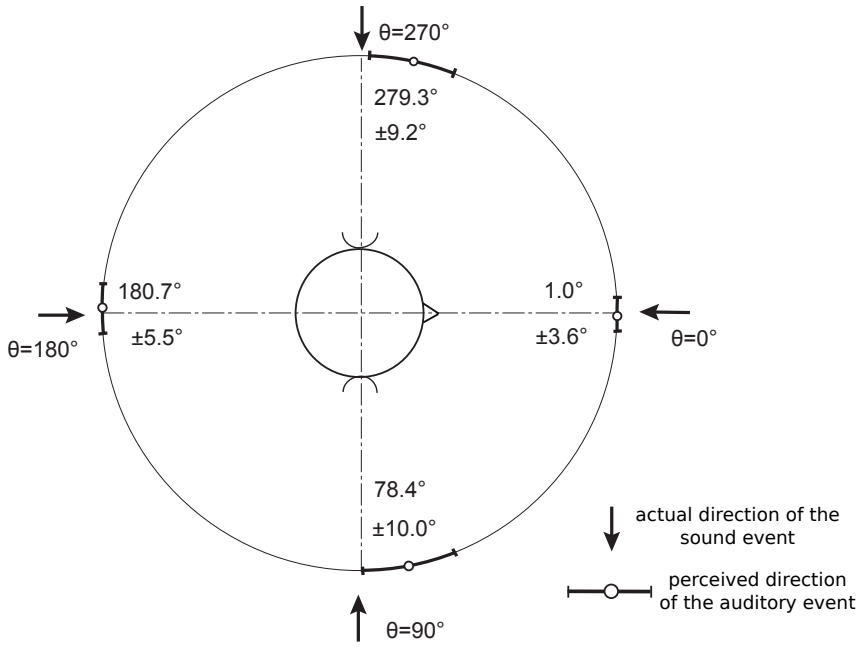


Figure 2.5: Localization blur in the horizontal plane for a 100 ms of white noise audio source. Arrows represent the actual direction of the sound event, the circles the average position reported by the listeners and the segments the perceived direction of the auditory event. Adapted from [25]

100 ms of white noise audio source. The smallest localization blur is roughly ±4° for the front direction (i.e., 0° azimuth) and the largest is about ±10° for lateral positions. Blauert also shows that the localization error varies according to the type of source, but the minimum localization blur is always in the frontal direction (ie 0° azimuth) reaching the lowest value for stimuli such as clicks (0.75°) and speech (1.5°).

On the other hand, the localization performance in the median plane is smaller than in the horizontal plane. In Figure 2.6, we can observe the results of the experiments under anechoic conditions performed in seven individuals in the median plane with a speech signal as input stimulus [25]. The smallest localization blur occurs at positions at the front or at small elevations where the location error is around ±10°. The localization blur and the localization error increase as the elevation increases until it reaches maximum values in the posterior hemisphere (i.e. greater than ±15°).

Head-Related Transfer Functions

The Head-Related Transfer Functions (HRTFs) are key components to analyze and synthesize binaurally rendered spatial audio. When filtering a sound through a pair of HRTFs, one for each ear, it is possible to place a sound in specific positions in space. In this section, we will describe the main characteristics of the HRTFs and their time counterpart, known as HRIRs (Head-Related Impulse response).

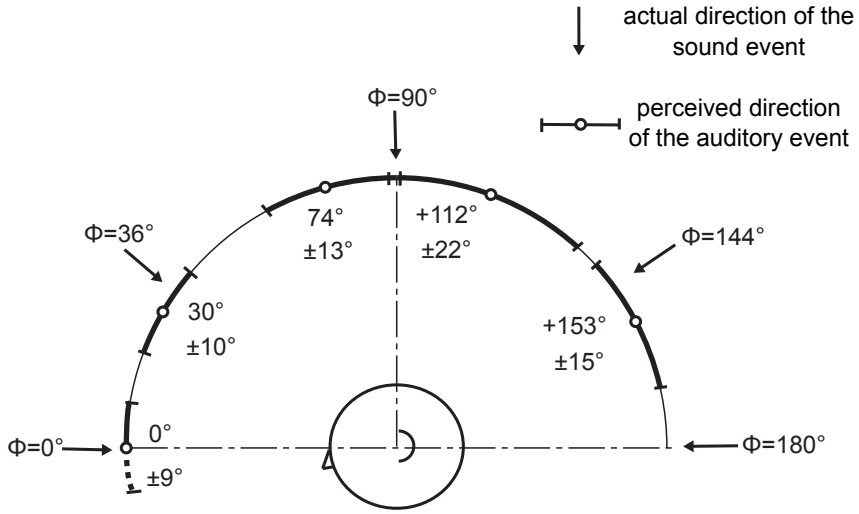


Figure 2.6: Localization blur in the median plane for a speech signal. Arrows represent the actual direction of the sound event, the circles the average position reported by the listeners and the segments the perceived direction of the auditory event. Adapted from [25]

Definition

The sound emitted by an audio source in free field reaches the two ears after interacting with the anatomical features of the person (i.e. head, torso and pinna). The resultant signal contains several cues described in Section 2.1.2 such as the ITD, the ILD and spectral cues, which together are modeled through the HRTFs. A pair of HRTFs for the left and right ear, H_L and H_R respectively, is defined by

$$\begin{aligned} H_L\left(r,\theta,\phi,f,a\right) &= \frac{P_L\left(r,\theta,\phi,f,a\right)}{P_0\left(r,f\right)}, \\ H_R\left(r,\theta,\phi,f,a\right) &= \frac{P_R\left(r,\theta,\phi,f,a\right)}{P_0\left(r,f\right)}, \end{aligned} \tag{2.1}$$

where, P_L and P_R represent the sound pressures in the frequency domain at the left and right ear, respectively, and P_0 represent the free field sound pressure in the frequency domain at the center of the head with the head absent [16]. Observe that P_0 must be generated by the same source with the same power as in the P_L and P_R measurements. The variable a depends on the anatomical features of each person and is normally represented by a set of anthropometric measurements from the head, torso and pinna. If the distance is roughly $r > 1$ m, the HRTFs are independent on the distance and are called *far-field HRTFs*. Otherwise, the HRTFs are distance-dependent and are called *near-field HRTFs*. As stated in Section 2.1 distance cues and therefore near-field HRTFs are beyond the scope of this study. Hence, when we refer to HRTFs we are referring to far-field HRTFs.

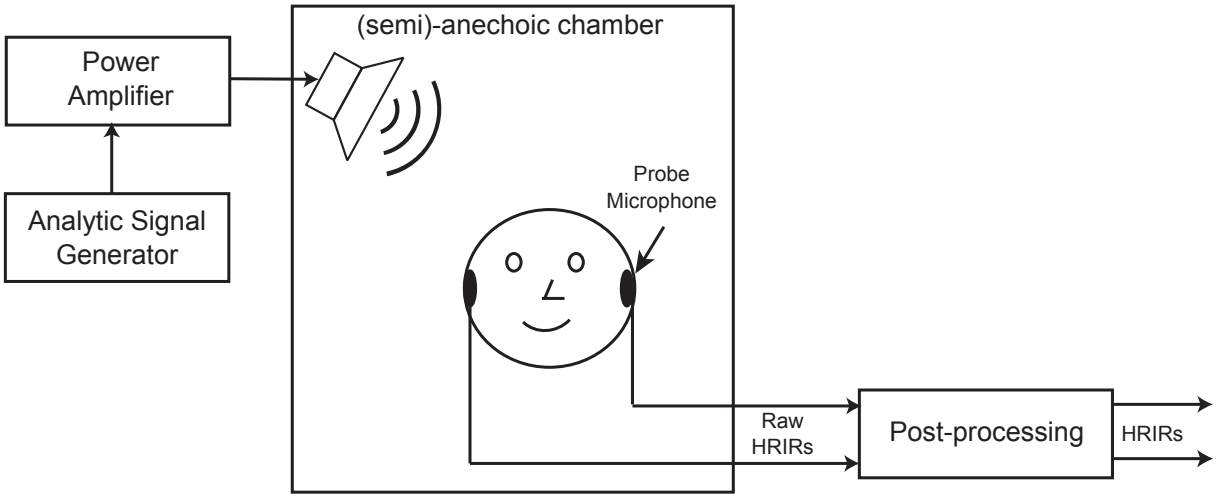


Figure 2.7: Block diagram for the HRIR measurement procedure

HRTF measurements

The HRTF of an individual is obtained by reproducing an *analytic signal* (i.e. a known synthetic signal) at the desired direction (at a distance at least 1 m from the source for far field HRTFs), and then measuring the *raw HRIR* using *probe microphones* located in the vicinity of the ear canal [16]. This procedure is repeated for each desired direction. The resulting raw HRIRs requires some post-processing stages before they can be used as filters as we will see later in this section.

Observe from Equation 2.1 that the HRTFs are continuous functions. However, in practice, the HRTFs are measured only for discrete positions in space. The HRTFs for unmeasured positions might be obtained by *spatial interpolation* [26, 27, 28, 29].

Figure 2.7 shows the block diagram for the HRTF measurement procedure whose details are considered below:

Input analytic signal: since it is very common the presence of noise in measurement systems, it is necessary that the characteristics of the input signal help to improve the signal-to-noise ratio (SNR). Although it is possible to increase the power level of the input signal, an excessive power increase could cause distortion in electro-acoustic systems such as loudspeakers and amplifiers. Thus, to increase the power level without violating the linearity of the electro-acoustic systems, an ideal input signal should have a low crest factor (ratio of peak values to the effective value) [30].

In the case of HRIRs, several types of signals have been employed such as deterministic signals (e.g. impulse [31], sinusoidal sweeps [32]), random signals (e.g. white noise, pink noise) and pseudorandom noise signals (e.g. maximal-length sequences [33], Golay codes [34, 35]).

From the aforementioned signals, nowadays the sinusoidal sweeps are the most widely used since they generate HRIRs with higher SNR and low crest factor [36]. For further

details, a comparative study of the different analytic signals to obtain impulse responses is given by Stan et al [30].

Probe microphone position: the microphone is commonly placed at a point near to the blocked ear canal entrance. This microphone position introduced by [37] has been widely used for its convenience and safety. Other positions are close to the unblocked ear canal entrance or still inside the ear canal near the eardrum.

Individuals: HRIRs are functions that depend on the anatomical characteristics that vary from person to person, making it difficult to have a set of general HRIRs. For this reason, HRIRs are measured for several subjects. As it is hard to keep people still during long periods of time for they tend to make small movements of the head and body, artificial heads and torsos such as KEMAR (Knowles Electronics Manikin For Acoustics Research) are commonly used [38]. KEMAR was designed based on the average dimensions of the anatomical characteristics according to the requirements of ANSI S3.36/ASA58- 1985 and IEC 60959:1990.

(Semi)-anechoic chamber: it is common for HRIR measurements to be made in an anechoic chamber in order to eliminate eventual reflections from the environment. However, due to technical difficulties and high costs of an anechoic chamber, HRIR measurements have been also made in semi-anechoic chambers [39]. In semi-anechoic chambers, the arrival time of the reflected waves can be controlled so that they arrive after the duration of the HRIRs (typically a few milliseconds). This can be achieved by placing acoustic absorbent material in the room [16]. Under these conditions, a time window is applied in the raw HRIRs to eliminate the unwanted reflections [39].

Post-processing of raw HRIRs: In addition to the aforementioned time truncation to eliminated reflections in semi-anechoic chambers, raw HRIRs are equalized to compensate for spectrum distortions caused by the electroacoustic system (i.e. microphones, speakers and amplifiers) [16] and to minimize changes in the subjective timbre of input signals [37]. The simplest approach is to measure the free field sound pressure $P_0(f)$ in frequency domain at the center of the head with the head absent. Let the transfer function of the electroacoustic system be $H_0(f)$, then the measured sound pressure is $H_0(f)P(\theta, \phi, f)$ and the sound pressure measured at the center of the head with the head absent is $H_0(f)P_0(f)$ (i.e. the subject is taken out from the chamber and the microphone is placed at the position corresponding to the head center). Thus, if we divide the two measured pressures $H_0(f)P(\theta, \phi, f)$ and $H_0(f)P_0(f)$, the unwanted effects of the electroacoustic system are eliminated, obtaining the definition of HRTF from Equation 2.1. The HRTFs obtained through this process are called measurement-equalized HRTFs [16].

Other approaches used are free field equalization and diffuse field equalization. The former is implemented with respect to one of the HRTFs measured in a specific direction [16], usually the frontal direction ($\theta = 0, \phi = 0$), and is defined as

$$H_{\text{free}}(\theta, \phi, f) = \frac{H(\theta, \phi, f)}{H(\theta = 0, \phi = 0, f)}. \tag{2.2}$$

On the other hand, the diffuse field equalization is performed with respect to root-mean-square value of HRTF magnitudes across all M directions (i.e. the diffuse field average) [16]:

$$H_{\text{diffuse}}(\theta, \phi, f) = \frac{H(\theta, \phi, f)}{\sqrt{\frac{1}{M} \sum_{i=0}^{M-1} |H(\theta_i, \phi_i, f)|^2}}. \tag{2.3}$$

HRTFs Databases

Using a procedure described in Section 2.1.4.2 or similar ones [32], several research teams have built publicly available HRTFs databases. For example, one of the most complete databases is the CIPIC database [39] because it has measurements for a large number of individuals (43 subjects and KEMAR) and their anthropometric data. 1250 measurements were performed in interaural coordinates for 25 non-uniformly distributed azimuths with a maximum resolution of 5° for locations near the median plane and 50 elevations with resolution 5.625° in the range −45° to 230.625°. A disadvantage of the CIPIC database is its low resolution in lateral directions (i.e. 15° to 20°). Another popular database is the ARI database [40]. The main advantages of ARI over the CIPIC database is that the former offers HRTFs from more subjects (about 150 subjects), and has better spatial resolution. There are also databases that only measured HRTFs for KEMAR [38] or similar mannequins. For a non-exhaustive list of HRTF databases, we refer to [41].

In an effort to facilitate the exchange of the data between researchers and users, the *Audio Engineering Society* (AES) has standardized the *Spatially Oriented Format for Acoustics* [42]. SOFA is a file format for storing and exchanging spatially oriented acoustic data like HRTFs. Currently, the most popular HRTF databases are publicly available for download in SOFA format [43], including the CIPIC and ARI databases.

Time-domain Features of HRIRs

The most important localization cue in time is the ITD. Although there are several methods to estimate ITD (see for a comprehensive review [18]), one straightforward approach is to calculate the difference between the onset delays between the left and right HRIR. In Figure 2.8, observe that at the beginning the HRIR is near zero due to the time

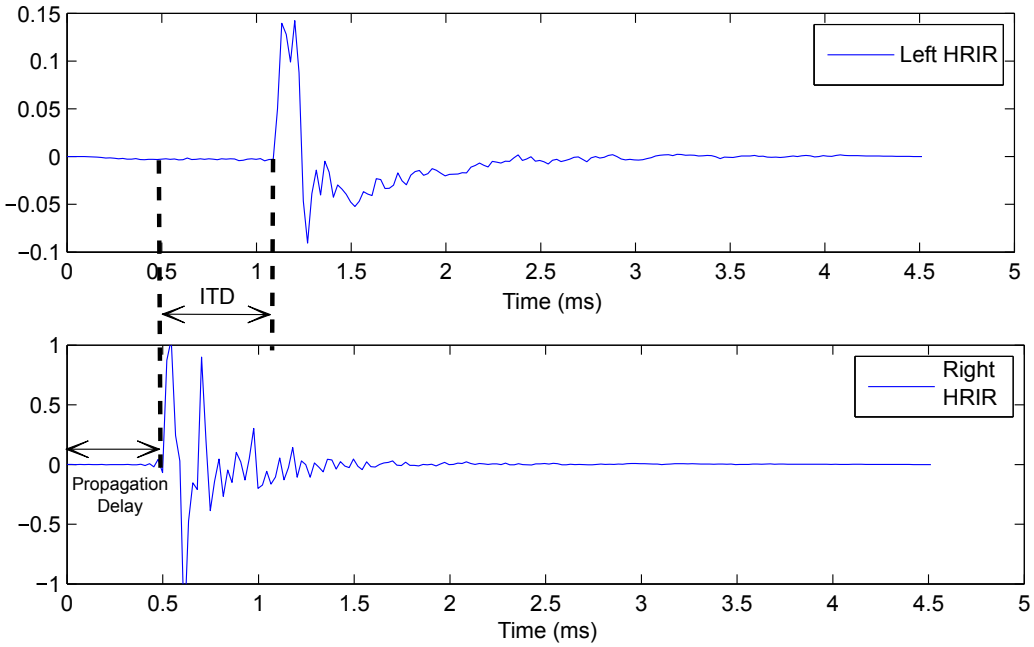


Figure 2.8: KEMAR HRIRs from CIPIC database for $\theta = 80°$, $\phi = 0°$. ITD calculated from the difference between the onset delays.

the sound takes to reach the ear during the measurement (i.e. the propagation time). Thus, the ITD is defined as the difference between the onset delays t_L and t_R, i.e.,

$$ITD(\theta, \phi) = t_L - t_R \tag{2.4}$$

Notice that to determine the onset delays a threshold needs to be chosen. In [44], they define the onsets as the instants at which the HRIRs reached 20% of their first maximum peak amplitudes. Observe also that the HRIR might be filtered prior to ITD estimation. For instance, Algazi et al [44] use a high-pass filter with cutoff frequency at 1500 Hz. On the other hand, Xie [41] uses a low-pass filter with cutoff frequency at 2.7 kHz to reduce the effects of the pinna. Finally, the HRIRs are often upsampled before ITD calculation to increase time resolution [16].

It is worth mentioning that in this dissertation we do not deal with ITD estimation for we focus only in spectral characteristics obtained from the HRTF magnitude. Even so, we included this section to give the reader a general idea of ITD in the context of spatial audio.

Frequency-domain features of HRTFs

In order to analyze the frequency characteristics of HRTFs, Figure 2.9 shows the HRTFs magnitudes for four azimuths in the horizontal plane where we can observe the following [1]:

1. Below roughly 200 Hz, there is a drop in the signal level. This drop occurs because the frequency response of the speakers used in the measurement of HRTFs is usually limited at low frequencies.

2. At frequencies below the 0.4-0.5 kHz band, the head attenuation effect is negligible so that the HRTFs magnitudes for both ears are around 0 dB and are roughly frequency-independent.

3. As the frequency increases above 1.5 kHz, the interaural level differences become more evident. Thus, the magnitude of the ipsilateral ear (e.g. see the magnitude of the left ear at the azimuth $\theta = 80°$ in Figure 2.9) is greater than the magnitude of the contralateral ear. It is worth noting that this difference increases with the azimuth angle. The difference of both magnitudes (i.e. the ILD) is defined as

$$ILD\left(\theta, \phi, f\right) = 20\log_{10}\left|\frac{H_R\left(\theta, \phi, f\right)}{H_L\left(\theta, \phi, f\right)}\right|. \tag{2.5}$$

4. Even though for $\theta = 0°$ the left and right HRTFs are similar, the small differences between them come from the fact that the head is not perfectly symmetric.

5. At high frequencies, from the 5-6 kHz band, the HRTFs magnitudes vary in a complex manner presenting several peaks and notches.

6. The peak near 4 kHz is due to the ear canal resonance.

Frequency-domain characteristics due to the pinna

At frequencies above 3 kHz, when the size of the pinna is comparable to the wavelength of the source, the characteristics of the pinna become indispensable to resolve front-back and up-down confusions as well as to our sound localization accuracy [23].

As discussed in Section 2.1.2, similar ITDs and ILDs cause ambiguities (i.e. front-back confusions) that are solved by the high-frequency characteristics due to the pinna. To understand how the pinna is able to provide spectral information to resolve such ambiguities, Figure 2.10 shows the HRTF magnitudes in the horizontal plane for the same ear at $\theta = 0°$ (i.e. front) and $\theta = 180°$ (i.e. back). Note that at high frequencies, the differences in magnitude from both directions are evident. These differences caused by the asymmetry of the head, the position of the ear and the pinna shape help to resolve front-back confusions [41].

Moreover, the HRTFs magnitudes are characterized by presenting peaks and notches at frequencies above roughly 5 kHz. The central frequency of the first of these notches is considered an important cue in vertical location [45]. Figure 2.11 shows the magnitude of

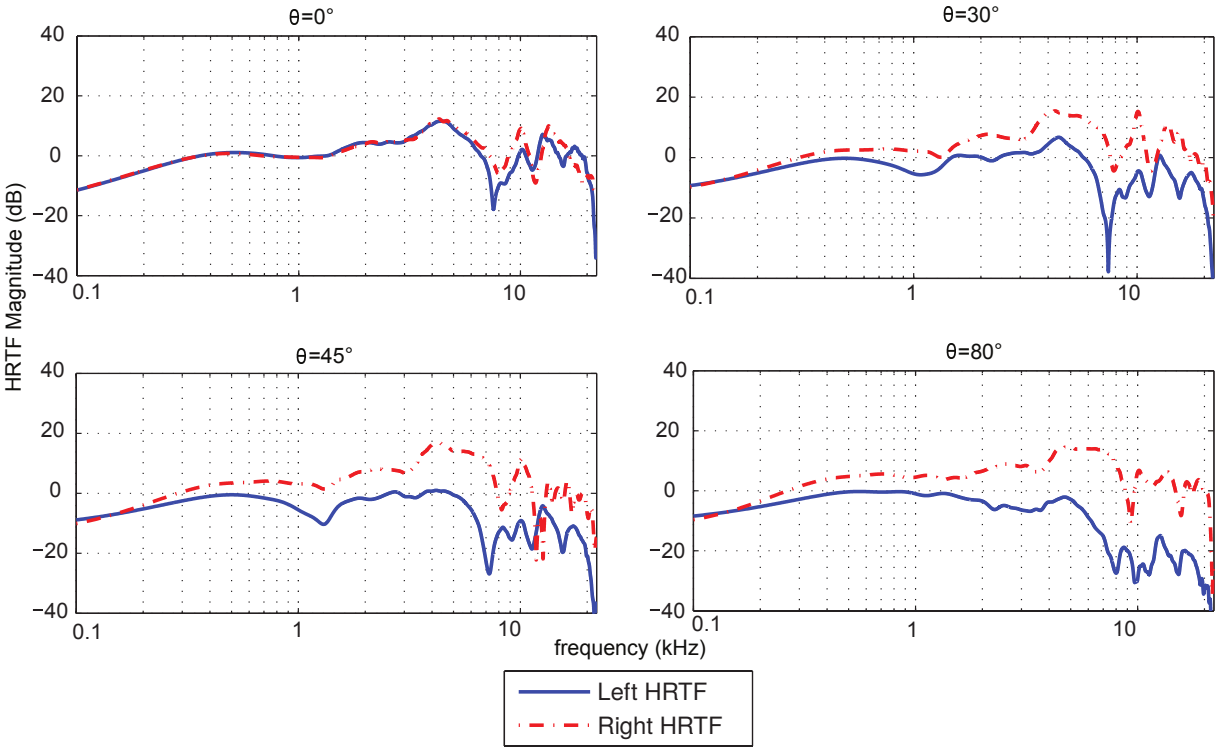


Figure 2.9: KEMAR's HRTFs magnitudes from CIPIC database for several azimuths in the horizontal plane.

several HRTFs in the median plane (i.e. $\theta = 0°$). Note that the center frequency of the first notch is approximately equal for both ears.

Another important factor regarding the frequency of the first notch is its variability among different individuals. Algazi et al. [39] averaged the center frequency of the first notch in 52 individuals from the CIPIC database in the direction $(\theta = 0°, \phi = 0°)$. The result obtained was 7.6 kHz with a standard deviation of 1050 Hz. Considering such a high standard deviation, the authors concluded that there is a high variability of the frequency of the first notch between individuals caused by the anthropometric characteristics of the pinna.

Figure 2.12 shows the left HRTF at a specific location for four different subjects. The inter-subject variability is especially notable for high frequencies (i.e. $f > 4$ KHz) where the monoaural cues introduced by the pinna are more prominent [1]. For this reason, the pinna is considered as the acoustic fingerprint of a subject. Various studies show a decrease in localization accuracy due to nonindividualized HRTFs [7], often producing front/back reversals, poor sound externalization (i.e. the subject perceives the sound inside his/her head) and incorrect elevation perception. Thus, it is necessary to personalize HRTFs to guarantee high-quality 3D sound perception.

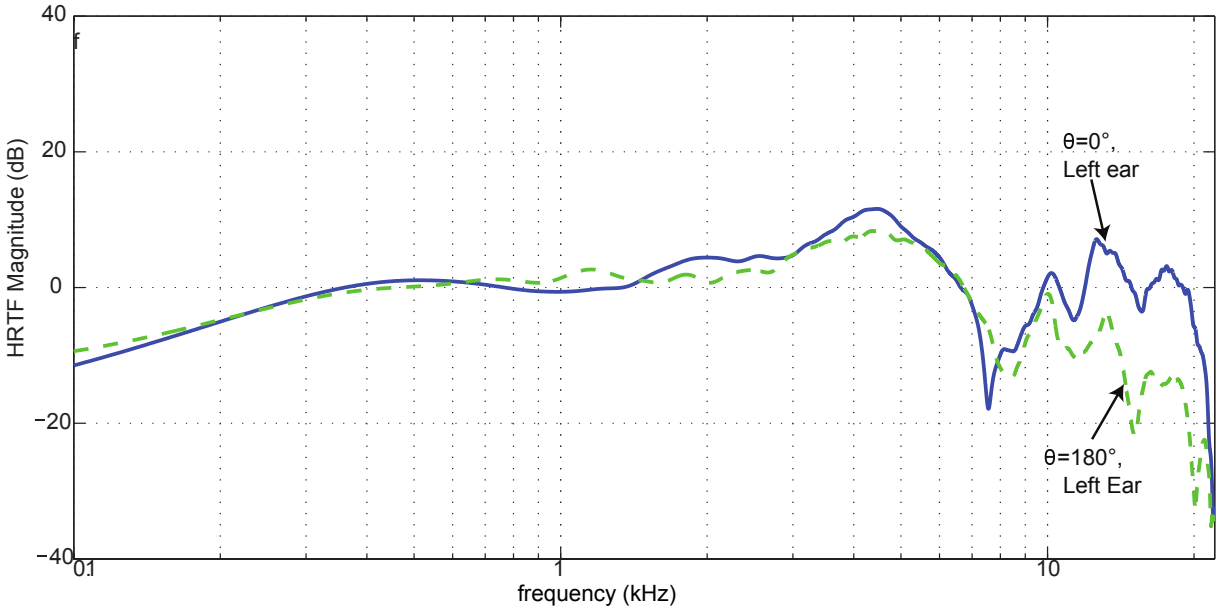


Figure 2.10: HRTFs magnitudes for azimuths 0° (i.e front) and 180° (i.e back) in the horizontal plane.

Minimum phase approximation of HRTFs

We use the minimum phase approximation of HRTFs in several of our contributions for it allows us to isolate the HRTF magnitude from the ITD. Since our work does not tackle the problem of ITD estimation, this decomposition lets us work only with the spectral characteristics of the HRTF magnitude.

In general, any transfer function and therefore any HRTF can be decomposed [46] as the product of its minimum phase function $H_{\min}(\theta, \phi, f)$, an all-pass function $\exp[j\psi_{all}(\theta, \phi, f)]$ and a linear phase function $\exp[-j2\pi fT(\theta, \phi)]$, i.e.

$$H(\theta, \phi, f) = H_{\min}(\theta, \phi, f) \exp[j\psi_{all}(\theta, \phi, f)] \exp[-j2\pi fT(\theta, \phi)], \tag{2.6}$$

where $T(\theta, \phi)$ is the time delay caused by the propagation of sound waves before reaching the ear and corresponds approximately to the propagation delay in Figure 2.8 [47]. On the other hand, we know that the magnitude of a transfer function and its corresponding minimum phase magnitude are equal [46], i.e. $|H| = |H_{min}|$ and

$$H_{\min}(\theta, \phi, f) = |H_{\min}(\theta, \phi, f)| \exp[j\psi_{\min}(\theta, \phi, f)], \tag{2.7}$$

where the phase of the minimum phase function and the logarithm of the magnitude are related by the Hilbert transform

$$\psi_{\min}(\theta, \phi, f) = -\frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{\ln|H(\theta, \phi, x)|}{f - x} dx. \tag{2.8}$$

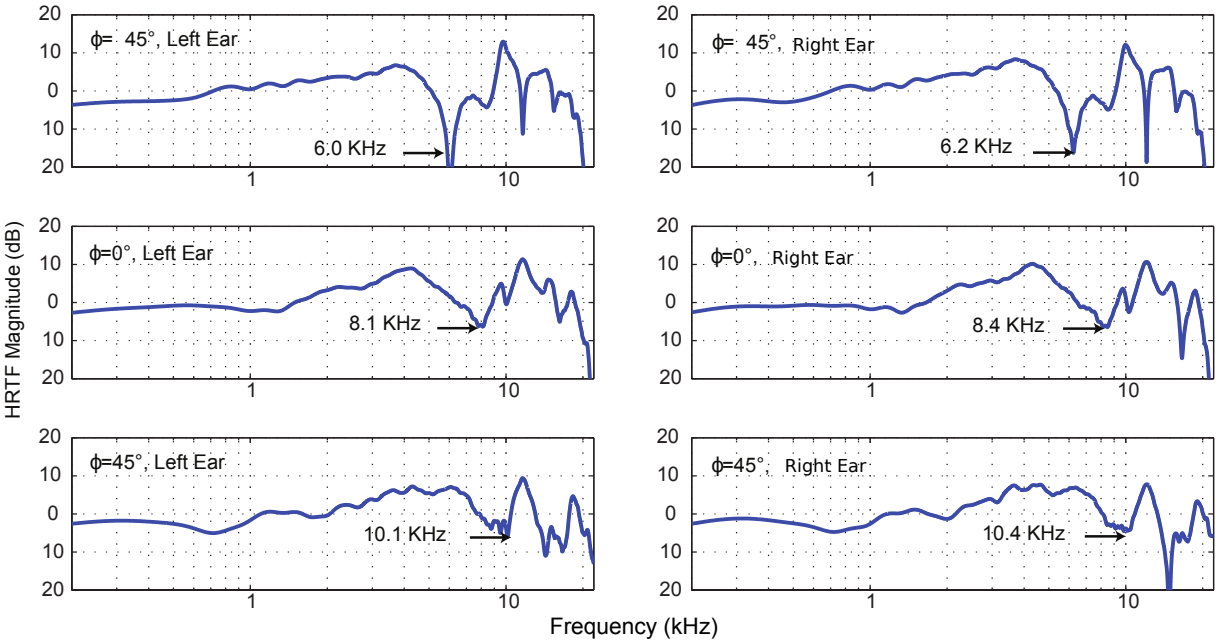


Figure 2.11: HRTFs magnitudes for several elevations in the median plane (i.e. $\theta = 0$). The arrows indicate the first notch.

From Equation 2.6, we deduce that the phase of an HRTF is given by

$$\psi(\theta, \phi, f) = \psi_{\min}(\theta, \phi, f) + \psi_{all}(\theta, \phi, f) - 2\pi f T(\theta, \phi). \tag{2.9}$$

If the phase component of the all-pass function ψ_{all} is negligible, an HRTF can be approximated as

$$H(\theta, \phi, x) \approx H_{\min}(\theta, \phi, f) \exp[-j2\pi f T(\theta, \phi)]. \tag{2.10}$$

Equation 2.10 is called the minimum phase approximation of an HRTF, in which the HRTF is approximated by its minimum phase function in cascade with a pure delay or linear phase $T(\theta, \phi)$ function [16]. In practice, this delay is just the ITD calculated from the method described in Section 2.1.4.4 or another one [18]. It is worth mentioning that in this case, the delay would be placed only on the HRTF (left or right) that guarantees the onset time is always positive. Studies carried out by [47] and [48] have demonstrated the validity of the minimum phase model of HRTFs. This approximation is important because it allows processing the HRTF using only its magnitude. Once the magnitude of the HRTF is processed, the complex HRTF can be reconstructed using Equation 2.10.

Manifold Learning

In manifold learning (also known as nonlinear dimensionality reduction), we are interested in discovering a smooth low dimensional surface or manifold, i.e., the intrinsic structure of a dataset embedded in a higher dimensional linear vector space [49, 50]. The main idea behind manifold learning is that if we can automatically discover the underlying

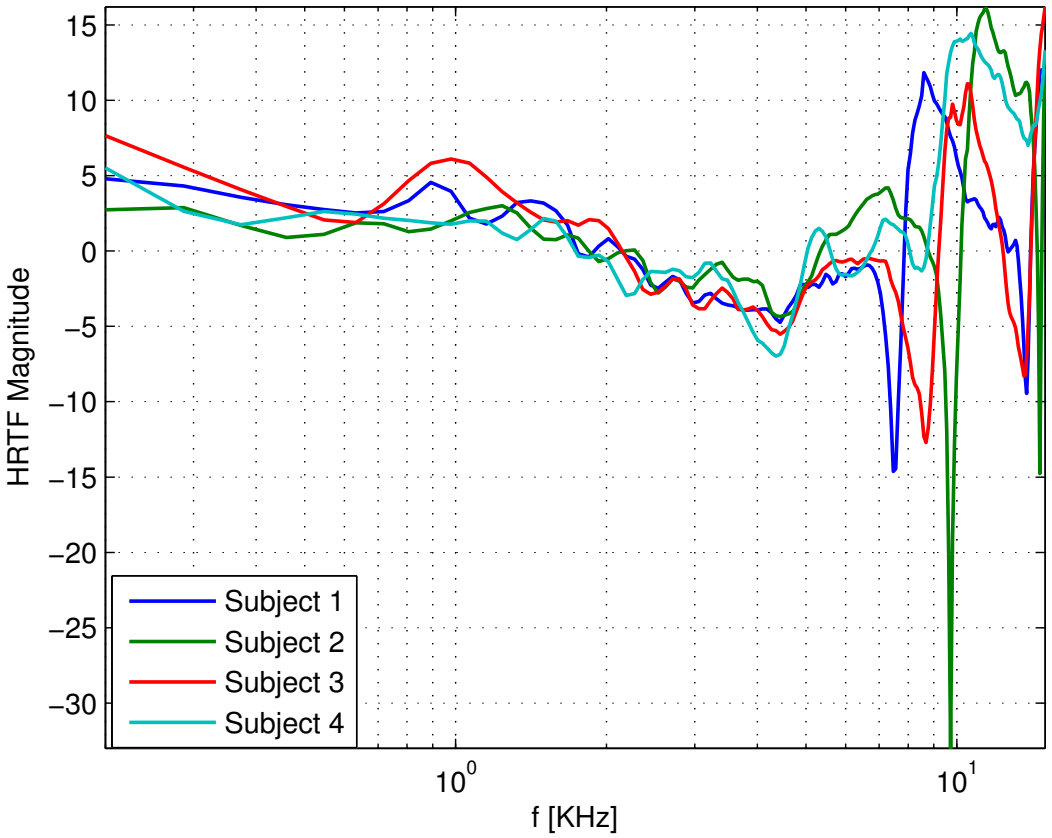


Figure 2.12: Left HRTFs at a specific location for four different subjects. The variability among subjects, more evident at high frequencies, is due to the monoaural cues introduced by the pinna.

manifold and "unfold" it, this may lead to easier data interpretation [50]. The manifold assumption has been a recurrent theme among the computational neuroscience scientists, who argue that manifolds are crucial for discovering and understanding the underlying nonlinear relationships of perception in the brain [11].

A popular manifold learning technique is "Isometric Feature Mapping" or Isomap. Classical linear dimensionality reduction methods such as Principal Component Analysis (PCA) and multidimensional scaling [51] (MDS) can only deal with flat euclidean structures, failing to discover nonlinear structures of the input data [52]. Unlike classical techniques, Isomap is capable of discovering the nonlinear degrees of freedom that underlie complex natural observations, such as human handwriting or images of a face under different viewing conditions [53]. Isomap is considered a global approach since it tries to map nearby high dimensional datapoints into nearby low dimensional datapoints. Likewise, faraway high dimensional datapoints are mapped into faraway low dimensional datapoints.

Figure 2.13 shows an illustrative example to intuitively explain how Isomap works on the "Swiss roll" dataset embedded in a 3D space. The two datapoints shown in Figure 2.13a appears to be very close according to their euclidean distance (blue dashed line) while

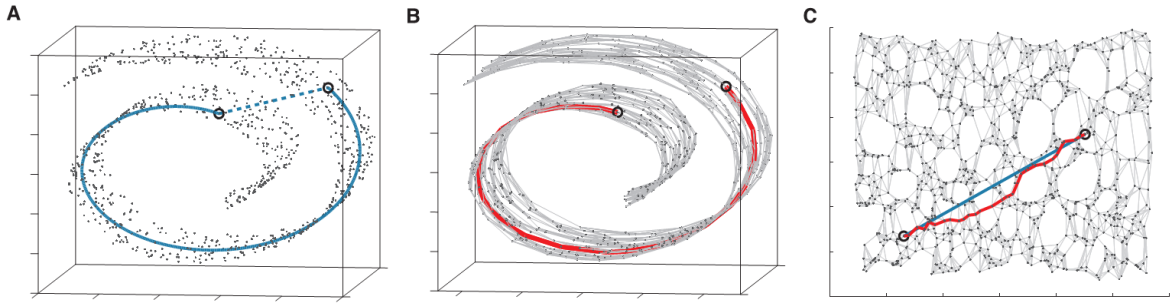


Figure 2.13: Illustrative example using the "swiss roll" dataset. Figure reproduced from [53]. (**a**) The Geodesic distance vs the Euclidean distance (dashed line) between two datapoints. (**b**) Geodesic distance estimation between two datapoints as the shortest path in a graph. (**c**) The unfold manifold recovered by Isomap.

their distance over the manifold (blue line) is much larger. This distance over the manifold is called the geodesic distance. Different from PCA that tries to preserve the euclidean distance, Isomap tries to preserve the geodesic distance. To achieve this goal, Isomap construct a graph over the manifold as shown in Figure 2.13b to estimate the true geodesic distance as the shortest path between two datapoints (red line). Observe that if we "unfold" the swiss roll (see Figure 2.13c), it can be intrinsically represented just in a 2D plane (i.e. the intrinsic dimensionality is two). In the unfolded dataset, it is more evident how Isomap approximates the true geodesic distance (blue line) using the shortest path between two datapoints (red line).

In the context of spatial audio and for the sake of clarity, we will reproduce Figure 1.1 again in Figure 2.14. In this intuitive example, since the only parameter we vary is elevation, we might assume that the intrinsic dimensionality is one and the manifold is embedded in a high dimensional 3D space. In contrast to the euclidean distance, the geodesic distance (red line) better reflects the fact that the two sources at $\phi = 45$ and $\phi = -45$ degrees are far away.

More formally, Isomap has three steps that can be summarized as follows:

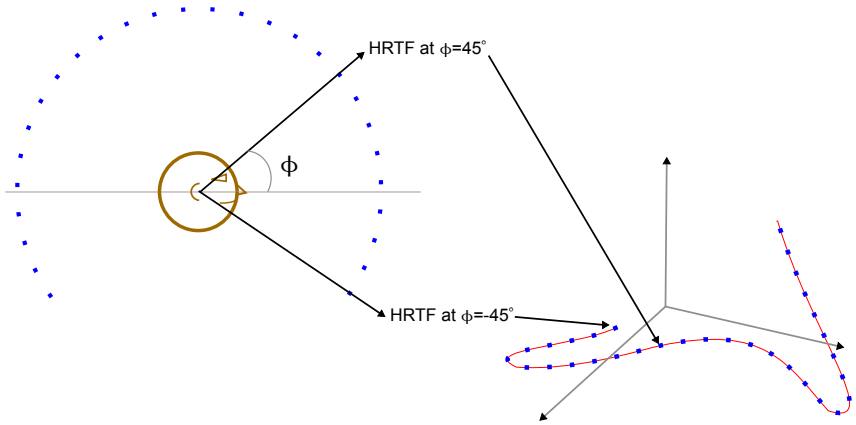


Figure 2.14: An intuitive example of a one-dimensional manifold embedded in a higher dimensional space. Figure inspired by [12].

1. A graph construction procedure where each sample is connected to some number of neighbors. Although this step might be as simple as selecting the nearest neighbors for each datapoint according to the euclidean distance, the graph construction procedure might be arbitrarily complex by taking into account the prior knowledge that we have of the problem [54].

2. We compute the pairwise estimated geodesic distance matrix. The estimated geodesic distance between each pair of datapoints is obtained along their corresponding shortest path. The Dijkstra's algorithm or the Floyd's Algorithm migh be used to calculate it.

3. We construct the low-dimensional embedding by applying classical multidimensional scaling (MDS) [51] to obtain the low dimensional datapoints.

The main drawback of Isomap is the possibility of erroneous connections in the graph due, among others, to noisy datasets [55]. This is called short-circuiting. Choosing the right number of neighbors is critical since even a single short-circuit can lead to drastically different (and incorrect) low-dimensional embedding [55]. In general, a large number of neighbors might cause short-circuiting and a small one might lead to a too sparse graph unable to approximate geodesic paths accurately. For instance, in Figure 2.13, if we increase too much the number of neighbors, at some point the two datapoints will be incorrectly connected by an edge (i.e. a short-circuit). Another drawback of Isomap is its computational complexity since one need to compute pairwise shortest paths between all sample pairs and perform an eigendecomposition over a non-sparse distance matrix [53]. Despite these weaknesses, Isomap has been successfully applied to several computer vision problem such as wood inspection [56] and head pose estimation [57]

There are several manifold learning techniques beyond Isomap. For example Laplacian Eigenmaps[58] which is similar to Isomap in that both construct a graph representation of the datapoints. In contrast to Isomap, Laplacian Eigenmaps attempts to preserve only local properties of the manifolds based on the pairwise distances between near neighbors. This is why it is considered a local approach in contrast to Isomap which is considered a global one. It means that Laplacian Eigenmaps tries only to map nearby high dimensional datapoints into nearby low dimensional datapoints. LLE [59] is another popular technique that, similar to Laplacian Eigenmaps, is local. For a comprehensive and intuitive review of linear and nonlinear dimensionality reduction techniques, we refer to reference [60].

Finally, in the next chapters of this thesis, we use the aforementioned manifold learning techniques to address the spatial audio challenges proposed in Chapter 1.

Chapter 3

Contribution I

Contribution I is published as:

L. Neto, F. Grijalva, V. Maike, L. Martini, D. Florencio, M. Baranauskas, A. Rocha and S. Goldenstein, "A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users," in *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 52-64, Feb. 2017. doi: 10.1109/THMS.2016.2604367

A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users

Laurindo Britto Neto[1], Felipe Grijalva[2], *Student Member, IEEE*, Vanessa Maike[3], Dinei Florencio[4], *Fellow Member, IEEE*, Luiz Martini[2], Cecília Baranauskas[3], Anderson Rocha[3], *Senior Member, IEEE*, and Siome Goldenstein[3], *Senior Member, IEEE*

Abstract

In this paper, we introduce a real-time face recognition (and announcement) system targeted at aiding the blind and low-vision people. The system uses a Microsoft Kinect sensor as a wearable device, performs face detection and uses temporal coherence along with a simple biometric procedure to generate a sound associated with the identified person, virtualized at his/her estimated 3D location. Our approach uses a variation of the K-Nearest Neighbors (K-NN) algorithm over histogram of oriented gradient (HOG) descriptors dimensionally reduced by principal component analysis (PCA). The results show that our approach, on average, outperforms traditional face recognition methods while requiring much less computational resources (memory, processing power and battery life) when compared to existing techniques in the literature, deeming it suitable for the wearable hardware constraints. We also show the performance of the system in the dark, using depth-only information acquired with Kinect's infrared camera. The validation uses a new dataset (that will be publicly-available upon acceptance), with 600 videos of 30 people, containing variation of illumination, background and movement patterns. Experiments with existing datasets in the literature are also considered. Finally, we conducted user experience evaluations on both blindfolded and visually-impaired users, showing encouraging results.

Introduction

According to the World Health Organization, 285 million people are estimated to be visually impaired worldwide: 39 million are blind and 246 have low vision [61]. Vision impairment is a hindrance on several daily activities, and there is a constant strive for new assistive devices [62].

[1]Department of Computing, Federal University of Piauí, Teresina, PI, Brazil.
[2]School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil.
[3]Institute of Computing, University of Campinas, Campinas, SP, Brazil.
[4]Microsoft Research, Redmond, WA, USA.

One of these difficult daily activities is people recognition and localization, as it plays a crucial role in social interaction at work and at home. It would be useful to know someone's location without having to hear him/her speak, so that the visually impaired person could turn to his/her interlocutor and engage him/her in conversation.

Such systems should have multiple design requirements with respect to hardware, face recognition algorithms as well as user interaction considerations. The hardware platform should be portable, preferably wearable, and cost-effective to eventually reach the vast majority of blind population, mainly living in developing countries [61]. Since wearable hardware resources are limited and the system's feedback revealing the person's name needs to be as immediate as possible, complex state-of-the-art face recognition approaches might not be a viable first choice technique. Instead, the face recognition algorithm should be fast enough to permit real-time performance on such hardware platforms while allowing a degree of robustness under different conditions (e.g., movement patterns, lighting and background variation). Furthermore, considering that the mean offline social network size of an individual with a visual disability ranges from 15 persons for western adolescents and young adults [63] to 31 persons for Chinese older adults [64], the number of faces to be recognized is considerably smaller than in other applications (e.g., security applications). This implies that the face recognition algorithm should be chosen and tested accordingly. Regarding the interaction with the visually impaired user, the system should provide an intuitive and limited feedback. Above all, it should avoid overwhelming the user's senses (i.e., auditory and touch).

In light of the aforementioned requirements, this paper presents a real-time wearable face recognition system aimed at individuals with visual disabilities. As soon as the system recognizes a face, it sends a 3D audio feedback to the visually impaired user. For example, after the system identifies a known person (say Jane), the visually impaired user will hear Jane's name as if coming from the exact location where Jane is.

This face recognition system is part of a wearable system to aid the blind and low-vision people. We named the wearable system as *Kinect Vision Blind* (KVB). Although still in its infancy, the KVB is a multi-year, long-term open-source project[5], with the bold goal of mobilizing multiple research groups around a common platform. It aims at designing distinct modes of operation, providing users with specialized functionalities such as navigation, people localization and recognition, object recognition, and textual information translation (e.g., signs, symbols and currency identification). In this paper, we show that it is possible to have a high-accuracy face recognition system in this limited-scope application with a simple yet computationally efficient approach that allows real-time performance on wearable devices by taking advantage of depth information and temporal coherence. Besides, we introduce a new RGB-D database composed of 600 videos from 30 people, where we performed extensive simulations, including experiments in

[5] http://revistapesquisa.fapesp.br/en/2015/06/14/recognizing-the-environment

the dark using Kinect's depth-only data. The results show accuracy rates with statistical significant differences ($p-value < 0.05$) compared to traditional face recognition methods publicly available. Finally, with the aim of evaluating user experience, we conducted tests of our system on both blindfolded and visually impaired users.

Related Work

Most of assistive systems aimed at visually impaired people use video processing to convert visual data into an alternative representation (e.g., auditory, haptic), and focus on daily tasks such as navigation [65, 66, 67, 68, 69, 70, 71] and object detection [72, 73].

On the other hand, there are few contributions addressing the face recognition task for visually impaired people. Previous studies [74, 75, 76] explored the feasibility of implementing a wearable face recognition aid for blind people based on computationally efficient algorithms such as robust correlation [74], image retrieval in conjunction with classification [75] and *Iterative Closest Points* (ICP) approach [76].

In light of this, several research groups have shifted focus towards the development of wearable face recognition prototypes for the blind community. For instance, Krishna et al. [77] developed a sunglasses with a pin hole camera, which uses the *Principal Component Analysis* (PCA) algorithm for face recognition [78]. The sunglasses system is validated with a highly controlled dataset, without disguises and with only variation in pose and illumination, which uses a precisely calibrated mechanism to provide a robust face recognition.

Moreover, using a camera atop a standard white cane, Astler et al. [79] implemented a face recognition software based on the commercial *Luxand Face SDK*[6] running on a server. The limitation of using this commercial solution is that it was not designed for running on low-power devices such as wearables.

More recently, although not designed with visually impaired users in mind, researchers [80] proposed a wearable face recognition system that uses a head mounted display. The system detects faces using the Viola-Jones technique [81] and employs a face recognition approach based on image sets [82].

On the other hand, various face recognition prototypes have been proposed that, despite being not wearable-friendly, are aimed at blind people. In this line, in [83], the authors used a smartphone to provide an audible feedback whenever a face from a database enters or exits the scene. Their detection algorithm runs the commercial *VeriLook*[7] face technology in a server. Similarly, Balduzzi et al. [84] employ a camera and a personal computer to process the incoming video stream using the *Local Binary Pattern* (LBP) approach [85], and to provide an audio feedback to the user whenever a face is recognized.

[6] https://www.luxand.com/facesdk/
[7] http://www.neurotechnology.com/verilook.html

Finally, as part of a modular software running on a handheld device equipped with an embedded RISC processor, called *Blind Assistant*, Battaglia et al. [86] proposed a face recognition module that detects faces using the Viola-Jones algorithm [81], recognizes them by means of the PCA algorithm [78], and provides an auditory feedback to the user.

Beyond face recognition, various studies have proposed wearable prototypes aimed at blind people to identify facial expressions [87, 88, 79, 89, 90, 91, 92], behavioral expressions (body mannerisms [88], head movements [92]) and other non-verbal communications cues (person localization [93, 94], number of people present, their age and gender distributions [95]).

Previous studies have also proposed wearable stereo vision systems to develop electronic travel aids for blind people [96, 97, 98]. Additionally, several works have demonstrated the feasibility of using the Microsoft Kinect as the main hardware component for wearable devices to assist visually impaired users in navigation tasks [99, 100, 101, 102, 103, 104, 105, 106, 107, 108] due to its capability of sensing depth information along with RGB data. Depth information has proven to be very useful in the aid of visually impaired people as it better enables the detection of objects in the scene along with their 3D spatial position. This confirms the potential of low-resolution 3D sensors for robust face detection and recognition. In particular, Ribeiro et al. [99] proposed a Kinect-based wearable prototype that, among other tasks, offers a module for face recognition based on a multiple-instance pruning cascade detector [109] and a learning-based descriptor [110]. However, since they focused mainly on navigation tasks, they did not report accuracy/performance results of the face recognition approach. Moreover, they did not perform user-experience experiments on their face recognition module.

Figure 3.1: Blindfolded user wearing our prototype.

Furthermore, although not focusing on visually impaired aids or wearable devices, previous Kinect-based face recognition studies [111, 112, 113, 114, 115, 116, 117, 118, 119] proposed complex frameworks (i.e., approaches not suitable for real-time systems on low-resource devices) using, for example, multi-modal sparse coding [118]. In contrast and as in previous works targeting visually impaired people [76, 74, 75, 77, 83, 84, 79,

86], we seek a computationally efficient face recognition algorithm capable of running in real-time on small, low-power devices such as wearables. As detailed above, several authors used traditional face detection [86, 80] and face recognition algorithms [77, 86, 84], or even commercial solutions running on servers that are not designed to work on low-resource devices [83, 79]. Here, we take advantage of the 3D sensor technology of Microsoft Kinect for accurately detecting faces, and we employ *Histogram of Oriented Gradients* (HOG) [120] and PCA combined with a variation of *K-Nearest Neighbor* (K-NN) algorithm, with the aim of achieving real-time performance by exploiting temporal coherence along contiguous video frames.

Finally, spatial audio, also known as 3D audio, has been successfully used on applications for visually impaired people such as navigational aids [121, 99, 122] and graphical computer interfaces [123]. In fact, it has proven to be superior to even simple spatial language ("left", "right", or "straight") for accomplishing tasks without vision when cognitive load is present [124]. Thus, instead of using monaural audio feedback as in previous works [77, 83, 84, 79, 86], we exploit the inherent spatial sense of the auditory system by using spatial audio to convey the location of a face in the environment.

System Description

The system we propose herein comprises a wearable prototype within the context of the KVB project. We are mainly interested in the RGB-Depth technology itself. In this vein, the Kinect is a good starting point for a rapid prototyping, as it is composed of an RGB camera to capture color images, an infrared (IR) emitter to emit infrared light beams and an IR depth sensor to read the IR beams reflected back to the sensor, and to compute the distance between an object and the sensor, enabling capture a depth image[8]. Thus, the hardware component of our prototype combines a Microsoft Kinect RGB-D sensor, an accelerometer/gyroscope/compass sensor, stereo headphones, and a portable computer. Although this is not envisioned as the final form factor, in this initial wearable prototype we attached the Kinect to the top of a skateboard helmet, and the notebook is carried in a backpack. We removed unnecessary parts of Kinect for the KVB system, such as its carcass, the microphone array and the tilt motor, making it lighter to transport. The Kinect's power supply adapter has been replaced by a rechargeable battery pack, and the headphones feature bone conduction technology, allowing the user to listen to the environment.

The diagram in Fig. 3.2 summarizes our system architecture that includes three modules: 1) face detection module (FDM), 2) face recognition module (FRM) and 3) 3D audio module (3DAM). When the system's Kinect sensor detects a face and estimates the

[8]https://msdn.microsoft.com/en-us/library/jj131033.aspx

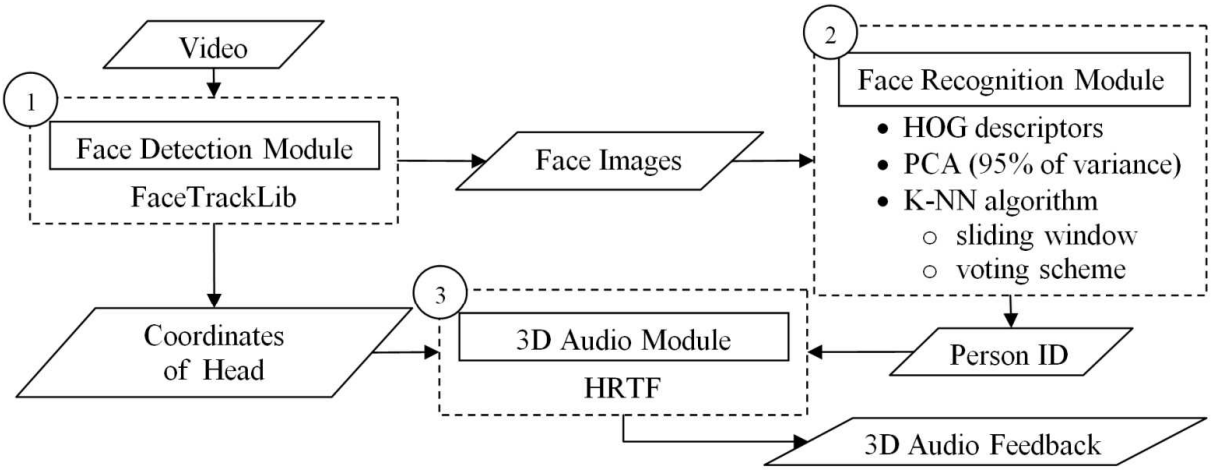


Figure 3.2: System architecture of our approach.

corresponding 3D position – in the Kinect's coordinate frame – of the detected person's head, it performs real-time face recognition, and finally it generates a virtual sound at the face location. For instance, if the system recognizes a known face, the user will hear the person's name virtualized at the location where that person is.

Face Detection Module – FDM

In the FDM, the system employs off-the-shelf face detection software (i.e., Microsoft Face Tracking SDK for Kinect for Windows – *FaceTrackLib*[9]). The *FaceTrackLib* uses color, depth and skeleton information to detect and track human faces in real time. Then, the FDM sends the bounding box around the detected face (i.e., face image – with average size of 64×72 pixels) to the FRM.

Face Recognition Module – FRM

In the FRM, we use 3,780-dimensional HOG descriptors (with window size = 64×128, block size = 16×16, block stride = 8×8, cell size = 8×8 and number of bins = 9), PCA and, for simplicity and efficiency, the K-NN classifier algorithm [125] using the euclidean distance metric. We know, from experience, that several other, perhaps more interesting, metrics could be used. However, we opted to use a standard one and not to fine tune the method to a specific metric. In addition, HOG descriptors and Euclidean metrics have been used before with relative success in the literature [126]. HOG descriptors showed good results to represent set of features for face identification [127]. Moreover, HOG has a controllable degree of invariance to local geometric transformations, providing invariance to translations and rotations smaller than the local spatial or orientation bin size [120].

Since the K-NN efficiency and performance is affected by a large number of attributes in the samples and the presence of redundant and irrelevant attributes [125], we perform

[9]https://msdn.microsoft.com/en-us/library/jj130970.aspx

a dimensionality reduction through PCA in the HOG descriptors keeping 95% of the total variance of the data.

To take advantage of temporal coherence, we varied the number of sequential video frames that are classified by K-NN. Thus, we classify each frame within the temporal sliding window, and the most voted person is the final label of the classification. For example, in a sliding window with size of 60 frames, 60 votes are computed, and only after that the system returns the most voted person as the final classification – it is returned every 60 frames of video.

Finally, it is worth mentioning that our recognition approach relies only on the RGB data as we seek a fast algorithm to allow real-time face recognition performance on wearable devices with limited hardware resources. Even so, we performed experiments using our approach over the depth-only data in order to analyze our technique's performance in the dark.

The Real-time Face Recognition System

In the Real-time Face Recognition System, we use the training set that produced the best accuracy rate in the results of experiments (see Section 3.5.1) – i.e., sliding window size of 60 frames, 48 samples/class in the training set and $K = 1$. First, for each frame with a detected face, we compute its HOG descriptor. Next, we center the descriptor data and multiply it by the PCA rotation matrix in order to reduce its dimensionality. Then, we use the K-NN algorithm to classify the sample, and finally we conduct the voting scheme. In this case, the voting scheme is performed only on the streaming video frames with a detected face (non-detection frames are ignored). The real-time system returns the 3D audio feedback of the identified person every 60 votes, achieving a frame rate close to 30 FPS in a high-performance laptop (Intel Core i7, 2.4GHz and 32GB RAM).

The system also identifies unknown people – i.e., people not registered in the system. We use a distance threshold for classifying a person as "not recognized," which was set empirically based on observed distance values. When the distance from the sample to each of its nearest neighbors is greater than a threshold distance, the unknown class wins a vote. So, if the unknown class wins the voting, the system will return the audio feedback "not recognized".

3D Audio Module – 3DAM

For rendering the 3D audio feedback to the visually impaired user, the 3DAM uses both the coordinates – in the Kinect's coordinate frame – of the detected person's head estimated by the FDM (i.e., Coordinates of Head) and the final classification result obtained by the FRM (i.e., Person ID). The 3DAM renders a virtual sound (i.e., the 3D audio feedback) at a given location by filtering a monaural sound through a pair of transfer

functions (one for each ear), called *Head-Related Transfer Functions* (HRTFs) [128]. The resultant 3D sound is presented to the visually impaired user through the bone conduction headphones, so that he/she will hear the sound as if coming from the coordinates of the detected person's head.

HRTFs are frequency-domain functions that depend on direction (i.e., azimuth and elevation angle), ear (i.e., left or right) and anthropometric features (i.e., head, torso and pinna). Since these anthropometric features differ widely among individuals, HRTFs need to be personalized [128]. Non-individualized HRTFs produce a decrease in localization accuracy (more prominent in elevation positions), often causing front/back reversals (i.e., perception of a sound source at a frontal location as though it were coming from a back location, or vice versa), and poor sound externalization (i.e., localization of sound inside the head) [128].

The most accurate approach of personalizing HRTFs is through direct measurements, but it involves expensive apparatus and complex procedures. Although several methods have been proposed to avoid such measurements[129, 130, 131], HRTF personalization remains an open problem. Despite the limitations of HRTFs, HRTF 3D audio rendering has proven to be an efficient way of conveying information to the visually impaired in navigation aids [121, 99, 122] since it is very efficient when cognitive load is present [124]. This is critical for the visually impaired people that rely heavily on their hearing sense. Therefore, for our initial prototype, we use HRTFs from the generic mannequin called KEMAR (*Knowles Electronic Manikin for Acoustic Research*), which is publicly available in the CIPIC database [132].

So far, we have described how the system will communicate the directional location of a face using 3D audio – i.e., its azimuth and elevation angle. In contrast to directional perception, distance perception is a far more complex process for rendering through headphones. Moreover, since our ability to estimate sound source distance is poor when compared to sound source directional estimation, we looked for simpler methods to encode distance. Therefore, after a first experiment with users where the encoding distance was used as a frequency variation (see Section 3.5.2.2), we decided to use spoken language – through messages like "John is three meters away".

Experimental Setup

In this section, we describe the experiments conducted to test and verify the accuracy of the system (Section 3.4.2) using our kinect-based dataset (Section 3.4.1.1) and a larger dataset (Section 3.4.1.2). In Section 3.4.3, we validate the user interaction aspects of the assistive wearable system.

Datasets

Unicamp Kinect Face Database

We built the Unicamp Kinect Face Database (UKFD) for this evaluation with proper authorization of Unicamp's Institutional Review Board. We collected videos from 30 subjects using the Kinect sensor and the tool *Microsoft Kinect Studio*. For each person, we captured 20 videos of 15 seconds with a frame rate of 30 FPS, giving a total of 600 videos. The dataset[10] comprises videos mainly by latinamerican subjects and few orientals, ten women and twenty men aged from 20 to 35 years. The videos were captured in an indoor environment. The Kinect was placed on a table at a distance of 3 meters from the bottom wall.

The participant was placed in front of the Kinect and with his/her back to the bottom wall, in the center of an area marked on the ground delimiting the Kinect's vision range where the people experience an optimal interaction (i.e., the optimum area for capturing depth data which is between 1.2 to 3.5 meters from the Kinect's position). We asked the participants to move in front of the Kinect freely within the mentioned Kinect's vision range area, but always trying to look at the kinect. The videos captured change in illumination (half the videos were recorded with full lighting and the others with reduced lighting), motion patterns (in half the videos the participants move in a natural way in the other at a faster speed), background (we use various colored cardstock to change the background color of the videos), with variation of facial expressions (the participants' facial expressions were not restrained – sometimes laughing, funny faces, serious faces, etc) and disguises (only with and without glasses). Fig. 3.3 depicts some representative sample faces from the dataset.

Unicamp Video-Attack Database

We also used the Unicamp Video-Attack Database (UVAD) [133]. The UVAD is a larger dataset that contains videos from 404 different subjects recorded with six different cameras and in two different sessions, considering the variation of background, lighting conditions and places (indoors and outdoors). This gives a total of 808 videos. All videos have Full HD quality, nine seconds of duration and a frame rate of 30 FPS.

Accuracy/Performance Experiments of the FRM

In the experiments, the 600 videos of the dataset were separated into five folds (120 videos/fold), each fold containing four videos per subject (4 videos/subject or 4 videos/class). For each video, we extracted the detected face images from the first 450 frames, obtaining 270,000

[10]This dataset will be made publicly-available upon acceptance of this paper. Meanwhile, research groups interested in this dataset may contact the laurindoneto@ufpi.edu.br

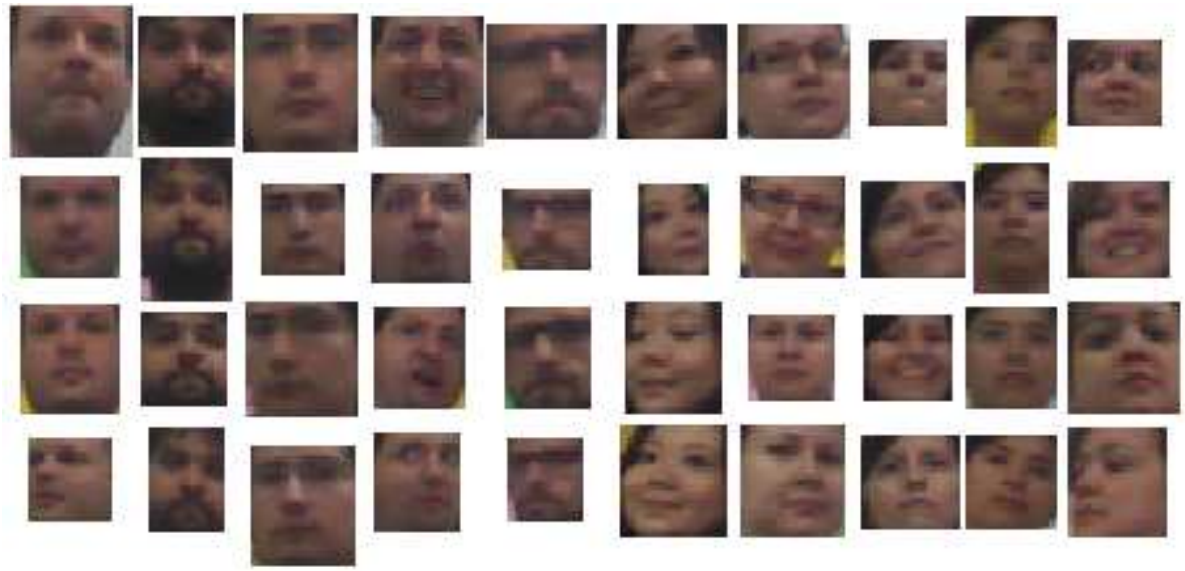


Figure 3.3: Some faces samples of our dataset.

samples for all five folds (54,000 samples/fold). Since in some of these frames there were no face detections, we computed them as non-detections. Next, we selected three random frames with face detection from each video and we created a sub-dataset of 1,800 samples (360 samples/fold). We repeated this selection process ten times, creating ten sub-datasets. For each sub-dataset, we performed a 5-fold cross-validation where four folds with 1,440 samples (48 samples/class) are used for training, and all samples from the remaining fold (54,000 samples) are used for testing. We trained over a small number of frames randomly selected from the four training folds and tested over all video frames from the test fold.

We performed a series of experiments described in Section 3.5 using the K-NN algorithm with $K = 3$, where we varied the samples per class, the sliding window size and the number of classes. Only in the last experiment we varied the hyperparameter K. These experiments serve both to find the best hyperparameters for our system and to compare it to other methods. Plots in Section 3.5 present the average accuracy rate over the ten sub-datasets for the mentioned experiments.

Moreover, we compared our approach to the Eigenfaces [78], Fisherfaces [134] and LBPH [85] methods. These traditional algorithms were chosen because they are publicly-available and they are as simple as our approach. Although state-of-the-art algorithms are not suitable for running on wearable devices, we also compared our approach to the current publicly-available state-of-the-art method, the VGG Convolutional Neural Network [135] (VGG CNN), in order to analyze the accuracy lost due to the wearable hardware constraints. It is worth noting that we used VGG as a feature extractor, and we have applied the same K-NN classifier and voting scheme described in Section 3.3.2 over the descriptors extracted from VGG. We used paired data in all the experiments and computed the p-value with both the Wilcox Signed Rank Test (WSR) and the Wilcox

Rank Sum Test (WRS) [136].

For a better understanding of the methods when dealing with more users, we considered an additional dataset (i.e., UVAD) in the experiments. We have computed the accuracy rate for UVAD, performing a 2-fold cross-validation with 30, 100 and 315 subjects (classes). We have selected the classes with at least 240 detected faces per video in both sessions (that is the reason we ended up with 315 classes). The face detection stage has been performed using the publicly-available Face++ API [11].

Finally, we performed experiments to evaluate an alternative people recognition approach in the dark, based on our original approach. In this alternative approach, instead of using RGB data for facial recognition, we used depth-only information.

User-Experience Experiments

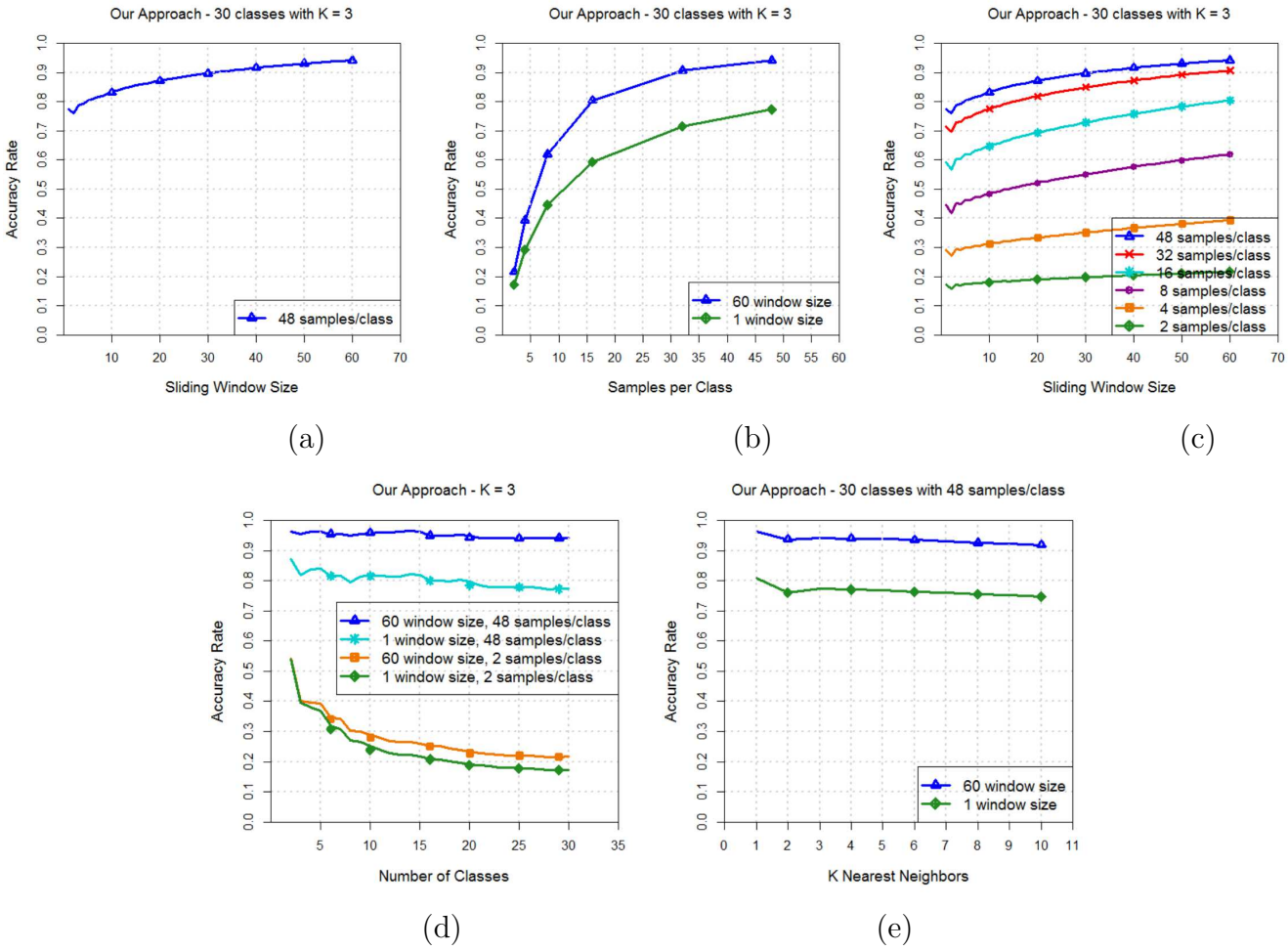


Figure 3.4: Results of the proposed approach: (a) sliding window size variation, (b) samples per class variation, (c) samples per class and sliding window size variation, (d) class number variation in the training base and (e) hyperparameter variation.

[11] http://www.faceplusplus.com/

Pilot Test

First, a pilot experiment (i.e., a first trial of the system within controlled experimental conditions) was conducted to test the KVB system. This pilot experiment allowed us not only to find critical problems (i.e., bugs and usability issues) in the system, but also, for further iterations, to refine it and improve the experiment itself. Design and usability heuristics were used as a way to formalize some of the impressions the participants had from the experiment. Participants were nine Human-Computer Interaction (HCI) researchers from the University of Campinas (São Paulo, Brazil), four of which were registered in the video dataset. The experiment was setup in rounds, each as the following:

(a) One participant volunteered to act as a blind user. After receiving brief instructions on how the system works, the participant was blindfolded and given the goal of both locating and reaching a target person;

(b) In silence, four other participants were placed in front of the blindfolded user. At least one of them was not registered in the dataset;

(c) The blindfolded user received the 3D audio feedback through the bone-conduction headphones, and through this feedback and body and/or head motion, scanned the room to locate the goal. For each person found, the user was asked to say who the system was telling that person was;

(d) The round was over either after two minutes or after the user signalized they had reached the goal.

Each participant went through two rounds and, from one round to the other, the group of four people changed.

Debriefing and Redesign

After the eighteen rounds, participants were gathered to discuss the experiment in a debriefing session. The objective was to talk about their impressions of the system, guided by a set of Natural User Interface (NUI) heuristics [137].

This discussion led to changes both in the KVB system's audio feedback information, and in the recognition algorithm (see the changes in Section 3.5.2.2). After these changes were made, a new experiment was conducted, this time with five adults that are actually blind. Three of them were born blind (one male, two females) and the other two (males) lost their sight recently.

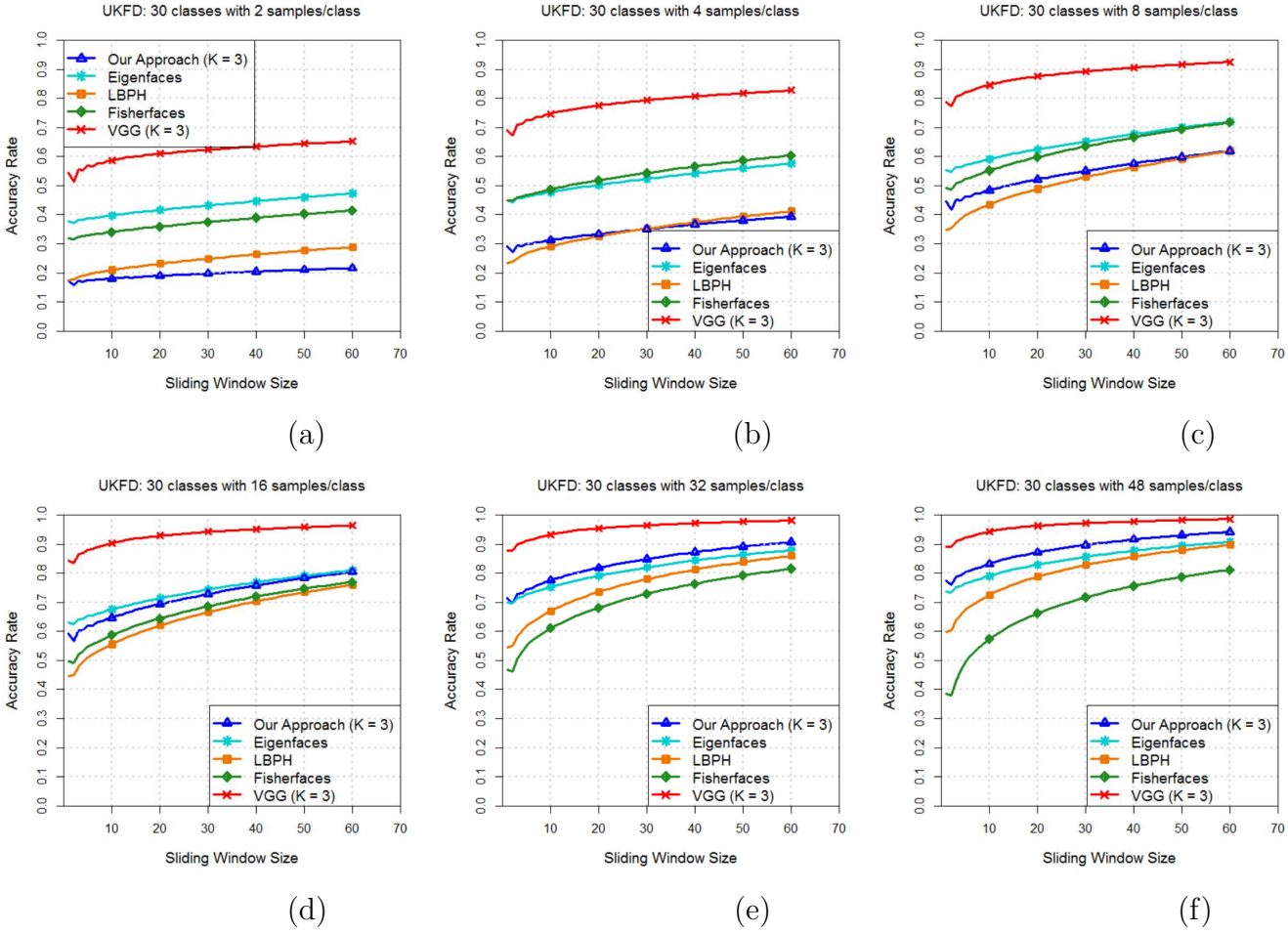


Figure 3.5: Comparing results of our approach to other methods, with: (a) 2 samples/class, (b) 4 samples/class, (c) 8 samples/class, (d) 16 samples/class, (e) 32 samples/class and (f) 48 samples/class.

Test with Visually Impaired Users

This time, instead of two rounds, each participant did one round of recognition. In addition, instead of four there were three people in front of them, all registered in the dataset. Then, instead of reaching one specific person, each blind participant had as goal locating and recognizing each of three people in front of them. After each round, participants were asked to respond to a Self Assessment Manikin (SAM) questionnaire [138], constructed with several layers of Ethylene-Vinyl Acetate (EVA) foam to make the manikins identifiable by touch. In addition, hidden behind the layers of EVA were RFID tags, one for each option of the questionnaire. Then, as participants approached an RFID reader to each manikin, a synthesized voice would describe aloud the selected option. If the last scanned choice was the one they wanted, participants would then scan a confirmation tag and their vote was registered.

Finally, we conducted a debriefing session with the blind participants, to understand how they deal with the task of people recognition without a system such as KVB, and

also to receive their feedback about the system.

Results

Next, we present the results of both Accuracy/Performance experiments and User Experience evaluations.

Accuracy/Performance Results

Sliding window size variation

In the first experiment, we dealt with window frame size variation (i.e., sliding window size). As expected, the higher the sliding window size, the greater the accuracy rate. The variation occurred in the range of 1 to 60 frames. Fig. 3.4(a) shows that the best accuracy rate is 94.26% with standard deviation $\sigma = 0.0031$ using a sliding window with 60 sequential frames. For the sliding window with 1 frame, the accuracy rate was 77.29% with $\sigma = 0.0030$.

Samples per class variation

In the second experiment, we varied the amount of samples per class of the training base. The best accuracy rate was 94.26% for a sliding window of size 60 frames and 48 samples/class. For the training base with 2 samples/class and a sliding window of size 60 frames, the accuracy rate was 21.64% ($\sigma = 0.016$), Fig. 3.4(b).

Samples per class and sliding window size variation

In the third experiment – a combination of the first two experiments – we varied both the amount of frames in the sliding window and the number of samples per class in the training base (Fig. 3.4(c)). We used the results of this experiment to compare our approach with other methods (see Section 3.6).

Class number variation in the training base

We varied the number of classes in the training base from 2 to 30 classes. We also varied the number of samples from 2 to 48 samples/class and the sliding window size from 1 to 60. In Fig. 3.4(d), we see that as the amount of samples per class increases and the number of classes decreases, there is an improvement on the accuracy rate.

Hyperparameter variation

In the last experiment, Fig. 3.4(e), we varied the sliding window size from 1 to 60, keeping the number of samples per class constant in 48 samples/class. We conducted a grid search for the best hyperparameter K for the K-NN algorithm, varying K from 1 to 10. The highest accuracy rate (96.44% with $\sigma = 0.0018$) was attained for $K = 1$.

Comparison using UKFD

We compared our approach to the Eigenfaces [78], Fisherfaces [134], LBPH [85], and VGG methods [135] through the experiment described in Section 3.5.1.3. Fig. 3.5 shows the results of this comparison for a sliding window size varying from 1 to 60, and samples per class going from 2 to 48. When compared to the other techniques, note that the accuracy rate of our approach performs the worst with two samples/class in the training base, and gradually improves with four, eight and 16 samples/class, and finally outperforms Eigenfaces, Fisherfaces and LBPH with 32 and 48 samples/class. As expected, VGG, the state-of-the-art approach, has the best accuracy rate (98.67%, $\sigma = 0.0017$) in all variations made. However, with 48 samples/class and window size of 60, the accuracy rate of our approach is 94.26% ($\sigma = 0.0031$), a competitive result in a real scenario.

Comparison using UVAD

Table 3.1 summarizes the results of the comparison using the UVAD database, showing the accuracy rates and standard deviations for a sliding window size of 60 frames and 48 samples/class in the training set of our approach to each of the other methods with 30, 100 and 315 classes. In the last experiment we did not consider Eigenfaces, Fisherfaces and LBPH as they had a poor performance with the experiment with 100 classes.

Table 3.1: Summary of comparison of accuracy rate and standard deviation (in parentheses) for UVAD with a sliding window size of 60 frames and 48 samples/class in the training set.

	Method				
# classes	**Our app.**	**VGG**	**Eigen**	**Fisher**	**LBPH**
30	86.84% (0.0641)	97.37% (0.0119)	19.34% (0.0674)	50.19% (0.1035)	81.41% (0.1321)
100	72.34% (0.015)	95.21% (0.006)	17.95% (0.033)	52.37% (0.0101)	51.84% (0.0757)
315	56.28% (0.0107)	89.71% (0.0091)	—	—	—

People recognition in the dark using depth-only data

We used 10 classes for this experiment, and we varied the sliding window size and the samples/class as depicted in Fig. 3.6. Observe that there was a large drop in accuracy

rate. The best accuracy rate was 56.23% ($\sigma = 0.0148$) and $K = 1$, using a sliding window size of 60 frames and 48 samples/class.

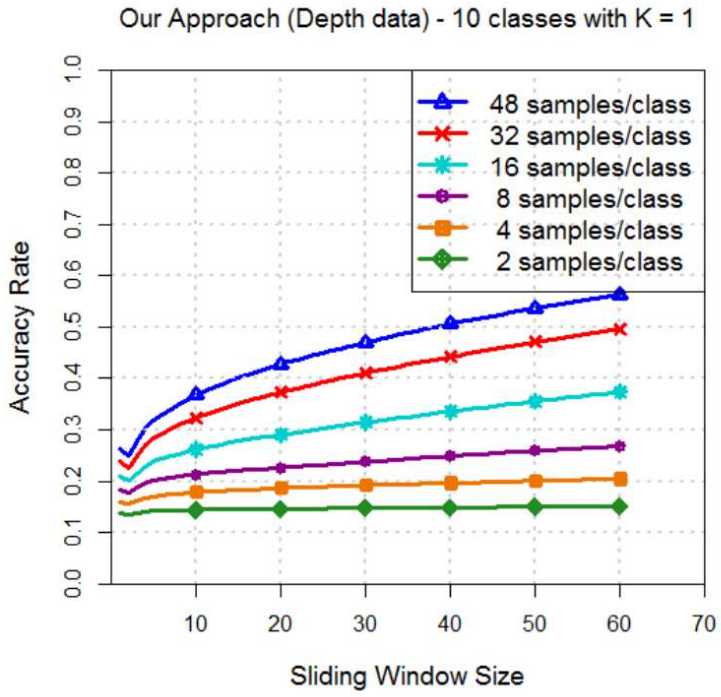


Figure 3.6: Results of the people recognition approach in the dark using depth-only data.

User-Experience Results

NUI Heuristics

The HCI specialists evaluated the KVB system through the 13 heuristics. First, specialists agreed that, on the compliance scale of -4 to 4, the KVB was a 4 on the [NH1] Operation Modes heuristic. They saw the user's displacement through the environment as an operation, which is a natural movement. Then, for the [NH2] "Interactability" heuristic, since the selectable objects are the people in front of the device, specialists also gave the maximum compliance grade. For the [NH3] Metaphor Adequacy heuristic, specialists understood that the system was a 3 on the scale because of the sound feedback that needed adjustments. It was indicating correctly how the user should move in terms of distance (front and back), but not in terms of sideways displacement (right and left). Finally, for the [NH7] Comfort heuristic, the compliance level was 4 because participants did not feel discomforts during the experiment, even with all the necessary equipment (helmet, backpack and headphones).

Interaction changes in the KVB system

In terms of audio feedback, there were three changes from the pilot to the second user-experience experiment.

First change is related to distance indication. In the pilot experiment the distance from the user to the person being recognized was translated into a high (if they were close) or low (if they were far apart) frequency of the person's name being repeated. This idea was based on parking sensors for cars, which emit beeps the closer the car is to colliding with something. However, in the context of the pilot experiment it did not work well, because the person's name is much larger than a beep, making it hard to map a frequency that both translates the distance correctly and maintains the name understandable. Hence, for the second user-experience experiment we changed the feedback to actually say how far someone is, e.g., "John is 3 meters away".

The second change in feedback is related to people recognition. In the pilot experiment the system would use the technical term "framing" to indicate it was recognizing someone, and the word "unknown" to show it found someone not registered in the dataset. The HCI experts found both words inadequate to nonprofessional users, so in the second experiment they were replaced by "trying to recognize" and "not recognized".

The third change was adding feedback for when the system is not sure if the person being recognized is actually registered in the database. In this case, the system says "It might be John", for instance. This audio feedback occurs only when the unknown class wins the voting, and the voting ratio between the second most voted class and the unknown class is greater than a threshold close to 1, indicating uncertainty between the two classes. There was a final change related to performance and accuracy. In the pilot, there was a delay when switching between identifying one person to identifying another, causing the feedback to still indicate the first person as the one being recognized. This delay was fixed for the second experiment.

Success rate of completed task

In the pilot experiment, where the user goal was of both locating and reaching a target person, in the 18 rounds most of the time the users recognized the people in front of him/her. However, only seven correctly reached the target person, resulting in a success rate of 39% in performing the task completely. After the changes made in the system, the next experiment, whose goal was of locating and recognizing each of three people in front of him/her, presented eight successful conclusions of the task from 15 rounds, resulting in a 53% success rate.

It should be noted that the environment where the second user-experience experiment was conducted had a lot of sunlight due to a background window (even with the curtains closed), which decreases the Kinect's performance. Although the conditions (different

task goals, different test environment) and subjects (blindfolded vs blind people) of both user experience experiments were different, the results suggest an improvement in the usability after the implementation of the changes described in Section 3.5.2.2.

SAM

In the Self Assessment Manikin questionnaire, participants rate their levels of Valence, Arousal and Dominance about their experience on a scale of 1 to 5. Table 3.2 shows their replies, and it is possible to see that the lowest grades were 3, and the mode for the five participants was 5 for each aspect (valence, arousal, dominance). In addition, four participants gave Dominance the maximum grade, showing most of them felt in control of the system.

Table 3.2: Results of the Self Assessment Manikin (SAM)

Participant	**Valence**	**Arousal**	**Dominance**
Participant 1	5	5	5
Participant 2	5	3	5
Participant 3	4	3	3
Participant 4	4	5	5
Participant 5	5	5	5
Mode	5	5	5

Blind user feedback

The participants from the second user-experience experiment thought that the voice that provided feedback was clear, and that the feedback messages made sense. Their only complaint in terms of ergonomy was the weight of the backpack, but they did not mind the helmet and suggested it could be stylish. However, they also showed real enthusiasm for the possibilities the system provides: autonomy and awareness of their surroundings. Without a system such as the KVB, participants said they usually have to rely on someone they trust (family member or friend) to describe or locate people around them. For recognizing familiar people,the blind subjects usually rely on their smell or voice. The participants told several stories of their daily life where they had problems because of their lack of awareness regarding the presence or absence of people in their surroundings. Therefore, the blind participants believe the system could be very helpful. In addition, they suggested that, instead of a helmet, the system could be placed on a cap or on glasses.

Discussion

The results of the experiments played a key role in the design of the real-time system. The sliding window size had to be chosen properly to obtain an acceptable accuracy rate without causing an excessive delay of the system's feedback. For instance, in a system

running at 20 FPS over a sliding window size of 60 frames, it will take roughly three seconds for the visually impaired user to hear the system's feedback, which is a reasonable delay time in this limited-scope application.

Moreover, the results show that, as the number of classes grows, our algorithm requires more samples/class in the training set to guarantee a high accuracy rate. In practice, this should not pose a problem since nowadays it is increasingly simple to obtain face images from family members (through social networks) and co-workers (through face recognition authentication systems, increasingly more used in workplaces).

Still, we used a small number of samples/class in the training set – up to 48 samples/class – across all experiments because our intention is to explore variations of this face recognition approach for other wearable devices with limited hardware resources, as we have already done on a smartwatch in Britto Neto et al. [139, 140].

Table 3.3 summarizes the results of the comparison, seen in Fig. 3.5(f), of our approach and each of the other methods, showing the accuracy rates for a sliding window size of 60 frames and 48 samples/class in the training set, and the p-value. Both in the WSR and in the WRS, the results of the p-values confirm that our approach performs the best among the compared traditional methods, showing a statistically significant improvement with 95% confidence. As expected, Table 3.3 also shows that VGG outperforms our approach by roughly 4% although being much more costly.

Table 3.3: Summary of comparison results for a sliding window size of 60 frames and 48 samples/class in the training set.

Method	**Accuracy**	σ	**WRS**	**WSR**
Our approach	94.26%	0.0031	—	—
Eigenfaces	90.83%	0.0044	0.0001197	1.776e-15
Fisherfaces	81.24%	0.0111	<2.2e-16	1.776e-15
LBPH	89.70%	0.0038	0.00126	5.862e-14
VGG	98.67%	0.0017	<2.2e-16	1.776e-15

As Table 3.1 shows, with 30 classes, the proposed approach still outperforms the traditional methods, keeping a high accuracy rate of 86.84%. Despite outperforming traditional methods, the performance of our approach decreases as the number of classes increases, reaching 56.28% for 315 classes.

Considering the lower accuracy rate for 315 classes in the UVAD database, one concern that might arise is the viability of implementing our approach in a real-life scenario. First, it should be kept in mind that this dataset is more challenging since it was collected using a variety of different cameras as well as at outdoors places, instead of a standard camera, which would be the camera of a given user. For instance, observe that the traditional approaches had a big drop in the accuracy rate even with 30 classes while our method was more resilient. Second and more important, considering that the mean offline social network size of an individual with a visual disability ranges from 15 persons for western

adolescents and young adults [63] to 31 persons for Chinese older adults [64], the number of faces to be recognized is considerably smaller than in other applications. In such scenario, our real-time approach still keeps a high accuracy rate of 86.84%, for UVAD, and over 94%, for UKFD, outperforming the traditional methods while requiring much less computational resources.

On the other hand, for generating a descriptor from a single video frame, observe, in Table 3.4, how the proposed approach uses less hardware resources than VGG and how fast its response time is. Moreover, Table 3.4 also shows that the memory consumption generated by storing the training bases in the proposed approach is much lower than VGG's.

Table 3.4: Computational time and memory footprint to generate a descriptor from a single video frame, and memory consumption of training bases with 315 classes, 48 samples/class, and components of the descriptors stored in variables with size of 4 bytes.

	Single video frame		**Training base**	
Method	**Time**	**RAM**	**Avg. desc. size**	**RAM**
Our approach	2.4ms	10.5MB	1090	78.59MB[a]
VGG (CPU)	347ms	663MB	4096	236.25MB
VGG (GPU)	48.1ms			

[a]It includes the size of the PCA rotation matrix: 3780x1090 = 15,72MB

Indeed, the VGG-CNN-based solution has shown interesting performance results, but for a real-time wearable system with hardware constrains, CNNs are not readily implementable in most of the existing Internet of Things (IoT) and wearable devices. For fast response times and high performance, CNNs need powerful graphics processing units which would overheat and drain the battery of a wearable device, currently curbing their use on resource-limited hardware.

In summary, considering the wearable hardware constraints and real-time requirements, our approach is still an efficient solution, showing a better performance than the traditional algorithms for this limited-scope application despite the accuracy loss when compared to state-of-the-art methods, that require much more memory and hardware resources.

With respect to the performance in the dark, although the results show that using depth-only information for people recognition is not an efficient approach, it proves to be a quite an interesting alternative when there is no light.

The User-Experience experiments allowed us to put the KVB system to the test with real potential users. The pilot experiment, although conducted with blindfolded users, allowed us to find critical performance, accuracy and usability issues.

The evaluation made by the HCI experts, guided by the NUI heuristics, showed a high level of compliance on the heuristics related to operation, interaction and comfort. This analysis also enabled us to find problems related to the audio feedback, either in its

contents or in its algorithm. Hence, the changes made in the KVB system after the pilot experiment provided a better experience for the users in the second experiment.

This improvement, also suggested by the increase in the success rate during the second user-experience experiment, is confirmed in the high levels of Valence, Arousal and Dominance reported by the blind users regarding their experience. The qualitative feedback given during the debriefing shows the excitement the visually impaired have towards technology that can provide them more autonomy and quality of life.

Conclusions

We described a wearable face recognition real-time system to aid the visually impaired. The system uses a Kinect sensor to acquire a RBG-D image, an efficient face recognition algorithm based on HOG, PCA and K-NN, and finally 3D audio to reproduce a virtualized sound at the detected position.

It is encouraging to verify that, for a limited-scope application, considering the wearable hardware constraints and real-time requirements, it is enough to use a simple and straightforward face recognition approach for achieving high accuracy rates. This shows that the complexity of the state-of-the-art in biometrics may be an overkill for limited-scope applications. We also demonstrated that, although using our approach over depth-only data is not an efficient algorithm, it has proved to be a quite an interesting alternative when there is no light. It would indeed be interesting to have an in-depth analysis on how to combine both information keeping a real-time performance to obtain better performance. We intend to explore these new ideas as a future work.

Our results confirm the potential of Microsoft Kinect sensor for accurate and fast face detection. The availability of depth and RGB data will be crucial for future work involving object recognition and navigation tasks using our KVB platform, since it permits a robust and rapid object detection.

The main limitation of Kinect's infrared technology is its lack of robustness under sunlight, turning kinect-based systems not suitable for outdoors environments. Moreover, according to our experience, even in indoor environments, sunlight coming in through the windows as well as the presence of reflective surfaces such as mirrors, hinder the proper operation of Kinect's infrared sensor. Still, note that the modules of our system might be easily adapted for running on other sensor technologies, such as stereo cameras, or even on other hardware platforms as we have already done in a smartwatch [139, 140]. Conveying the directional location of a face in the environment using 3D Audio proved to be an efficient feedback that does not overwhelm the person's auditory sense. On the other hand, distance location appears to be more challenging to communicate. After the pilot experiment, we passed from a frequency variation approach to a spoken language approach. Further research should explore alternative forms to encode distance (e.g.,

musical tones).

Another concern that might arise is the relatively low success rate in the user experiments. Although the users were very confident when using the prototype, we realized that some training time is needed to increase the spatial awareness of the blind user wearing the system. Moreover, generic HRTFs, as used in this initial stage, tend to hinder the user's audio localization capability. Even so, this disadvantage of spatial audio is overshadowed by its versatility of not overwhelming the user's hearing sense, which is crucial for blind people, especially when cognitive load is present [124]. All these factors together affect the overall performance of the system. During the development of the KVB system, we are still working to overcome these challenges in future work.

Another interesting point is the effect of a moving Kinect on the system's performance. Although the Kinect was reasonably stationary when the videos of the database were collected, we asked the people to move in front of the camera. Still, it is worth keeping in mind that sometimes during the experiments with people, the head movements might not be as smooth as in the dataset videos. A moving camera tends to capture blurry and noisy images because the lack of focus caused by its movement [141] which, in turn, are more challenging to recognize. These factors affect any camera in movement and, in our case, certainly might reduce the recognition performance during the real-life experiments. In a future work, it would interesting to consider deblurring techniques [142] or digital image stabilization [143] to address this issue.

The experiment with blind users showed us that there is an actual social demand for a real-time face recognition system for the visually impaired, and that the KVB system is on the right track to meet this demand. The feedback from the users provided us with valuable information regarding the usual strategies blind people use to recognize people in their daily lives, which, in turn, points to interesting future work. For instance, giving the user a full description of an unknown person, such as gender, height, age and clothes.

Finally, for future work, we intend to perform experiments with other available benchmarks and datasets, such as the Lock3DFace [144], which comprises comprehensive and difficult situations complementary to those considered in this paper, e.g., face occlusion with glasses, left or right face profiles, and different periods of acquisition over time.

Chapter 4

Contribution II

Contribution II is published as:

F. Grijalva, L. Martini, D. Florencio and S. Goldenstein, "A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 559-570, March 2016. doi: 10.1109/TASLP.2016.2517565

A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features

Felipe Grijalva[1], *Student Member, IEEE,* Luiz Martini[1],
Dinei Florencio[2], *Senior Member, IEEE,*
and Siome Goldenstein[3], *Senior Member, IEEE*

Abstract

We present a new anthropometry-based method to personalize Head-Related Transfer Functions (HRTFs) using manifold learning in both azimuth and elevation angles with a single nonlinear regression model. The core element of our approach is a domain-specific nonlinear dimensionality reduction technique, denominated Isomap, over the intraconic component of HRTFs resulting from a spectral decomposition. HRTF intraconic components encode the most important cues for HRTF individualization, leaving out subject-independent cues. First, we modify the graph construction procedure of Isomap to integrate relevant prior knowledge of spatial audio into a single manifold for all subjects by exploiting the existing correlations among HRTFs across individuals, directions, and ears. Then, with the aim of preserving the multifactor nature of HRTFs (i.e. subject, direction and frequency), we train a single artificial neural network to predict low-dimensional HRTFs from anthropometric features. Finally, we reconstruct the HRTF from its estimated low-dimensional version using a neighborhood-based reconstruction approach. Our findings show that introducing prior knowledge in Isomap's manifold is a powerful way to capture the underlying factors of spatial hearing. Our experiments show, with p-values less than 0.05, that our approach outperforms using, either a PCA linear reduction, or the full HTRF, in its intermediate stages.

Introduction

The *Head-Related Transfer Functions* (HRTFs) encode audio localization cues such as *Interaural Time Difference* (ITD), *Interaural Level Difference* (ILD) and spectral coloring, caused by sound scattering around the head, pinna and torso before it reaches the eardrum [145].

Since HRTFs differ widely among individuals, it is necessary to personalize them to ensure high-quality spatial audio. Nonindividualized HRTFs hinder localization accuracy, often causing front-back and up-down confusions [146].

[1]School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil
[2]Microsoft Research, Redmond, WA, USA
[3]Institute of Computing, University of Campinas, Campinas, SP, Brazil.

The unsolved problem of HRTF customization is increasingly subject of much research due to the growing importance of auditory augmented reality applications [147, 130]. The most accurate approach to personalizing HRTFs is through direct measurements. However, this is a complex, time-consuming, expensive, and not scalable procedure [148].

In light of this, several alternative methods aimed at avoiding measuring HRTFs have been proposed, including the estimation of HRTFs from a small set of measurements [149]. Furthermore, there are several theoretical models (e.g. spherical head model [150], snowman model [151], structural models [152, 153]) which attempt to approximate the complicated human anatomy. Additionally, several numerical methods (e.g. boundary element method [154, 155], finite-difference time-domain method [156]) have been proposed. However, they require expensive acquisition hardware and are computationally intensive. Pursuing a different direction, several authors have proposed perceptual-based methods, where subjects choose their HRTFs through listening tests by tuning some parameters until they achieve an acceptable spatial accuracy [157, 158]. Moreover, Sunder et al. proposed an individualization method in the horizontal plane that uses a frontal projection headphone to introduce idiosyncratic pinna cues [159].

Alongside the aforementioned methods, HRTFs can also be customized from anthropometric measurements. Anthropometry-based regression methods predict individualized HRTFs using a model derived from a baseline database. It is precisely this kind of individualization methods that this work focuses on.

This paper introduces a new customization method to personalize HRTFs using Isomap, a nonlinear dimensionality reduction technique. Here, we extend for all directions the ideas of our preliminary study in the horizontal plane [160]. Our main contribution is our graph construction procedure for learning a single Isomap manifold for all subjects that incorporates important prior knowledge of spatial audio to exploit the correlation existing among HRTFs across individuals, directions and ears. Besides, instead of personalizing the HRTFs directly, we customize the *intraconic* component of HRTFs resulting from a spectral decomposition [161]. The intraconic component of HRTFs aims at providing the most important cues for individualization, leaving out subject-independent cues. Finally, our approach constructs a single regression model using an artificial neural network that does not break the inherent multifactor nature of HRTFs (i.e. frequency, direction and subject factors).

Related Work

Anthropometry matching is the most straightforward way to personalize HRTFs from anthropometric data. In this context, various approaches in the literature [162, 163, 164] customize HRTFs by finding the best match in a baseline database of anthropometric features.

Middlebrooks [165] introduced an anthropometry-based method that uses frequency scaling of HRTFs based on the assumption that inter-subjects difference in anatomy features produce a frequency shift in individualized HRTFs.

On the other hand, anthropometric regression methods predict the individualized HRTFs of a new subject using a model derived from a baseline database. Linear dimensionality reduction techniques such as *Principal Component Analysis* (PCA) [166] and *Independent Component Analysis* (ICA) [167] have been widely used prior to customization. There are several HRTF customization methods to map anthropometric features to low-dimensional HRTFs previously calculated with PCA [168, 169, 170, 171].

Due to the inability of linear regression methods to predict the complex relationship between anthropometric features and low-dimensional HRTFs, various authors introduced nonlinear regression techniques such as *Artificial Neural Networks* (ANN) [172] and *Support Vector Regression* (SVR) [167, 173] in conjunction with PCA [172, 173] or ICA [167]. Moreover, because SVR is only capable of training a multiple-to-one regression model (i.e. SVR needs to train a separate model for each dimension of low-dimensional HRTFs), Wang et al. [174] proposed a *joint SVR* to exploit the correlation between components of low-dimensional HRTFs.

The anthropometry-based regression methods described so far construct a model for each direction, which in turn means that the inherent multi-factor nature of HRTFs (i.e. frequency, direction and subject) is broken [175]. To overcome this problem, Grindlay et al. [176] introduced a three-mode (i.e. frequency, direction and subject mode) multilinear tensor representation for HRTFs . A single linear regression model is used for mapping anthropometric features to a five-dimensional vector obtained by means of *N-mode Singular Value Decomposition* (N-mode SVD) and which represents the subject mode in the tensor. A similar tensor-based approach is used in [175] and [177], but to construct the regression model, they employed an ANN and high-order partial least squares, respectively. More recently, Bilinski et al. [178] used a HRTF tensor representation to learn a sparse vector of a subject's anthropometric features as a linear superposition of the anthropometric features of a training subset. They applied the same sparse vector to synthesize the HRTF of a subject.

In addition to linear representations, nonlinear dimensionality reduction techniques have been also applied to HRTFs. Duraiswami et al. [179] applied *Locally Linear Embedding* (LLE) [180] to learn the nonlinear manifold structure in median plane HRTFs of the same subject. They also proposed a new HRTF interpolation method that estimates an HRTF as a linear combination of its neighbors on the low-dimensional manifold.

Furthermore, Kapralos et al. compared PCA, Isomap [181] and LLE through correlation analysis [182] and subjective experiments [183], concluding that Isomap and LLE outperform PCA in finding the underlying factors of spatial hearing.

Based on the results of Duraiswami et al. [179] and Kapralos et al. [182, 183] using

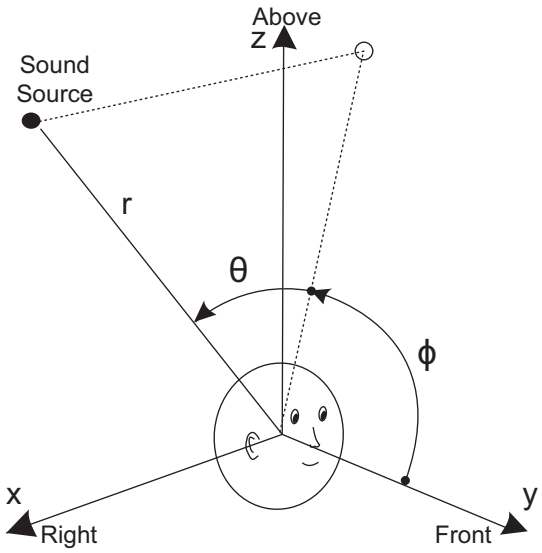


Figure 4.1: The interaural coordinate system as described in [187], where azimuth is defined in the range $-90° \leq \theta \leq 90°$ and elevation in $-90° < \phi \leq 270°$.

LLE and Isomap for HRTF interpolation and dimensionality reduction, in our previous work [160], we proposed a novel technique for customizing horizontal plane HRTFs using Isomap.

All aforementioned manifold learning studies support the idea suggested by Seung et al. [184] that nonlinear manifold techniques are crucial for understanding how perception arises from the dynamics of neural networks in the brain.

In this paper, we extend the ideas of our previous customization method [160] for locations beyond the horizontal plane. In this line, we apply Isomap over HRTFs to construct a manifold structure and then we employ an artificial neural network to predict the HRTFs for a new subject based on his anthropometric parameters.

As in previous works [169, 172, 176, 160], we work with the minimum phase assumptions of HRTFs [185], i.e., a minimum-phase function cascaded with a pure delay. In practice, the pure delay is the ITD and it is commonly cascaded in either the left or right HRTF of each left-right HRTF pair. It is important to stress that the calculation of ITD is beyond the scope of this work. Various studies address the estimation of ITD, notably in [150, 186]. Besides, unlike previous works [169, 172, 176, 160], we do not personalize the HRTFs or the directional transfer functions (i.e. mean removed HRTFs [166]) directly. Instead, we customize the intraconic component of HRTFs resulting from a spectral decomposition of HRTFs magnitude as suggested by Romigh and Simpson [161]. Here, we focus only on the spectral features of the intraconic component of HRTFs magnitude and, unless otherwise stated, when we refer to HRTF we are referring to its intraconic portion. Finally, in this work, we only use the interaural coordinate system [187] depicted in Figure 4.1.

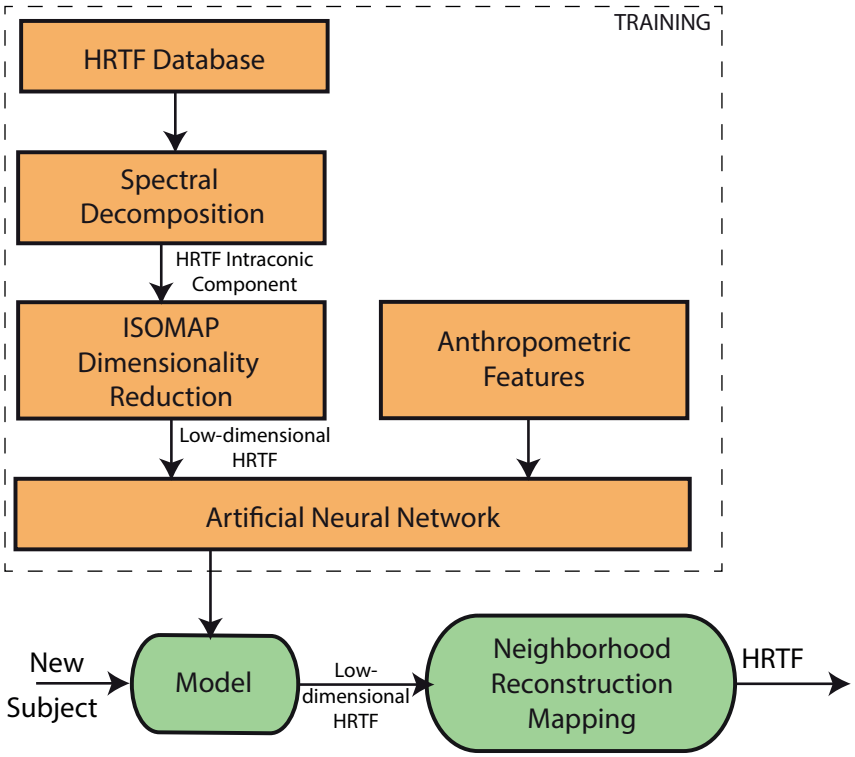


Figure 4.2: Pipeline of our HRTF customization approach.

Methodology

Figure 4.2 summarizes the pipeline of our HRTF customization approach. First, the extraction of the intraconic portion from full HRTFs aims at providing the most important cues for individualization, leaving out subject-independent cues. Then, Isomap with a custom graph construction procedure performs a nonlinear mapping of the intraconic component of HRTFs to a low-dimensional space. Subsequently, an ANN learns a regression model from a training dataset to relate anthropometric features to low-dimensional HRTFs. Finally, for a new subject with known anthropometric parameters, the model predicts his low-dimensional HRTFs which in turn are mapped back to the high-dimensional space by means of a neighborhood-based reconstruction approach.

We performed simulations of our approach on the CIPIC HRTF database [187]. We estimated the performance of such simulations using k-fold cross-validation and spectral distortion as metric. For comparison, we implemented PCA instead of Isomap for dimensionality reduction, and we tested the full HRTFs instead of their intraconic component. In summary, four conditions were tested: Isomap over full HRTFs, Isomap over the intraconic portion of HRTFs, PCA over full HRTFs and PCA over the intraconic portion of HRTFs. We also performed paired t-tests between the aforementioned conditions.

We chose only PCA for comparison because in this work we aim at exploring, first, whether it is worth using more complex techniques in the dimensionality reduction stage of anthropometry-based methods. Therefore, here we preferred to focus on how to construct

and interpret a single manifold for all subjects, which had also not been addressed by prior works using manifold learning. Finally, we use a spectral distortion metric, as widely used in similar studies.

HRTF Personalization

Spectral Decomposition

In a recent study, Romigh and Simpson [161] decomposed the HRTF at each location as the sum of average, lateral and intraconic spectral components. First, they obtained directional spectra by subtracting the mean across all locations (i.e. the average component) from each HRTF. Then, they calculated the lateral component for each azimuth angle as the median spectrum of all directional spectra measured at that azimuth angle. Lastly, they computed the intraconic component by subtracting the corresponding lateral component from the directional spectra at each location [161]. In order to recover the original HRTF spectrum at each location, they added together the corresponding average, lateral and intraconic components. Finally, the complex-valued HRTF were recovered using minimum phase assumptions.

After a series of psychoacoustic experiments where a listener's component were swapped out for the corresponding KEMAR's component, Romigh and Simpson found that the intraconic component encodes the most important cues for HRTF individualization and localization is only minimally affected by introducing non-individualized cues into the other HRTF components [161].

Based on these results, we used the intraconic spectral components as ground-truth HRTFs instead of the full ones. For simplicity, we will use the term intraconic HRTF when referring to its intraconic component[4].

Dimensionality Reduction using Isomap

Let $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^D$ be a high-dimensional dataset in a $D \times N$ matrix of N sample vectors $\mathbf{x}_i$ and $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \subset \mathbb{R}^d$ be a corresponding low-dimensional representation in a $d \times N$ matrix of N sample vectors $\mathbf{y}_i$, where $d < D$.

Isomap is a nonlinear dimensionality reduction technique first introduced in [181] that provides a method for reducing $\mathbf{X}$ into a low-dimensional embedding $\mathbf{Y}$. Linear dimensionality reduction methods such as PCA attempt to preserve pairwise Euclidean distances by retaining most variance as possible [181]. However, such techniques does not take into account the datapoint neighborhood [188].

[4] To recover the full HRTF, the intraconic, lateral and average components should be added together.

On the other hand, Isomap aims to maintain the intrinsic geometry of data (i.e. the datapoint neighborhood relationships) by preserving the pairwise geodesic distances (i.e. the distance over the manifold) [181]. For example, in nonlinear manifolds such as in the Swiss Roll dataset [188], PCA might map two datapoints as near points as measured by the Euclidean distance, while their geodesic distance is much larger.

Isomap can be summarized in three steps. The first step is to construct a graph $G(V, E)$ on the high-dimensional dataset $\mathbf{X}$. Each sample $\mathbf{x}_i \in \mathbf{X}$ is represented by a node $v_i \in V$, and two nodes v_i and v_j are connected by an edge $(v_i, v_j) \in E$ with length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ is one of the K nearest neighbor of $\mathbf{x}_j$. The edge length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ is given by some distance metric between $\mathbf{x}_i$ and $\mathbf{x}_j$ [181]. A common metric and the one used in this paper is the Euclidean distance.

In the second step, we calculate the geodesic distance between each pair of points by computing the shortest path between these two nodes in G. Then, after calculating the geodesic distances between all datapoints in $\mathbf{X}$, they are stored pairwise in a matrix $\mathbf{D}_G$. The pairwise geodesic distance matrix $\mathbf{D}_G$ represents the geodesic distances between all samples on the manifold [189].

The third and final step is to construct the d-dimensional embedding by applying multidimensional scaling (MDS) on $\mathbf{D}_G$ [188]. Formally, the eigenvectors of the double-centered matrix $\tau(\mathbf{D}_G)$ are calculated, where $\tau(\mathbf{D}_G) = -\mathbf{H}\mathbf{S}_G\mathbf{H}/2$, $\{S_G\}_{ij} = (\{D_G\}_{ij})^2$ (i.e. $\mathbf{S}$ is the matrix of squared distances) and $H_{ij} = \delta_{ij} - 1/N$ (i.e. $\mathbf{H}$ is the centering matrix). Recall that N is the number of sample points and δ is the Kronecker delta function. Finally, let λ_p be the p^{th} eigenvalue (in decreasing order) of the matrix $\tau(\mathbf{D}_G)$, and v_p^i be the i^{th} component of the p^{th} eigenvector. Then set the p^{th} component of the d-dimensional coordinate vector $\mathbf{y}_i$ equal to $\sqrt{\lambda_p} v_p^i$ [181].

In the first step of Isomap, we need to construct a graph, i.e., we need to select a number of neighbors for each high-dimensional point. Common approaches construct the graph by finding the K nearest neighbors or all neighbors within a specified radius r of each data point. In general, neighborhood selection in Isomap presents an opportunity to incorporate a priori knowledge from data [190]. With this in mind, we aim at constructing the graph G by taking advantage of the existing correlations among the HRTFs at different directions, frequencies, and individuals. One of our contributions is our graph G construction procedure:

Criterion 1. *if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent HRTFs of the same location and ear but different subject, then connect them.*

In previous studies [169, 170, 167, 172, 174], they performed dimensionality reduction separately for each direction. Here, instead of applying Isomap separately for each location and ear, with this criterion, we tried to exploit the correlation of HRTFs among subjects across same directions. Using this criterion, $P - 1$ neighbors were obtained, where P is

the number of subjects in the dataset $\mathbf{X}$.

Criterion 2. *Let (θ_i, ϕ_i) and (θ_j, ϕ_j) be interaural coordinates (azimuth θ and elevation ϕ) of HRTFs represented by $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. Regardless of the subject, if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent HRTFs of same elevation (i.e. $\phi_i = \phi_j$) and opposite ears, and θ_i is the mirror horizontal azimuth of θ_j (i.e. $\theta_i = -\theta_j$), then connect $\mathbf{x}_i$ and $\mathbf{x}_j$.*

The intuition behind this criterion was to take advantage of the correlation existing due to left-right symmetry of HRTFs at frequencies below 5.5 kHz [191]. Applying this criterion, P neighbors were obtained.

Criterion 3. *Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be HRTFs of the same subject and ear. If $\mathbf{x}_j$ is one of the eight HRTFs surrounding $\mathbf{x}_i$, then connect them.*

The intuition behind this criterion was to emphasize the similarities between spatially close HRTFs of the same subject and ear. Using this criterion, eight neighbors were obtained.

With the aim of clarifying how the above mentioned criteria were applied, Figure 4.3 shows an illustrative example. Note that with our criteria, it is straightforward to prove that the constructed graph G is always connected.

Before applying Isomap, we first need to select the number of neighbors, K, and the intrinsic dimensionality, d. Due to the criteria proposed for the graph construction explained earlier, the number of neighbors was set to $K = 2P+7$, i.e., $P-1$ from Criterion 1,

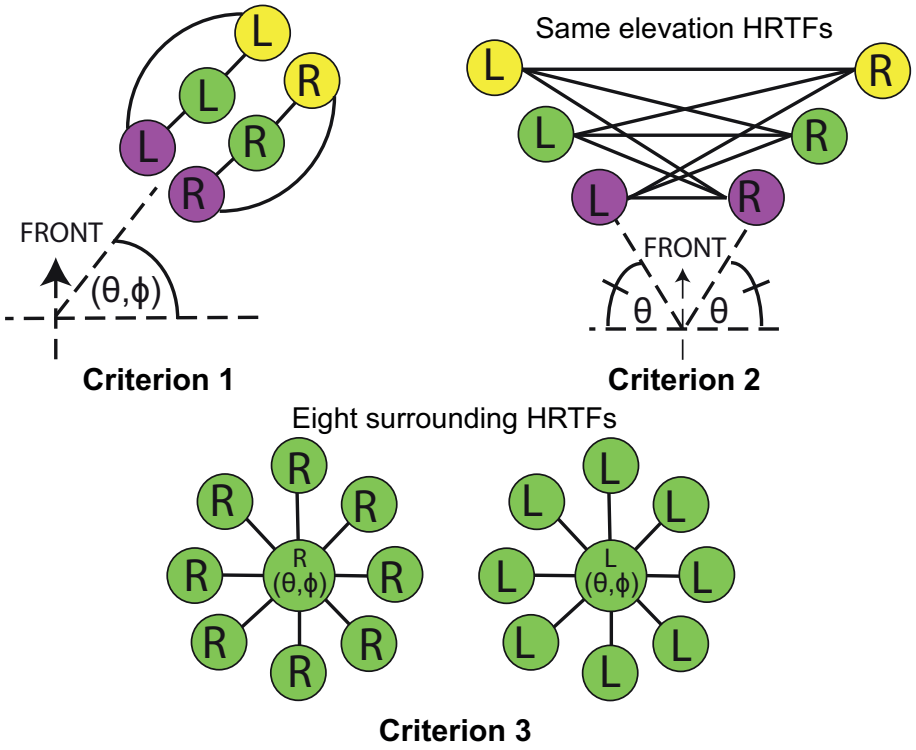


Figure 4.3: Illustrative example of the criteria to construct the Isomap's graph for $P = 3$ subjects. Color represents HRTFs of the same subject, and (θ, ϕ) represents the azimuth and elevation in the interaural coordinate system. L=Left ear and R=Right ear.

P from Criterion 2 and eight from Criterion 3. We determined the intrinsic dimensionality by means of the maximum likelihood intrinsic dimensionality estimator [192]. This dimensionality estimator attempts to reveal the intrinsic geometric structure of the observed data and it has demonstrated to be a good choice in manifold learning problems [188, 193].

Finally, note that, unlike previous works [167, 169, 170, 171, 172, 174], we applied dimensionality reduction only once, over the entire dataset, for HRTFs of all subjects, directions and ears. This way, as tensor-based approaches [175, 176, 177, 178] do, we tried to preserve the multi-factor (i.e. frequency, direction and subject) nature of HRTFs.

Regression using an Artificial Neural Network

Artificial Neural Networks (ANN) are systems capable of approximating nonlinear functions of their inputs. Since the relationship between HRTFs and anthropometric parameters is very complex, a nonlinear predictor is suitable for this task. Here, we used a back propagation ANN[5] whose inputs are s anthropometric parameters, the azimuth angle, the elevation angle, and the ear (Left=1, Right=-1). The outputs of the ANN are the low-dimensional HRTFs obtained by Isomap. Besides, the ANN uses sigmoid activation functions in the hidden layer and a linear activation function in the output layer.

We trained the ANN using Levenberg-Marquardt optimization and an early stopping approach for improving generalization and to avoid overfitting. This way, we used a training subset for updating the network parameters. We also monitored a validation subset during the training process. When the validation error increased for 10 iterations, the training was stopped and the network parameters at the minimum of the validation error were returned.

We varied the number of hidden layer units and selected 35 hidden nodes that produced the lowest mean squared validation error. With this network topology, we achieved a mean squared validation error of 0.0078 that corresponds to a 0.91 coefficient of determination (R^2-value).

After the regression model is learned, the individual HRTF on the low-dimensional space for a new subject can be predicted by his anthropometric parameter measurements.

Finally, since our approach trains only one ANN for all HRTF data, the ANN exploits the relationships between low-dimensional components of HRTFs across directions and ears.

[5] A Multilayer Perceptron with a single hidden layer

Neighborhood Reconstruction Mapping

Unlike linear reduction techniques, Isomap produce a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \in \mathbb{R}^d$ from the samples in $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^D$ without generating an explicit map [188]. As we are interested in high-dimensional HRTFs, we need to project a low-dimensional HRTF predicted by the ANN back into the original space. Since Isomap assumes that a sample and its neighbors are locally linear, we can perform the mapping using a linear combination of a sample's K neighbors. Thus, the reconstructed HRTF, $\hat{H}$, is

$$\hat{H} = \sum_{i=1}^{K} w_i \mathbf{x}_i. \tag{4.1}$$

To calculate the weights w_i, we followed Brown et al. [189], and chose w_i to be the inverse Euclidean distance between the sample and the neighbor i in the low-dimensional embedding.

Experiments

We implemented the proposed personalization method according to the block diagram in Figure 4.2. Next, we define the elements and conditions of the simulations.

HRTF Database We used the publicly available CIPIC database [187] which contains Head-Related Impulse Responses (HRIRs) of both ears measured for 45 subjects at 25 azimuths and 50 elevations (i.e. $M = 1250$ locations per subject and ear) in the interaural coordinate system. We selected only the subjects whose anthropometric features were complete (i.e. 35 subjects). Because not all anatomical features of CIPIC database are relevant for HRTF individualization, we selected $s = 8$ anthropometric parameters according to [194]: head depth, pinna offset back, cavum concha width, fossa height, pinna height, pinna width , pinna rotation angle and pinna flare angle. For selecting those parameters, the authors in [194] performed a statistical analysis in the entire virtual auditory space (i.e. azimuth and elevation) based on PCA, Pearson's product-moment correlation coefficient analysis and multiple linear regression analysis. Note that the only parameter related to head dimensions is head depth. Although both head depth and head width are important for ITD estimation [150] (head height is less relevant), keep in mind that our work does not deal with ITD but with the spectral features of minimum phase's magnitude of HRTFs as stated in Section 4.2. Thus, the fact that we use head depth as the only parameter related to head dimensions in our study does not pose a critical problem.

HRIR pre-processing Each HRIR from CIPIC database has roughly 4.5 ms (i.e. 200 samples long) for a frequency sampling of 44.1 kHz and 16 bit resolution. First, we transformed each HRIR into a HRTF by means of a 512-point FFT. In order to reduce the effects of the limitations in the frequency response of the equipments utilized for HRIR measurement, we filtered the HRTFs to retain frequencies between 200 Hz and 15 kHz, leaving 172 magnitude coefficients (i.e. each HRTF is a point in a $D = 172$ dimensional space)[6]. Finally, we applied the spectral decomposition described in Section 4.4.1 to the filtered HRTFs, preserving the intraconic component. In order to analyze the effects of personalizing the intraconic HRTFs instead of the full ones, we performed the same simulation for both conditions.

k-fold Cross-Validation We divided the dataset into five folds of seven subject each. Then, we applied k-fold cross-validation, using four folds for training (i.e. $P = 28$ subjects) and one for testing. So, we estimated the model from $N = 2 \cdot P \cdot M = 70,000$ HRTFs.

Dimensionality Reduction For Isomap, the intrinsic dimensionality was estimated by means of the maximum likelihood intrinsic dimensionality estimator [192]. So, we reduced the N HRTFs of dimension $D = 172$ to $d = 5$ dimensions. Since the number of subjects is $P = 30$[7], for the Isomap graph construction, each HRTF is connected to $K = 2P + 7$ neighbors. On the other hand, instead of Isomap, we implemented also PCA for comparing both methods. For PCA, we used $d = 5$ components that correspond to 88% of variance retained, which is in line with previous studies showing that five PCA components (approximately 90% of variance) capture the most perceptually relevant properties of HRTFs [166, 195]. Finally, k-fold cross-validation was also applied to evaluate the PCA performance. Both Isomap and PCA were implemented using Matlab Dimensionality Reduction Toolbox [188].

Neural Network As explained in Section 4.4.3, the inputs of the artificial neural network are the $s = 8$ anthropometric parameters, the ear (Left=-1, Right=1), the azimuth and the elevation. The outputs are the low-dimensional HRTFs. We used Matlab Neural Network Toolbox to implement the ANN.

Performance Metric As an error metric, we chose the mean spectral distortion in dB defined by

[6]It is worth noticing that our method does not predict frequency bins for $f < 200$ Hz. Low-frequency components for $f < 200$ Hz might be recovered by extrapolation after the personalization procedure.

[7]Errata: It is $P = 28$

$$SD_M = \sqrt{\frac{1}{N_f}\sum_{k=1}^{N_f}\left(20\log_{10}\frac{|H(k)|}{\left|\hat{H}(k)\right|}\right)^2}, \tag{4.2}$$

where H and $\hat{H}$ represent the measured and reconstructed HRTF, respectively, and N_f is the number of frequency points. The reconstructed HRTF, $\hat{H}$, was calculated using Equation 4.1.

In summary, we tested the following four conditions:

1. Intraconic HRTFs as ground-truth and Isomap as dimensionality reduction method, labeled as Intra-ISO.

2. Full HRTFs as ground-truth and Isomap as dimensionality reduction method, labeled as Full-ISO.

3. Intraconic HRTFs as ground-truth and PCA as dimensionality reduction method, labeled as Intra-PCA.

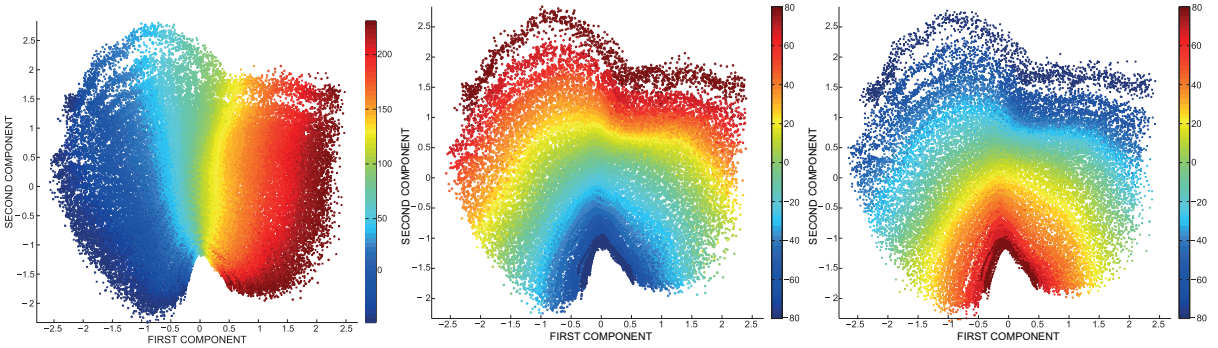
4. Full HRTFs as ground-truth and PCA as dimensionality reduction method, labeled as Full-PCA.

Analysis and Results

Isomap Manifold Analysis

It is important to analyze how the Isomap embedded components relate to source azimuth and elevation. Figure 4.4a shows the two-dimensional manifold (i.e. first embedded dimension vs second one) where the color represents the elevation angle. We observe that the first component of Isomap embedding roughly increases with elevation. That tendency is confirmed by the correlation coefficient between elevation angle and the first component value, which is 0.94.

Figures 4.5a through 4.5c present the first component as a function of source location. In Figure 4.5a is evident the strong correlation between elevation and first component. Besides, first component's value is negative for front locations (i.e. $\phi < 90°$) while it is positive for rear positions. This pattern suggests that the first component can distinguish front locations from back ones. Notice also in the same figure that there is a tendency for the first component to increase in magnitude as the source moves from the frontal plane (i.e. $\phi = 90°$). On the other hand, in Figures 4.5b and 4.5c, there is no clear pattern between first component and azimuth as in the case of elevation. In fact, error bars tend to increase as source moves towards contralateral locations and their mean value (i.e. datapoints in plots) keep roughly constant with respect to azimuth angle.



a) Both Ears. Color represents Elevation. b) Left Ear. Color represents Azimuth. c) Right Ear. Color represents Azimuth.

Figure 4.4: Two-dimensional Manifold. All Isomap components are normalized to have zero mean and unit variance.

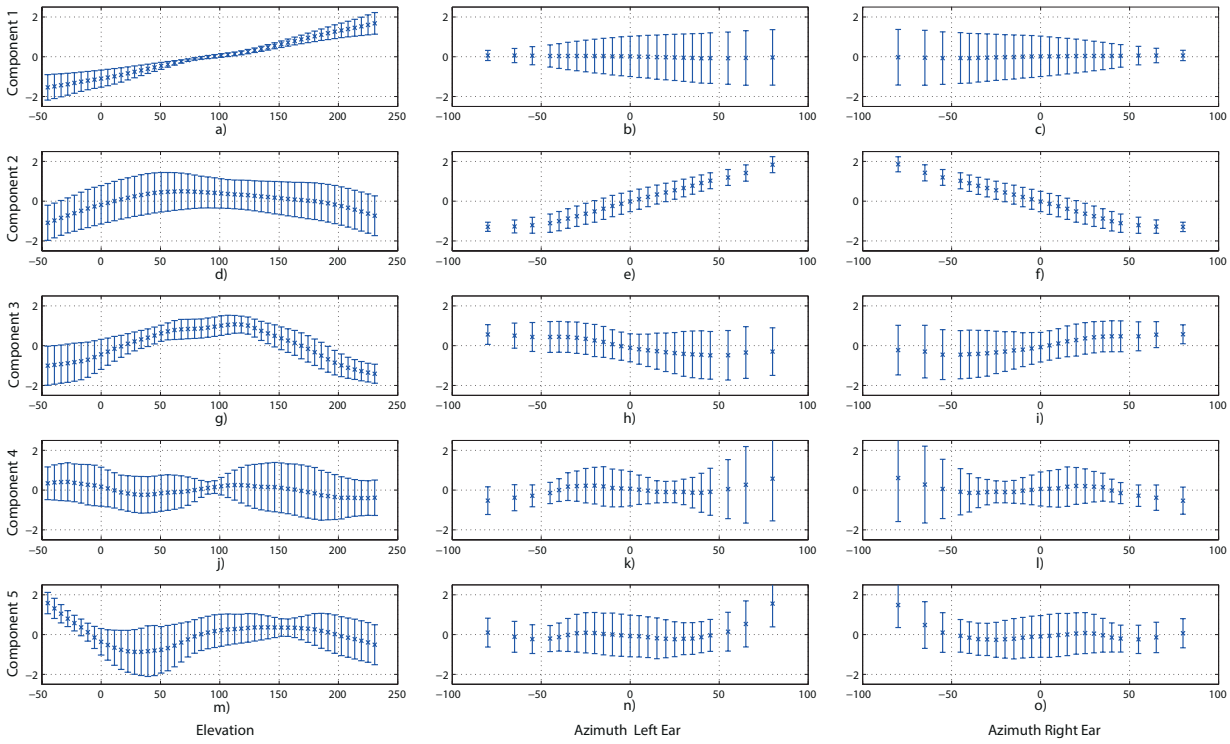


Figure 4.5: Isomap components as a function of location. In all plots, datapoints represent the mean across all subjects for a specific location and error bars correspond to a ± 1 standard deviation interval. For azimuth data (second and third column plots), we separated left and right ear plots to put in evidence ipsilateral and contralateral variability across individuals. All Isomap components are normalized to have zero mean and unit variance.

Figure 4.4b and 4.4c show the same two-dimensional manifold but this time the color represents azimuth angles. We plotted separately low-dimensional HRTFs of each ear to put in evidence the symmetry introduced in the graph construction procedure proposed in section 4.4.2. Observe that the second component of Isomap embedding roughly increases with azimuth for left ear. This trend, as expected because the left/right simmetry, is inverted (i.e. Isomap second component decreases with azimuth) for low-dimensional HRTFs of the right ear. Moreover, the correlation coefficient between azimuth angle and

Isomap second component is 0.879 for left ear and −0.880 for right ear.

Figures 4.5e and 4.5f present the second component as a function of source azimuth where it is evident the strong correlation between azimuth and Isomap second component. Observe also that the component's value tends to be roughly positive for contralateral locations and negative for ipsilateral directions. On the other hand, Figure 4.5d shows that the second component encodes also some elevation cues, although this relationship is not as strong as in the case of first component. Notice that the second component only tends to decrease when the source moves from frontal plane (i.e. $\phi = 90°$). However, it is not capable of distinguishing front locations from back ones since it exists front and back elevation angles that produce the same component's value. Furthermore, error bars are larger when compared to Figure 4.5a which suggests that second component varies widely across subjects at a specific elevation angle.

Figure 4.5g through 4.5i show the third Isomap component as a function of direction. In Figure 4.5g, the third component tends to decrease as the source move from frontal plane, which reveals that the third component is capable of encoding some elevation information. However, this component can't resolve front/back ambiguities because back elevation angles can produce the same component's value as front ones. Thus far, the third component behavior is similar to the second component. Nonetheless, Figures 4.5h and 4.5i show no clear pattern between azimuth and the third component as in the case of the second component.

So far, we have mainly analyzed directional relationships of Isomap components. With respect to inter-subject differences captured by Isomap, they are far more complex to visualize than directional ones due to its non-linear nature. Still, observe the variability of the black points in Figure 4.6. These are low-dimensional HRTFs of same direction but different subject in the two-dimensional manifold where most of variance is captured (i.e. first and second dimensions). Although they are relatively close to each other, as expected because of our graph construction procedure, their high variability is due to inter-subject differences.

Moreover, in general, error bars in Figure 4.5 for all Isomap components increase for contralateral locations and the same pattern is observed when the source moves from the frontal plane. This trend confirms that the head shadowing effect and vertical cues (mostly introduced by the pinna) causes wide variations in HRTFs.

In summary, we found that the Isomap first component is strongly correlated with elevation and the second one with azimuth. Although the second and third components also encode some elevation cues in a lesser degree, they are not able to distinguish front locations from back ones. The pattern of the remaining two dimensions shown in Figure 4.5 is considerably more complicated. Still, in general, for all components inter-subject variability increases for contralateral locations and when sources move away from frontal plane.

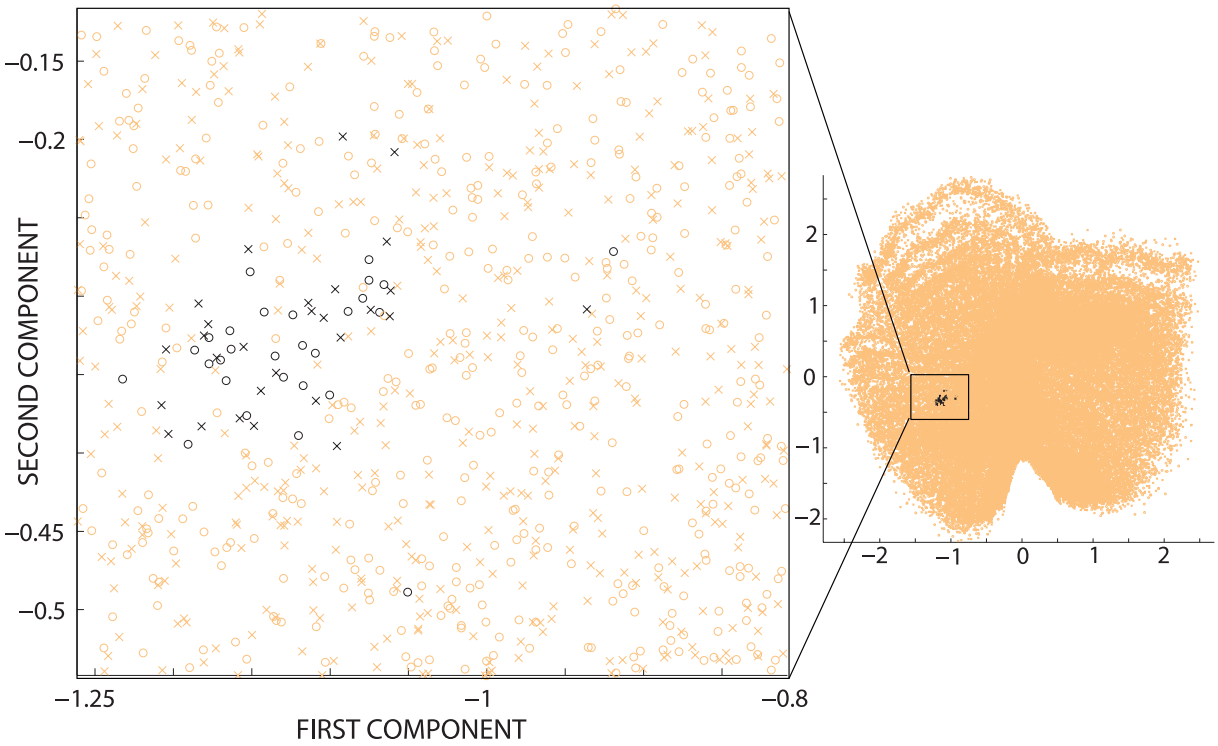


Figure 4.6: Two-dimensional manifold. Black datapoints are low-dimensional HRTFs of same direction but different subject. Left and right ear are represented by circle and cross markers, respectively.

Table 4.1: Paired t-test for different frequency bands. All bold entries refer to a statistically significant difference at a 95% confidence level. The lower the p-value the better. P-values are shown up to the third decimal place.

Paired t-test	0.2 - 1.0 kHz	1.0 - 2.0 kHz	2.0 - 4.0 kHz	4.0 - 8.0 kHz	8.0 - 15.0 kHz
Intra-ISO with respect to Full-ISO	**0.001**	**0.007**	**0.008**	**0.046**	**0.003**
Intra-ISO with respect to Intra-PCA	**0.001**	**0.000**	**0.004**	**0.002**	**0.000**
Full-ISO with respect to Full-PCA	**0.000**	**0.000**	**0.000**	**0.002**	**0.003**
Intra-PCA with respect to Full-PCA	**0.000**	**0.001**	**0.029**	0.175	0.365

Spectral Distortion Analysis

As stated in Section 4.5, we performed simulations for four conditions: Intra-ISO, Full-ISO, Intra-PCA and Full-PCA.

Figure 4.7 shows the mean spectral distortion (MSD) for four frequency bands. In the same figure, error bars represent 95% confidence intervals ($\pm 2\sigma$). Observe that, as expected, the MSD increases with frequency but Isomap performed better than PCA, specially in frequencies above 4 kHz that normally are harder to predict because of their high inter-subject variability. Still, in Intra-ISO and Full-ISO conditions, Isomap manages to keep MSD roughly below 4 dB and 6 dB for $4-8$ kHz and $8-15$ kHz bands, respectively. Moreover, note that the confidence interval shows that in the frequency ranges 2 to 4 kHz and 4 to 8 kHz the Isomap conditions have much less variability than their PCA counterparts.

Table 4.1 summarizes the results of a series of paired t-tests along different frequency

Table 4.2: Mean Spectral Distortion in dB for different frequency bands. Bold entries refer to the lower MSD in a specific frequency band.

Band [kHz]	Intra-ISO	Full-ISO	Intra-PCA	Full-PCA
0.2 - 1.0	**1.2572**	1.3231	1.6134	2.0795
1.0 - 2.0	**1.7882**	1.8873	2.2548	2.7443
2.0 - 4.0	**2.2142**	2.2972	2.8642	3.1020
4.0 - 8.0	**3.4673**	3.5218	4.6730	4.9446
8.0 - 15.0	**5.8007**	5.9760	7.6836	7.5635

Table 4.3: Paired t-test for Vertical Mean Spectral Distortion. All bold entries refer to a statistically significant difference at a 95% confidence level. The lower the p-value the better. P-values are shown up to the third decimal place.

Paired t-test	Left	Right
Intra-ISO with respect to Full-ISO	**0.000**	**0.000**
Intra-PCA with respect to Full-PCA	0.351	0.279

bands, where bold entries indicate a statistically significant difference at a 95% confidence level ($p < 0.05$). We found that for all frequency bands, both Isomap conditions data (i.e. Intra-ISO and Full-ISO) come from a population with a mean less than its corresponding PCA condition (i.e. Intra-PCA and Full-PCA), confirming that Isomap performs better than PCA. Moreover, for all frequency bands, Intra-ISO shows a small but statistically significant improvement over Full-ISO. On the other hand, for PCA, although Intra-PCA presents a more evident improvement over Full-PCA, this improvement is only statistically significant in low frequency bands. Still, observe that the error bars, particularly for high frequency bands, are in general smaller for both intraconic conditions.

Observe that for the intraconic conditions in Table 4.1, although the relatively high

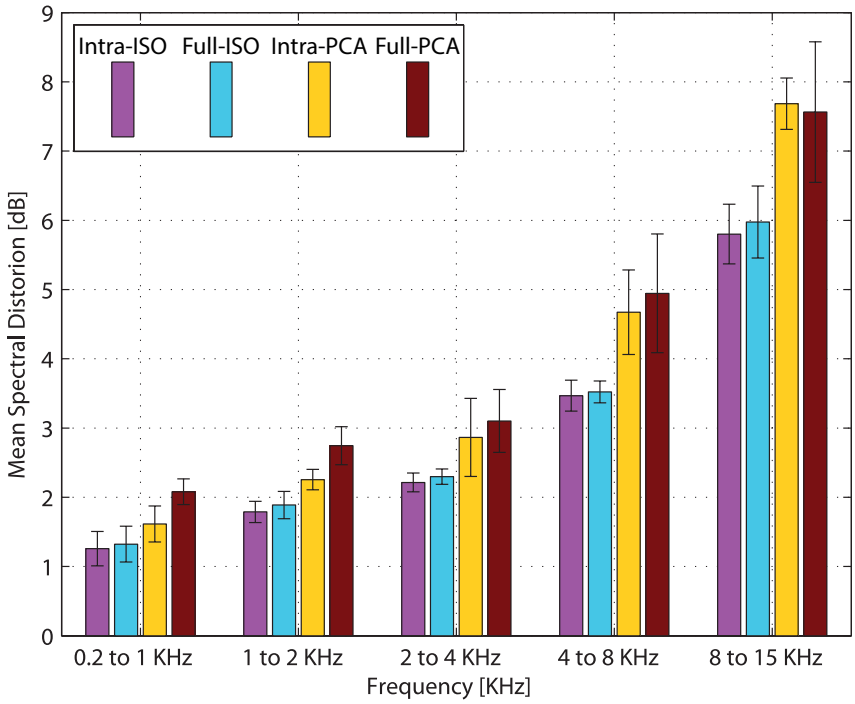


Figure 4.7: Mean Spectral Distortion for different frequency bands. Error bars represent 95% confidence intervals ($\pm 2\sigma$). MSD values are tabulated in Table 4.2.

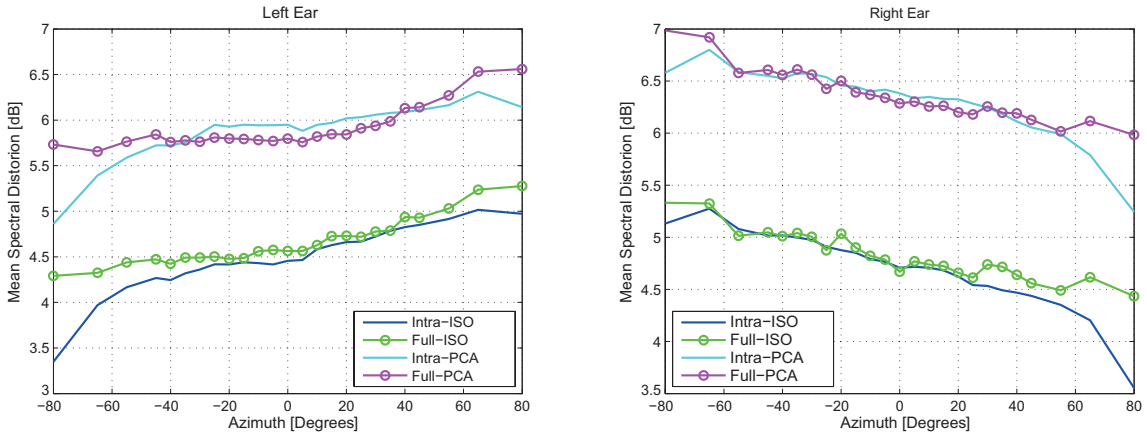


Figure 4.8: Vertical Mean Spectral Distortion.

Table 4.4: Paired t-test for Lateral Mean Spectral Distortion. All bold entries refer to a statistically significant difference at a 95% confidence level. The lower the p-value the better. P-values are shown up to the third decimal place.

Paired t-test	Left	Right
Intra-ISO with respect to Full-ISO	**0.000**	**0.000**
Intra-PCA with respect to Full-PCA	0.368	0.271

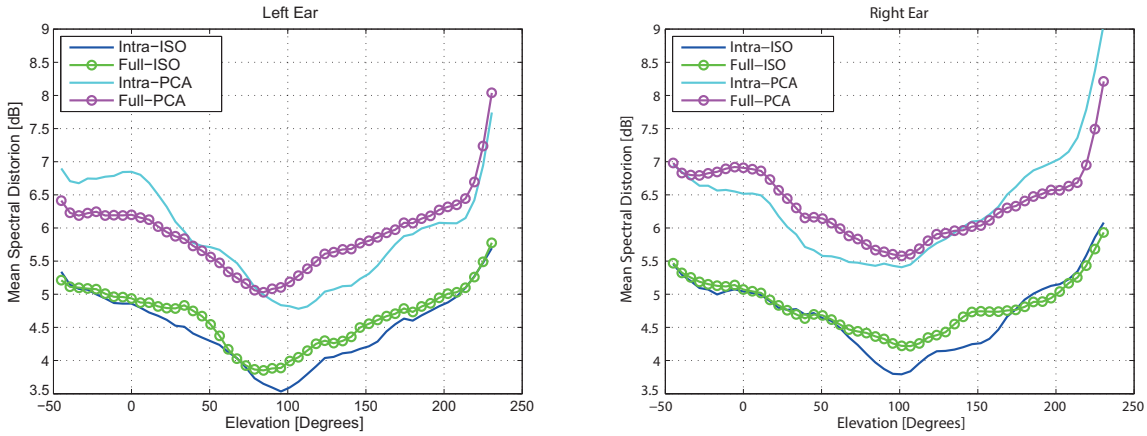


Figure 4.9: Lateral Mean Spectral Distortion.

p-value of 0.046 at the $4-8$ kHz band corresponds to a statistically significant difference at a 95% confidence level, they do not differ in a statistically significant way at a 99% confidence level. However, observe also that MSD sub-band analysis has a strong propensity to hide some important causes of distortion, e.g, MSD at ipsilateral and contralateral locations tend to cancel each other out due to the averaging across ears and directions. For the aforesaid reasons, it is convenient to analyze the MSD of each ear separately as a function of sound source position.

Figure 4.8 shows the MSD as a function of azimuth. Because we calculate this MSD across all elevations for a specific azimuth (i.e. the MSD in the cone of confusion), we refer it as vertical MSD. As before, both Isomap conditions present less vertical MSD than the

PCA ones. On the other hand, in a general way, Intra-ISO condition performs better than Full-ISO. This is especially notable for ipsilateral locations where Intra-ISO reaches up to 1 dB improvement over Full-ISO. For contralateral locations, due to the head shadowing effect, the vertical MSD tends to increase for both Full-ISO and Intra-ISO conditions.

Furthermore, in a paired t-test performed for each ear separately (refer to Table 4.3), the Intra-ISO condition showed a statistically significant improvement ($p < 0.05$) over the Full-ISO condition. Nonetheless, in a similar paired t-test between Intra-PCA and Full-PCA, no statistically significant improvement was found. This last result is not surprising taking into account that we found no statistically significant difference (refer to Table 4.1) between PCA conditions for high-frequency bands which in turn are the major contributors to elevation perception. Moreover, we expected some improvement of Intra-ISO over Full-ISO because the information lost after the spectral decomposition (i.e. the lateral and average components) is less perceptually relevant for HRTF personalization [161].

Figure 4.9 shows the MSD as a function of elevation. Because we calculate this MSD across all azimuths for a specific elevation, we refer it as lateral MSD. In this figure, it is clear that Isomap performs better than PCA in all conditions. Observe that, for all conditions, the lateral MSD decreases as the sound source moves toward the frontal plane (i.e. $\phi = 90°$), reaching a minimum around top directions. Note also that the Full-ISO and Intra-ISO lateral MSD stays roughly below 5.5 dB, except for very low elevations at back locations where the lateral MSD reaches up to 6 dB. This increase of lateral MSD confirms that complex scattering of sound waves coming from low elevations are harder to predict. Besides, the Intra-ISO lateral MSD shows a modest improvement over Full-ISO that is more prominent for back locations closer to the frontal plane.

Again, we performed paired t-tests for each ear separately between Intra-ISO and Full-ISO, and between Intra-PCA and Full-PCA conditions (refer to Table 4.4). We found that Intra-ISO data comes from a population with a MSD less than Full-ISO condition ($p < 0.05$). However, we did not found statistically significance difference between PCA conditions. Taking into account that high-frequency cues are needed for front/back discrimination, this last result is in accordance with the lack of statistically significance difference found in high-frequency bands for PCA conditions in Table 4.1.

Although prior works have performed experiments on different baseline datasets, anthropometric features, frequency bands and spatial locations, we would like to make a reasonable comparison with the approaches used in those works in terms of the spectral distortion reported by them. However, it should be kept in mind that most works do not report standard deviation values, which makes a fair comparison harder. Personalization methods based on linear dimensionality reduction techniques in conjunction with linear regressors [169, 170] report MSD scores across all frequencies near 6 dB, which is higher than our MSD (4.6 dB, $\sigma = 0.15$). On the other hand, MSD across all frequencies on customization techniques using linear dimensionality reduction together with nonlinear

regressors ranges from roughly 3 [172] to 5 dB [174]. Although our results are slightly better than [174], they are lower than [172], which is –according to our research– the smallest score reported among studies using MSD. Finally, in general, tensor-based approaches [176, 175, 177, 178] perform better than PCA-based methods, reaching their best performance at 3.5 dB in the frequency band 50 Hz – 8 kHz [178], which is comparable to our results in the frequency band 0.2 – 8 kHz (2.9 dB, $\sigma = 0.0735$).

So far, we have restricted our comparison to methods using MSD as metric. We considered relevant to analyze our data using the variance metric proposed by Middlebrooks [165], which produced very similar results to those using MSD in the sense that Intra-ISO presented the best performance, while both PCA conditions performed the worst. The error of Intra-ISO for frequency bands up to 4 kHz is less than 3.58 dB^2, grows in the 4 to 8 kHz band (8.9 dB^2, $\sigma = 0.7$) and reaches its maximum for frequencies above 8 kHz (24.96 dB^2, $\sigma = 1.13$). As a reference, using a frequency scaling approach, Middlebrooks [165] found that the 95 percentile of inter-subject spectral difference (measured by the variance metric) across 990 pairs of subjects was 9.3 dB^2. This confirms what we found using MSD with respect to the weaker performance of our method in frequency bands above 8 kHz.

Conclusion and Future Work

The findings of this paper show that Isomap has proven to be a powerful technique to discover the manifolds of spatial hearing. By incorporating important prior knowledge, Isomap was capable of explaining the directional factor (i.e. azimuth and elevation) of spatial audio. Even though no Isomap component alone explains inter-subject differences, the wide inter-dimension variability observed confirms its nonlinear behavior. Hence the importance of nonlinear regressors such as Artificial Neural Networks (ANN) to map anthropometric features into low-dimensional HRTFs. Unlike regression techniques such as Support Vector Regression, ANN is a multiple output predictor that permits to exploit the correlations between Isomap components (i.e. inter-dimension correlations). Moreover, instead of constructing one regression model per direction, our approach lets to construct a single model that does not break the inherent multifactor nature of HRTFs (i.e. frequency, direction and subject factors).

In all simulations performed, the results show that Isomap has a better performance and less variability than PCA as measured by the mean spectral distortion (MSD) with 95% confidence intervals. Furthermore, our results put in evidence that Isomap can capture high-frequency cues from intraconic HRTFs where PCA does not. Thus, we confirmed that the intraconic representation effectively encodes the most important cues for individualization of HRTFs.

On the other hand, the main weakness of Isomap is the lack of an explicit mapping

function [188] to project new high-dimensional datapoints into an existing low-dimensional embedding (i.e. out-of-sample extension), and to reconstruct a low-dimensional datapoint into a high-dimensional representation(i.e. back-projection). Out-of-sample extension might be performed by means of the Nystrom approximation [196, 197], so that, for new datapoints, there is no need to recalculate the entire manifold.

The back-projection is a more challenging problem to overcome. Here, we have reconstructed high-dimensional HRTFs using a linear combination of its neighbors (i.e neighborhood-based reconstruction). It should be observed that the main weakness of this reconstruction is that its accuracy depends on how dense the initial database is. This problem might be addressed using some spatial HRTF interpolation before HRTF personalization to guarantee a more populated manifold. However, note that if the initial database is not sampled adequately in space, the resultant interpolated HRTFs will not be suitable to reconstruct the personalized HRTFs. In this sense, although the CIPIC dataset is one the most complete publicly available HRTF datasets including anatomical measurements, we expect that our method could perform better given a more suitable input dataset (i.e. higher spatial resolution, more subjects and better quality anatomical measurements). Lastly, although we chose the reconstruction weights to be the corresponding neighbor's euclidean inverse distances, there is an alternative approach where the weights are determined in a least-squares optimization. However, this approach proved to produce larger spectral distortion.

One question that arises for practical use is whether our method will produce an acceptable perceptual result. Since low-frequency distortion is low, we expect that cues acting on those bands will not be affected in listening tests. On the other hand, the error in high-frequency bands is relatively high, which will affect elevation perception. However, it should be kept in mind that we demonstrated that an important part of this distortion is due to contralateral and low elevation errors. Moreover, previous studies concluded that the spectral detail of HRTFs at high frequency is inaudible [198], which in turn, implies that the high contralateral error is likely to be, to some extent, perceptually irrelevant. Thus, in listening tests, we expect the localization accuracy to be good at lateral locations, reasonable at vertical directions but poor in low elevations.

Another problem that might arise is how to apply Isomap when the graph has two or more connected components. In such case, the resultant components would lie on different manifolds. Further studies might address this problem using techniques to merge multiple manifolds as proposed, e.g., in [199]. It should be noted that, since our approach guarantees a connected graph, we do not address the non-connected graph case.

In future work, we plan to explore a subband representation for HRTFs in conjunction with manifold learning. Since different localization cues act in different frequency bands, a subband representation would permit a more flexible way to construct the manifold structure. In this context, it would be possible to incorporate prior knowledge in the

subband where this prior is effectively valid. For example, we could introduce left/right symmetry only in subbands where symmetry is more prominent.

We also plan to explore a multi-task learning (MTL) approach to learn the regression model. MTL learns multiple related tasks simultaneously using a shared representation aimed at improving generalization [200]. In the HRTF personalization context, a task could be, e.g., learning a regression model per direction. The MTL approach might preserve the multi-factor nature of HRTFs by using a shared representation instead of learning a single regression model as we do in this paper.

Chapter 5

Contribution III

Contribution III is published as:

F. Grijalva, L. C. Martini, D. Florencio and S. Goldenstein, "Interpolation of Head-Related Transfer Functions Using Manifold Learning," in *IEEE Signal Processing Letters*, vol. 24, no. 2, pp. 221-225, Feb. 2017. doi: 10.1109/LSP.2017.2648794

Interpolation of Head-Related Transfer Functions using Manifold Learning

Felipe Grijalva[1], *Student Member, IEEE,* Luiz Martini[1],
Dinei Florencio[2], *Fellow Member, IEEE,*
and Siome Goldenstein[3], *Senior Member, IEEE*

Abstract

We propose a new Head-Related Transfer Function (HRTF) interpolation method using Isomap, a nonlinear dimensionality reduction technique. First, we construct a single manifold for all subjects across both azimuth and elevation angles through the construction of an Intersubject Graph (ISG) that includes important prior knowledge of the HRTFs such as correlations across individuals, directions and ears. Then, for a new direction, we predict its corresponding low-dimensional HRTF by interpolating over same subject low-dimensional measured HRTFs. Finally, we use a Local Neighborhood Mapping (LNM) in the manifold to reconstruct the high-dimensional HRTF from measured HRTFs of all subjects. We show that a single manifold representation obtained through the ISG is a powerful way to allow measured HRTFs from different subjects to contribute for reconstructing the HRTFs for new directions. Moreover, our results suggest that a small number of spatial measurements capture most of acoustical properties of HRTFs. Lastly, our approach outperforms other linear and nonlinear dimensionality reduction techniques such as Principal Component Analysis (PCA), Locally Linear Embedding (LLE), and Laplacian Eigenmaps (LEM).

Introduction

In spatial audio, *Head-Related Transfer Functions* (HRTFs) measurement is a complex, time-consuming and expensive procedure. In this context, HRTF interpolation techniques can be applied to achieve high spatial resolution with as few measurements as possible. A small set of measurements might considerably reduce the overall measurement time as well as the equipment costs. Moreover, a high spatial resolution effectively minimize undesired audio artifacts (e.g. clicks), which is very critical for the increasing spatial audio applications (e.g. virtual reality [201], assistive technologies [202]).

[1]School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil
[2]Microsoft Research, Redmond, WA, USA
[3]Institute of Computing, University of Campinas, Campinas, SP, Brazil.

The most straightforward interpolation methods estimate an HRTF at an unknown position as a weighted average of nearby measured HRTFs. Different schemes to choose the weights have been used such as linear interpolation [203], bilinear interpolation [204], cubic splines [205], among other variants [206, 207, 208, 209, 210]. Other methods such as parametric approaches [211, 212, 213] interpolate low-order representations of the HRTFs.

On the other hand, since HRTFs varies rapidly at high frequencies [16], a better performance has been attained by interpolating HRTFs from a weighted sum of basis functions, which can be performed either in the spatial or frequency domain. Among spatial domain approaches, there are methods based on azimuthal harmonics [214, 215], spherical harmonics [216, 217, 218, 219, 220, 221, 222], and interpolation from a small set of measurements [223, 224, 225]. With respect to frequency domain methods, most works use PCA [226, 227, 228, 229, 230] to perform interpolation in a low-dimensional linear space. To exploit the inherent nonlinearity nature of HRTFs, Duraiswami et al. proposed to use the nonlinear dimensionality reduction technique *Locally Linear Embedding* (LLE) [231] to construct a manifold for a single subject to interpolate HRTFs only on the median plane. However, they found that their algorithm is not very stable when using all directions from one subject.

In this letter, we propose an HRTF interpolation method based on the manifold learning technique Isomap [232]. Our approach captures important prior knowledge of HRTFs across subjects, directions, and ears to obtain the Isomap's manifold through a custom graph construction, which will be referred to as the *Intersubject Graph* (ISG). Similar to the common PCA representation in [225], the ISG considers HRTFs from all subjects to construct a more expressive low-dimensional embedding. Moreover, unlike [231], we construct a single manifold for all subjects, ears, and locations to perform interpolation in a low-dimensional nonlinear space. A single manifold obtained through the ISG allows measured HRTFs from different subjects to contribute for reconstructing the HRTFs for new directions. Finally, similar to [223, 224, 225], our technique only requires a small set of measurements to obtain higher resolution HRTFs.

We use the interaural coordinate system as in [132]. When we refer to HRTF we are referring to its magnitude, i.e., we use the HRTF minimum phase approximation and we do not deal with *Interaural Time Difference* (ITD) interpolation.

HRTF Interpolation using Isomap

First, we construct a single manifold from measured HRTFs of all subjects using Isomap with a custom graph-construction procedure (i.e. the ISG). Then, for a new direction (i.e. outside of the directions used for constructing the manifold), we predict its low-dimensional HRTF by interpolating over same subject low-dimensional HRTFs. Finally, we use a *Local Neighborhood Mapping* (LNM) in the manifold to obtain the high-

dimensional HRTFs from their low-dimensional representation, i.e., we reconstruct the HRTF for a new direction not only from same subject HRTFs but from measured HRTFs of all subjects.

Manifold Construction using Isomap

Isomap is a nonlinear dimensionality reduction technique [232] that reduces a high-dimensional dataset $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^D$ represented by a $D \times N$ matrix of N sample vectors $\mathbf{x}_i$ into a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \subset \mathbb{R}^d$ represented by a $d \times N$ matrix of N sample vectors $\mathbf{y}_i$, where $d < D$. Isomap takes into account the datapoint neighborhood relationships by preserving the pairwise geodesic distances (i.e. the distance over the manifold) to maintain the intrinsic geometry of the data [233].

Isomap has three steps: first, we construct a graph $G(V, E)$ from $\mathbf{X}$, where each sample $\mathbf{x}_i \in \mathbf{X}$ represents a node $v_i \in V$. Two nodes v_i and v_j are connected by an edge $(v_i, v_j) \in E$ with length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ is one of the K neighbors of $\mathbf{x}_j$. The edge length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ is given by some distance metric between $\mathbf{x}_i$ and $\mathbf{x}_j$. Here we use the Euclidean distance. Second, we estimate the geodesic distances on the manifold between each pair of points in $\mathbf{X}$ by computing the shortest path between each corresponding pair of nodes in G. We store these distances in the pairwise geodesic distance matrix $\mathbf{D}_G$. Finally, we construct the d-dimensional embedding by applying multidimensional scaling [234] (MDS) on $\mathbf{D}_G$ to find the d-dimensional coordinate vectors $\mathbf{y}_i$.

Isomap's first step selects a number of neighbors for each data point. Neighborhood selection presents an opportunity to incorporate a priori knowledge from data [235]. Thus, instead of using common neighborhood selection approaches (e.g. K nearest neighbors) and inspired by our previous work on HRTF personalization [131], we construct the graph G (i.e. the ISG) by exploiting the correlations among the HRTFs across directions, ears, and subjects, according to the following criteria:

Criterion 1. *if* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *represent HRTFs of the same location and ear but different subject, then connect them. Instead of applying Isomap separately for each subject as in [231], with this criterion, we tried to exploit the correlation of HRTFs among subjects*

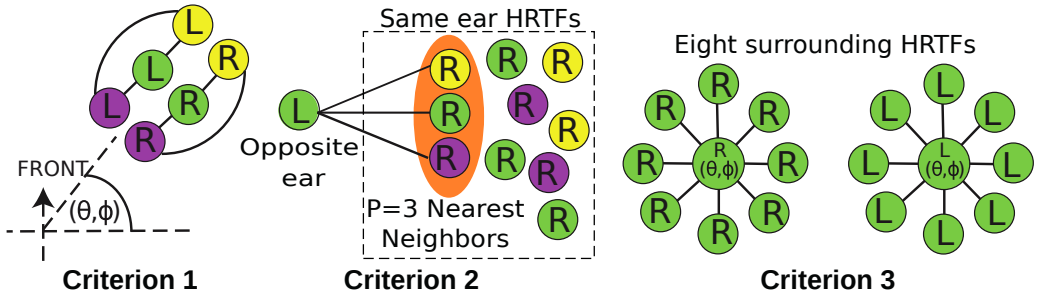


Figure 5.1: Illustrative example of the ISG for $P = 3$ subjects. Color represents same subject HRTFs, and (θ, ϕ) represents the azimuth and elevation. L=Left ear and R=Right ear.

across same directions. Using this criterion, $P-1$ neighbors were obtained, where P is the number of subjects.

Criterion 2. *Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be HRTFs of different ears, and $\mathbf{X}_{ear}$ be a subset of $\mathbf{X}$ containing all HRTFs of the same ear as $\mathbf{x}_j$ (i.e. the opposite ear of $\mathbf{x}_i$). If $\mathbf{x}_j$ is one of the P nearest neighbors of $\mathbf{x}_i$ in $\mathbf{X}_{ear}$, then connect them. The P neighbors obtained from this criterion attempt to incorporate the correlation existing due to left-right symmetry of HRTFs at frequencies below 5.5 kHz [41].*

Criterion 3. *Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be HRTFs of the same subject and ear. If $\mathbf{x}_j$ is one of the eight HRTFs surrounding $\mathbf{x}_i$, then connect them. The eight neighbors obtained from this criterion emphasize the similarities between spatially close HRTFs of the same subject and ear.*

It is straightforward to prove that the ISG is always connected (see Fig. 5.1). Isomap takes as parameters the number of neighbors, K, and the intrinsic dimensionality, d. Due to our ISG, the number of neighbors is fixed to $K = 2P + 7$, i.e., $P - 1$ from Criterion 1, P from Criterion 2 and eight from Criterion 3. To estimate the intrinsic dimensionality we use the maximum likelihood intrinsic dimensionality estimator [236], which has been also successfully used in other manifold learning problems [233, 193]. This estimator tries to reveal the intrinsic geometric structure of the observed data.

Interpolation and Local Neighborhood Mapping

For an unknown direction, it is possible to estimate its low-dimensional HRTF using some interpolation method. Here, for each low-dimensional embedding's component separately, we construct a linear interpolant function $y = F(\theta, \phi, e)$ per subject from the low-dimensional component's values y at known positions (θ, ϕ) and ears e. For an unknown position and ear (θ_q, ϕ_q, e_q), the linear interpolant F produce an interpolated value y_q. It is worth noting that we linearly interpolate only from same subject low-dimensional HRTFs. Since we want to recover the high dimensional HRTFs, we need a way to reconstruct it from its low-dimensional version. Different from PCA, Isomap and other manifold learning techniques (e.g. LLE) do not offer an explicit mapping between low and high dimensional datapoints. However, we can take advantage of the Isomap's locally linear assumptions to calculate the reconstructed high-dimensional HRTF, $\hat{H}$, as a linear combination of its neighbors (i.e. a Local Neighborhood Mapping, LNM) in the low-dimensional embedding using

$$\hat{H} = \sum_{i=1}^{K} w_i \mathbf{x}_i, \text{ where } \hat{H} \subset \mathbb{R}^D. \tag{5.1}$$

The weights w_i should capture the intrinsic geometry of the corresponding neighborhood in the low-dimensional space. Thus, inspired by LLE, we calculate the weights by solving a constrained least squares problem, where the constrained weights w_i that best reconstruct each low-dimensional datapoint $\mathbf{y}_i$ from its K neighbors in the low-dimensional space must sum up to one [237]. Observe that, according to Eq. 5.1, the same weights reconstruct the high-dimensional HRTFs.

Moreover, note that the LNM is performed from HRTFs of all subjects (i.e. in the manifold)[4]. Finally, although the LNM reconstructs an HRTF as a linear combination of its neighbors, we compute the weights in a new and nontrivial way.

Simulations

HRTF Database We used the publicly available CIPIC database [132] which contains *Head-Related Impulse Responses* (HRIRs) for 45 subjects at 25 azimuths and 50 elevations (i.e. 1250 locations) in the interaural coordinate system. HRIRs in the CIPIC database were sampled at 44.1 kHz for 16 bit resolution, and have 200 samples long.

HRIR pre-processing First, we transformed the HRIRs into HRTFs through a 512-point FFT. Then, we preserved the frequencies from 200 Hz to 14 kHz for reducing the effects of the limitations in the frequency response of the equipments used during HRIR measurement. The resultant HRTFs have 160 frequency bins (i.e. $D = 160$). Finally, because the spectral detail of HRTFs at high frequency is inaudible [198], we smoothed[5] the HRTFs using a bandwidth of 1 ERB (equivalent rectangular bandwidth) and 2 ERB for low ($f < 5$ kHz) and high ($f \geq 5$ kHz) frequencies respectively.

Training and test sets Since directional resolution in CIPIC database is approximately the same (i.e. 5° along most azimuths and 5.625° along all elevations), we constructed the training set by uniformly sampling every 20° in azimuth (i.e. 9 azimuths) and every 22.5° in elevation (i.e. 14 elevations) for all $P = 45$ subjects, giving a total training set size of $N = 2 \cdot P \cdot M = 11340$, where $M = 9 \cdot 14 = 126$ locations per subject and ear (i.e. 10.08% of the 1250 available locations). According to this sampling strategy, we should have included the azimuths $-60°$ and 60° in the training set. However, since the CIPIC database does not include these measurements, we included the azimuths $-65°$ and 65° instead. The remaining directions (i.e. 1124 locations per subject and ear) were used for constructing the testing set.

[4] It is worth emphasizing that even though we interpolate only from same-subject low-dimensional HRTFs, this operation is just to calculate the interpolated low-dimensional HRTF .The reconstructed high-dimensional HRTF is obtained as a linear combination from all subjects in the manifold.

[5] We filtered the HRTFs using a filter bank as explained in [198].

Table 5.1: VM for several manifold learning techniques. Asterisks show a statistically significant difference (p-value<0.01)

Manifold Learning Technique	VM
ISOMAP-ISG*	$2.986, \sigma = 0.65$
ISOMAP	$3.595, \sigma = 0.73$
LEM-ISG	$4.537, \sigma = 0.94$
LEM	$9.375, \sigma = 1.53$
LLE-ISG	$13.396, \sigma = 1.57$
LLE	$4.553, \sigma = 0.81$

Dimensionality Reduction Besides PCA with 95% of variance retained, we compared Isomap against two similar manifold learning techniques that also assume local linearity according to a graph construction: LLE [237] and *Laplacian Eigenmaps* [238](LEM). We estimated the intrinsic dimensionality for these manifold learning approaches using the maximum likelihood estimator [236]. So, each HRTF was embedded in a $d = 5$ dimensional space. Since the number of subjects $P = 45$, according to our ISG each HRTF is connected to $K = 2P + 7$ neighbors. To analyze how our ISG affects the interpolation error, we implemented these manifold learning techniques with the ISG (labeled as ISOMAP-ISG, LEM-ISG, and LLE-ISG) and without it (labeled as ISOMAP, LEM, and LLE). For the conditions without the ISG, we chose the number of neighbors so that at least 99.9% of data points were embedded (i.e. 30 neighbors).

Performance Metric We chose a distortion metric in dB^2 based on variance introduced by [239] that will be referred to as the *variance metric* (VM). The VM is defined for each pair of measured and reconstructed HRTF (H and $\hat{H}$ respectively) as the variance of the difference spectrum

$$\text{VM} = \text{var}\left(20\log_{10}|H| - 20\log_{10}\left|\hat{H}\right|\right) \tag{5.2}$$

For our analysis, we average the results from Eq. 5.2 across bands, directions or subjects, as needed. As noted by [239], this metric is a better choice than the spectral distortion because the latter will be non-zero if there are differences due to the overall gain (i.e. when there is a constant intensity offset) between the measured and reconstructed HRTF despite their similar shapes.

Results and Discussion

Table 5.1 shows the value of VM for several manifold learning techniques with and without ISG. ISOMAP-ISG outperforms the others with statistical significance (p-value<0.01). Observe that our ISG improves the performance of ISOMAP and LEM but not of LLE. The reminder of our analysis will focus on the best results from Table 5.1 for each manifold learning technique (i.e. ISOMAP-ISG, LEM-ISG, and LLE) and PCA.

Fig. 5.2 shows the VM averaged across subjects. Observe that ISOMAP-ISG outperforms the other methods with statistical significance (p-value<0.01). For most subjects,

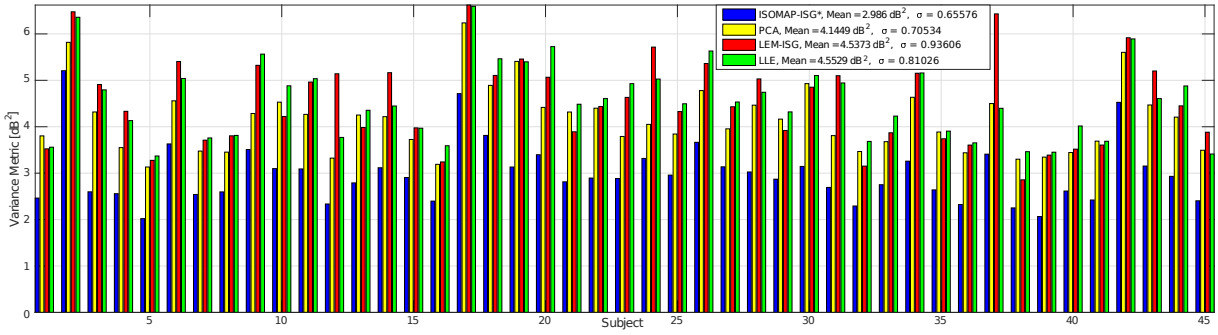


Figure 5.2: Variance Metric averaged across subjects.

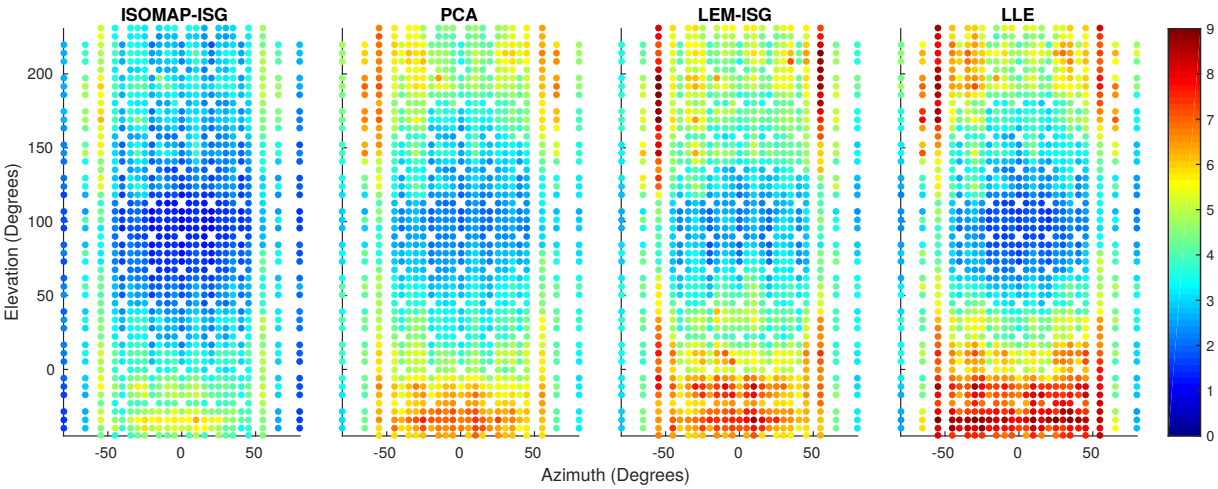


Figure 5.3: Variance metric as a function of direction.

ISOMAP-ISG keeps the VM below 3 dB^2.

The analysis of the VM as a function of direction or frequency band often yields more insights into the distortion behavior. In Fig. 5.3, it is evident the superiority of ISOMAP-ISG with respect to the other techniques, especially at low elevation and extreme lateral locations. However, even ISOMAP-ISG struggles to keep the VM below 5 dB^2 at frontal low-elevations. Fig. 5.4a shows the VM for ipsilateral, contralateral and median plane positions. Error bars represent 95% confidence intervals and asterisks indicate a statistically significant difference. ISOMAP-ISG outperforms the other techniques with statistical significance (p-value<0.01). As expected, the contralateral VM increases for all methods but ISOMAP-ISG keeps the error below 4 dB^2. Fig. 5.4b shows the VM for different bands. ISOMAP-ISG outperforms the others at high frequencies with statistical significance, keeping distortion below roughly 3 dB^2. However, for low frequencies LEM-ISG produces less distortion than the other techniques.

Note that in all the simulations aforementioned, LLE and LLE-ISG[6] performed the worst. Although in a previous work [231] the LLE interpolation performance was not as weak as here, they only considered a manifold for a single user on the median plane

[6]Errata: It is LEM-ISG, not LLE-ISG

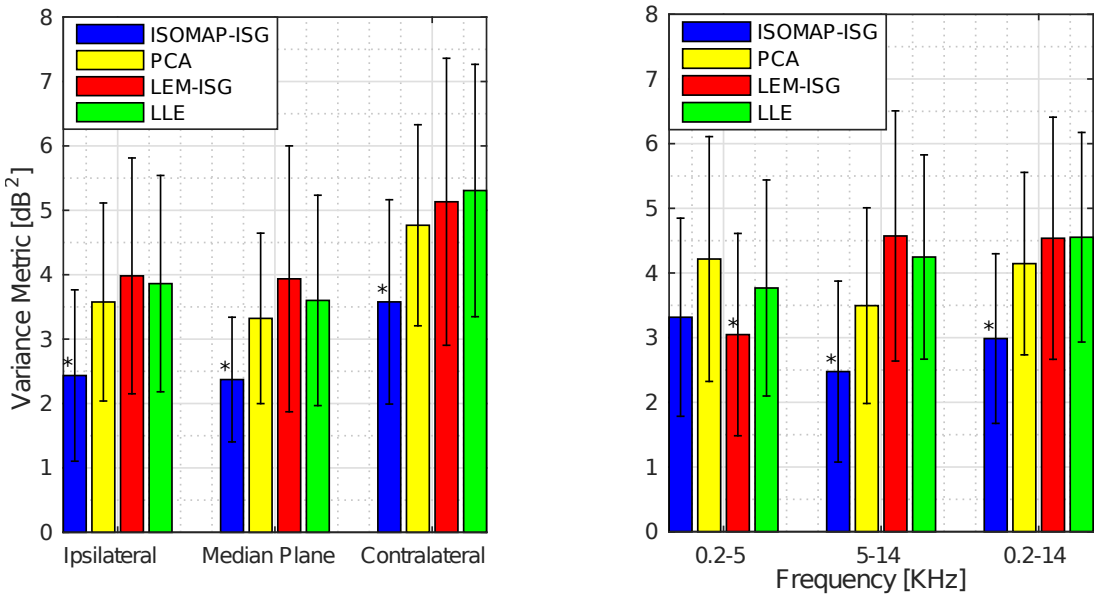


Figure 5.4: a) VM at ipsilateral, median plane, and contralateral locations. b) VM by frequency band. Asterisks indicate a statistically significant difference (p-value<0.01).

and all the elevations except one were used to construct the training set. In contrast, we perform a more challenging experiment constructing a single manifold for all users and directions with a training set of 10.08% of all available HRTFs. As noted in [233], a possible explanation for this loss in performance in some applications (e.g. biomedical datasets [240], derivation of perceptual-motor actions [241]) is that LLE struggles to find a suitable embedding in manifolds containing gaps and in the tendency of LLE to collapse a lot of data points very near in the low-dimensional space.

Conclusion

We introduced ISOMAP-ISG, an HRTF interpolation technique based on Isomap that constructs a single manifold for all subjects, directions and ears through the construction of the Intersubject Graph (ISG), which explores relevant prior knowledge of the HRTFs. We show that ISOMAP-ISG outperforms PCA and other manifold learning approaches.

Although perceptual experiments are necessary, our findings suggest that a small number of spatial measurements capture most of acoustical properties of HRTF as also noted by [242, 225, 226, 218]. Moreover, a single manifold representation obtained through the ISG has proven to be a powerful way to allow measured HRTFs from different subjects to contribute for reconstructing the HRTFs for new directions.

One question that might arise is how to include in the manifold a new subject outside of the training set without having to recalculate the entire manifold which is a computationally intensive task. Without an explicit mapping, HRTFs for a new subject in Isomap can be mapped using the Nystrom approximation [243]. This procedure might be used

for personalizing a subject's HRTF from a small set of measurements as in [225].

Observe that our approach might be not adequate for real-time interpolation since it requires information from all subjects in the database, which in turn, may be computationally prohibitive on platforms with limited hardware resources. For real-time tasks, interpolation based on continuous models [244] might be a more suitable choice.

Besides the constrained least squares approach for the LNM, we performed simulations with a inverse distance approach as used by [245, 131, 246] but it produced larger distortion. We also performed simulations increasing the training set, where in all cases ISOMAP-ISG outperforms the other methods, and the distortion reduces as expected. For instance, if we construct a training set by uniformly sampling every 11.25° in elevation, the VM for ISOMAP-ISG is 1.8 dB^2 with $\sigma = 0.3$.

Future research might focus in other forms of choosing the LNM weights and constructing the ISG. For instance, although we use eight neighbors for Criterion 3 to keep the graph as sparse as possible, a future work might choose other values than eight, or even to use an adaptive approach according to the spatial resolution of the HRTF dataset. Future research might also compare the manifold learning methods against other families of methods (e.g. parametric approaches, weighted sum approaches). Finally, it would be interesting to apply Isomap for HRTF extrapolation and near-field HRTF interpolation.

Chapter 6

Contribution IV

Contribution IV is published as:

F. Grijalva, L. C. Martini, B. Masiero and S. Goldenstein, "A Recommender System for Improving Median Plane Sound Localization Performance Based on a Nonlinear Representation of HRTFs," in IEEE Access, vol. 6, pp. 24829-24836, 2018. doi: 10.1109/ACCESS.2018.2832645

A Recommender System For Improving Median Plane Sound Localization Performance Based on a Nonlinear Representation of HRTFs

Felipe Grijalva[1], *Student Member, IEEE,* Luiz Martini[1],
Bruno Masiero[1], *Member, IEEE,*
and Siome Goldenstein[2], *Senior Member, IEEE*

Abstract

We propose a new method to improve median plane sound localization performance using a nonlinear representation of head-related transfer functions (HRTFs) and a recommender system. First, we reduce the dimensionality of an HRTF dataset with multiple subjects using manifold learning in conjunction with a customized intersubject graph (ISG) which takes into account relevant prior knowledge of HRTFs. Then, we use a sound localization model to estimate a subject's localization performance in terms of polar error (PE) and quadrant error rate (QE). These metrics are merged to form a single rating per HRTF pair that we feed into a recommender system. Finally, the recommender system takes the low-dimensional HRTF representation as well as the ratings obtained from the localization model to predict the best HRTF set, possibly constructed by mixing HRTFs from different individuals, that minimizes a subject's localization error. The simulation results show that our method is capable of choosing a set of HRTFs that improves the median plane localization performance with respect to the mean localization performance using non-individualized HRTFs. Moreover, the localization performance achieved by our HRTF recommender system shows no significant difference to the localization performance observed with the best matching non-individualized HRTFs but with the advantage of not having to perform listening tests with all individuals' HRTFs from the database.

Introduction

As augmented reality applications become more relevant, there is an increasing effort in 3D audio research and specifically in head-related transfer functions (HRTFs) to obtain high quality spatial audio. HRTFs are the main component of binaurally rendered 3D audio and are used to simulate sound sources as if they were coming from arbitrary positions in space [16]. HRTFs are complex-valued frequency functions that model the

[1]School of Electrical and Computer Engineering, University of Campinas, Campinas, SP, Brazil
[2]Institute of Computing, University of Campinas, Campinas, SP, Brazil.

relationships between human anatomy and a sound source before reaching the ears. These functions ideally should be measured for each subject individually to avoid poor localization performance due to mismatch of spatial cues contained in the HRTFs [247].

However, HRTF measurement [32, 248] is a complex procedure that requires an expensive apparatus (e.g. a (semi-)anechoic chamber, in-ear microphones, and a loudspeaker array). Moreover, it is usually time-consuming for high-spatial resolutions and tiring for the participants.

In order to avoid such measurements, several alternatives have been proposed, including theoretical [249], numerical [250], and inference methods [131, 246]. In contrast to the above mentioned physically-based techniques, in perceptual-based techniques the subjects have an active role during the personalization process by tuning some parameters (e.g. PCA weights [251]) for several target directions until they achieve an acceptable spatial accuracy. However, this procedure might also be time-consuming, depending on the ability of the human listener and the number of parameters and target directions. An alternative approach is to optimize these parameters through the use of a machine learning algorithm where the listener is required to localize a sound source with [252] or without [253] knowledge of the target directions. There are also database matching techniques [254] where the listener selects the best HRTFs among a set of HRTFs from other subjects. Although there is no need to tune any parameter, these methods still require the listener to perform listening tests. In order to speed up these techniques, it is desirable to find a way to reduce the number of listening trials while still minimizing the localization error.

In the light of facts exposed above, we propose the use of a recommender system to find the best HRTF set, with HRTFs for each direction selected from a larger HRTF dataset constituted by HRTFs from multiple subjects, in order to improve the listener's localization performance in the median plane. Our system recommends the best mixed HRTF set by estimating a subject's localization performance through a human sound-source localization model [255]. Moreover, with a small number of listening tests, our HRTF recommender system achieves a performance statistically similar to the best performance with non-individualized HRTFs but without having to perform listening trials for every individual HRTF in the database.

Inspired by our previous works [246, 131, 256], the recommender's input feature vectors are low-dimensional HRTFs that we obtain using manifold learning in conjunction with a customized intersubject graph (ISG) aiming to capture relevant prior knowledge of HRTFs. The outputs of our recommender system are the ratings that we obtain through the sound-source localization model proposed by Baumgartner et al. [255] (henceforth called the Baumgartner model).

Note that our approach is not an individualization technique such as [252, 253]. Moreover, different from [252], in our approach the listener is not aware of the target direction

as in [253]. We also use a human sound-source localization model [255] which is more suitable than the regression model used by [253] since it takes into account psychoacoustic factors.

The remaining of the manuscript is organized as follows. We describe our recommender system in Section 6.2. We present the conditions of our simulations in Section 6.3 and we analyze the results in Section 6.4. Finally, we conclude in Section 6.5.

Recommender System

Recommender systems are widely used to predict the preference that a user would give to an item (e.g. books, movies) [257]. Here, we are specifically interested in content-based recommender systems (see Figure 6.1) where a feature vector is available for each item (HRTF) and each rating made by the users (localization accuracy). In Section 6.2.1, we describe how we obtain such feature vectors using a nonlinear representation of HRTFs. Next, in Section 6.2.2, we show how the ratings were calculated through a sound-source localization model. Finally, in Section 6.2.3, we describe mathematically the problem of content-based recommender systems and how spatial audio fits into it.

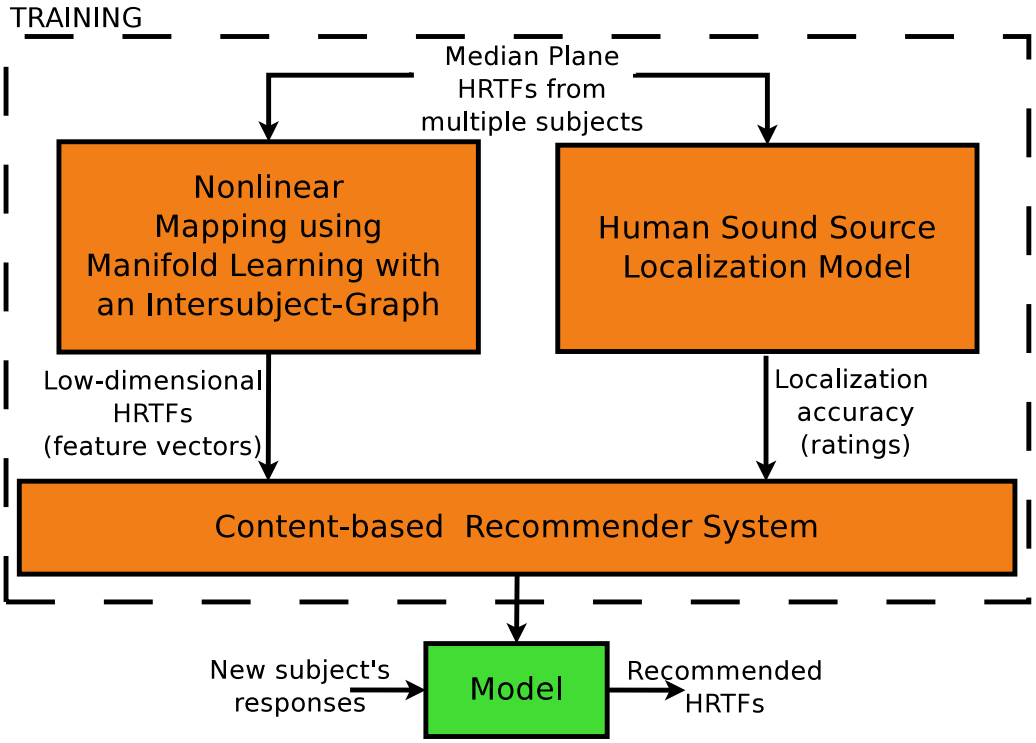


Figure 6.1: We construct a model by training a recommender system using low-dimensional median plane HRTFs as feature vectors and localization accuracy as ratings, obtained through a nonlinear mapping and a human sound localization model, respectively. For a new subject's responses, our system recommends the best mixed HRTF set constituted by HRTFs from multiple subjects.

Nonlinear representation of HRTFs

Nonlinear dimensionality reduction techniques (i.e. manifold learning) reduce a high-dimensional dataset $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathbb{R}^D$ represented by a $D \times N$ matrix of N sample vectors $\mathbf{x}_i$ into a low-dimensional embedding $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \subset \mathbb{R}^d$ represented by a $d \times N$ matrix of N sample vectors $\mathbf{y}_i$, where $d < D$. Here, a datapoint $\mathbf{x}_i$ is the vector resulting from the concatenation of the left and right Directional Transfer Function (DTF) magnitudes, and $\mathbf{y}_i$ are the feature vectors used in the recommender system. A DTF is the component of an HRTF that is specific to sound source localization. It is obtained by dividing an HRTF by its direction-independent common component (i.e. the component including spectral features such as the ear canal resonance and microphone response [225]), which in turn is calculated by averaging all HRTFs from a specific individual [258].

A well-known manifold learning technique is Isomap [232], which attempts to preserve the pairwise geodesic distance (i.e. the distance over the manifold) in order to maintain the intrinsic geometry of the data unlike PCA that retains most variance and attempts to preserve pairwise Euclidean distances. For example, in nonlinear manifolds such as in the Swiss Roll dataset [233], PCA might map two datapoints as near points (measured by the Euclidean distance), while their geodesic distance is much larger.

Isomap has three steps. First, it takes into account the datapoint neighborhood relationships by constructing a graph $G(V, E)$ from $\mathbf{X}$, where each sample $\mathbf{x}_i \in \mathbf{X}$ represents a node $v_i \in V$. Two nodes v_i and v_j are connected by an edge $(v_i, v_j) \in E$ with length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ is one of the K neighbors of $\mathbf{x}_j$. The edge length $d_{\mathbf{X}}(\mathbf{x}_i, \mathbf{x}_j)$ is given by some distance metric between $\mathbf{x}_i$ and $\mathbf{x}_j$ (e.g. Euclidean distance). Then, we estimate the geodesic distances on the manifold between each pair of points in $\mathbf{X}$ by computing the shortest path between each corresponding pair of nodes in G. We store these distances in the pairwise geodesic distance matrix $\mathbf{D}_G$. Finally, we construct the d-dimensional embedding by applying multidimensional scaling [234] (MDS) on $\mathbf{D}_G$ to find the d-dimensional coordinate vectors $\mathbf{y}_i$.

Since neighborhood selection presents an opportunity to incorporate prior knowledge [54], instead of using common approaches (e.g. K nearest neighbors) and inspired by our previous works on HRTF personalization [246, 131] and interpolation [256], we construct the graph G by exploiting the correlations among the HRTFs across directions and subjects (we named it the Intersubject Graph, ISG), according to the following criteria:

Criterion 1. *if $\mathbf{x}_i$ and $\mathbf{x}_j$ represent datapoints of the same location but different subject, then connect them. Instead of applying Isomap separately for each subject as in [231], with this criterion, we tried to exploit the correlation of HRTFs among subjects across same directions. Using this criterion, $P-1$ neighbors were obtained, where P is the number of subjects.*

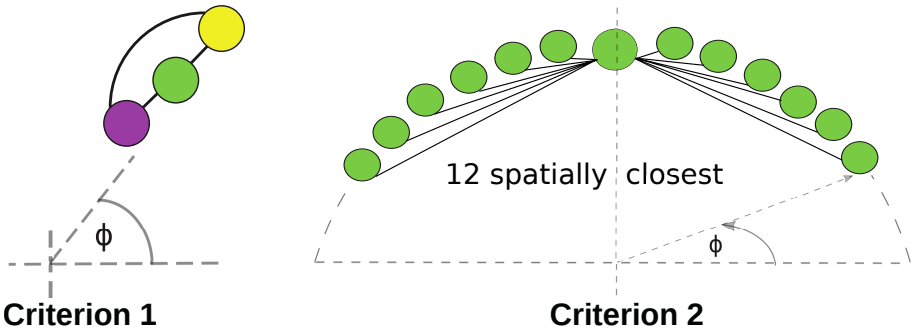


Figure 6.2: Illustrative example of the ISG for $P = 3$ subjects and $k_s = 12$. Color represents same subject HRTFs, and ϕ represents elevation.

Criterion 2. *Let* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *be datapoints of the same subject. If* $\mathbf{x}_j$ *is one of the* k_s *datapoints spatially closest to* $\mathbf{x}_i$*, then connect them. The* k_s *neighbors obtained from this criterion emphasize the similarities between spatially close HRTFs of the same subject.*

It is straightforward to prove that the ISG is always connected (see Fig. 6.2 for an illustrative example). Isomap takes as parameters the number of neighbors, K, and the intrinsic dimensionality, d. Due to our ISG, the number of neighbors is fixed to $K = P+k_s-1$, i.e., $P-1$ from Criterion 1 and k_s from Criterion 2. To estimate the intrinsic dimensionality we use the maximum likelihood intrinsic dimensionality estimator [236]. This estimator has been previously employed in other manifold learning problems [233, 193] and tries to reveal the intrinsic geometric structure of the observed data.

Note that there are other manifold learning methods that could be used with our ISG procedure. For example Laplacian Eigenmaps [238] which is similar to Isomap in that both construct a graph representation of the datapoints. In contrast to Isomap, Laplacian Eigenmaps attempts to preserve only local properties of the manifolds based on the pairwise distances between near neighbors [233].

Ratings from localization model

We use the model for sound-source localization in sagittal planes proposed by Baumgartner et al. [255]. Although the model is applicable to several sagittal planes within the lateral range $\pm 30°$, we only focus on the median plane responses. In this model, it is possible to predict the listener performance in terms of localization error, which, in turn, can be interpreted as the rating a subject would give to certain HRTF, i.e., the localization accuracy obtained with that HRTF. Specifically, we use the model to predict the localization error of listening through non-individualized HRTFs in the median plane. Therefore, we run a series of virtual psychoacoustic experiments to measure a subject localization performance using others' instead of their own ears [255].

The model, that requires a listener-specific calibration, is based on the comparison of an internal sound representation with a template obtained from human listeners' HRTFs.

Since it returns a probabilistic prediction of a polar angle response, we are able to predict the localization performance through local polar error (PE) and quadrant error rate (QE). For both, we follow Middlebrooks [258] and define the PE as the RMS average of polar errors that were less than 90° in magnitude, and the QE as polar errors expressed in percentage form that were larger than 90° .

We normalize the QE and PE to $[0, 1]$ interval, where 1 represents the lowest error (i.e. better localization performance). In order to obtain a single rating to use on the recommender system, we calculate the rating $z^{(i,j)}$ by subject j using HRTF i as the minimum between the normalized PE and QE. We decided to use the minimum because if one of the normalized metrics is low, the overall localization performance is degraded, which in turn means that the corresponding low-rated HRTF is not suitable for the listener.

Content-based recommender system

In content-based recommender systems, we have a d-dimensional feature vector $\mathbf{y}^{(i)} \in \mathbb{R}^{d+1}$ (i.e. including the intercept or bias term) for each item i (e.g. movie, book). We also have a set of ratings on certain scale (e.g. 5-star rating scale) given by user j over a part of the i items we want to recommend.

The goal is to predict user j ratings using a separate linear regression model per user $\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)}$, where $\boldsymbol{\theta}^{(j)}$ is a parameter vector for user j. More formally, we want to learn $\boldsymbol{\theta}^{(j)}$ by minimizing the following linear regression problem per user

$$J = \min_{\boldsymbol{\theta}^{(j)}} \frac{1}{2} \sum_{i:r(i,j)=1} \left(\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} - z^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{k=1}^{d} {\theta_k}^{(j)}, \tag{6.1}$$

where $r\,(i, j) = 1$ if user j has rated item i (0 otherwise), $z^{(i,j)}$ is the rating by user j on item i (if defined), and ${\theta_k}^{(j)}$ is the k-th parameter from the parameter vector $\boldsymbol{\theta}^{(j)}$. The last term from Eq. 6.1 is an L1 regularizer which reduces overfitting and encourages sparsity.

Since we are interested in more than one user, we can reformulate Eq. 6.1 to include n_u users as follows

$$J_m = \min_{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(n_u)}} \frac{1}{2} \sum_{j=1}^{n_u} \sum_{i:r(i,j)=1} \left(\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} - z^{(i,j)} \right)^2 + \frac{\lambda}{2} \sum_{j=1}^{n_u} \sum_{k=1}^{d} {\theta_k}^{(j)} \tag{6.2}$$

The cost function J_m from Eq. 6.2 can be optimized by means of Gradient Descent or similar algorithms. Prior to the optimization, we perform a mean normalization operation on the ratings by subtracting the corresponding average item rating μ_i. Hence, to make

predictions of ratings, we need to add back the corresponding mean, i.e.,

$$\left(\boldsymbol{\theta}^{(j)}\right)^T \mathbf{y}^{(i)} + \mu_i \tag{6.3}$$

In the context of binaural spatial audio, the items i that we want to recommend are HRTFs. The feature vector $\mathbf{y}^{(i)} \in \mathbb{R}^{d+1}$ is the low-dimensional HRTF representation as explained in Section 6.2.1. The rating $z^{(i,j)}$ by subject j on HRTF i is the predicted localization performance obtained by the localization model as described in Section 6.2.2.

Since in practice it is unfeasible that a user has rated all the HRTFs in a database, we randomly selected only a limited number of ratings per direction from the test subject's ratings. In real psychoacoustic experiments, it would be equivalent to a user that performs a limited number of trials per direction. Note that 0 trials per direction means that the listener has not rated any HRTF. In this case, the recommender system just returns the average rating for each HRTF as stated in Eq. 6.3. Finally, our method recommends the HRTFs with the highest predicted rating (i.e. lowest error) per direction. Note that this implies that our method might recommend HRTFs from different subjects, to be combined to form a new HRTF set.

Simulations

Database and Localization model Since the Baumgartner localization model (implemented in the Auditory Modeling Toolbox [259] as `baumgartner2014`) requires a listener-specific calibration [255], the model is only available for 23 subjects, which are included in the 97 subjects from the ARI database[3]. We used the localization model to calculate the ratings for the 23 subjects. So, each of the 23 subjects has ratings for all HRTFs from the whole ARI database, including ratings for the listener's own HRTFs. We only selected 44 directions corresponding to median plane HRTFs.

Pre-processing and Dimensionality reduction We filtered the HRTFs to preserve frequencies between 200 Hz and 18 kHz and calculated the DTFs. We then concatenated the left and right ear DTFs into a single feature vector per direction, as explained in Section 6.2.1. Finally, we reduced the dimensionality of the z-score scaled feature vectors (i.e. normalized to have zero mean and unit variance). These low-dimensional vectors serve as HRTF feature vectors for the recommender system. We compared several linear and nonlinear methods implemented in the *Matlab Dimensionality Reduction* Toolbox [233]. With respect to linear methods, we used PCA with 95% of variance retained. For nonlinear methods, we implemented Isomap and Laplacian Eigenmaps with our ISG (labeled as Isomap-ISG and LEM-ISG) and without it (labeled as Isomap and LEM). For all nonlinear

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

methods, the maximum likelihood estimator [236] established the intrinsic dimensionality to $d = 13$. With respect to the ISG parameter k_s, it should be chosen according to the spatial resolution (5° for the ARI database in almost the entire median plane) and the localization blur in the median plane which varies from ±9° to ±22° [25]. For instance, we chose $k_s = 12$ since the 12 spatially closest sampling points in ARI database cover a ±30° region, i.e, it covers the entire region of the maximum localization blur in the median plane at the 5° spatial resolution of the ARI database.

Recommender system predictions We used a leave-one-out cross validation scheme [260] to test the performance of our method. To do so, we use the localization model of 23 listeners obtained from the ARI database. For each test subject we select its HRTF feature vectors and train our model with the HRTF feature vectors from the remaining 22 subjects. This is done independently for each one of the 23 subjects and the results are latter combined. The recommender was constructed using the ratings from all these 23 subjects on the HRTF feature vectors of the remaining $P = 96$ subjects (i.e. including the test subject's ratings but excluding its HRTF feature vectors). Since in practice a subject performs a limited number of psychoacoustic experiments per direction, we randomly selected only 0 to 8 ratings per direction from the test subject's ratings. Finally, the regularization term λ was selected using a grid search.

Metrics Once we have the HRTFs with the highest predicted rating per direction, we can evaluate them for a specific listener using its Baumgartner model to estimate the listener's localization performance in terms of QE and PE.

Results and Discussion

Figure 6.3 shows the two-dimensional manifold (i.e. first embedded dimension vs second one) recovered with Isomap using our ISG. Observe that there is a strong correlation (the correlation coefficient is 0.98) between the first component and the elevation angle. A similar correlation coefficient (0.99) is found for Laplacian Eigenmaps with our ISG whereas for PCA the correlation coefficient is much lower (0.82).

Before analyzing the localization performance using our recommender, we first analyze the localization performance without it. For example, Fig. 6.4 shows the performance predicted by the Baumgartner model for subject NH12 when using different HRTF sets from ARI database. As expected, the best performance is obtained using the subject's own HRTFs. Observe also that the best performance with someone else's HRTFs is attained using the HRTFs of subject NH93. Ideally, we expect that our recommender system achieves this performance (i.e. the best performance with others' HRTFs) which is better than the mean performance with others' HRTFs (doted line in Fig. 6.4). It is

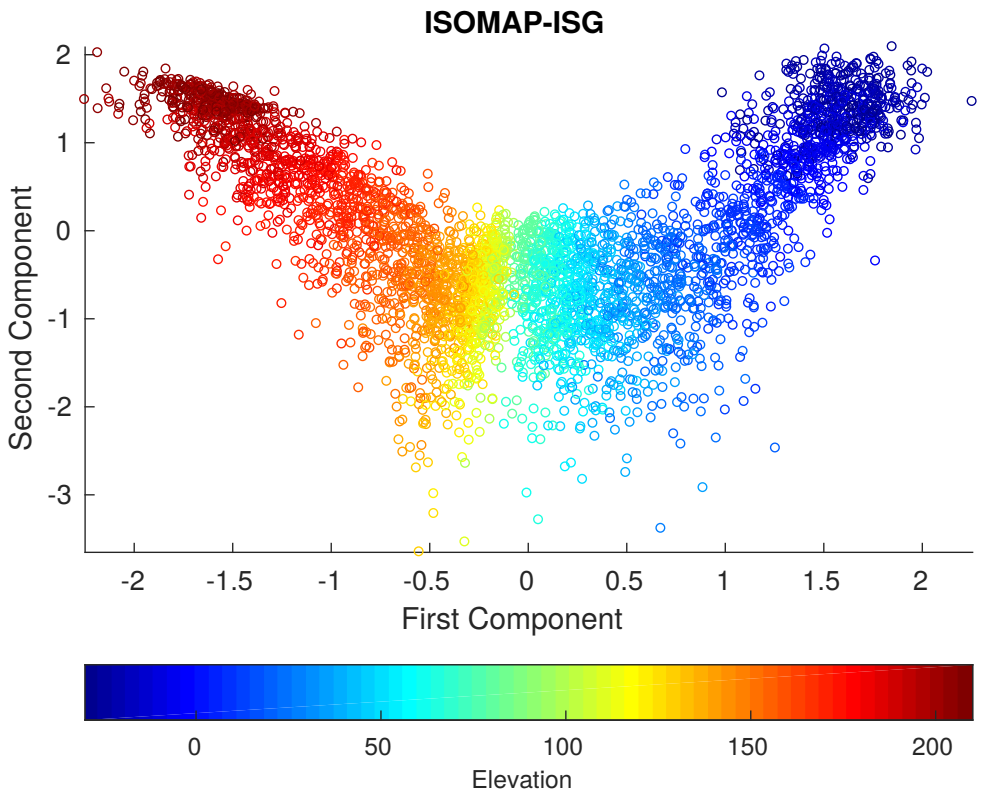


Figure 6.3: Two-dimensional manifold recovered with Isomap using our ISG. All components are normalized to have zero mean and unit variance. Color represents elevation.

worth mentioning that for subject NH12 to achieve the best performance with others' HRTFs (i.e. with NH93's HRTFs) without our recommender, the listener should perform listening tests with the HRTFs from all 96 subjects, which is unfeasible in practice. For instance, to find the best performance with others' HRTFs by carrying out an exhaustive search with three trials for each of the 44 median plane directions on every ARI's subject, a listener should perform $44 \times 3 \times 96 = 12672$ listening tests.

In contrast, Fig. 6.5 presents the localization error using our recommender as a function of the number of trials per direction for subject NH12. Note that ISO-ISG and LEM-ISG outperform the other methods even with only one trial per direction, reaching a better performance than the mean and best performance with others' HRTFs. Although there is some minor improvement when increasing the number of trials beyond three for both ISG conditions, the largest improvement occurs during the first three trials per direction.

Since the ISG conditions have outperformed the others, the remaining analysis will focus on LEM-ISG with three trials per direction. In Fig. 6.6, for LEM-ISG, we show the localization performance relative to the listener-specific performance with its own HRTFs (i.e. the PE and QE variation). For example, the 8° PE variation for NH16 means that the recommended HRTFs provide localization performance that is 8° worse than its own HRTFs. In general, the proposed method has a tendency to reduce the localization error with respect to the mean performance with others' HRTFs and in many cases the error reduction is better than that achieved with the best HRTFs from others. Note also that

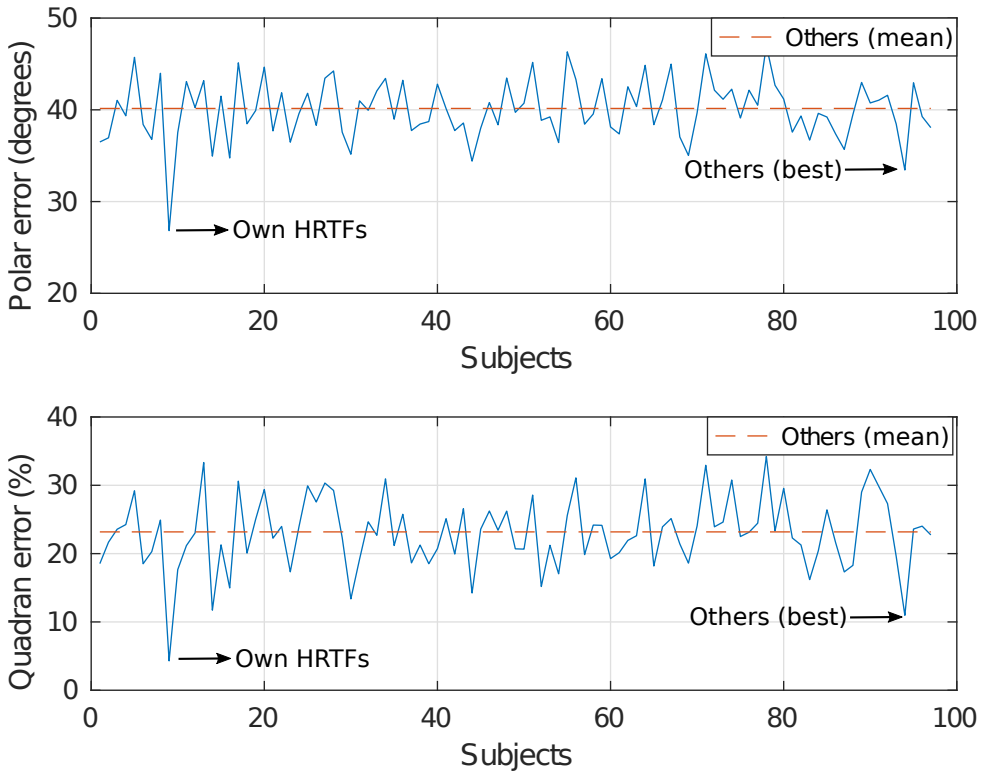


Figure 6.4: Localization performance for subject NH12 when using different HRTFs sets from ARI database.

there are a few negative variations. For instance, the negative PE variation for NH39 means that the recommended HRTFs provide localization performance that is roughly 1° better than its own HRTFs. On the other hand, for NH42 the performance using our recommender is worst when compared to the other subjects. This might be due to the fact that even the best performance variation with others' HRTFs is relatively high with respect to the other individuals.

Finally, Fig. 6.7 shows the localization error for LEM-ISG averaged across all subjects. The bars represent 95% confidence intervals. Paired t-tests confirm that the recommended HRTFs reduce the PE and QE errors with respect to the mean performance with others' HRTFs. Moreover, there is no statistical significance between the performance with the recommended and the best performance with others' HRTFs, which confirms that our recommender system actually improves the localization performance without having to subject the user to perform listening tests on HRTFs from all other subjects on the ARI database aiming at finding the best performance with HRTFs from someone else. For instance, to achieve a similar performance to the best performance with others' HRTFs, our recommender would only need 44 directions $\times$ 3 trials/direction = 132 listening tests, in contrast to the 12672 required by an exhaustive search. On the other hand, the performance with the subject's own HRTFs is still better than the performance with the recommended HRTFs.

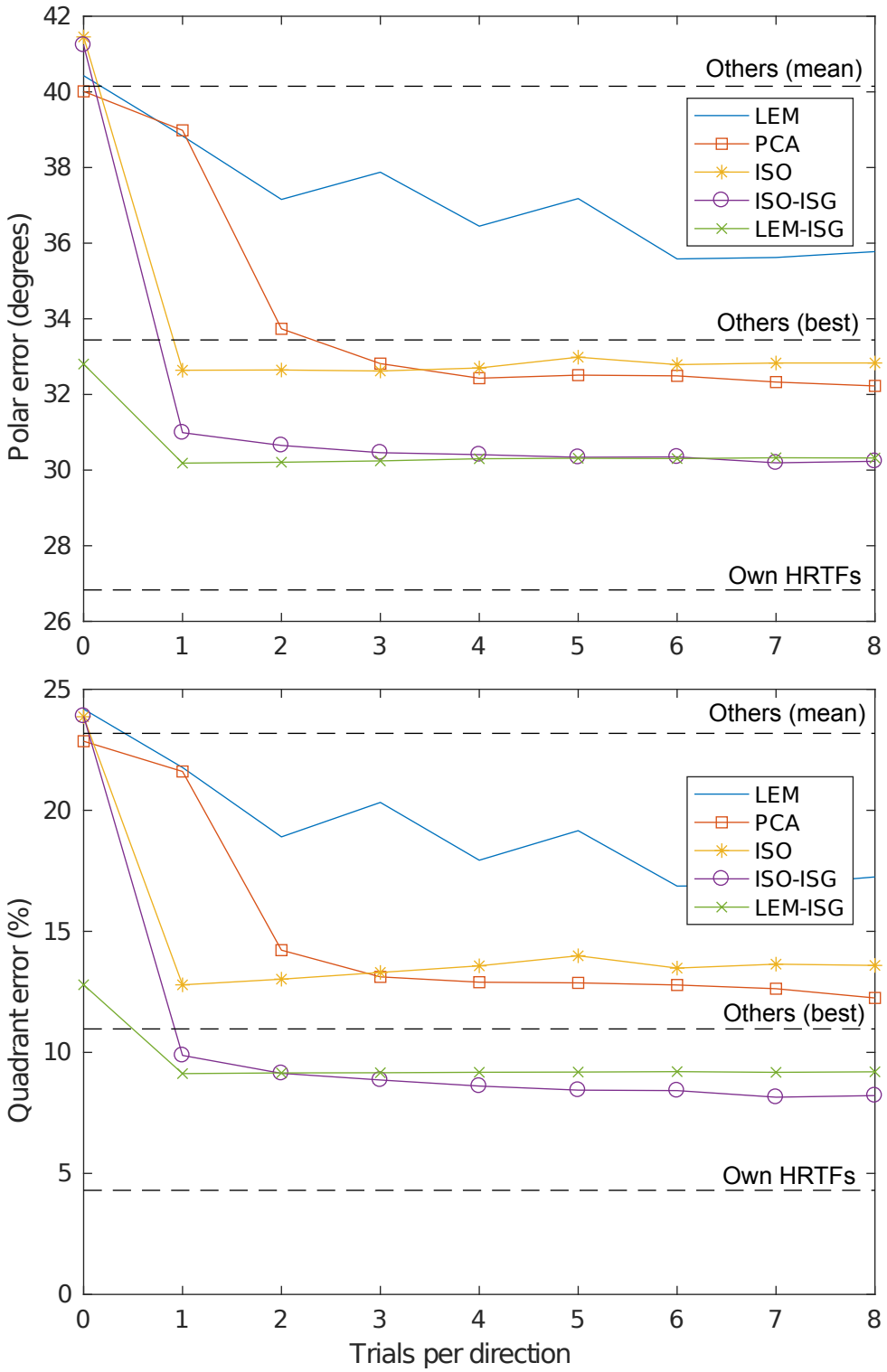


Figure 6.5: Localization performance using our recommender as a function of the number of trials per direction for subject NH12.

Conclusion

We show that although the performance with the subject's own HRTFs is still better than the performance with the recommended HRTF set constructed by combining HRTFs from different individuals, our HRTF recommender can actually reduce the localization

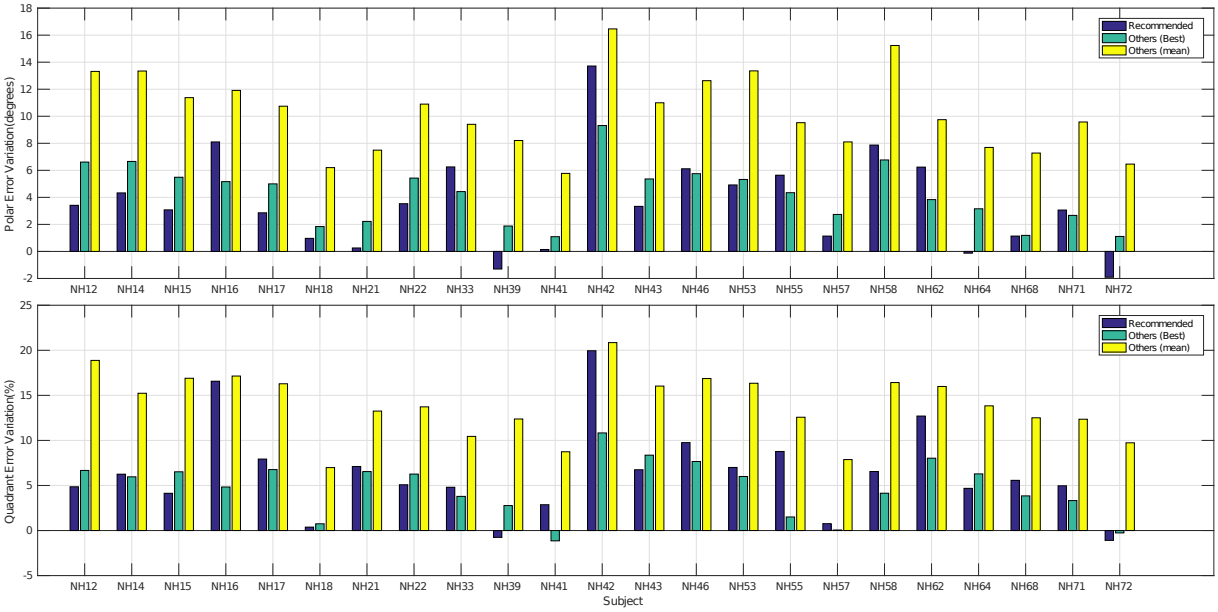


Figure 6.6: Localization performance relative to the listener-specific performance with its own HRTFs (i.e. PE and QE variation) for three trials per direction using LEM-ISG.

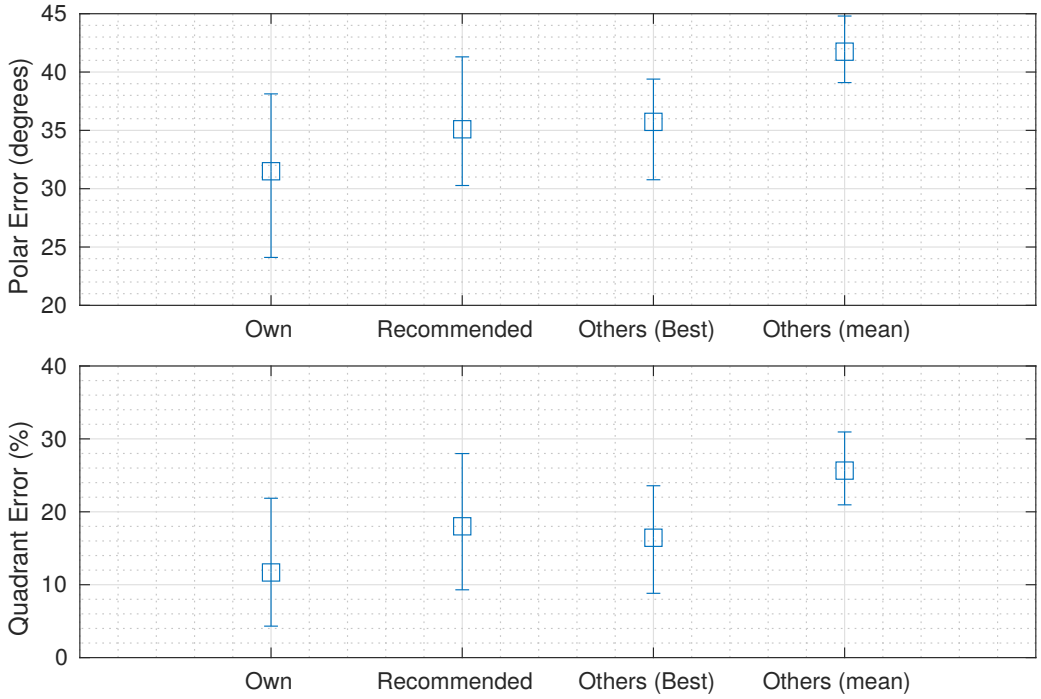


Figure 6.7: Localization performance averaged across subjects for three trials per direction using LEM-ISG. Bars represent 95% confidence intervals.

error with respect to the mean performance with others' HRTFs. Moreover, our technique achieves a performance statistically similar to the best performance with others' HRTFs but with the advantage of not having to perform long and tiring listening test on multiple subjects' datasets looking for the best performance with HRTFs from other listener. We also demonstrate that our ISG on manifold learning techniques such as Isomap and LEM can reduce the error with a small number of trials, outperforming PCA, Isomap and LEM.

Although three trials per direction seems to be too much, note that in practice the number of directions can be reduced if the recommender system is used in conjunction with some interpolation technique [256, 225]. For instance, if the listener performs three trials per direction every 20° instead of every 5°, the number of total trials would reduce drastically. Then, an interpolation method might be used to increase the spatial resolution.

Future works might try different criteria to construct the manifold. For example, instead of taking only neighbors from the same location in Criterion 1, we can select more HRTFs from the vicinity in sagittal planes adjacent to the median plane since it might occur that two subjects are not perfectly aligned during measurement. Furthermore, in a future work, it would be interesting to use more complex recommender algorithms such as [261] to try to obtain a larger improvement when increasing the number of trials beyond three.

Chapter 7

Discussion

So far we have analyzed the results of our four contributions independently of each other. In this chapter, we want to discuss our results as a whole to give a context to our contributions. Note that the analysis presented here have been written after publishing our contributions so that it contains several ideas that appeared during the reviewing processes. In Section 7.1 we discuss several aspects regarding our intersubject graph and the manifold construction. Section 7.2 analyzes the different metrics that we used along our contributions. In Section 7.3, we discuss why there is a need of larger HRTF databases.

Intersubject graph and manifold construction

In contrast to prior works, a recurrent element of our last three contributions has been the construction of a single manifold aiming at incorporating prior knowledge of HRTFs across subjects and directions. We named the graph associated to the manifold the intersubject graph (ISG)[1]. The advantage of constructing a single manifold using our ISG is that we are able to use information from several individuals at once to improve the overall performance of the techniques herein proposed and to preserve the multifactor nature of HRTFs (i.e. subject, direction and frequency). For instance, in our HRTF personalization and interpolation contribution, a single manifold lets us to reconstruct HRTFs by incorporating the information from different subjects. In the same line, it allows to recommend a HRTF set composed by HRTFs from different individuals. With this in mind, observe that we construct the ISG according to criteria derived from each specific problem that we are tackling (i.e. personalization in Chapter 4, interpolation in Chapter 5 and HRTF recommender in Chapter 6) as well as the locations used.

A common criterion along all of our contributions was to connect HRTFs from the same location and ear but different subject (see the Criterion 1 from all contributions).

[1]Observe that even though our graph construction procedure was first introduced in Contribution II, we later adopted the term ISG from our Contribution III onwards.

It is worth mentioning that although this criterion exploits the correlation of HRTFs among subjects across same directions, it does not consider, for instance, the fact that two different subjects are not perfectly aligned during measurement. Furthermore, another common criterion was to connect spatially close HRTFs from the same subject and ear. Beyond the expected similarities due to their similar spatial positions, this criterion alleviates the fact that a subject might not be perfectly aligned during measurement. Note that although we fixed the number of HRTFs that we connect under this criterion, the number of connected neighbors must be thought as a parameter depending on the spatial resolution of the HRTF dataset and the localization blur. With respect to spatial resolution, it implies that this parameter might be even an adaptive parameter when the spatial resolution is not uniform. Moreover, it might be also related to the localization blur as we do in our recommender system in Contribution IV.

On the other hand, there were also different criteria used across our contributions. Consider our second contribution regarding HRTF personalization whose Criterion 2 attempts to include the spatial left/right symmetry of HRTFs. Although this criterion led to good results in the case of HRTF personalization, we found out in our third contribution that the results were not as good as expected for HRTF interpolation. Therefore, we decided to relax this criterion for our HRTF interpolation contribution. However, observe that in both cases the idea is still to exploit the same knowledge, i.e., HRTFs have a strong left/right symmetry. In fact, we can think of Criterion 2 from Contribution III as a relaxed symmetry criterion.

However, note that our ISG does not take into account prior information on a per band basis. For instance, we consider left/right symmetry as though it were valid along all frequencies while the evidence shows that it is only valid up to roughly 5.5 kHz [41]. This surely contributed to larger spectral distortion at high frequencies in our interpolation and personalization techniques.

One concern that has arisen along all of our contributions was how to interpret the dimensions of the resultant manifold. In our second contribution, we demonstrated that the first and second manifold components are strongly correlated with azimuth and elevation using our intersubject graph construction. With respect to inter-subject variability, we found that no specific component(s) correlate(s) to inter-subject variation, at least in a clear way as directional factors do, confirming the need of nonlinear regression techniques such as neural networks to relate anthropometric features with the manifold components.

Even though in our HRTF personalization contribution we used Isomap, it is also possible to use other manifold learning techniques provided that they are based on neighborhood graphs such as Laplacian Eigenmaps, Locally Linear Embedding, and Maximum Variance Unfolding (see [60] for a comprehensive review of these and other dimensionality reduction techniques). In fact, we applied several of these manifold learning techniques when constructing the manifold for our HRTF interpolation method as well as our recom-

mender system in Contribution IV. Furthermore, the manifold learning techniques that we used can only embed data that gives rise to a connected neighborhood graph. If the neighborhood graph contains more than one connected component, the resultant components would lie on different manifolds. It is worth mentioning here that all of our contributions using manifold learning guarantee a connected graph so that we do not address the non-connected graph case.

Beyond the non-connected graph disadvantage, the main weakness of these manifold learning techniques based on neighborhood graph is the lack of an explicit mapping function [60] to project new high-dimensional datapoints into an existing low-dimensional embedding (i.e. out-of-sample extension), and to reconstruct a low-dimensional datapoint into a high-dimensional representation (i.e. back-projection). Out-of-sample extension might be performed by means of the Nystrom approximation [262, 263], so that, for new datapoints, there is no need to recalculate the entire manifold. This is crucial for computationally intensive techniques like Isomap that require performing an eigendecomposition of a full pairwise distance matrix [53].

The back-projection is a more challenging problem to overcome. For our HRTF personalization contribution, we have reconstructed high-dimensional HRTFs using a linear combination of its neighbors (i.e. neighborhood-based reconstruction) since we are assuming local linearity. It should be observed that the main weakness of this reconstruction is that its accuracy depends on how dense the initial database is. This problem might be addressed using some spatial HRTF interpolation before HRTF personalization to guarantee a more populated manifold. However, note that if the initial database is not sampled adequately in space, the resultant interpolated HRTFs will not be suitable to reconstruct the personalized HRTFs. Even though we chose the reconstruction weights to be the corresponding neighbor's euclidean inverse distances, there is an alternative approach where the weights are determined in a least-squares optimization. In fact, observe that in our HRTF interpolation contribution, we have calculated the weights by solving a constrained least-squares problem. Finally, note that our HRTF recommender system does not require back-projection because in that contribution we are not interested in reconstructing HRTFs.

HRTF similarity criterion

In our HRTF personalization work we used, as a similarity criterion or metric, the spectral distortion which has been widely employed in prior works to compare predicted HRTFs against the ground truth ones. Contrastingly, in our Contribution III (i.e. HRTF interpolation), we used a new metric introduced by Middlebrooks [264] (i.e. the variance metric) instead of the spectral distortion. We opted for the variance metric because the main drawback of the spectral distortion is that if there is, e.g, a constant intensity offset

between a similar-shape pair of HRTFs to be compared, the spectral distortion will be nonzero while its variance metric will not.

It is worth noting that there are several similarity metrics in addition to the ones used in this thesis such as the signal to distortion ratio (see [16, Ch. 5] for a non-exhaustive list). However, observe that finding a relevant HRTF similarity criterion is not trivial and is still an open problem since a distance criterion should consider objective as well as subjective aspects (i.e. perceptual aspects) [265]. It means that a relevant distance criterion should consistently predict audible differences in listening tests [265]. For instance, in our HRTF personalization method the high contralateral error as measured by the spectral distortion does not necessarily imply a perceptually relevant error. In fact, previous studies concluded that the spectral detail of HRTFs at high frequency and contralateral directions is to some extent inaudible [266].

HRTF databases

Currently, one critical problem when trying to personalize HRTFs from anatomical features using data-driven techniques is the lack of sufficiently large HRTF datasets (i.e. both with respect to the number of subjects and reliable anthropometric data) in publicly available databases. To the best of our knowledge, the largest publicly available HRTF database including anthropometric data for 51 subjects is the LISTEN database [267]. Although other databases offer HRTFs for over a hundred of subjects (e.g. 105 for the RIEC database [268], 120 for the ARI database [40]), they have anthropometric data only for close to 50 subjects, which is a number similar to the CIPIC database [39] and the ITA database [269].

Most databases [39, 267, 40] have only measured certain anthropometric features that are believed to be relevant to our sound perception (e.g. head width and head depth correlates strongly with ITD). However, it should be noted that anthropometric parameter selection for predicting personalized HRTFs, especially pinna-related parameters, continues to be an open problem. For instance, in our second contribution we chose the anthropometric parameters according to [270] because, from our point of view, they perform the most complete statistical analysis of anthropometric feature selection in the entire virtual auditory space (i.e. azimuth and elevation) based on PCA, Pearson's product-moment correlation coefficient analysis and multiple linear regression analysis.

In order to obtain more reliable anthropometric data, more recently, several research laboratories [268, 269] have measured 3D scans of the head and/or pinna in addition to anthropometric measurements. Due to the increasing spatial audio applications, especially in virtual and augmented reality, we expect that larger HRTF databases along with morphological data will be freely available, which in turn will allow to properly apply more complex data-driven algorithms. For example, in a recent ongoing project [271], the

authors claim to have begun to measure HRTFs and 3D scans of humans into a publicly available database.

Beyond anthropometric data, Xie et al. [272], after comparing the similarity between the Chinese HRTFs and those of Western subjects, concluded that Chinese and Western HRTFs differ significantly on statistics. This implies that HRTF personalization techniques should consider not only the number of subjects but also people from different populations.

As it can be inferred so far, the lack of HRTF data is a great constraint to apply data-driven techniques such as our proposed HRTF personalization method. For this reason, we opted for exploring perceptual-based methods along with machine learning techniques in our Contribution IV. It is worth noting that we incorporated our previous ideas from our HRTF personalization and interpolation contributions such as the intersubject graph into the recommender system. Even though our HRTF recommender still requires the subjects to perform listening tests, our HRTF interpolation contribution might be used in conjunction with the recommender system to drastically reduce the number of total trials required to improve the localization performance.

Finally, in our Contribution I regarding the application of spatial audio for conveying spatial information to visually impaired people through a proof of concept wearable device, we used the HRTFs from KEMAR. Although we did not perform formal listening tests, the visually impaired participants reported that sometimes the perceived sounds were as though they were coming from behind during the usability tests. This was expected due to the non-individualized HRTFs. Despite this disadvantage, spatial audio is still attractive in this application because it does not overwhelm the user's hearing sense.

Chapter 8

Conclusion and Future Work

Conclusions

In this thesis, we investigated how to incorporate spatial audio prior knowledge using manifold learning to tackle several challenges of broad interest among the spatial audio community such as HRTF personalization, HRTF interpolation and sound localization improvement through perceptual tests. Moreover, we presented a practical application of spatial audio aimed at visually impaired people for identifying the location of known faces.

For all of our contributions using manifold learning, the construction of a single manifold across subjects through the Intersubject Graph (ISG) has proven to lead to a powerful HRTF representation capable of incorporating prior knowledge of spatial audio and capturing the underlying factors of spatial hearing. Moreover, the advantage of constructing a single manifold using our ISG is the use of information from other individuals to improve the overall performance of the techniques proposed in our contributions. The results showed that our ISG-based techniques have outperformed other linear and nonlinear methods on the three spatial audio problems addressed by this thesis, i.e., HRTF personalization, HRTF interpolation and sound localization improvement based on perceptual tests.

In our Contribution II proposing an HRTF personalization technique based on anthropometric measurements, we found that Isomap in conjunction with our ISG was capable of explaining the directional factors of HRTFs (i.e. azimuth and elevation) while the intersubject differences were reflected in the wide inter-dimension variability. Moreover, instead of constructing one regression model per direction, our HRTF personalization approach allows the construction of a single model that does not break the inherent multifactor nature of HRTFs (i.e. frequency, direction and subject factors) using a multiple output Artificial Neural Network that exploits the correlations between the manifold components. Our experiments showed, with p-values less than 0.05, that our approach outperforms us-

ing, either a PCA linear reduction, or the full HTRF, in its intermediate stages. This confirmed that the intraconic representation effectively encodes the most important cues for individualization of HRTFs. Furthermore, we expect that our HRTF personalization contribution might perform better given a more suitable input dataset (i.e. higher spatial resolution, more subjects and better quality anatomical measurements). Finally, provided that spectral detail of HRTFs at high frequency is inaudible, in listening tests we expect the localization accuracy to be good at lateral locations, reasonable at vertical directions but poor in low elevations.

In our Contribution III exploring an HRTF interpolation method, we showed that a single manifold representation allows measured HRTFs from different subjects to contribute for reconstructing the HRTFs for new directions. Moreover, our results suggest that a small number of spatial measurements capture most of the acoustical properties of HRTFs. Our interpolation approach outperforms other linear and nonlinear dimensionality reduction techniques such as principal component analysis, locally linear embedding, and Laplacian eigenmaps. Observe that our interpolation technique might not be adequate for real-time interpolation since it requires information from all subjects in the database, which in turn, may be computationally prohibitive on platforms with limited hardware resources.

Since the lack of HRTF data is a great constraint to apply data-driven techniques such as our proposed HRTF personalization method, we opted for exploring perceptual-based methods along with machine learning techniques in our Contribution IV. The simulation results show that our HRTF recommender is capable of choosing a set of HRTFs that improves the median plane localization performance with respect to the mean localization performance using non-individualized HRTFs. Moreover, the localization performance achieved by our HRTF recommender system shows no significant difference to the localization performance observed with the best matching non-individualized HRTFs, but with the advantage of not having to perform listening tests with all individuals' HRTFs from the database. We also demonstrate that our ISG in conjunction with manifold learning techniques such as Isomap and LEM can reduce the error with a small number of trials, outperforming PCA, Isomap and LEM. Although three trials per direction seems to be too much, it is pertinent to emphasize that in practice the number of directions can be reduced if the recommender system is used in conjunction with some interpolation technique, reducing drastically the number of total trials. For instance, with our HRTF interpolation contribution, we showed that we can reconstruct median plane HRTFs by sampling every 22.5°.

Finally, in our Contribution I where we used spatial audio to indicate a face location to visually impaired people, the results revealed that our approach, on average, outperforms traditional face recognition methods while requiring much less computational resources, making it suitable for the wearable hardware constraints and real-time requirements. Con-

veying the directional location of a face in the environment using 3-D Audio proved to be an efficient feedback that does not overwhelm the person's auditory sense. On the other hand, distance location appears to be more challenging to communicate. After the pilot experiment, we passed from a frequency variation approach to a spoken language approach. Moreover, generic HRTFs, as used in this prototype, tend to hinder the user's audio localization capability. Even so, this disadvantage of spatial audio is counterbalanced by its versatility of not overwhelming the user's hearing sense, which is crucial for blind people, especially when cognitive load is present.

Future Works

In this section, we conjecture on some future research directions to be considered based on the problems addressed in this thesis

1. In this work, we proposed the idea of a single manifold across subjects without considering that different cues act in different frequency bands. Future works might explore subband representation for HRTFs in conjunction with manifold learning. This would permit to incorporate prior knowledge in the subband where this prior is effectively valid.

2. A future work might explore a multi-task learning (MTL) approach to learn the regression model. MTL learns multiple related tasks simultaneously using a shared representation aimed at improving generalization. In the HRTF personalization context, a task could be, e.g., learning a regression model per direction. The MTL approach might preserve the multi-factor nature of HRTFs by using a shared representation instead of learning a single regression model as we do in our HRTF personalization contribution. We expect that this approach would reduce the spectral distortion at high frequencies.

3. In this work, we employed several distance metrics such as the spectral distortion and the variance metric. Beyond using just different metrics, a future work might use metric learning to find a new relevant distance metric that introduces perceptual information.

4. An immediate direction to be followed would be to apply manifold learning for HRTF extrapolation and near-field HRTF interpolation.

5. In this work, we used a content-based recommender system. Future contributions might explore other recommender systems techniques such as collaborative filtering or hybrid approaches.

6. The graph construction procedure offers a great flexibility to introduce prior knowledge of spatial audio. Future works might try different criteria to construct the manifold. For example, instead of taking only neighbors from the same location in Criterion 1 of all our contributions, we can select more HRTFs from the vicinity in sagittal planes adjacent to the median plane since it might occur that two subjects are not perfectly aligned during measurement.

7. If larger HRTF databases were available, including 3D models of human anatomy, it would be to possible personalize HRTFs using deep neural networks which require large amounts of data to work properly.

8. In our practical application of spatial audio to assist visually impaired people, we encoded distance using a frequency variation and a spoken language approaches. A future work should explore alternative forms to communicate distance (e.g., musical tones).

References

[1] Durand Begault. *3D Sound for Virtual Reality and Multimedia*. AP Professional, 1994.

[2] Aki Härmä, Julia Jakka, Miikka Tikander, Matti Karjalainen, Tapio Lokki, Jarmo Hiipakka, and Gaëtan Lorho. Augmented Reality Audio for Mobile and Wearable Appliances. *J. of the Audio Eng. Soc.*, 52(6), 2004.

[3] Ronald Azuma, Mike Daily, and Jimmy Krozel. Advanced human-computer interfaces for air traffic management and simulation. In *Flight Simulation Technologies Conf.*, Reston, Virigina, July 1996. American Institute of Aeronautics and Astronautics.

[4] Shraga Shoval, Iwan Ulrich, and Johann Borenstein. NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired]. *Robotics & Automation Magazine, IEEE*, 10(1), 2003.

[5] Brian F. G. Katz, Slim Kammoun, Gaëtan Parseihian, Olivier Gutierrez, Adrien Brilhault, Malika Auvray, Philippe Truillet, Michel Denis, Simon Thorpe, and Christophe Jouffrais. NAVIG: Augmented reality guidance system for the visually impaired. *Virtual Reality*, 16(4):253–269, June 2012.

[6] V. Valimaki, A. Franck, J. Ramo, H. Gamper, and L. Savioja. Assisted Listening Using a Headset: Enhancing audio perception in real, augmented, and virtual environments. *IEEE Signal Processing Magazine*, 32(2):92–99, March 2015.

[7] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[8] W. Zhixin and C. Cheung-Fat. Continuous Function Modeling of Head-Related Impulse Response. *IEEE Signal Processing Letters*, 22(3):283–287, March 2015.

[9] Mitsuo Matsumoto, Mikio Tohyama, and Hirofumi Yanagawa. A method of interpolating binaural impulse responses for moving sound images. *Acoustical Science*

and Technology, 24(5):284–292, 2003.

[10] Jeroen Breebaart. Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis. *The Journal of the Acoustical Society of America*, 133(1):EL1–EL6, January 2013.

[11] H. Sebastian Seung and Daniel D. Lee. The manifold ways of perception. *science*, 290(5500):2268–2269, 2000.

[12] Ramani Duraiswami and Vikas C. Raykar. The manifolds of spatial hearing. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 3, pages iii–285. IEEE, 2005.

[13] A. Kohlrausch, J. Braasch, D. Kolossa, and J. Blauert. An Introduction to Binaural Processing. In Jens Blauert, editor, *The Technology of Binaural Listening*, pages 1–32. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[14] Jan C. Schacher and Philippe Kocher. Ambisonics spatialization tools for max/msp. *Omni*, 500(1), 2006.

[15] Hagen Wierstorf and Sascha Spors. Sound field synthesis toolbox. In *Audio Engineering Society Convention 132*. Audio Engineering Society, 2012.

[16] Bosun Xie. *Head-Related Transfer Function and Virtual Auditory Display*. J Ross, Plantation, FL, USA., 2013.

[17] Lord Rayleigh. XII. On our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[18] Brian F. G. Katz and Markus Noisternig. A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, 135(6):3530–3540, 2014.

[19] Ewan A. Macpherson and John C. Middlebrooks. Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236, 2002.

[20] John C Middlebrooks and David M Green. Sound localization by human listeners. *Annual review of psychology*, 42(1):135–159, 1991.

[21] Frederic L. Wightman and Doris J. Kistler. 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, 1999.

[22] Stephen Perrett and William Noble. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332, 1997.

[23] V Ralph Algazi, Carlos Avendano, and Richard O Duda. Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3):1110–1122, 2001.

[24] Gregg H Recanzone. Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences*, 95(3):869–875, 1998.

[25] Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT press, 1997.

[26] German Ramos and Maximo Cobos. Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications. *The Journal of the Acoustical Society of America*, 134(3):1735–1738, 2013.

[27] German Ramos, Maximo Cobos, Balázs Bank, and Jose A. Belloch. A Parallel Approach to HRTF Approximation and Interpolation Based on a Parametric Filter Model. *IEEE Signal Processing Letters*, 24(10):1507–1511, 2017.

[28] Hannes Gamper. Head-related transfer function interpolation in azimuth, elevation, and distance. *The Journal of the Acoustical Society of America*, 134(6):EL547–EL553, 2013.

[29] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein. Interpolation of Head-Related Transfer Functions Using Manifold Learning. *IEEE Signal Processing Letters*, 24(2):221–225, February 2017.

[30] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262, 2002.

[31] Shouichi Takane, Daisuke Arai, Tohru Miyajima, Kanji Watanabe, Yôiti Suzuki, and Toshio Sone. A database of Head-Related Transfer Functions in whole directions on upper hemisphere. *Acoustical science and technology*, 23(3):160–162, 2002.

[32] Piotr Majdak, Peter Balazs, and Bernhard Laback. Multiple exponential sweep method for fast measurement of head-related transfer functions. *Journal of the Audio Engineering Society*, 55(7/8):623–637, 2007.

[33] Ning Xiang and Manfred R. Schroeder. Reciprocal maximum-length sequence pairs for acoustical dual source measurements. *The Journal of the Acoustical Society of America*, 113(5):2754–2761, 2003.

[34] Pavel Zahorik. Limitations in using Golay codes for head-related transfer function measurement. *The Journal of the Acoustical Society of America*, 107(3):1793–1796, 2000.

[35] Bin Zhou, David M. Green, and John C. Middlebrooks. Characterization of external

ear impulse responses using Golay codes. *The Journal of the Acoustical Society of America*, 92(2):1169–1171, 1992.

[36] Swen Müller and Paulo Massarani. Distortion immunity in impulse response measurements with sweeps. In *International Congress on Sound & Vibration*, volume 18, 2011.

[37] Henrik Møller. Fundamentals of binaural technology. *Applied acoustics*, 36(3-4):171–218, 1992.

[38] William G. Gardner and Keith D. Martin. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[39] V. Ralph Algazi, Richard O. Duda, Dennis M. Thompson, and Carlos Avendano. The cipic HRTF database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 99–102. IEEE, 2001.

[40] Piotr Majdak. ARI HRTF Database. https://www.kfs.oeaw.ac.at/, 2017.

[41] BoSun Xie, XiaoLi Zhong, Dan Rao, and ZhiQiang Liang. Head-related transfer function database and its analyses. *Science in China Series G: Physics, Mechanics and Astronomy*, 50(3):267–280, June 2007.

[42] Piotr Majdak, Yukio Iwaya, Thibaut Carpentier, Rozenn Nicol, Matthieu Parmentier, Agnieszka Roginska, Yôiti Suzuki, Kankji Watanabe, Hagen Wierstorf, and Harald Ziegelwanger. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[43] Sofa. SOFA (Spatially Oriented Format for Acoustics) - Sofaconventions. https://www.sofaconventions.org/mediawiki/index.php/Main_Page, 2018.

[44] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. Estimation of a spherical-head model from anthropometry. *Journal of the Audio Engineering Society*, 49(6):472–479, 2001.

[45] Erno HA Langendijk and Adelbert W. Bronkhorst. Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4):1583–1596, 2002.

[46] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2009.

[47] Doris J. Kistler and Frederic L. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3):1637–1647, 1992.

[48] A. Kulkarni, S. K. Isabelle, and H. S. Colburn. On the minimum-phase approxima-

tion of head-related transfer functions. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995.,*, pages 84–87. IEEE, 1995.

[49] Jose A. Costa and Alfred O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

[50] Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision*, 7(2–3):81–227, 2012.

[51] Warren Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[52] Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.

[53] Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[54] Lawrence K. Saul and Sam T. Roweis. Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. *J. Mach. Learn. Res.*, 4:119–155, December 2003.

[55] Mukund Balasubramanian and Eric L. Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7–7, 2002.

[56] Matti Niskanen and Olli Silvén. Comparison of dimensionality reduction methods for wood surface inspection. In *Sixth International Conference on Quality Control by Artificial Vision*, volume 5132, pages 178–189. International Society for Optics and Photonics, 2003.

[57] Ahmed Elgammal and Chan-Su Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[58] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[59] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[60] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: A comparative. *J Mach Learn Res*, 10:66–71, 2009.

[61] World Health Organization. Visual impairment and blindness: Fact sheet number 282, August 2014. Accessed: 2015-10-10. Available: http://www.who.int/mediacentre/factsheets/fs282/en/.

[62] S.L. Joseph, Jizhong Xiao, Xiaochen Zhang, B. Chawda, K. Narang, N. Rajput, S. Mehta, and L.V. Subramaniam. Being Aware of the World: Toward Using Social Media to Support the Blind With Navigationd. *IEEE Trans. Hum.-Mach. Syst.*, 45(3):399–405, 2015.

[63] Sabina Kef, Joop Hox, and H. Habekothe. Social networks of visually impaired and blind adolescents. Structure and effect on well-being. *Soc. Netw.*, 22(1):73–91, 2000.

[64] Chong-Wen Wang, Cecilia Chan, Andy Ho, and Zhifan Xiong. Social networks and health-related quality of life among Chinese older adults with vision impairment. *J. Aging Health*, 20(7):804–823, 2008.

[65] T. Gallagher, E. Wise, H.C. Yam, B. Li, E. Ramsey-Stewart, A.G. Dempster, and C. Rizos. Indoor navigation for people who are blind or vision impaired: Where are we and where are we going? *J. Locat. Based Serv.*, 8(1):54–73, 2014.

[66] Roberto Manduchi and James Coughlan. (Computer) Vision Without Sight. *Commun. ACM*, 55(1):96–104, 2012.

[67] D. Dakopoulos and N.G. Bourbakis. Wearable obstacle avoidance electronic travel aids for blind: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 40(1):25–35, 2010.

[68] U.R.a Roentgen, G.J.b Gelderblom, M.e Soede, and L.P.c de Witte. Inventory of electronic mobility aids for persons with visual impairments: A literature review. *J. Vis. Impair. Blind.*, 102(11):702–724, 2008.

[69] Thierry Pun, Patrick Roth, Guido Bologna, Konstantinos Moustakas, and Dimitrios Tzovaras. Image and Video Processing for Visually Handicapped People. *J. Image Video Process.*, 2007(5):1–12, 2007.

[70] J. Xiao, S.L. Joseph, X. Zhang, B. Li, X. Li, and J. Zhang. An Assistive Navigation Framework for the Visually Impaired. *IEEE Trans. Hum.-Mach. Syst.*, 45(5):635–640, 2015.

[71] B. Ando, S. Baglio, V. Marletta, and A. Valastro. A Haptic Solution to Assist Visually Impaired in Mobility Tasks. *IEEE Trans. Hum.-Mach. Syst.*, 45(5):641–646, 2015.

[72] Rabia Jafri, Syed Ali, Hamid Arabnia, and Shameem Fatima. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *Vis. Comput.*, 30(11):1197–1222, 2014.

[73] Xiaodong Yang, Shuai Yuan, and YingLi Tian. Assistive Clothing Pattern Recognition for Visually Impaired People. *IEEE Trans. Hum.-Mach. Syst.*, 44(2):234–243, 2014.

[74] C Iordanoglou, Kenneth Jonsson, Josef Kittler, and Jiri Matas. Wearable face recognition aid. In *IEEE ICASSP*, 2000.

[75] Giovanni Fusco, Nicoletta Noceti, and Francesca Odone. Combining retrieval and classification for real-time face recognition. In *IEEE AVSS*, 2012.

[76] Wei Li, Xudong Li, Martin Goldberg, and Zhigang Zhu. Face Recognition by 3D Registration for the Visually Impaired Using a RGB-D Sensor. In *ECCV*. Springer, 2014.

[77] S. Krishna, G. Little, J. Black, and S. Panchanathan. A wearable face recognition system for individuals with visual impairments. In *ACM ASSETS*, 2005.

[78] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3(1):71–86, 1991.

[79] D.a Astler, H.a Chau, K.a Hsu, A.a Hua, A.a Kannan, L.a Lei, M.a Nathanson, E.a Paryavi, M.a Rosen, H.a Unno, C.a Wang, K.a Zaidi, X.a Zhang, and C Tang. Increased accessibility to nonverbal communication through facial and expression recognition technologies for blind/visually impaired subjects. In *ACMASSETS*, 2011.

[80] Y. Utsumi, Y. Kato, K. Kunze, M. Iwamura, and K. Kise. Who are you? - A wearable face recognition system to support human memory. In *ACM AH*, 2013.

[81] P. Viola and M.J. Jones. Robust real-time face detection. *Int. J. Computer Vision*, 57(2):137–154, 2004.

[82] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *IEEE CVPR*, 2010.

[83] K.M. Kramer, D.S. Hedin, and D.J. Rolkosky. Smartphone based face recognition tool for the blind. In *IEEE EMBC*, 2010.

[84] L Balduzzi, G Fusco, F Odone, S Dini, M Mesiti, A Destrero, and A Lovato. Low-cost face biometry for visually impaired users. In *IEEE BioMS*, 2010.

[85] T. Ahonen, A. Hadid, and M. Pietikainen. Face Description with Local Binary Patterns: Application to Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2037–2041, 2006.

[86] Filippo Battaglia and Giancarlo Iannizzotto. An open architecture to develop a handheld device for helping visually impaired people. *IEEE Trans. Consum. Electron.*, 58(3):1086–1093, 2012.

[87] Shafiqur Réhman and Li Liu. iFeeling: Vibrotactile Rendering of Human Emotions on Mobile Phones. In *Mobile Multimedia Process.* Springer, 2010.

[88] S. Krishna and S. Panchanathan. Assistive technologies as effective mediators in interpersonal social interactions for persons with visual disability. In *Comput. Helping People with Special Needs*. Springer, 2010.

[89] A.K.M.M. Rahman, M.I. Tanveer, A.S.M.I. Anam, and M. Yeasin. IMAPS: A smart phone based real-time framework for prediction of affect in natural dyadic conversation. In *Vis. Commun. and Image Process.* IEEE, 2012.

[90] M.I. Tanveer, A.S.M. Iftekhar Anam, M. Yeasin, A.K.M. Mahbubur Rahman, and S. Ghosh. FEPS: A sensory substitution system for the blind to perceive facial expressions. In *ACM ASSETS*, 2012.

[91] M.I.a Tanveer, A.S.M.I.b Anam, M.b Yeasin, and M.c Khan. Do you see what I see? Designing a Sensory Substitution Device to access non-verbal modes of communication. In *ACM ASSETS*, 2013.

[92] A.S.M.I. Anam, S. Alam, and M. Yeasin. Expression: A dyadic conversation aid using Google Glass for people who are blind or visually impaired. In *MobiCASE*. IEEE, 2014.

[93] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan. Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In *IEEE HAVE*, 2008.

[94] L. Gade, S. Krishna, and S. Panchanathan. Person localization using a wearable camera towards enhancing social interactions for individuals with visual impairment. In *ACM SIGMM*, 2009.

[95] M.I. Tanveer and M.E. Hoque. A google glass app to help the blind in small talk. In *ASSETS*. ACM, 2014.

[96] G. Balakrishnan, G. Sainarayanan, R. Nagarajan, and Sazali Yaacob. Wearable Real-Time Stereo Vision for the Visually Impaired. *Eng. Lett.*, 14(2):6–14, 2007.

[97] Hugo Fernandes, Paulo Costa, Vitor Filipe, L. Hadjileontiadis, and João Barroso. Stereo vision in blind navigation assistance. In *World Automation Congr.* IEEE, 2010.

[98] Kai Lin, Tak Lau, Chi Cheuk, and Yunhui Liu. A wearable stereo vision system for visually impaired. In *IEEE ICMA*, 2012.

[99] F.a Ribeiro, D.b Florencio, P.A.b Chou, and Z.b Zhang. Auditory augmented reality: Object sonification for the visually impaired. In *IEEE MMSP*, 2012.

[100] A. Khan, F. Moideen, J. Lopez, W.L. Khoo, and Z. Zhu. KinDectect: Kinect

detecting objects. In *Comput. Helping People with Special Needs*. Springer, 2012.

[101] V.a Filipe, F.b Fernandes, H.b Fernandes, A.c Sousa, H.d Paredes, and J.e Barroso. Blind navigation support system based on Microsoft Kinect. In *Conf. Software Develop. for Enhancing Accessibility and Fighting Info-exclusion*. Elsevier, 2012.

[102] Titus Tang and Wai Li. An assistive EyeWear prototype that interactively converts 3D object locations into spatial audio. In *ISWC*. ACM, 2014.

[103] Yingli Tian. RGB-D Sensor-Based Computer Vision Assistive Technology for Visually Impaired Persons. In *Comput. Vision and Mach. Learning with RGB-D Sensors*. Springer, 2014.

[104] Juan Gomez, Guido Bologna, and Thierry Pun. See ColOr: an extended sensory substitution device for the visually impaired. *J. Assist. Technol.*, 8(2):77–94, 2014.

[105] A.a Bhowmick, S.a Prakash, R.a Bhagat, V.a Prasad, and S.M.b Hazarika. IntelliNavi: Navigation for blind based on kinect and machine learning. In *Multi-disciplinary Trends in Artificial Intell.* Springer, 2014.

[106] C.a Stoll, R.a Palluel-Germain, V.c Fristot, D.c Pellerin, D.a Alleysson, and C.a Graff. Navigating from a depth image converted into sound. *Appl. Bionics Biomech.*, 2015:1–9, 2015.

[107] Hotaka Takizawa, Shotaro Yamaguchi, Mayumi Aoyagi, Nobuo Ezaki, and Shinji Mizuno. Kinect cane: an assistive system for the visually impaired based on the concept of object recognition aid. *Pers. Ubiquitous Comput.*, 19(5):955–965, 2015.

[108] Y.H. Lee and G. Medioni. Wearable RGBD indoor navigation system for the blind. In *Comput. Vision - ECCV 2014 Workshops*. Springer, 2015.

[109] Cha Zhang and Paul Viola. Multiple-instance pruning for learning efficient cascade detectors. In *Advances in Neural Inform. Process. Syst.*, 2008.

[110] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *IEEE CVPR*, 2010.

[111] Billy Li, Ajmal Mian, Wanquan Liu, and Aneesh Krishna. Face recognition based on Kinect. *Pattern Anal. Appl.*, pages 1–11, 2015.

[112] Munawar Hayat, Mohammed Bennamoun, and Amar El-Sallam. An RGB–D based image set classification for robust face recognition from Kinect data. *Neurocomputing*, 171:889–900, 2016.

[113] S. Elaiwat, M. Bennamoun, F. Boussaid, and A. El-Sallam. A Curvelet-based approach for textured 3D face recognition. *Pattern Recognit.*, 48(4):1235–1246, 2015.

[114] João Cardia Neto and Aparecido Marana. 3DLBP and HAOG fusion for face recog-

nition utilizing Kinect as a 3D scanner. In *ACM SIGAPP*, 2015.

[115] Stefano Berretti, Naoufel Werghi, Alberto del Bimbo, and Pietro Pala. Selecting stable keypoints and local descriptors for person identification using 3D face scans. *Vis. Comput.*, 30(11):1275–1292, 2014.

[116] Gee Hsu, Yu Liu, Hsiao Peng, and Po Wu. RGB-D-Based Face Reconstruction and Recognition. *IEEE Trans. Inf. Forensics Secur.*, 9(12):2110–2118, 2014.

[117] Cesare Ciaccio, Lingyun Wen, and Guodong Guo. Face recognition robust to head pose changes based on the RGB-D sensor. In *IEEE ICB*, 2013.

[118] B. Li, A. Mian, Wanquan Liu, and A. Krishna. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In *WACV*. IEEE, 2013.

[119] G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh. On RGB-D face recognition using Kinect. In *IEEE ICB*, 2013.

[120] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE CVPR*, 2005.

[121] J.M.a Loomis, R.G.b Golledge, and R.L.c Klatzky. Navigation system for the blind: Auditory display modes and guidance. *Presence Teleoperators Virtual Environ.*, 7(2):193–203, 1998.

[122] B.F.G.a Katz, S.b Kammoun, G.a Parseihian, O.b Gutierrez, A.b Brilhault, M.a Auvray, P.b Truillet, M.a Denis, S.c Thorpe, and C.b Jouffrais. NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Real.*, 16(4):253–269, 2012.

[123] Stuart Goose and Carsten Möller. A 3D Audio Only Interactive Web Browser: Using Spatialization to Convey Hypermedia Document Structure. In *ACM MM*, 1999.

[124] Roberta Klatzky, James Marston, Nicholas Giudice, Reginald Golledge, and Jack Loomis. Cognitive load of navigating without vision when guided by virtual sound versus spatial language. *J. Exp. Psychol. Appl.*, 12(4):223, 2006.

[125] Christopher Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[126] Oscar Déniz, Gloria Bueno, Jesús Salido, and Fernando De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognit. Lett.*, 32(12):1598–1603, 2011.

[127] W.R. Schwartz, Huimin Guo, Jonghyun Choi, and L.S. Davis. Face Identification Using Large Feature Sets. *IEEE Trans. Image Process.*, 21(4):2245–2255, 2012.

[128] Bosun Xie. *Head-related transfer function and virtual auditory display*. J Ross,

2013.

[129] Felipe Grijalva, Luca Martini, Siome Goldenstein, and Dinei Florencio. Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane. In *IEEE ICASSP*, pages 4473–4477, 2014.

[130] K. Sunder, Jianjun He, Ee Tan, and Woon-Seng Gan. Natural Sound Rendering for Headphones: Integration of signal processing techniques. *IEEE Signal Process. Mag.*, 32(2):100–113, 2015.

[131] F. Grijalva, L. Martini, D. Florencio, and S. Goldenstein. A Manifold Learning Approach for Personalizing HRTFs from Anthropometric Features. *IEEE ACM Trans. Audio Speech Lang. Process.*, 24(3):559–570, 2016.

[132] V. Algazi, Richard Duda, Dennis Thompson, and Carlos Avendano. The cipic hrtf database. In *IEEE WASPAA*, 2001.

[133] A. Pinto, W. Robson Schwartz, H. Pedrini, and A. Rocha. Using Visual Rhythms for Detecting Video-Based Facial Spoof Attacks. *IEEE Trans. Inf. Forensics Secur.*, 10(5):1025–1038, 2015.

[134] P.N. Belhumeur, J.P. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):711–720, 1997.

[135] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *BMVC*, 1(3):6, 2015.

[136] Patrick Bridge and Shlomo Sawilowsky. Increasing Physicians' Awareness of the Impact of Statistics on Research Outcomes: Comparative Power of the t-test and Wilcoxon Rank-Sum Test in Small Samples Applied Research. *J. Clin. Epidemiol.*, 52(3):229 – 235, 1999.

[137] VanessaReginaMargarethLima Maike, Laurindo Britto Neto, SiomeKlein Goldenstein, and MariaCecíliaCalani Baranauskas. Heuristics for NUI Revisited and Put into Practice. In *Human-Computer Interaction: Interaction Technologies*. Springer Int. Publishing, 2015.

[138] Margaret Bradley and Peter Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry*, 25(1):49–59, 1994.

[139] L.S. Britto Neto, V.R.M.L. Maike, F.L. Koch, M Baranauskas, A.R. Rocha, and S.K. Goldenstein. A Wearable Face Recognition System Built into a Smartwatch and the Visually Impaired User. In *ICEIS*, 2015.

[140] L.S. Britto Neto, V.R.M.L. Maike, F.L. Koch, M. Baranauskas, A.R. Rocha, and S.K. Goldenstein. A Wearable Face Recognition System Built into a Smartwatch

and the Blind and Low Vision Users. In *Enterprise Information Systems*, number 241, pages 515–528. Springer Int. Publishing, 2015.

[141] Hsueh Wang, Yafim Landa, Maurice Fallon, and Seth Teller. Spatially Prioritized and Persistent Text Detection and Decoding. In *Camera-Based Document Analysis and Recognition*, pages 3–17. Springer Int. Publishing, 2013.

[142] Seung Jung, Tae Kim, and Sung Ko. A novel multiple image deblurring technique using fuzzy projection onto convex sets. *IEEE Signal Process. Lett.*, 3(16):192–195, 2009.

[143] S. Kumar, H. Azartash, M. Biswas, and T. Nguyen. Real-Time Affine Global Motion Estimation Using Phase Correlation and its Application for Digital Image Stabilization. *IEEE Trans. Image Process.*, 20(12):3406–3418, 2011.

[144] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *IEEE ICB*, 2016.

[145] Durand R. Begault. *3D Sound for Virtual Reality and Multimedia*. AP Professional, Cambridge, 1994.

[146] Elizabeth M. Wenzel, Marianne Arruda, Doris J Kistler, and Frederic L Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Amer.*, 94(1):111–123, July 1993.

[147] Vesa Valimaki, Andreas Franck, Jussi Ramo, Hannes Gamper, and Lauri Savioja. Assisted Listening Using a Headset: Enhancing audio perception in real, augmented, and virtual environments. *IEEE Signal Process. Mag.*, 32(2):92–99, March 2015.

[148] Henrik Möller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3-4):171–218, 1992.

[149] BoSun Xie. Recovery of individual head-related transfer functions from a small set of measurements. *J. Acoust. Soc. Amer.*, 132(1):282–94, July 2012.

[150] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, 49(6):472–479, June 2001.

[151] V. Ralph Algazi, Richard O. Duda, Ramani Duraiswami, Nail A. Gumerov, and Zhihui Tang. Approximating the head-related transfer function using simple geometric models of the head and torso. *J. Acoust. Soc. Amer.*, 112(5):2053, October 2002.

[152] C.P. Brown and R.O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Process.*, 6(5):476–488, 1998.

[153] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In

2013 18th Int. Conf. Digital Signal Process., pages 1–8, 2013.

[154] Makoto Otani and Shiro Ise. Fast calculation system specialized for head-related transfer function based on boundary element method. *J. Acoust. Soc. Amer.*, 119(5):2589, May 2006.

[155] Yuvi Kahana and Philip Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *J. Sound and Vibration*, 300(3):552–579, 2007.

[156] Hironori Takemoto, Parham Mokhtari, Hiroaki Kato, Ryouichi Nishimura, and Kazuhiro Iida. Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *J. Acoust. Soc. Amer.*, 132(6):3832–3841, December 2012.

[157] Sungmok Hwang, Youngjin Park, and YounSik Park. Modeling and Customization of Head-Related Impulse Responses Based on General Basis Functions in Time Domain. *Acta Acustica united with Acustica*, 94(6):965–980, November 2008.

[158] Kimberly J. Fink and Laura Ray. Individualization of head related transfer functions using principal component analysis. *Applied Acoustics*, 87:162–173, January 2015.

[159] Kaushik Sunder, Ee-Leng Tan, and Woon-Seng Gan. Individualization of Binaural Synthesis Using Frontal Projection Headphones. *J. Audio Eng. Soc.*, 61(12):989–1000, December 2013.

[160] Felipe Grijalva, Luiz Martini, Siome Goldenstein, and Dinei Florencio. Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane. In *Internacional Conference on Acoustics, Speech, and Signal Processing*, pages 4473–4477, Florence, Italy, 2014. IEEE.

[161] Griffin D Romigh and Brian D Simpson. Do you hear where I hear?: isolating the individualized sound localization cues. *Frontiers in Neuroscience*, 8(370), December 2014.

[162] D.N. Zotkin, R. Duraiswami, and L.S. Davis. Rendering Localized Spatial Audio in a Virtual Auditory Space. *IEEE Trans. Multimedia*, 6(4):553–564, August 2004.

[163] Edgar Torres, Felipe Orduña, and Fernando Arámbula. Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database. *Applied Acoustics*, 97:84–95, April 2015.

[164] Kazuhiro Iida, Yohji Ishii, and Shinsuke Nishioka. Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener's pinnae. *J. Acoust. Soc. Amer.*, 136(1):317–33, July 2014.

[165] John C Middlebrooks. Individual differences in external-ear transfer functions re-

duced by scaling in frequency. *J. Acoust. Soc. Amer.*, 106(3):1480–1492, 1999.

[166] D Kistler and F Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *J. Acoust. Soc. Amer.*, 91(3):1637–47, March 1992.

[167] Q.H. Huang and Q.L. Zhuang. HRIR personalisation using support vector regression in independent feature space. *Electronics Letters*, 45(19):1002, 2009.

[168] C. Jin, P. Leong, J. Leung, A. Corderoy, and S. Carlile. Enabling individualized virtual auditory space using morphological measurements. In *First Pacific-Rim Conf. on Multimedia (2000 Int. Symposium on Multimedia Information Process.)*, pages 235–238. IEEE, 2000.

[169] Takanori Nishino, Kazuhiro Iida, Naoya Inoue, Kazuya Takeda, and Fumitada Itakura. Estimation of HRTFs on the horizontal plane using physical features. *Applied Acoustics*, 68(8):897–908, 2007.

[170] Song Xu, Zhizhong Li, and Gavriel Salvendy. Improved method to individualize head-related transfer function using anthropometric measurements. *Acoustical Science and Technology*, 29(6):388–390, 2008.

[171] David Schönstein and B Katz. HRTF selection for binaural synthesis from a database using morphological parameters. In *Proc. 20th Intl. Congr. Acoust.(ICA)*, pages 1–6, Aug 2010.

[172] Hongmei Hu, Lin Zhou, Hao Ma, and Zhenyang Wu. HRTF personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics*, 69(2):163–172, February 2008.

[173] Q Huang and Y Fang. Modeling personalized head-related impulse response using support vector regression. *Journal of Shanghai University*, 13(6):428–432, 2009.

[174] Zhixin Wang and Cheung Fat Chan. HRIR customization using common factor decomposition and joint support vector regression. In *European Signal Process. Conf.*, pages 1–5. IEEE, 2013.

[175] Lin Li and Qinghua Huang. HRTF Personalization Modeling based on RBF Neural Network. In *Internacional Conference on Acoustics, Speech, and Signal Processing*, pages 3707–3710. IEEE, 2013.

[176] Graham Grindlay and M. Alex O. Vasilescu. A Multilinear (Tensor) Framework for HRTF Analysis and Synthesis. In *Internacional Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 161–164. IEEE, 2007.

[177] Qinghua Huang and Lin Li. Modeling individual HRTF tensor using high-order partial least squares. *EURASIP Journal on Advances in Signal Processing*, 2014(1):58,

2014.

[178] Piotr Bilinski, Jens Ahrens, Mark Thomas, Ivan Tashev, and John Platt. HRTF magnitude synthesis via sparse representation of anthropometric features. In *Internacional Conference on Acoustics, Speech, and Signal Processing*, pages 4468–4472. IEEE, 2014.

[179] R. Duraiswami and VC Raykar. The Manifolds of Spatial Hearing. In *Internacional Conference on Acoustics, Speech, and Signal Processing*, pages 285–288. IEEE, 2005.

[180] S Roweis and L Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, December 2000.

[181] J Tenenbaum, V de Silva, and J Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–23, December 2000.

[182] Bill Kapralos and Nathan Mekuz. Application of dimensionality reduction techniques to HRTFs for interactive virtual environments. In *Int. Conf. Advances Comput. Entertainment Technol.*, pages 256–257. ACM, 2007.

[183] Bill Kapralos, Nathan Mekuz, Agnieszka Kopinska, and Saad Khattak. Dimensionality reduced HRTFs: a comparative study. In *Int. Conf. Advances Comput. Entertainment Technol.*, page 59. ACM, December 2008.

[184] HS Seung and DD Lee. The manifold ways of perception. *Science*, 290:2268–2269, 2000.

[185] A. Kulkarni, S.K. Isabelle, and H.S. Colburn. On the minimum-phase approximation of head-related transfer functions. In *Workshop on Applications in Signal Processing to Audio and Acoustics*, pages 84–87. IEEE, 1995.

[186] Kanji Watanabe, Kenji Ozawa, Yukio Iwaya, Yo Iti Suzuki, and Kenji Aso. Estimation of interaural level difference based on anthropometry and its effect on sound localization. *J. Acoust. Soc. Amer.*, 122(5):2832–41, November 2007.

[187] V Algazi, R Duda, D Thompson, and C Avendano. The CIPIC HRTF database. In *Workshop on Applications in Signal Processing to Audio and Acoustics*, pages 99–102. IEEE, 2001.

[188] Laurens Van Der Maaten, Eric Postma, and Jaap Van Den Herik. Dimensionality reduction: A comparative review. *J. Machine Learning Research*, 10:1–41, 2009.

[189] W Michael Brown, Shawn Martin, Sara N Pollock, Evangelos A Coutsias, and Jean-Paul Watson. Algorithmic dimensionality reduction for molecular structure analysis. *J. Chemical Physics*, 129(6), August 2008.

[190] K Saul Lawrence and T Roweis Sam. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. *J. Machine Learning Research*, 4:119–155, 2002.

[191] BoSun Xie, XiaoLi Zhong, Dan Rao, and ZhiQiang Liang. Head-related transfer function database and its analyses. *Science in China Series G: Physics, Mechanics and Astronomy*, 50(3):267–280, June 2007.

[192] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*. NIPS Foundation, 2005.

[193] Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(5):796–809, 2008.

[194] M. Zhang, R Kennedy, T Abhayapala, and W. Zhang. Statistical method to identify key anthropometric parameters in hrtf individualization. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, pages 213–218. IEEE, 2011.

[195] Jeroen Breebaart. Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis. *J. Acoust. Soc. Am.*, 133(1):EL1–EL6, January 2013.

[196] Yoshua Bengio, Jean François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems*. NIPS Foundation, 2004.

[197] John Platt. Fastmap, metricmap, and landmark mds are all nystrom algorithms. In *Int. Workshop on Artificial Intelligence and Statistics*, pages 261–268, 2005.

[198] Bosun Xie and Tingting Zhang. The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency. *Acta Acustica united with Acustica*, 96(2):328–339, March 2010.

[199] Heeyoul Choi, Seungjin Choi, Anup Katake, Yoonseop Kang, and Yoonsuck Choe. Manifold alpha-integration. In *PRICAI 2010: Trends in Artificial Intelligence*, pages 397–408. Springer, 2010.

[200] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[201] Ramani Duraiswami and Dmitry Zotkin. Efficient physics based simulation of spatial audio for virtual and augmented reality. *The Journal of the Acoustical Society of America*, 140(4):2999–3000, 2016.

[202] L. Neto, F. Grijalva, V. Maike, L. Martini, D. Florencio, M. Baranauskas, A. Rocha, and S. Goldenstein. A Kinect-Based Wearable Face Recognition System to Aid Visually Impaired Users. *IEEE Trans. Hum.-Mach. Syst.*, 2016. to be published.

[203] Flemming Christensen, Henrik Møller, Pauli Minnaar, Jan Polgsties, and S. Olesen. Interpolating Between Head-Related Transfer Functions Measured with Low-Directional Resolution. In *Audio Eng. Soc. Conv.* AES, 1999.

[204] Frederic Wightman, Doris Kistler, and Marianne Arruda. Perceptual consequences of engineering compromises in synthesis of virtual auditory objects. *J. Acoust. Soc. Am.*, 92(4):2332–2332, 1992.

[205] T. Nishino, S. Kajita, K. Takeda, and F. Itakura. Interpolating head related transfer functions in the median plane. In *Workshop on Applicat. of Signal Process. to Audio and Acoust.*, pages 167–170. IEEE, 1999.

[206] Corey Cheng and Gregory Wakefield. Spatial frequency response surfaces: An alternative visualization tool for head-related transfer functions (HRTFs). In *Int. Conf. Acoust., Speech, Signal Process.*, pages 961–964. IEEE, 1999.

[207] Kanji Watanabe, Shouichi Takane, and Yôiti Suzuki. Interpolation of head-related transfer functions based on the common-acoustical-pole and residue model. *Acoust. Sci. Technol.*, 24(5):335–337, 2003.

[208] Fábio Freeland, Luiz Biscainho, and Paulo Diniz. Interpositional transfer function for 3D-sound generation. *J. Audio Eng. Soc.*, 52(9):915–930, 2004.

[209] Fakheredine Keyrouz and Klaus Diepold. A new HRTF interpolation approach for fast synthesis of dynamic environmental interaction. *J. Audio Eng. Soc.*, 56(1/2):28–35, 2008.

[210] Ryouichi Nishimura, Hiroaki Kato, and Naomi Inoue. Interpolation of head-related transfer functions by spatial linear prediction. In *Int. Conf. Acoust., Speech, Signal Process.*, pages 1901–1904. IEEE, 2009.

[211] D. Breebaart, Fabian Nater, and A. Kohlrausch. Parametric binaural synthesis: Background, applications and standards. In *Proc. NAG-DAGA*, pages 172–175, 2009.

[212] Marcelo Queiroz and Gustavo de Sousa. Efficient binaural rendering of moving sound sources using hrtf interpolation. *J. New Music Res.*, 40(3):239–252, 2011.

[213] German Ramos and Maximo Cobos. Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications. *J. Acoust. Soc. Am.*, 134(3):1735–1738, 2013.

[214] T. Ajdler, Luciano Sbaiz, and Martin Vetterli. The Plenacoustic Function and Its Sampling. *IEEE Trans. Signal Process.*, 54(10):3790–3804, 2006.

[215] Thibaut Ajdler, Christof Faller, Luciano Sbaiz, and Martin Vetterli. Sound field analysis along a circle and its applications to HRTF interpolation. *J. Audio Eng.*

Soc., 56(3):156–175, 2008.

[216] Michael Evans, James Angus, and Anthony Tew. Analyzing head-related transfer function measurements using surface spherical harmonics. *J. Acoust. Soc. Am.*, 104(4):2400–2411, 1998.

[217] R. Duraiswaini, Dmitry Zotkin, and Nail Gumerov. Interpolation and range extrapolation of hrtfs [head related transfer functions]. In *Int. Conf. Acoust., Speech, Signal Process.*, volume 4, pages 45–48. IEEE, 2004.

[218] Russell Martin and Ken McAnally. Interpolation of head-related transfer functions. Technical Report DSTO-RR-0323, Air Operations Division Defence Science and Technology Org., 2007.

[219] W. Zhang, R. Kennedy, and T. Abhayapala. Efficient Continuous HRTF Model Using Data Independent Basis Functions: Experimentally Guided Approach. *IEEE Trans. Audio Speech Lang. Process.*, 17(4):819–829, 2009.

[220] Wen Zhang, Thushara Abhayapala, Rodney Kennedy, and Ramani Duraiswami. Insights into head-related transfer function: Spatial dimensionality and continuous representation. *J. Acoust. Soc. Am.*, 127(4):2347–2357, 2010.

[221] Wen Zhang, Mengqiu Zhang, Rodney A. Kennedy, and Thushara D. Abhayapala. On high-resolution head-related transfer function measurements: An efficient sampling scheme. *Audio Speech Lang. Process. IEEE Trans. On*, 20(2):575–584, 2012.

[222] Matthieu Aussal, François Alouges, and Brian Katz. A study of spherical harmonics interpolation for hrtf exchange. In *Proc. Meetings on Acoust.*, page 050010. ASA, 2013.

[223] V. Lemaire, F. Clerot, S. Busson, R. Nicol, and V. Choqueuse. Individualized HRTFs from few measurements: A statistical learning approach. In *Proc. Int. Joint Conf. Neural Networks*, pages 2041–2046. IEEE, 2005.

[224] Pierre Guillon, Rozenn Nicol, and Laurent Simon. Head-related transfer functions reconstruction from sparse measurements considering a priori knowledge from database analysis: A pattern recognition approach. In *Audio Eng. Soc. Conv.* AES, 2008.

[225] Bo-Sun Xie. Recovery of individual head-related transfer functions from a small set of measurementsa). *J. Acoust. Soc. Am.*, 132(1):282–294, 2012.

[226] S. Carlile, C. Jin, and V. Van Raad. Continuous virtual auditory space using HRTF interpolation: Acoustic and psychophysical errors. In *1st Pacific-Rim Conf. on Multimedia*, pages 220–223. IEEE, 2000.

[227] J. Chen, B. Veen, and K. Hecox. Synthesis of 3D virtual auditory space via a spatial

feature extraction and regularization model. In *Virtual Reality Annual Int. Symp.*, pages 188–193. IEEE, 1993.

[228] F. Keyrouz and K. Diepold. A Rational Hrtf Interpolation Approach for Fast Synthesis of Moving Sound. In *IEEE 12th Digital Signal Process. Workshop 4th Signal Process. Educ. Workshop*, pages 222–226, 2006.

[229] Jiashu Chen, Barry D. Van Veen, and K. E. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *J. Acoust. Soc. Am.*, 97:439 – 452, 1995.

[230] Lin Wang, Fuliang Yin, and Zhe Chen. Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights. *Acoust. Sci. Technol.*, 30(6):395–403, 2009.

[231] Ramani Duraiswami and Vikas Raykar. The manifolds of spatial hearing. In *Int. Conf. Acoust., Speech, Signal Process.*, volume 3, pages 285–288. IEEE, 2005.

[232] Joshua Tenenbaum, Vin De Silva, and John Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[233] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: A comparative. *J. Mach. Learning Research*, 10:66–71, 2009.

[234] Warren Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4):401–419, 1952.

[235] Lawrence Saul and Sam Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *J. Mach. Learning Research*, 4:119–155, 2003.

[236] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 777–784, 2004.

[237] Sam Roweis and Lawrence Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[238] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14, pages 585–591, 2001.

[239] John C. Middlebrooks. Individual differences in external-ear transfer functions reduced by scaling in frequency. *J. Acoust. Soc. Am.*, 106(3):1480–1492, 1999.

[240] Ik Lim, Pablo de Heras Ciechomski, Sofiane Sarni, and Daniel Thalmann. Planar arrangement of high-dimensional biomedical data sets by isomap coordinates. In *IEEE Symp. Comput.-Based Medical Syst.*, pages 50–55, 2003.

[241] Odest Jenkins and Maja Mataric. Deriving action and behavior primitives from

human motion data. In *Int. Conf. Intelligent Robots and Syst.*, pages 2551–2556. IEEE, 2002.

[242] G. Romigh, D. Brungart, R. Stern, and B. Simpson. Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions. *IEEE J. Sel. Top. Signal Process.*, 9(5):921–930, 2015.

[243] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems*, volume 16, pages 177–184. NIPS Foundation, 2004.

[244] W. Zhixin and C. Cheung-Fat. Continuous Function Modeling of Head-Related Impulse Response. *IEEE Signal Process. Lett.*, 22(3):283–287, March 2015.

[245] W. Michael Brown, Shawn Martin, Sara N. Pollock, Evangelos A. Coutsias, and Jean-Paul Watson. Algorithmic dimensionality reduction for molecular structure analysis. *J. Chem. Phys.*, 129(6):064118, 2008.

[246] Felipe Grijalva, Luiz Martini, Siome Goldenstein, and Dinei Florencio. Anthropometric-based customization of head-related transfer functions using Isomap in the horizontal plane. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 4473–4477, 2014.

[247] Elizabeth M. Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. Localization using nonindividualized head-related transfer functions. *J. Acoust. Soc. Am.*, 94(1):111–123, 1993.

[248] Martin Pollow, Bruno Masiero, Pascal Dietrich, Janina Fels, and Michael Vorländer. Fast measurement system for spatially continuous individual HRTFs. In *Audio Engineering Society Conference: UK 25th Conference: Spatial Audio in Today's 3D World*. Audio Engineering Society, 2012.

[249] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. Mixed structural modeling of head-related transfer functions for customized binaural audio delivery. In *Digital Signal Processing (DSP), 2013 18th International Conference on*, pages 1–8. IEEE, 2013.

[250] Makoto Otani and Shiro Ise. Fast calculation system specialized for head-related transfer function based on boundary element method. *J. Acoust. Soc. Am.*, 119(5):2589–2598, 2006.

[251] Kimberly J. Fink and Laura Ray. Individualization of head related transfer functions using principal component analysis. *Appl. Acoust.*, 87:162–173, 2015.

[252] Kazuhiko Yamamoto and Takeo Igarashi. Fully Perceptual-Based 3D Spatial Sound

Individualization with an Adaptive Variational AutoEncoder. *ACM Trans Graph*, 2017.

[253] Yuancheng Luo, Dmitry N. Zotkin, and Ramani Duraiswami. Virtual autoencoder based recommendation system for individualizing head-related transfer functions. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

[254] Brian FG Katz and Gaëtan Parseihian. Perceptually based head-related transfer function database optimization. *J. Acoust. Soc. Am.*, 131(2):EL99–EL105, 2012.

[255] Robert Baumgartner, Piotr Majdak, and Bernhard Laback. Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2):791–802, August 2014.

[256] F. Grijalva, L. C. Martini, D. Florencio, and S. Goldenstein. Interpolation of Head-Related Transfer Functions Using Manifold Learning. *IEEE Signal Process. Lett.*, 24(2):221–225, February 2017.

[257] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to Recommender Systems Handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, Boston, MA, 2011.

[258] John C. Middlebrooks. Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3):1493–1510, August 1999.

[259] P. L. Søndergaard and P. Majdak. The Auditory Modeling Toolbox. In *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pages 33–56. Springer, Berlin, Heidelberg, 2013.

[260] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[261] Paulo Chiliguano and Gyorgy Fazekas. Hybrid music recommender using content-based and social information. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2618–2622. IEEE, 2016.

[262] Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *Advances in Neural Information Processing Systems*, volume 16, pages 177–184. NIPS Foundation, 2004.

[263] John Platt. FastMap, MetricMap, and Landmark MDS are all Nystrom Algorithms. In *AISTATS*, 2005.

[264] John C. Middlebrooks. Individual differences in external-ear transfer functions re-

duced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3):1480–1492, 1999.

[265] Ki-Seung Lee and Seok-Pil Lee. A relevant distance criterion for interpolation of head-related transfer functions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1780–1790, 2011.

[266] Bosun Xie and Tingting Zhang. The Audibility of Spectral Detail of Head-Related Transfer Functions at High Frequency. *Acta Acustica united with Acustica*, 96(2):328–339, March 2010.

[267] O. Warusfel. LISTEN HRTF DATABASE. http://recherche.ircam.fr/, 2017.

[268] Kanji Watanabe, Yukio Iwaya, Yôiti Suzuki, Shouichi Takane, and Sojun Sato. Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoustical science and technology*, 35(3):159–165, 2014.

[269] Ramona Bomhardt, Matias de la Fuente Klein, and Janina Fels. A high-resolution head-related transfer function and three-dimensional ear model database. In *Proceedings of Meetings on Acoustics 172ASA*, volume 29, page 050002. ASA, 2016.

[270] M. Zhang, R. A. Kennedy, T. D. Abhayapala, and Wen Zhang. Statistical method to identify key anthropometric parameters in HRTF individualization. In *2011 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pages 213–218. IEEE, 2011.

[271] Rahulram Sridhar, Joseph G. Tylka, and Edgar Choueiri. A Database of Head-Related Transfer Functions and Morphological Measurements. In *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017.

[272] Bosun Xie, Xiaoli Zhong, and Nana He. Typical data and cluster analysis on head-related transfer functions from Chinese subjects. *Applied Acoustics*, 94:1–13, 2015.

Appendix A

Permission Grants

DECLARAÇÃO

As cópias dos documentos da minha autoria ou coautoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Tese de Doutorado intitulada “Manifold Learning for Spatial Audio Rendering (Aprendizado de Variedades para a Síntese de Áudio Espacial)” não infringem os dispositivos da Lei n. 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 04 de abril de 2018

Felipe Leonel Grijalva Arévalo
RNE: V8213918

Luiz César Martini
RG: 5446578-3

Appendix B

Institutional Review Board approval

Contribution IV of this work was part of two projects approved by Unicamp Institutional Review Board *CAAE 15641313.7.0000.5404* and *CAAE 31818014.0.0000.5404*. Next, we include the IRB's approvals to conduct the experiments with human subjects described in Chapter 3.

FACULDADE DE CIENCIAS MEDICAS - UNICAMP (CAMPUS CAMPINAS)

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Criação de uma Base de Vídeos para o Desenvolvimento de métodos de Reconhecimento Facial Utilizando Câmeras de Profundidade

Pesquisador: Laurindo de Sousa Britto Neto

Área Temática:

Versão: 3

CAAE: 15641313.7.0000.5404

Instituição Proponente: Instituto de Computação

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 368.619

Data da Relatoria: 19/08/2013

Apresentação do Projeto:

Trata-se de um projeto de pesquisa que visa a construção de uma base de vídeos coloridos de profundidade para uso científico, com o registro de uma quantidade significativa de pessoas adultas (n=30). Estas pessoas deverão exibir suas faces em diferentes condições de movimentação e iluminação, além de certa variação dos cenários de fundo.

Objetivo da Pesquisa:

Criar uma base de vídeos para identificação e reconhecimento de faces, que contenha informações tipicamente registradas por câmeras de profundidade.

Avaliação dos Riscos e Benefícios:

Não há riscos previsíveis, tampouco benefícios diretos aos sujeitos da pesquisa. Entretanto, os resultados do estudo poderão gerar importantes benefícios sociais e acadêmicos para a instituição e para a área de conhecimento.

Comentários e Considerações sobre a Pesquisa:

Projeto de pesquisa bem apresentado, em todos os seus itens. Em resposta ao questionamento emitido em parecer anterior, os autores esclareceram sobre a forma de recrutamento dos voluntários e esclareceram que estes não deverão pertencer, obrigatoriamente, a qualquer grupo vulnerável, incluindo alunos e funcionários da instituição proponente.

Endereço: Rua Tessália Vieira de Camargo, 126
Bairro: Barão Geraldo **CEP:** 13.083-887
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-8936 **Fax:** (19)3521-7187 **E-mail:** cep@fcm.unicamp.br

FACULDADE DE CIENCIAS MEDICAS - UNICAMP (CAMPUS CAMPINAS)

Continuação do Parecer: 368.619

Considerações sobre os Termos de apresentação obrigatória:

Foram apresentados: folha de rosto, projeto de pesquisa original, formulário gerado pela Plataforma Brasil e TCLE. Todos estes documentos estão de acordo com as regras do sistema CEP/CONEP e atendem aos preceitos da Resolução 466/2012-CNS,MS.

Recomendações:

Não há.

Conclusões ou Pendências e Lista de Inadequações:

Projeto aprovado.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

Cabe ao pesquisador desenvolver o projeto conforme delineado, elaborar e apresentar os relatórios parciais e final, bem como encaminhar os resultados para publicação, com os devidos créditos aos pesquisadores associados e ao pessoal técnico participante do projeto (Resolução 466/2012 CNS/MS).

CAMPINAS, 22 de Agosto de 2013

Assinador por:
Fátima Aparecida Bottcher Luiz
(Coordenador)

Endereço: Rua Tessália Vieira de Camargo, 126
Bairro: Barão Geraldo **CEP:** 13.083-887
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-8936 **Fax:** (19)3521-7187 **E-mail:** cep@fcm.unicamp.br

FACULDADE DE CIENCIAS MEDICAS - UNICAMP (CAMPUS CAMPINAS)

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Investigando Natural User Interfaces (NUIs) - Tecnologias e Interação em Contexto de Acessibilidade

Pesquisador: Vanessa Regina Margareth Lima Maike

Área Temática:

Versão: 1

CAAE: 31818014.0.0000.5404

Instituição Proponente: Instituto de Computação

Patrocinador Principal: Universidade Estadual de Campinas - UNICAMP

DADOS DO PARECER

Número do Parecer: 709.828

Data da Relatoria: 24/06/2014

Apresentação do Projeto:

Trata-se de um projeto de pesquisa, no âmbito de uma tese de doutorado, que visa estudar, analisar, implementar e avaliar tecnologias que utilizam interfaces naturais (NUIs) para aumentar a acessibilidade. A pesquisadora pretende implementar tecnologias assistivas (como um capacete usando a tecnologia Kinect) para aumentar a interatividade de portadores de doença visual. Os testes serão realizados no Instituto de Computação (IC) da Unicamp. Serão utilizados 30 voluntários de pesquisa (20 usuários sem deficiência e 10 usuários com deficiência). O projeto conta com financiamento através de uma bolsa de doutorado da CAPES, e tem início previsto de avaliação das tecnologias desenvolvidas em 01/2015.

Objetivo da Pesquisa:

Propor melhores maneiras de construir tecnologias com interfaces naturais – Natural User Interfaces (NUIs) – no contexto de acessibilidade. Como sub-objetivos estão o design e a avaliação, dentro do contexto de acessibilidade, de novas tecnologias que também pertençam ao paradigma NUI.

Avaliação dos Riscos e Benefícios:

A pesquisadora descreve que não há riscos previsíveis durante os testes com os usuários. Quanto aos benefícios diretos aos voluntários, a pesquisa prevê que os mesmos possam aprender a

Endereço: Rua Tessália Vieira de Camargo, 126
Bairro: Barão Geraldo **CEP:** 13.083-887
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-8936 **Fax:** (19)3521-7187 **E-mail:** cep@fcm.unicamp.br

FACULDADE DE CIENCIAS MEDICAS - UNICAMP (CAMPUS CAMPINAS)

utilizar tecnologias digitais que não fazem parte do seu cotidiano.

Comentários e Considerações sobre a Pesquisa:

Considero a pesquisa interessante e valida para aplicação em seres humanos. Os resultados podem ter uma alta relevância na área de tecnologias assistivas. O projeto também é claro e sucinto.

Considerações sobre os Termos de apresentação obrigatória:

Foram apresentados: 1) projeto de pesquisa detalhado; 2) folha de rosto, devidamente preenchida, datada e assinada pelo diretor associado do Instituto de Comunicação/UNICAMP, (instituição proponente); 3) termo de consentimento livre e esclarecido (TCLE), no modelo proposto pelo CEP/UNICAMP, de acordo com a Res. CNS-MS 466/12.

Recomendações:

Conclusões ou Pendências e Lista de Inadequações:

Documentos de acordo com a Res. CNS-MS 466/12. Os objetivos e a metodologia com os voluntários é clara. Considero o projeto aprovado.

Projeto aprovado em reunião do colegiado, em 24-06-2014.

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

- Se o TCLE tiver mais de uma página, o sujeito de pesquisa ou seu representante, quando for o caso, e o pesquisador responsável deverão rubricar todas as folhas desse documento, apondo suas assinaturas na última página do referido termo (Carta Circular nº 003/2011/CONEP/CNS).

- Cabe ao pesquisador desenvolver a pesquisa conforme delineada no protocolo aprovado, elaborar e apresentar os relatórios parciais e final, bem como encaminhar os resultados para publicação com os devidos créditos aos pesquisadores associados e ao pessoal técnico participante do projeto (Resolução 466/2012 CNS/MS). Os relatórios deverão ser enviados através da Plataforma Brasil- ícone Notificação.

- Eventuais modificações ou emendas ao protocolo deverão ser apresentadas ao CEP de forma

Endereço: Rua Tessália Vieira de Camargo, 126
Bairro: Barão Geraldo **CEP:** 13.083-887
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-8936 **Fax:** (19)3521-7187 **E-mail:** cep@fcm.unicamp.br

FACULDADE DE CIENCIAS MEDICAS - UNICAMP (CAMPUS CAMPINAS)

Continuação do Parecer: 709.828

clara e sucinta, identificando a parte do protocolo a ser modificada (com destaque) e suas justificativas. As modificações deverão ter parecer de aprovação deste CEP antes de serem implementadas.

CAMPINAS, 05 de Julho de 2014

Assinado por:
Fátima Aparecida Bottcher Luiz
(Coordenador)

Endereço: Rua Tessália Vieira de Camargo, 126
Bairro: Barão Geraldo **CEP:** 13.083-887
UF: SP **Município:** CAMPINAS
Telefone: (19)3521-8936 **Fax:** (19)3521-7187 **E-mail:** cep@fcm.unicamp.br