Washington University in St. Louis
Washington University Open Scholarship

Engineering and Applied Science Theses & Dissertations

McKelvey School of Engineering

Summer 8-15-2017

Improving Pure-Tone Audiometry Using Probabilistic Machine Learning Classification

Xinyu Song
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds

Part of the Biomedical Engineering and Bioengineering Commons, Computer Sciences Commons, and the Psychology Commons

Recommended Citation

Song, Xinyu, "Improving Pure-Tone Audiometry Using Probabilistic Machine Learning Classification" (2017). *Engineering and Applied Science Theses & Dissertations*. 310.
https://openscholarship.wustl.edu/eng_etds/310

This Dissertation is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in Engineering and Applied Science Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Sciences
Department of Biomedical Engineering

Dissertation Examination Committee:
Dennis L. Barbour, Chair
Steven E. Petersen
Baranidharan Raman
Mitchell S. Sommers
Kurt A. Thoroughman

Improving Pure-Tone Audiometry Using Probabilistic Machine Learning Classification

by
Xinyu Song

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2017
St. Louis, Missouri

© 2017, Xinyu David Song

Table of Contents

List of Figures .... **v**
List of Tables .... **vii**
List of Abbreviations .... **viii**
Acknowledgments .... **ix**
Abstract of the Dissertation .... **xiv**
Chapter 1: Introduction **1**
1.1. Hearing Loss 1
1.2. Methods for Hearing Evaluation 3
1.2.1. Pure-Tone Audiometry 3
1.2.2. Alternatives to Manual Pure-Tone Audiometry 8
1.3. Psychometric Functions 12
1.4. Inference for Psychometric Functions 16
1.4.1. Fitting Psychometric Functions 17
1.4.2. Sampling for Psychometric Functions 19
1.5. Concluding Remarks 21
Chapter 2: Machine Learning Background **24**
2.1. Gaussian Processes 24
2.1.1. Supervised Learning 24
2.1.2. Bayesian Inference 25
2.1.3. Gaussian Process Regression (GPR) 27
2.1.4. Gaussian Process Classification (GPC) 29
2.1.5. Covariance Functions 32
2.1.6. Hyperparameters 34
2.2. Active Sampling 35
2.2.1. Definition of Active Sampling 36
2.2.2. Uncertainty Sampling 40
2.2.3. Bayesian Active Learning by Disagreement 41
2.3. Concluding Remarks 42

Chapter 3: Automated Estimation of Human Threshold Audiograms Using Active Machine Learning 43

3.1. Introduction 43

3.2. Methodology 44

3.2.1. Machine Learning Algorithm 44

3.2.2. Participants 48

3.2.3. Experimental Procedure 48

3.2.4. Data Analysis 52

3.3. Results 53

3.4. Discussion 62

3.5. Concluding Remarks 68

Chapter 4: Uni- and Multidimensional Audiometric Function Estimation Using Gaussian Process Classification 69

4.1. Introduction 69

4.2. Methodology: 1D Psychometric Function 71

4.2.1 Simulation Details 72

4.2.2 Gaussian Process Construction 72

4.2.3. Evaluation 74

4.3. Methodology: 2D Psychometric Function 76

4.3.1 Simulation Details 77

4.3.2. Gaussian Process Construction 78

4.3.3. Evaluation 80

4.4. Results: 1D Psychometric Function 81

4.5. Results: 2D Psychometric Function 87

4.6. Discussion 91

4.7. Concluding Remarks 94

Chapter 5: Estimation of Multidimensional Audiometric Functions Using Active Gaussian Process Classification 95

5.1. Introduction 95

5.2. Methodology 96

5.2.1. Simulation Details 97

5.2.2. Gaussian Process Construction 98

5.2.3. Sampling Methods 99
5.2.4. Evaluation 101
5.3. Results 102
5.4. Discussion 109
5.5. Concluding Remarks 111
Chapter 6: Summary & Future Direction 112
6.1. Summary of Findings 112
6.2. Recommendations for Future Direction 113
6.2.1. Human Studies 114
6.2.2. Psychometric Extensions 115
6.2.3. Efficiency Improvements 117
6.2.4. Large-Scale Distribution 121
6.3. Concluding Remarks 122
References 124
Appendix 1: Supplemental Figures 141
A1.1. Numerical Estimates for 1D Psychometric Functions 141
A2.1. Numerical Estimates for 2D Psychometric Functions 143
Curriculum Vitae 145

List of Figures

Figure 1.1: A standard clinical audiometer. 4

Figure 1.2: Example of a clinical air conduction threshold audiogram. 5

Figure 1.3: Example of the clinical modified Hughson-Westlake procedure. 6

Figure 1.4: Example of a psychometric function. 14

Figure 2.1: Example of Bayesian inference. 26

Figure 2.2: Illustration of a latent function passed through a sigmoidal likelihood. 31

Figure 2.3: Illustration of the effects of a length scale hyperparameter. 34

Figure 2.4: A diagram of a standard active learning cycle. 38

Figure 3.1: Illustration of the machine learning (ML) audiogram technique. 48

Figure 3.2: Photo of the sound isolation booth used to conduct experiments. 49

Figure 3.3: Diagram of the automated ML audiogram procedure. 51

Figure 3.4: Sample plots of HW, ML1, and ML2 audiogram results. 55

Figure 3.5: Agreement between ML and HW results for all valid ears. 56

Figure 3.6: Sample HW and ML results for 3 distinct ears. 57

Figure 3.7: Direct comparisons of ML1, ML2, and HW threshold estimates. 60

Figure 3.8: Agreement between ML and HW as a function of algorithm iteration. 61

Figure 3.9: Normalized posterior variance as a function of algorithm iteration. 62

Figure 3.10: Data from Listener 17, who fell asleep during ML estimation. 65

Figure 4.1: Plot of 4 human audiometric phenotypes. 77

Figure 4.2: Examples of 1D PF estimation using PR and GPC. 82

Figure 4.3: Error in 1D α and β estimates for PR and GPC across all conditions. 83

Figure 4.4: Error in 1D PR and GPC α and β estimates for different β values. 84

Figure 4.5: Error in 1D PR and GPC α and β estimates for different sampling distributions. 85

Figure 4.6: Direct comparison of 1D PR and GPC α and β estimates. 86

Figure 4.7: Sample posterior surface with samples and threshold/spread estimates. 88

Figure 4.8: Representative posterior distributions for four audiogram phenotypes. 89

Figure 5.1: Representative example of GPC inference using active sampling. 104

Figure 5.2: Representative posterior distributions for each sampling method. 105

Figure 5.3: Mean acquisition maps for each sampling technique. 106

Figure 5.4: Summary of active GP performance across all conditions. 107

Figure 5.5: Summary of active GP performance by audiogram phenotype. 108

Figure 5.6: Summary of active GP performance by β (spread) value. 109

Figure 6.1: Left- and right-ear GP human audiogram estimates. 119

Figure 6.2: Screenshot of the current website implementation for remote audiometric testing. 122

Figure A1.1: Error in 1D 50% point and IQR estimates for PR and GPC across all conditions. 141

Figure A1.2: Error in 1D PR and GPC 50% point and IQR estimates for different β values. 142

Figure A1.3: Error in 1D PR and GPC 50% point and IQR estimates by sample split. 142

List of Tables

Table 3.1: Summary of delivered samples for both HW and ML procedures. 54

Table 3.2: Differences between the ML audiogram estimate and the HW estimate. 58

Table 3.3: Test-retest reliability of ML audiogram. 59

Table 4.1: Absolute errors of 2D GP α and β estimates by number of samples. 90

Table 4.2: Absolute errors of 2D GP α and β estimates by value of β. 90

Table 4.3: Absolute errors of 2D GP α and β estimates by audiogram phenotype. 90

Table A1.1: Absolute errors of 2D GP 50% point and IQR estimates by number of samples... 143

Table A1.2: Absolute errors of 2D GP 50% point and IQR estimates by value of β. 143

Table A1.3: Absolute errors of 2D GP 50% point and IQR estimates by phenotype. 144

List of Abbreviations

1D: Unidimensional

2D: Two-dimensional

BALD: Bayesian active learning by disagreement

dB: Decibels

dB HL: Decibels hearing level

dB SPL: Decibels sound pressure level

GP: Gaussian process

GPC: Gaussian process classification

GPR: Gaussian process regression

HW: Modified Hughson-Westlake procedure

Hz: Hertz (cycles per second)

K-S: Kolmogorov-Smirnov test

ML: Machine learning

MLAG: Machine learning audiogram

PF: Psychometric function

PR: Probit regression

Acknowledgments

As I reflect back at the end of my long PhD adventure, I want to offer my gratitude to the many individuals without whom I could never have made it this far. First, I owe a huge debt of gratitude to my advisor, Dr. Barbour. I joined your lab in 2012 with many ideas but uncertain about the way forward. You helped me select and design projects that represented compelling technological and scientific advances, but took special care to ensure they were in line with my interests. Throughout my PhD trajectory, you've trusted me to forge my own path forward in research, but offered much-needed incentives and structure when appropriate. You have always been supportive not only of my scholarly endeavors, but also of my many interests outside of the laboratory; I've very much enjoyed sharing them with you. I am incredibly fortunate for the opportunity to work with and learn from a supportive, compassionate mentor like you.

The Barbour Lab has been filled with individuals who have had a profound impact on me not only as a scientist, but also as a person. To my labmates and fellow PhD candidates Jeff, Ruiye, and Wensheng: We all started at Washington University around the same time, and now, after completing our PhD journeys together, I feel like we are family. It goes without saying that you've been brilliant scientific peers, but I've also really appreciated our many non-work-related conversations on anything from politics and technology to entertainment and crazy start-up ideas (and at times, just listening to me complain). Your support and comradery have meant so much to me, and you have truly brightened my graduate school experience. To our post-docs, Noah and Ammar, and our lab technician, Kim: I've always looked up to you as role models of how to conduct scientific research, and I'm very thankful for all of the help and advice you've provided during my time here. Of course, I would be remiss not to mention the energy and humor you've brought to the lab, always making this workplace a genuinely fun place to be.

I've been incredibly fortunate to have the support of some wonderful faculty members through my graduate school experience: Dr. Petersen, Dr. Raman, Dr. Sommers, and Dr. Thoroughman. From my qualifications exam (when I really had no idea what I was doing) to my Project Building presentations (when I was way too ambitious with my proposed research) to my thesis proposal (when I really thought I finally had my project figured out) and now, my defense, you have been there with me every step of the way. As my doctoral project has evolved over the years, your encouragement, insight and advice have always illuminated the way forward. I would not be where I am today without your gracious guidance and support.

This work could not have been completed without the help of two computer science professors. To Dr. Weinberger: I still vividly remember that particular machine learning class in spring 2014 when you first introduced Gaussian processes and active learning. I had originally thought that applying those techniques to audiometry would be a small side project, but it quickly took on a life of its own. To Dr. Garnett: I am truly fortunate for the opportunity to work with and learn from you. Your Bayesian methods courses have helped immensely in my basic grasp of the concepts, and whenever I still didn't understand something (which was often!), you've always taken the time to patiently and thoroughly offer your insight.

Over my 6 years as a PhD student, Washington University has offered a positive, collaborative environment for my studies. In particular, the highly interdisciplinary Cognitive, Computational and Systems Neuroscience (CCSN) pathway has been instrumental in influencing how I think about science, and I am very thankful for my professors and fellow students in that pathway. I would also like to acknowledge my funding sources: the American Hearing Research Foundation (AHRF), the Center for Integration of Medicine and Innovative Technology (CIMIT), and Washington University in St. Louis. The research in this thesis would not have been possible without your generous support, for which I will always be grateful.

Within the BME department, I would like to thank Dr. Yin and Dr. George, the two department chairs during my time here: you have always been encouraging and have helped shape this department into the supportive, nurturing place it is has been. I owe so much gratitude to the department staff, Karen, Kate, Glen, and Amanda: I've lost track of how many times I've run into your office desperately needing help on anything from a room reservation to a last-minute grant submission; you've never hesitated to assist me, no matter how busy you were.

Over its approximate three years of existence, this machine learning audiogram project has involved many brilliant minds across various disciplines: Kiron, James, Braham, and Jacob in computer science; Brittany and Rebecca in audiology; Nikki, Katherine, and Joey in engineering, and Danish in ENT. Each of you has brought unique skills and expertise to the project (in addition to your lively personalities), and I've learned a great deal from working with you all. I truly would not have gotten where I am now without all of your help along the way; thank you.

I would like to thank my teammates on the audiogame projects: Tommy, Qihan, Sumeet, Jeff, Brittany, David, Danny, and Alice, as well as my collaborators in the Central Institute for the Deaf: Brent, Elizabeth, Cathy, Shannon, and Nancy. While ultimately that work didn't make it into my main thesis, we accomplished some very exciting things and I've learned a great deal through working together (including how to program in Android!). I have no doubt that the hard work we put into this project has set the groundwork for exciting research to come.

A special thank you to my friends in the online music and local game development communities. There is no doubt that, apart from my PhD work, making music and making games have been the biggest parts of my life for these past six years. During times when my research felt tedious or frustrating, I could always turn to these endeavors as a productive outlet for my creative energy. You guys have been amazingly supportive of everything that I do, and I've found (and continue to find) so much inspiration from your kindness, talent, and determination.

These acknowledgments would not be complete without thanking my family: my mom, my dad, my grandparents on both sides of the family, my brothers Eddie and Michael, and many others. From a young age, you have instilled in me the creativity, curiosity, discipline, and will to succeed that have shaped me into the scientist and individual I am today. No matter what paths I have pursued in life, you have never stopped believing in me. Your unwavering faith in my ability and your unconditional love are the greatest sources of inspiration in my life, and this doctoral work is truly the fruit of your selfless devotion over my 27 years.

Xinyu David Song

Washington University in St. Louis

August 2017

Dedicated to my family.

ABSTRACT OF THE DISSERTATION

Improving Pure-Tone Audiometry Using Probabilistic Machine Learning Classification

by

Xinyu D. Song

Doctor of Philosophy in Biomedical Engineering

Washington University in St. Louis, 2017

Professor Dennis L. Barbour, Chair

Hearing loss is a critical public health concern, affecting hundreds millions of people worldwide and dramatically impacting quality of life for affected individuals. While treatment techniques have evolved in recent years, methods for assessing hearing ability have remained relatively unchanged for decades. The standard clinical procedure is the modified Hughson-Westlake procedure, an adaptive pure-tone detection task that is typically performed manually by audiologists, costing millions of collective hours annually among healthcare professionals. In addition to the high burden of labor, the technique provides limited detail about an individual's hearing ability, estimating only detection thresholds at a handful of pre-defined pure-tone frequencies (a threshold audiogram). An efficient technique that produces a detailed estimate of the audiometric function, including threshold and spread, could allow for better characterization of particular hearing pathologies and provide more diagnostic value. Parametric techniques exist to efficiently estimate multidimensional psychometric functions, but are ill-suited for estimation of audiometric functions because these functions cannot be easily parameterized.

The Gaussian process is a compelling machine learning technique for inference of nonparametric multidimensional functions using binary data. The work described in this thesis utilizes Gaussian process classification to build an automated framework for efficient, high-resolution estimation of the full audiometric function, which we call the machine learning audiogram (MLAG). This Bayesian technique iteratively computes a posterior distribution describing its current belief about detection probability given the current set of observed pure tones and detection responses. The posterior distribution can be used to provide a current point estimate of the psychometric function as well as to select an informative query point for the next stimulus to be provided to the listener. The Gaussian process covariance function encodes correlations between variables, reflecting prior beliefs on the system; MLAG uses a composite linear/squared exponential covariance function that enforces monotonicity with respect to intensity but only smoothness with respect to frequency for the audiometric function.

This framework was initially evaluated in human subjects for threshold audiogram estimation. 2 repetitions of MLAG and 1 repetition of manual clinical audiometry were conducted in each of 21 participants. Results indicated that MLAG both agreed with clinical estimates and exhibited test-retest reliability to within accepted clinical standards, but with significantly fewer tone deliveries required compared to clinical methods while also providing an effectively continuous threshold estimate along frequency. This framework's ability to evaluate full psychometric functions was then evaluated using simulated experiments. As a feasibility check, performance for estimating unidimensional psychometric functions was assessed and directly compared to inference using standard maximum-likelihood probit regression; results indicated that the two methods exhibited near identical performance for estimating threshold and spread. MLAG was then used to estimate 2-dimensional audiometric functions constructed using existing audiogram phenotypes. Results showed that this framework could estimate both threshold and spread of the full audiometric function with high accuracy and reliability given a sufficient sample count; non-active sampling using the Halton set required between 50-100 queries to reach clinical reliability, while active sampling strategies reduced the required number to around 20-30, with Bayesian active leaning by disagreement exhibiting the best performance of the tested methods. Overall, MLAG's accuracy, reliability, and high degree of detail make it a promising method for estimation of threshold audiograms and audiometric functions, and the framework's flexibility enables it to be easily extended to other psychophysical domains.

Chapter 1: Introduction

"In general and irrespective of the age at which it develops, disabling hearing impairment has devastating consequences for interpersonal communication, psychosocial well-being, quality of life and economic independence."

−Quotation from (Olusanya et al., 2014)

"The deployment of accurate, automated [audiometric] methods to allow reallocation of time toward doctoral level activities is not only desirable, it is imperative."

−Quotation from (Margolis and Morgan, 2008)

1.1. Hearing Loss

Hearing loss is a critical public health concern. Over 360 million individuals worldwide are estimated to have disabling hearing loss (Pascolini and Smith, 2009; World Health Organization, 2012), accounting for approximately 5% of the world's population. For individuals 65 years and above, the proportion of affected individuals rises to 1 in 3. In the United States, approximately 37.5 million adults 18 and older, 15%, report some degree of hearing loss (Blackwell *et al.*, 2014; NIDCD, 2014), making it the most prevalent neurological disorder in the country.

Typical cases of hearing disability can be classified into two main categories: conductive and sensorineural hearing loss (Sataloff and Sataloff, 2005). Conductive hearing loss describes hearing loss that results from an interference of sound transmission through the external/middle to the inner ear, while sensorineural hearing loss describes hearing loss that results from damage in the inner ear (particularly hair cells) and/or the auditory nerve. There are numerous factors that lead to hearing loss, including diseases such as otosclerosis (De Souza and Glasscock, 2003) and Ménière's disease (Ménière and Atkinson, 1961), ototoxic drugs or chemicals (Schacht and Hawkins, 2006), noise exposure (Rabinowitz, 2000), trauma (Fitzgerald, 1996), and age-related degeneration (presbycusis) (Robinson and Sutton, 1979).

Hearing loss represents a large worldwide burden of disease (Mathers *et al.*, 2000; Cruickshanks *et al.*, 2003; Olusanya *et al.*, 2014) and can have a dramatic impact on quality of life (Mulrow *et al.*, 1990; Dalton *et al.*, 2003). For adults, hearing impairment can have detrimental impacts on relationships, social function, cognitive ability, emotional well-being, physical ability, and career trajectory (Weinstein and Ventry, 1982; Thomas *et al.*, 1983; Chen, 1994; Wallhagen *et al.*, 1996; Mohr *et al.*, 2000; Strawbridge *et al.*, 2000; Helvik *et al.*, 2006; Helvik *et al.*, 2009). Hearing disability can be even more detrimental for children, including the estimated 7.5 million affected children 5 years or younger (World Health Organization, 2012). Hearing loss in children can interfere with speech and language development (Yoshinaga-Itano *et al.*, 1998; Blamey *et al.*, 2001; Briscoe *et al.*, 2001; Yoshinaga-Itano, 2003), making detection even more critical.

Conductive hearing loss is often correctable via surgical or pharmaceutical intervention, while sensorineural hearing loss is currently not reversible (Sataloff and Sataloff, 2005). However, treatments can often dramatically improve hearing ability and significantly enhance quality of life for impacted individuals (Appollonio *et al.*, 1996; Cohen *et al.*, 2004; Vermeire *et al.*, 2005; Chisolm *et al.*, 2007). Perhaps the most common treatment for hearing loss is the use of a hearing aid, a device that amplifies and processes environmental sound that can be specifically tuned for individual hearing losses (Sataloff and Sataloff, 2005; Katz *et al.*, 2009). More severe sensorineural hearing losses are sometimes treated using cochlear implants, devices that replace cochlear hair cells with direct electrical stimulation of the auditory nerve (Clark *et al.*, 1979; Bond *et al.*, 2009). Cochlear implantation has been particularly effective for improving language development in deaf children (Svirsky *et al.*, 2000; Sharma *et al.*, 2002). Research has also demonstrated promising results for using auditory or speech training to improve listening ability in hearing-impaired individuals (Sweetow and Palmer, 2005; Burk and Humes, 2008; Fu and Galvin, 2008; Henshaw and Ferguson, 2013; Tye-Murray, 2014).

Despite these treatment options, however, hearing loss remains an underdiagnosed condition, partially owing to the lack of hearing screening procedures in standard clinics (Bogardus Jr *et al.*, 2003; Yueh *et al.*, 2003), which leads to lack of awareness in affected individuals. Other factors include the social stigma associated with and the financial burden of hearing treatment (Kochkin, 1993; 2007); this is reflected in the very low prevalence of hearing aid use among individuals who could benefit from them (Popelka *et al.*, 1998; Chien and Lin, 2012).

1.2. Methods for Hearing Evaluation

1.2.1. Pure-Tone Audiometry

Perhaps the most common method for clinical hearing evaluation is pure-tone audiometry, which serially presents pure-tones of varying frequency and intensity to locate subjects' auditory thresholds, or the lowest intensities at which an individual can detect a pure tone for a given frequency (Stach, 2010). A set of pure-tone thresholds given for a select number of frequencies collectively form an audiogram, which acts as a summary of an individual's overall hearing ability. In this thesis, we will refer to an audiogram (in the traditional clinical sense) as a *threshold audiogram* because they are inherently comprised of only threshold information. **Figure 1.1** shows a photograph of a typical audiometer used to estimate a threshold audiogram.

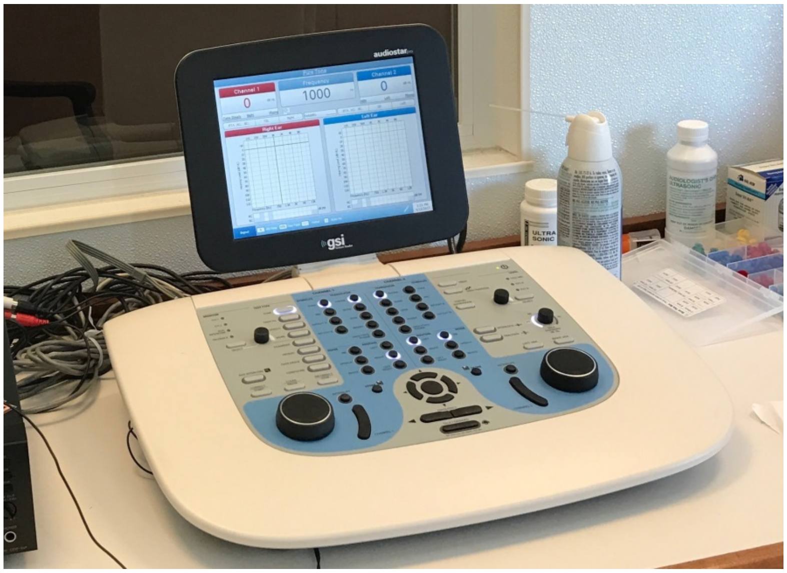

Figure 1.1: A standard clinical audiometer. This device allows for manual delivery of highly-calibrated pure tones at standard audiological frequencies and a variety of sound levels.

A threshold audiogram typically consists of auditory thresholds provided at a select number of pure-tone frequencies: 250 to 8000 Hz in octave intervals, with intermediate or edge frequencies sometimes included (American National Standards Institute, 2004a; American Speech-Language-Hearing Association, 2005); this frequency range is similar to the frequency range important for speech (French and Steinberg, 1947) Thresholds are most commonly reported in decibels hearing level (dB HL), which are sound level measures relative to what is considered to be normal hearing (Katz *et al.*, 2009; Stach, 2010). Relative to physical sound pressure level (dB SPL), each frequency has a different correction term in order to convert between SPL and HL (American National Standards Institute, 2004b). This conversion allows for threshold audiograms of normal-hearing individuals to be represented by a horizontal line at 0 dB HL.

An example of a threshold audiogram can be seen in **Figure 1.2**, in this example providing thresholds for 6 audiogram frequencies. Thresholds for left and right ears, given by blue X and red O marks, respectively, are conducted separately for each individual. The pure-tone average, or the mean of the measured thresholds at 500, 1000, and 2000 Hz, is often used as a quantitative summary of overall hearing ability. Categories for describing degree of hearing loss, including normal, mild, moderate, severe, and profound loss, are typically defined relative to the pure-tone average, although they may differ slightly depending on the reference (Goodman, 1965; Jerger and Jerger, 1980; Katz *et al.*, 2009).

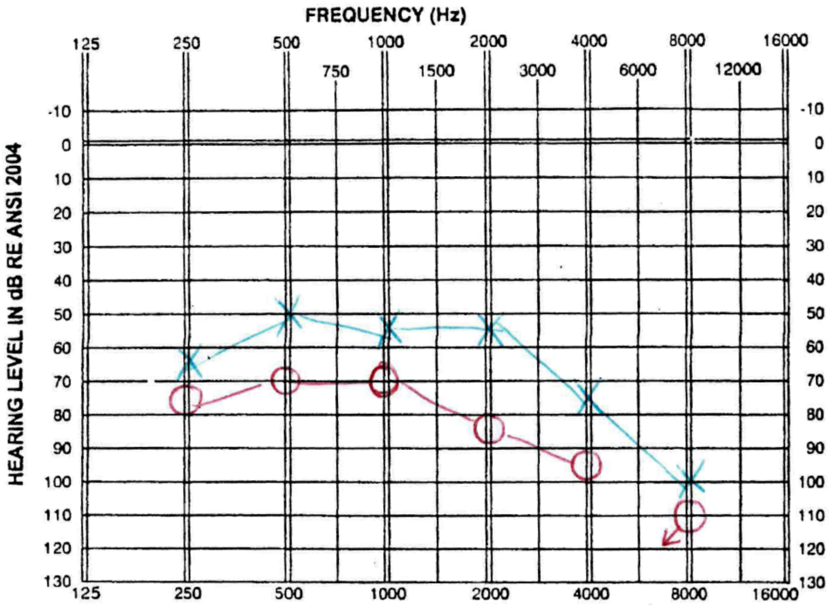

Figure 1.2: Example of a clinical air conduction threshold audiogram. Hearing thresholds in the right and left ears are denoted by red O and blue X marks, respectively. The red mark at 8 kHz was a no-response, wherein the subject did not detect the stimulus provided even at the loudest intensity (110 dB HL in this case).

In accordance with established guidelines (American National Standards Institute, 2004a; American Speech-Language-Hearing Association, 2005), pure-tone audiometry is typically performed in clinic following the modified Hughson-Westlake procedure (Hughson and Westlake, 1944; Carhart and Jerger, 1959). The modified Hughson-Westlake (HW) procedure is an variant of the method of limits (Levitt, 1971; Kingdom and Prins, 2010; Gescheider, 2015) that adaptively selects pure tones to deliver in order to rapidly achieve an estimate for threshold.

Note that all references to the Hughson-Westlake procedure or HW throughout this thesis refer to this *modified* Hughson-Westlake procedure (Carhart and Jerger, 1959).

In the HW procedure, listeners are asked to indicate when they detect (even if minimally) the presence of a tone, typically by raising their hands or by pressing a button. Testing proceeds on a per-frequency basis beginning at 1 kHz. The intensity of the initial tone is chosen to be a sound level well above putative threshold (with additional steps taken if the initial tone is not detected). Thereafter, each time the listener detects a presented tone, its intensity is decreased by 10 dB for the subsequent presentation; each time the listener does not detect a tone, its intensity is raised by 5 dB for the subsequent presentation. This adaptive rule is followed for a certain number of reversals; threshold for the current frequency is considered to be the lowest intensity at which the subject perceives the tone approximately 50% of the time. This procedure is then repeated from the beginning additional frequencies and for the contralateral ear. **Figure 1.3** shows an example HW run for detecting audiometric threshold at a single frequency.

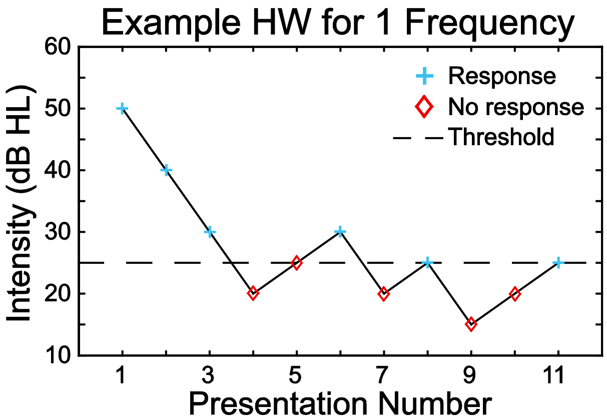

Figure 1.3: Example of the clinical modified Hughson-Westlake procedure. This figure shows an example procedure at one frequency; this process is typically repeated at all 6-9 standard audiogram frequencies.

As a “2-up, 1-down” task, the threshold returned by this procedure corresponds to the 70.7% detection probability point on a listener’s psychometric function (Levitt, 1971). The HW method is employed for both air conduction and bone conduction audiograms (Franks, 2001; Katz *et al.*, 2009), which test distinct mechanisms of sound transmission and differ primarily in the type of transducer used. Masking of the contralateral ear is sometimes used for individuals who exhibit large inter-ear threshold audiogram differences, which helps to account for high-intensity tones being detected by the non-test ear (Katz *et al.*, 2009; Stach, 2010).

Pure-tone threshold audiograms provide data that enables healthcare professionals to diagnose specific disorders (which often show loss localized to certain frequency ranges), screen for hearing disability, and monitor hearing changes over time, among many other applications (Katz *et al.*, 2009; Stach, 2010). The adaptive Hughson-Westlake procedure has been a staple for hearing assessment for decades; its steps are easy to follow and it can be quite efficient in the hands of experienced audiologists. It has an accepted test-retest reliability of approximately 5 dB HL (Jerlvall and Arlinger, 1986; Fausti *et al.*, 1990; Stuart *et al.*, 1991; Schmuziger *et al.*, 2004; Katz *et al.*, 2009), which is reasonably high for most screening purposes.

However, HW pure-tone audiometry has several disadvantages. Perhaps the most striking is that these manually conducted pure-tone audiograms are highly time- and labor-intensive for medical professionals as a whole. It is estimated that annually, audiologists collectively spend around 2 million hours performing pure-tone audiometry alone (Margolis and Morgan, 2008), a rote task that does not leverage audiologists' considerable expertise. Furthermore, the pure-tone audiogram provides only threshold data at any queried frequencies, with no information provided for intermediate frequencies. Although threshold audiograms such as Figure 1.2 linearly interpolate between measured thresholds for display purposes, no systematic estimate is provided for intermediate frequencies. The HW algorithm, while designed to quickly converge on threshold, exhibits inefficiencies such as multiple high-probability stimuli being presented for each frequency (although this can be mitigated somewhat by a skilled audiologist) and identical stimuli presented repeatedly near threshold. Finally, the HW procedure is highly predictable, which facilitates the intentional subversion of test results by noncooperative listeners.

1.2.2. Alternatives to Manual Pure-Tone Audiometry

Automated Audiometry

In parallel to the development of adaptive conventional approaches like the one described above, automated audiometry methods have been developed for clinical audiometry, with the earliest form designed by Georg von Békésy in the late 1940s (von Békésy, 1947). Many computerized audiometric methods designed to ensure consistency and save labor have been developed, with some employing a method of adjustment similar to Békésy's technique but most recent methods using a method of limits resembling the HW algorithm (Ho *et al.*, 2009; Margolis *et al.*, 2010; Swanepoel *et al.*, 2010; Mahomed *et al.*, 2013). Even with ready access to powerful digital computing technology today, however, automated audiometry sees relatively little use in clinical diagnostic settings, with most audiograms still obtained manually (Vogel *et al.*, 2007).

A recent exhaustive review and meta-analysis was conducted of techniques developed for automated threshold audiometry (Mahomed *et al.*, 2013). A wide range of automated techniques produced audiograms generally comparable to manual audiograms, with an absolute average difference of 4.2 dB HL and a standard deviation of 5.0 dB HL (n = 360). Test-retest reliability among these automated methods demonstrated an absolute average difference of 2.9 dB HL and a standard deviation of 3.8 dB HL (n = 80). As a comparison, manual threshold audiometry in the reported studies produced an absolute average difference of 3.2 dB HL and a standard deviation of 3.9 dB HL (*n* = 80). These studies indicate that computerized automation of pure-tone audiometry procedures yields threshold audiograms comparable in value and test-retest reliability to conventional manual procedures. However, the majority of automated techniques reviewed utilized adaptive techniques (including automated versions of the HW algorithm itself), which share many limitations with traditional pure-tone audiometry, particularly in performing inference one frequency at a time and reporting thresholds only at those points.

Sweep-Based Audiometry

Sweep-based audiometry is one alternative technique to adaptive methods that addresses the disadvantage of having data only at discrete frequencies. The first application of sweep-based audiometry was known as Békésy audiometry (von Békésy, 1947) and was application of the method of adjustment (Levitt, 1971; Gescheider, 2015). Listeners control the intensity of a pure-tone stimulus and are instructed to repeatedly increase its intensity until audible, then decrease its intensity until inaudible. The tone gradually sweeps across frequency in the meantime, allowing the final estimate to trace continuously across the threshold of hearing (von Békésy, 1947; Stach, 2010). A more recent implementation of sweep-based audiometry is Audioscan®, which uses a series of iso-intensity sweeps across frequency at varying intensities to trace out a high-resolution threshold curve (Meyer-Bisch, 1996; Ishak *et al.*, 2011).

Because these sweep-based techniques trace out relatively continuous threshold curves along the frequency dimension, they have been shown to successfully identify various hearing pathologies that have been difficult to detect using discrete pure-tone audiometric approaches (Jerger, 1960; Zhao *et al.*, 2002; Zhao *et al.*, 2014). Despite this advantage, however, these techniques do not currently see substantial use in the clinic. One major reason is the lengthened testing time required compared to conventional PTA, particularly with sweep rates that are sufficiently slow to be comfortable for listeners (Ishak *et al.*, 2011). Furthermore, substantial engagement by the listener is required, which could lead to inefficient acquisition, inaccuracies, and/or intentional misrepresentation.

Bayesian Audiometric Techniques

Several more recent methods have taken a Bayesian approach to estimating pure-tone threshold audiograms, incorporating prior information from existing threshold audiogram shapes to inform optimal sequential selection of tones for efficient audiogram estimation (Özdamar *et al.*, 1990; Stadler, 2009). The first approach uses a small database of weighted candidate audiometric patterns and iteratively selects the next tone at the frequency exhibiting maximum variance among patterns (Özdamar *et al.*, 1990). Initial pattern probabilities are chosen according to prevalence of that pattern in the population, and probabilities are updated each iteration. A more recent method uses a Gaussian mixture model and a chosen utility function to select the optimal query point by maximizing the expected utility (Stadler, 2009). Similarly to the previous method, the model is initially trained using prior data (in this case, a database of 100000 threshold audiograms), and model parameters are updated after each observation.

Both Bayesian techniques have demonstrated efficiency and accuracy in estimating audiograms for simulated and human listeners (Özdamar *et al.*, 1990; Eilers *et al.*, 1993; Stadler, 2009; Guan, 2011). However, like traditional adaptive methods, these methods limit the frequencies queried to the standard 6-9 audiogram frequencies, with no systematic estimates provided for intermediate frequencies. An extension of these Bayesian techniques to form a more continuous threshold estimate across frequency could provide the high resolution associated with sweep-based techniques while maintaining the stimulus selection efficiency of adaptive techniques.

Self-Diagnostic Tools

In more recent years, particularly with the rise of mobile smartphone technology, a number of non-clinical diagnostic tools for the end user have emerged. These tools have been deployed for many platforms, including landline phone (Watson *et al.*, 2012; Williams-Sanchez *et al.*, 2014), Internet browser (Bexelius *et al.*, 2008; Molander *et al.*, 2013), and the increasingly common mobile electronic device (Szudek *et al.*, 2012; Handzel *et al.*, 2013; Swanepoel *et al.*, 2014; Saliba *et al.*, 2016). These self-diagnostic tools are much more accessible than traditional forms of hearing diagnosis; individuals are able to utilize these tests on their own personal computers or smartphones without the need to visit a specialty clinic. However, studies on these techniques have primarily demonstrated their utility as a screening rather than detailed diagnostic tool, making their current implementations unlikely to replace standard clinical methods.

Physiological Measures

In addition to psychophysical tests, certain physiological measures are sometimes recorded in-clinic to gauge hearing ability. One widely used physiological measure is the auditory brainstem response (ABR), a subclass of auditory evoked potentials and a neurophysiological response that can be detected with scalp electrodes (Jewett *et al.*, 1970; Hecox and Galambos, 1974). The ABR has been shown to reflect the pure-tone threshold audiogram in certain frequency ranges and is particularly useful in diagnosis of certain functional disorders, such as acoustic tumors (Hecox and Galambos, 1974; Selters and Brackmann, 1977; Stapells and Oates, 1997). A second common measure is otoacoustic emissions (OAEs), which are low-intensity, frequency-specific sounds generated from the cochlea that occur both spontaneously and in response to delivered stimuli (Kemp, 1978; Probst *et al.*, 1987; Probst *et al.*, 1991). Because they reflect the integrity of hair cells, OAEs reflect overall hearing ability to some extent, and are useful in monitoring of potentially ototoxic treatment (Gorga *et al.*, 1997; Dorn *et al.*, 1999).

Both the ABR and otoacoustic emissions have proven particularly successful in assessing the hearing ability and sensitivity of young children who do not have the capability of performing pure-tone audiometry (Katz *et al.*, 2009; Stach, 2010). However, these physiological measures have seen comparatively lower employ relative to pure-tone audiometry due to lower sensitivity, higher test complexity, and increased testing time.

1.3. Psychometric Functions

Psychophysics describes the relationship between physical and perceptual processes, quantifying a subject's perception while a sensory stimulus feature is systematically altered (Fechner, 1860). This relationship is traditionally described using a psychometric function (PF), which describes a subject's task performance as a function of a physical variable or variables. For instance, in pure-tone audiometry, the sensory domain consists of pure-tone auditory stimuli, and stimulus features being manipulated are the frequency and intensity of these pure tones. The full PF across the variable space captures not only a subject's thresholds, but also the degree of uncertainty of a subject's performance around those thresholds. For instance, research has hypothesized higher levels of internal noise in children versus adults for pure-tone detection and discrimination tasks (Allen and Wightman, 1994; Bargones *et al.*, 1995; Buss *et al.*, 2006; 2008); this effect across frequency/intensity space could be captured using the full audiometric PF.

Psychophysical tasks can be assigned a threshold below which successful task performance is considered unreliable and above which performance is considered reliable, such as the 70.7% audiometric threshold returned by the HW procedure. However, psychophysical responses are not absolute; for instance, a listener may still detect a tone played slightly below the reported threshold. To describe this inherent uncertainty, we typically assign a detection *probability* to each stimulus describing an individual's performance at that stimulus.

For certain stimulus parameters, the subject's detection probability increases with increasing value. For example, as the intensity (sound level) of a pure-tone stimulus increases, a listener will detect it with higher probability. The relationship between a psychometric variable (for which performance increases with increasing value) and a subject's response probability can be described using a unidimensional (1D) PF. For a PF on a psychometric variable, we typically model the response probability as a sigmoidally increasing function with stimulus value (Klein, 2001; Kingdom and Prins, 2010; Gescheider, 2015).

A 1D psychometric function is typically characterized using two main parameters: threshold α and spread β. The threshold α corresponds to the point of inflection and describes the stimulus level for which performance is halfway between the highest and lowest values. The spread β characterizes the degree of uncertainty around threshold; a higher value of β lengthens the transition region where response probabilities are not at the minimum or maximum values. (Note that the definition of β varies between sources; β is sometimes used to describe slope, the inverse of spread. In this thesis, we consistently use β to describe spread.)

An example of a psychometric function can be seen in **Figure 1.4**. For a detection task, in which subjects respond when they detect a stimulus but are not forced to make a choice at each presentation (Fechner, 1860; Kingdom and Prins, 2010), the idealized minimum and maximum response probabilities take values of 0 and 1, respectively. Threshold α describes the point of maximum uncertainty, where detection probability is 0.5. Spread β quantifies the change in stimulus level required to produce a particular change in probability. For non-detection tasks such as *n*-alternative forced choice (Fechner, 1860; Kingdom and Prins, 2010) or to account for or lapse or guess rates in detection tasks (Klein, 2001; Wichmann and Hill, 2001a), additional parameters λ and γ are sometimes added. However, because pure-tone audiometry is inherently a detection task, we will focus our development of PFs on the idealized detection case.

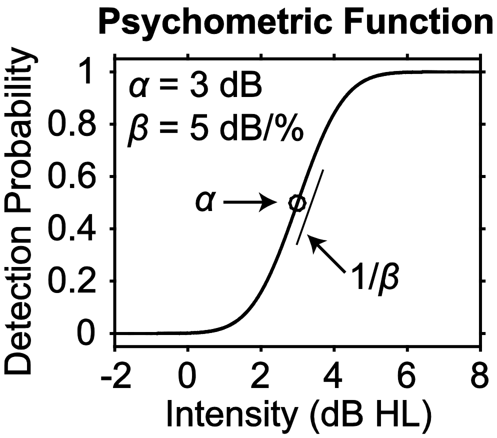

Figure 1.4: Example of a psychometric function. Threshold α corresponds to the point of inflection (at which detection probability is 0.5) and spread β quantifies the amount of response uncertainty around the threshold. For this example, the cumulative Gaussian function (Equation 1.2) was used to generate the curve.

Many sigmoidal functions have been used to model PFs, two of the most common being the logistic and cumulative Gaussian functions, shown in Equations 1.1 and 1.2, respectively (Klein, 2001; Falmagne, 2002; Kingdom and Prins, 2010; Gescheider, 2015):

$$\psi\left(x\right) = \frac{1}{1 + \exp\left(-\frac{x-\alpha}{\beta}\right)}, \tag{1.1}$$

$$\psi\left(x\right)=\frac{1}{\beta\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\tfrac{1}{2}\left(\frac{z-\alpha}{\beta}\right)^{2}\right)dz, \tag{1.2}$$

where x is stimulus level, $\psi\left(x\right)=p\left(y=1\middle|x\right)$ is detection probability, and z is a variable of integration. Both equations constrain the detection probability $\psi\left(x\right)$ to the probabilistic range $\left[0,1\right]$, with $\psi\left(x\right)$ monotonically increasing with increasing x. The α parameter determines the location of the 50% point and the β parameter adjusts the relative shallowness of the sigmoid. Standard parametric choices of 1D PF models are a subset of generalized linear models (McCullagh and Nelder, 1989), which are comprised of a linear predictor transformed with a monotonic link function.

Many if not most real-world psychophysical phenomena of interest, however, are inherently multidimensional, with more than one variable that effects change in subjects' performance. In pure-tone audiometry, for instance, listener performance is affected by both the frequency and intensity of the delivered tones. In addition to psychometric input variables, multidimensional PFs often include one or more non-psychometric variables, against which detection probability does not systematically increase. For pure-tone audiometry, the non-psychometric dimension is frequency; unlike with intensity, increasing values for frequency do not systematically result in higher detection probabilities and in fact, describing the effect of frequency on a listener's performance is a goal of audiometric testing. A limited number of multidimensional PFs have been characterized, including auditory filters (Patterson, 1976; Shen *et al.*, 2014), external contrast noise functions (Lesmes *et al.*, 2006) and visual fields (Heijl and Krakau, 1975; Bengtsson *et al.*, 1997). The PFs in these cases are parameterized, though the mechanistic justification for doing so could be limited. However, the pure-tone audiogram PF includes a non-psychometric frequency input variable for which any particular parametric justification is weak.

One elegant conceptualization of psychometric functions was proposed in (Kuss *et al.*, 2005), in which any parametric 1D PF formulation can be decomposed into core and sigmoid functions. The core function contains the psychometric parameters α and β and is related to the detection probability; large positive core values produce detection probabilities close to 1; large negative core values produce detection probabilities close to 0, and core values close to 0 produce detection probabilities near 0.5. The core function is often a linear function $\left(x-\alpha\right)/\beta$ for psychometric variables to capture monotonicity, although it can take other parametric forms (e.g. logarithmic or polynomial) to account for non-psychometric variables. The sigmoid function is a nonlinear transformation of the core function that "squashes" core values, which span $\left(-\infty,\infty\right)$, into the probabilistic range $\left[0,1\right]$. In this framework, the logistic function (1.1) can be decomposed into core function $\left(x-\alpha\right)/\beta$ and sigmoid $1/\left(1+e^{-x}\right)$. However, a limitation of this framework is that core functions must be specified parametrically, limiting its utility in cases for which parametric justification on a particular domain is weak or nonexistent.

1.4. Inference for Psychometric Functions

Psychometric functions have been a subject of study for decades. Research regarding PFs can be categorized into two broad topics: 1) methods to effectively estimate a PF from a set of data, and 2) methods to efficiently sample psychometric space to quickly arrive at an estimate. While there are now relatively standardized methods for fitting PFs, the question of how to most efficiently sample, particularly for multidimensional PFs, is still an active area of research.

1.4.1. Fitting Psychometric Functions

The problem of fitting psychometric functions to some set of observed data has overwhelmingly focused on the unidimensional case for psychometric variables (for which detection probability increases monotonically). Techniques for estimating 1D PFs typically assume a parametric form for the PF, typically a sigmoidal function such as (1.1) or (1.2).

Perhaps the most common method for fitting psychometric functions is maximum-likelihood regression (Morgan, 1992; Collet, 2003; Kingdom and Prins, 2010). In this method, binary repetitions at identical stimulus values are typically collapsed into a single proportion. Given the observed binomial data $\mathcal{D}$ and a set of parameters θ, typically $\theta = \left\{\alpha, \beta\right\}$ for this application, we choose the set of parameters $\hat{\theta}$ that maximizes the likelihood of observing these data given the parameters $p\left(\mathcal{D}\middle|\theta\right)$. Often this maximum cannot be solved for analytically, so optimization methods such as the Nelder-Mead simplex method (Nelder and Mead, 1965; Kingdom and Prins, 2010) are employed to numerically locate the maximum.

Estimation of psychometric functions has also utilized Bayesian approaches (Bayes and Price, 1763; Jaynes, 2003), which accounts for prior beliefs on parameter distributions via a prior distribution (see Section 2.1.2). A Bayesian variant on the maximum-likelihood procedure, referred to as the constrained maximum-likelihood method (Treutwein and Strasburger, 1999), places prior distributions (in this particular work, beta distributions) on parameter values to obtain a point estimate of maximum-likelihood parameter values. Later work (Kuss *et al.*, 2005) proposed a fully Bayesian treatment of 1D PF estimation, in which a full probability distribution can be constructed over the parameters and point estimates can be obtained by minimizing some loss function, following standard Bayesian decision theory (Berger, 1985).

A limited number of parameter-free methods for PF estimation have been developed, which do not assume a particular parametric model as the form of the PF. One nonparametric technique uses the Spearman-Kärber method (Spearman, 1908; Kärber, 1931; Morgan, 1992), which can numerically estimate the moments of the PF (Klein, 2001; Miller and Ulrich, 2001). A second, more recent technique uses local linear fitting (Fan *et al.*, 1995), which locally approximates a function using a Taylor expansion (Zychaluk and Foster, 2009). Although somewhat sensitive to method specifications, both techniques have demonstrated high accuracy and reliability for estimation of 1D PFs, even when compared to the appropriate parametric models (Klein, 2001; Miller and Ulrich, 2001; Zychaluk and Foster, 2009). However, nonparametric models for 1D PF estimation have seen very limited use in practice compared to parametric methods.

Frameworks for estimation of multidimensional PFs, which typically include at least one non-psychometric dimension, have been considerably scarcer. Certain parametric models have been developed for specific psychometric spaces, including the auditory filter (Patterson, 1974; 1976; Shen and Richards, 2013), the visual field (Heijl and Krakau, 1975; Bengtsson *et al.*, 1997), the external noise contrast function (Lesmes *et al.*, 2006), and elliptical-threshold PFs such as color difference detection (Kujala and Lukka, 2006). More flexible frameworks for multidimensional PF estimation have more recently been proposed, which specify particular parameterizations of threshold functions across non-psychometric dimensions (Vul *et al.*, 2010; DiMattina, 2015; Watson, 2017). However, an incorrect choice of model parameterization can result in errors, and cases for which parametric justification is weak or nonexistent are not covered (Vul *et al.*, 2010).

1.4.2. Sampling for Psychometric Functions

A related question to psychometric function inference is how best to select samples to efficiently produce accurate estimates of the PF. Traditionally, sampling for 1D PFs is accomplished using the method of constant stimuli, first proposed by Gustav Fechner (Fechner, 1860). In this method, a set of stimuli of varying stimulus levels are chosen such that they straddle a putative threshold value. Stimuli from this set are then delivered to the subject in a random order, with *m* repetitions delivered at *n* different stimulus values overall. When executed properly, the method of constant stimuli provides a well-spaced set of stimulus presentations that can capture psychometric behavior within a range of interest. However, downsides to this technique include sensitivity of PF estimation results to both the number of distinct stimulus levels delivered and to the number of repetitions at each level (often necessitating large numbers of samples to produce an accurate estimate), and the need for an additional procedure (e.g. a method of limits run) to determine the proper range for delivered stimulus values, and (Levitt, 1971).

To counteract the efficiency limitations of the method of constant stimuli, a number of adaptive procedural techniques have been developed for estimation of threshold only. A particularly well-established set of procedural methods is up-down methods, also called staircase methods (Dixon and Mood, 1948; Levitt, 1971; Kingdom and Prins, 2010). Up-down techniques follow the simple rule that if a stimulus is detected, the stimulus level for the next presentation should be decreased, and vice versa, until the test terminates. The standard up/down method uses the same step size for both stimulus increases and decreases, returning the 50% detection probability point (Dixon and Mood, 1948; Kingdom and Prins, 2010). Test termination typically occurs after a certain number of reversals, with the threshold estimate computed as the mean stimulus intensity across the last few trials containing a reversal (García-Pérez, 1998; Kingdom and Prins, 2010).

Modifications to the up-down method include transformed methods, in which some number of consecutive identical responses must be observed before adjusting stimulus level (for instance, 2 detections in a row before stimulus level is decreased) (Wetherill and Levitt, 1965; Levitt, 1971), and weighted methods, in which unequal step sizes are used for up vs. down (Kaernbach, 1991); these modifications correspond to different detection probabilities on the curve. A further refinement of up-down techniques was parameter estimation by sequential testing (PEST), which adaptively narrowed the step size of stimulus level changes to efficiently determine threshold (Taylor and Creelman, 1967; Findlay, 1978). Although these procedural techniques are designed to estimate thresholds only, a parametric fit can be performed to the set of stimuli and responses collected during the procedure using a fitting technique (Section 1.4.1); this strategy has been called a "hybrid adaptive procedure" (Hall, 1981; Leek *et al.*, 1992).

A second class of adaptive techniques seeking 1D thresholds utilizes methods that have control over the exact stimulus placement for each delivery, rather than relying only on step size changes (Treutwein, 1995; Leek, 2001; Kingdom and Prins, 2010). The general framework involves at every iteration computing an estimate based on the data observed so far, then choosing the next sample point based on some value derived from the current estimate, typically to maximize some information measure. For example, the "best PEST" technique places each subsequent stimulus delivery at the current estimate's 50% threshold value (Pentland, 1980). A Bayesian variant of this technique was proposed with QUEST, which uses the set of all trials collected so far as well as prior information to construct a posterior distribution and uses its mode, median or mean as the point estimate for threshold, which is sampled in the subsequent iteration (Watson and Pelli, 1983; King-Smith *et al.*, 1994). Various updates on and variants of these "maximum-likelihood" adaptive procedures have been developed for different stimulus spaces and test designs (Green, 1993; Dai, 1995; He *et al.*, 1998; Leek *et al.*, 2000; Leek, 2001; Linschoten *et al.*, 2001).

The previously described adaptive techniques focus mainly on threshold, but a number of adaptive techniques for estimation of the entire PF (in particular, spread) have been developed. An early technique is adaptive probit estimation, which alternates between computing a probit fit given the current set of data (Finney, 1971) and selecting a new block of stimuli to deliver given the current estimate (Watt and Andrews, 1981). The modified ZEST technique, a Bayesian method, iteratively updates a 2-dimensional posterior probability distribution on threshold and spread parameters and selects the next stimulus level that minimizes the expected variance upon observing that trial (King-Smith and Rose, 1997).

Perhaps the most well-known technique for full 1D PF estimation is the Ψ-method (Kontsevich and Tyler, 1999). Like the modified ZEST technique, the Ψ-method iteratively updates posterior probabilities for threshold and spread. It then selects the stimulus intensity that maximizes the expected information gain after that sample is observed (i.e. minimizes the expected entropy). More recent techniques have refined this adaptive framework to better characterize optimal selection of points and to account for lapse and guess rates (Brand and Kollmeier, 2002; Shen and Richards, 2012; Shen *et al.*, 2015). Extensions of these adaptive 1D PF estimation methods, particularly the Ψ-method, have also been applied to estimate multidimensional PFs (Kujala and Lukka, 2006; Lesmes *et al.*, 2006; Lesmes *et al.*, 2010; Vul *et al.*, 2010; DiMattina, 2015).

1.5. Concluding Remarks

Clinical assessment of hearing ability is overwhelmingly conducted using manual HW pure-tone audiometry, costing millions of hours in labor each year for audiologists collectively. Automated techniques have existed for decades but have primarily been variants on the HW procedure itself; other methods that have been developed, such as Bayesian techniques, are highly efficient but nonetheless produce estimates only at standard audiogram frequencies. Audiometric threshold estimates that are relatively continuous across frequency have been estimated using sweep-based methods, but such methods tend to be time-consuming and tedious for the listener.

Additionally, the full audiometric function (pure-tone multidimensional psychometric function) is almost never estimated in any capacity, which limits researchers' ability to fully characterize certain forms of hearing disability for better diagnostic insight. Accurate and efficient techniques for estimating multidimensional PFs exist but all assume that the non-psychometric dimensions or threshold shapes can be parameterized in some form. However, the threshold curves for audiometric functions cannot be justifiably parameterized, as the particular shape of a threshold audiogram is itself indicative of the type of hearing disability.

In the work described by this thesis, we propose a method for estimating arbitrarily shaped audiometric functions using Bayesian machine learning classification, which we call the machine learning audiogram (MLAG). This technique does not require explicit parameterization of the audiometric function, instead encoding relationships between input points along any stimulus dimension, and can perform inference using binary response data typical of pure-tone detection tasks. This inherent flexibility allows for a variety of shapes to be encoded, including full audiometric functions and by implication, the associated threshold audiograms. While this thesis focuses on the particular application for pure-tone audiometry, the technique is general-purpose and can be extended to other psychophysical domains in a straightforward manner.

Chapter 2 introduces the concepts from probability theory and machine learning relevant to the work in this thesis. Chapter 3 describes MLAG applied to estimation of threshold audiograms in human listeners, comparing the results of our technique to traditional clinical methods. Chapter 4 characterizes the ability of MLAG to estimate multidimensional audiometric functions, including both threshold and spread, in simulated experiments. Finally, Chapter 5 extends the work of Chapter 4 to incorporate active sampling techniques, with the goal of improving efficiency.

Chapter 2: Machine Learning Background

2.1. Gaussian Processes

Parametric models, which have a deterministic form given a set of parameter values, are often employed to describe natural phenomena. However, many natural phenomena, such as weather patterns, heart rate over time, or neuronal firing activity, have a degree of randomness that are not well-described by deterministic models. Such phenomena are often better characterized using a stochastic process (Gubner, 2006), a collection of random variables that can describe the evolution of an inherently random process as a function of some independent variable.

The work in this thesis heavily employs the Gaussian process (GP), a mathematically convenient subclass of stochastic processes that encodes particular relationships between function values. While the GP has been conceptualized for decades, particularly in the geostatistics field where it has been referred to as “kriging” (Krige, 1951; Cressie, 1990), it has recently been adapted and further developed for machine learning applications (Rasmussen, 1996; Williams and Rasmussen, 1996; Gibbs, 1998; Williams and Barber, 1998; Rasmussen and Williams, 2006). This section provides a formal definition of the GP and describes how to use GPs for inference, as well as some details relating to GP construction.

2.1.1. Supervised Learning

Supervised learning describes a subset of machine learning techniques that aims to perform inference on some system after observing a set of training data $\mathcal{D} = \left\{ \left(\mathbf{x}_i, y_i \right)_{i=1}^{n} \right\}$, where $\mathbf{x}_i$ is a feature vector (input location) at observation i and y_i is a measurement or known value corresponding to location $\mathbf{x}_i$ (Bishop, 2006; Hastie *et al.*, 2009; Murphy, 2012). In supervised learning, we use the observed values $\mathcal{D}$ to train some model, encoding the system properties provided within the observations. We then use the trained model to make predictions on a vector of unobserved features $\mathbf{X}^*$. For instance, the system we wish to perform inference on may be some unknown latent function $f\left(x\right)$, with y_i corresponding to a noisy function observation, i.e. $y_i = f\left(\mathbf{x}_i\right) + \varepsilon$. Supervised learning techniques stand in contrast to unsupervised learning techniques, which infer structure from "unlabeled" data that do not have corresponding categorical or numerical values.

A variety of techniques fall under the umbrella of supervised learning, including fully parametric estimators such as linear or logistic regression, statistical classifiers such as linear discriminant analysis, and nonparametric models such as *k*-nearest neighbor, support vector machines or neural networks. Gaussian process inference is a probabilistic method, which provides not only a point estimate of a function values at each test point in $\mathbf{X}^*$, but a probability distribution describing its belief and uncertainty about the corresponding function value.

2.1.2. Bayesian Inference

A statistical framework that is particularly synergistic with the concept of supervised learning is the Bayesian technique. Broadly, the Bayesian framework encodes a set of beliefs about a system that can be updated in light of observed data. In function space, Bayesian inference begins by assigning a prior probability over functions that describes a belief about which functions are expected *a priori*, as well as some likelihood (observation model) that describes how observations are generated from the underlying function (Gibbs, 1998; Jaynes, 2003; Xiang and

Fackler, 2015). Given some observed data, Bayes's Theorem can be applied to derive a posterior probability, which describes the updated beliefs about the function considering both the observed responses and the prior belief (Bayes and Price, 1763; Jaynes, 2003). Relative to supervised learning, the prior can be conceptualized as the initial state of the model, the observations upon which the posterior is conditioned as the training data, and the posterior distribution on some test set $\mathbf{X}^*$ as the post-training model prediction.

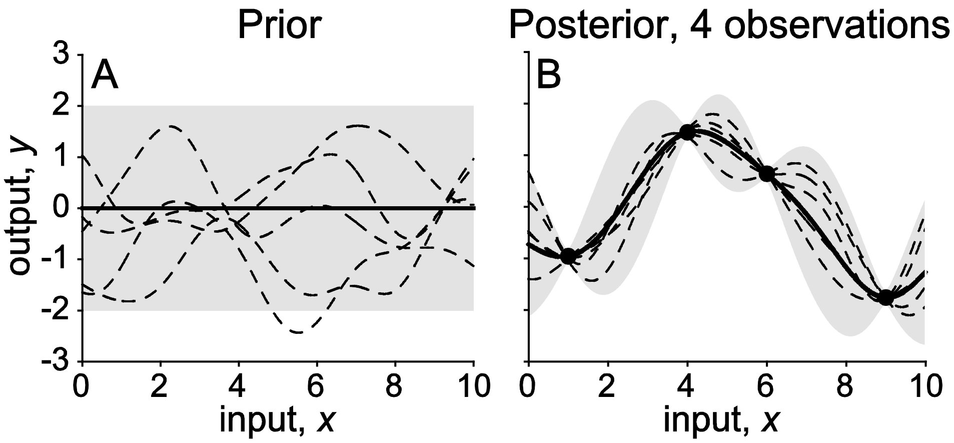

Figure 2.1: Example of Bayesian inference. (A) 5 sample functions drawn from a prior distribution on functions. (B) The posterior distribution after 4 observations are made, as well as 5 sample draws from this distribution. Solid lines denote the mean, shaded gray areas denote 2 standard deviations about the mean, and dotted lines denote single draws from the corresponding distributions.

Figure 2.1 shows an example of inference using the Bayesian method. The prior distribution over functions, along with several draws from the distribution, is shown in Figure 2.1A. Figure 2.1B shows the posterior distribution after making several observations, coupled with a Gaussian error likelihood. Notably, the uncertainty of the posterior distribution is considerably decreased around the observations, limiting the space of possible functions to the subset that passes near the observed points. The 5 draws from the posterior distribution show functions from this subset.

2.1.3. Gaussian Process Regression (GPR)

Let $f\left(x\right)$ be a latent function on an arbitrary input space $x \in \mathcal{X}$ that we wish to model. A GP is a mathematically convenient mechanism to encode prior knowledge about f, which we can update in light of observed data via Bayesian inference. Formally, a GP is defined as a collection of random variables (stochastic process), any finite subset of which jointly form a Gaussian distribution. A GP is a natural extension of the multivariate Gaussian distribution $\mathcal{N}\left(\mu, \Sigma\right)$ to infinite domains; like the multivariate Gaussian distribution, it is fully specified by its first two moments: a mean function $\mu\left(x\right)$ and a positive semidefinite covariance function $K\left(x, x'\right)$. The mean function accounts for the central tendency of the latent function while the covariance function accounts for the correlation structure of the latent function. Given μ and K, the latent function f can be endowed with a GP prior distribution:

$$p\left(f\right) = \mathcal{GP}\left(\mu\left(x\right), K\left(x, x'\right)\right). \tag{2.1}$$

Consider a set of observations $\mathcal{D} = \left\{\mathbf{X}, \mathbf{y}\right\}$. Our GP prior on f implies a multivariate Gaussian distribution for the corresponding latent function values $\mathbf{f} = f\left(\mathbf{X}\right)$, but does not specify how these latent function values are related to our observations $\mathbf{y}$. We must therefore use a *likelihood function* (observation model) to describe the relationship between the latent function and our observations, or $p\left(\mathbf{y} \middle| \mathbf{f}\right)$. The likelihood function can be any arbitrary model, but in the standard regression case, the latent function f can be viewed as the underlying "true" behavior from which each observation is derived with some residual error. Under this model, each observed value y_i is realized by independently corrupting the true value of the latent function $f\left(\mathbf{x}_i\right)$ by zero-mean additive Gaussian noise with variance σ_ε^2, allowing us to write the following expression for the likelihood:

$$p\left(\mathbf{y}\middle|\mathbf{f},\sigma^2\right) = \mathcal{N}\left(\mathbf{y};\mathbf{f},\sigma_\varepsilon^2\mathbf{I}\right). \tag{2.2}$$

Given a GP prior on f (Equation 2.1) and some observations over the input space, prediction can be performed about the behavior of f for unobserved inputs using Bayesian inference. Consider a set of *n* training observations $\left\{\left(\mathbf{x}_i, y_i\right)_{i=1}^n\right\} = \left(\mathbf{X},\mathbf{y}\right)$ and a set of unobserved test inputs $\mathbf{X}^*$ on which we wish to perform inference. We can use Bayes's Theorem to produce an expression for the joint posterior of the latent function at training and test inputs given the training observations:

$$p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{y},\mathbf{X}^*\right) = \frac{p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{X}^*\right)p\left(\mathbf{y}\middle|\mathbf{f},\mathbf{X}\right)}{p\left(\mathbf{y}\middle|\mathbf{X}\right)}. \tag{2.3}$$

The predictive posterior distribution can be determined by marginalizing out the nuisance training set latent variables and substituting Equation 2.2:

$$p\left(\mathbf{f}^*\middle|\mathbf{X},\mathbf{y},\mathbf{X}^*\right) = \int p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{y},\mathbf{X}^*\right)d\mathbf{f} = \frac{1}{p\left(\mathbf{y}\middle|\mathbf{X}\right)}\int p\left(\mathbf{y}\middle|\mathbf{f},\mathbf{X}\right)p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{X}^*\right)d\mathbf{f}. \tag{2.4}$$

By definition of the GP, the joint probability $p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{X}^*\right)$ is multivariate Gaussian:

$$p\left(\mathbf{f},\mathbf{f}^*\middle|\mathbf{X},\mathbf{X}^*\right) = \mathcal{GP}\left(\begin{bmatrix}\mu\left(\mathbf{X}\right)\\ \mu\left(\mathbf{X}^*\right)\end{bmatrix},\begin{bmatrix}K\left(\mathbf{X},\mathbf{X}\right) & K\left(\mathbf{X},\mathbf{X}^*\right)\\ K\left(\mathbf{X}^*,\mathbf{X}\right) & K\left(\mathbf{X}^*,\mathbf{X}^*\right)\end{bmatrix}\right). \tag{2.5}$$

The joint probability $p\left(\mathbf{f}, \mathbf{f}^* \middle| \mathbf{X}, \mathbf{X}^*\right)$ is Gaussian for a GP (2.5), and the likelihood function $p(\mathbf{y} | \mathbf{f}, \mathbf{X})$ is Gaussian by design (2.2). In the regression case, we can solve the integral in (2.4) in closed form, producing the posterior belief $p\left(\mathbf{f}^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{X}^*\right)$ about $\mathbf{f}^*$ analytically. Gaussian residual error is a maximum entropy likelihood and a commonly employed objective likelihood function (Xiang and Fackler, 2015). Furthermore, this GP regression framework can be viewed as an extension of the well-established Bayesian linear regression (Box and Tiao, 1992) to regression with an infinite number of basis functions (Williams, 1998).

The posterior belief about f is in this case a new GP with parameters that have been updated to reflect the information contained in the new inputs $\mathbf{X}^*$ and the previous observations $\mathcal{D}$. The posterior mean reflects our updated beliefs about f, weighing both our prior knowledge and the information contained in the observations. The posterior covariance encodes the remaining uncertainty about the latent function, with the diagonal entries (the posterior variance) encoding the marginal uncertainty remaining about a latent function value at a point.

2.1.4. Gaussian Process Classification (GPC)

The GP *regression* framework described in Section 2.1.3 assumes that the output values $\mathbf{y}$ are continuous and real-valued. However, c*lassification* problems, in which each input in a set $\mathbf{X}$ can be assigned to one of a finite set of N classes $C_1, C_2, \ldots C_N$, represent another important type of function approximation particularly relevant for psychometrics. A special case of classification problems are binary classification problems with output values assigned to one of only two classes. Many psychometric tasks are designed with two possible responses, including the detection task used for the standard threshold audiogram in which subjects indicate when they hear a tone (Carhart and Jerger, 1959; Kingdom and Prins, 2010). We therefore focus our current treatment on binary classification, although GP classification can readily be extended to multiple classes (Williams and Barber, 1998; Rasmussen and Williams, 2006).

For GP classification, we also place a GP prior on the latent function: $p\left(f\right)=\mathcal{GP}\left(\mu,K\right)$ (2.1). GP classification differs from GP regression primarily in the choice of the likelihood function $p\left(\mathbf{y}\middle|\mathbf{f}\right)$. In the regression case, we assumed that the observed values y_i were simply the latent function values f_i corrupted by additive zero-mean Gaussian noise with variance σ_ε^2 (2.2). In the case of binary classification, however, observed outputs y_i can take on one of only two class identities: either 1 (success) or 0 (failure). The latent function f is not directly observed but is instead a hidden function whose value is related to the degree of class membership, where larger values of f generate higher probabilities of success. To obtain the probabilistic distribution $p\left(y=1\middle|f\right)$, we transform f using a monotonically increasing sigmoid function Φ to constrain the resulting values to the range $\left[0,1\right]$. For a binary observation $y_i\in\mathbf{y}$ associated with a multidimensional input $\mathbf{x}_i\subset\mathbf{X}$, we assign the following likelihood:

$$p\left(y_i=1\middle|f_i\right)=\Phi\left(f_i\right). \tag{2.6}$$

Some convenient choices of Φ include the logistic (logit) function $\Phi\left(f_i\right)=1\big/\left(1+e^{-f_i}\right)$ or the cumulative Gaussian (probit) function $\Phi\left(f_i\right)=\int_{-\infty}^{f_i}\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz$. These functions constrain the range of outputs to the probabilistic range $\left[0,1\right]$, with large positive f values producing output values near 1 and large negative f values producing output values near 0; the effect of Φ can be seen in the example in **Figure 2.2**. Furthermore, both functions are consistent with longstanding psychometric function approximation, with the logistic and probit functions being popular models for psychometric behavior (Wichmann and Hill, 2001a; Kuss *et al.*, 2005; Kingdom and Prins, 2010). Following Bayesian decision theory, class membership can be predicted by thresholding (Berger, 1985). Like GP and Bayesian regression, GP classification can be viewed as a generalization of Bayesian logistic/probit regression (Rasmussen and Williams, 2006).

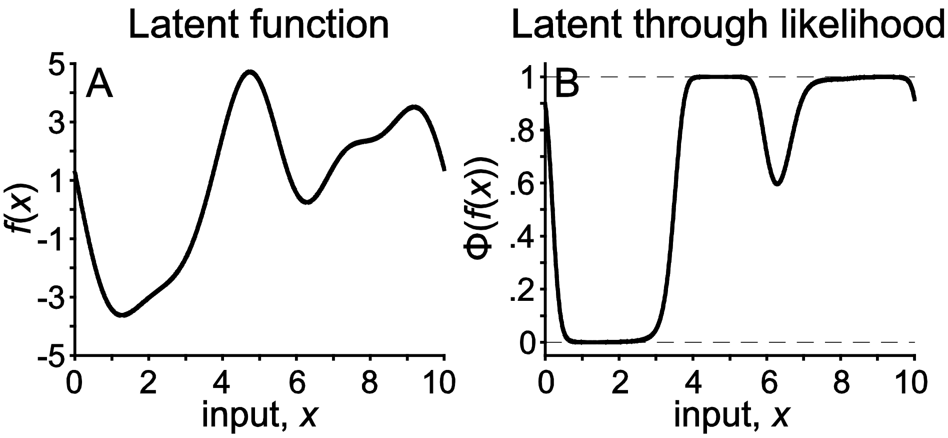

Figure 2.2: Illustration of a latent function passed through a sigmoidal likelihood. (A) A sample latent function $f\left(x\right)$ drawn from a Gaussian process. (B) The class probability $\Phi\left(f\left(x\right)\right)$ obtained by "squashing" this sample function through the logistic likelihood $\Phi\left(z\right)=1/\left(1+e^{-z}\right)$.

As with the regression case, the predictive posterior distribution $p\left(f^{*}\left|\mathbf{X},\mathbf{y},\mathbf{x}^{*}\right.\right)$ can be expressed using (2.4), but now with an explicit sigmoidal likelihood function:

$$\begin{aligned} p\left(f^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right) &= \frac{1}{p\left(\mathbf{y} \middle| \mathbf{X}\right)} \int p\left(\mathbf{y} \middle| \mathbf{f}\right) p\left(\mathbf{f}, f^* \middle| \mathbf{X}, \mathbf{x}^*\right) d\mathbf{f} \\ &= \frac{1}{p\left(\mathbf{y} \middle| \mathbf{X}\right)} \int \prod_{i=1}^{n} \Phi\left(f_i\right) p\left(\mathbf{f}, f^* \middle| \mathbf{X}, \mathbf{x}^*\right) d\mathbf{f} \end{aligned} . \tag{2.7}$$

Rather than the posterior distribution on the underlying latent function $p\left(f^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right)$, however, the classification scheme is often interested in the posterior distribution of positive response probability $p\left(y^* = 1 \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right)$. Combining Equations 2.6 and 2.7 yields the probability of class identity for a test observation y^*:

$$\begin{aligned} p\left(y^* = 1 \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right) &= \int p\left(y^* = 1 \middle| f^*\right) p\left(f^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right) df^* \\ &= \int \Phi\left(f^*\right) p\left(f^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right) df^* \end{aligned} . \tag{2.8}$$

Unlike in the regression case defined by the Gaussian likelihood (2.2), the sigmoidal likelihood for classification in (2.6) makes the integrals in (2.7) and (2.8) analytically intractable. Instead, this posterior distribution must be estimated, either by sampling methods such as Markov Chain Monte Carlo (Neal, 1993; Andrieu *et al.*, 2003) or by using a Gaussian approximation to the posterior distribution. Some common Gaussian approximations are the Laplace approximation, which uses a second-order Taylor expansion to match the curvature of the distribution at the mode (Williams and Barber, 1998), and expectation propagation, which attempts to match the first two moments of the distribution (Minka, 2001).

2.1.5. Covariance Functions

The GP covariance function describes correlations between latent function values at different inputs and can be used to represent structure in information about f. While we can always numerically specify a covariance matrix K for a finite set of observations, a covariance function (or kernel) $K\left(x, x'\right)$ provides a general framework for encoding relationships between function values on an unrestricted input domain. The covariance function does not typically specify an exact form for the latent function f, but it can encode systematic relationships between function values for specific sets of inputs. Importantly, several off-the-shelf covariance functions can be effectively used to model behavior for a wide range of latent function shapes. One such example is the squared exponential (SE) covariance function:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = s^2 \left[\frac{-\left(\mathbf{x} - \mathbf{x}'\right)^2}{2\ell^2} \right], \tag{2.9}$$

where s^2 represents the maximum covariance and ℓ represents a length constant. The covariance function value $K\left(\mathbf{x}, \mathbf{x}'\right)$ in this case is large when the values of $\mathbf{x}$ and $\mathbf{x}'$ are close and falls off with the square of the distance between them. The length constant ℓ acts as a normalization term to determine the distance needed for a particular change in covariance, effectively representing a measure of smoothness. The SE covariance function is very flexible because it simply specifies that, relative to ℓ, function values at points near one another are highly correlated, while those at points far away are not. Therefore, it is able to represent a wide range of latent function shapes so long as they are generally smooth. Other covariance functions exist that can capture other aspects of latent function behavior, such as linearity, periodicity, or chaotic behavior (Rasmussen and Williams, 2006; Duvenaud, 2014).

2.1.6. Hyperparameters

Mean and covariance functions often have associated parameters, such as the s and ℓ terms in Equation 9. These parameters of the moment functions (and not of the latent function itself) are called hyperparameters. Different choices of hyperparameters can have significant effects on latent function behavior. For instance, a large ℓ in Equation 9 will result in a posterior that favors much smoother latent functions compared to a small ℓ. **Figure 2.3** shows three different unidimensional GP regression predictions for the same observed data set. Each GP utilizes a SE covariance function (2.9) but differs in length scale ℓ. Note that short length scales allow for more local fluctuations while long length scales enforce more global smoothness.

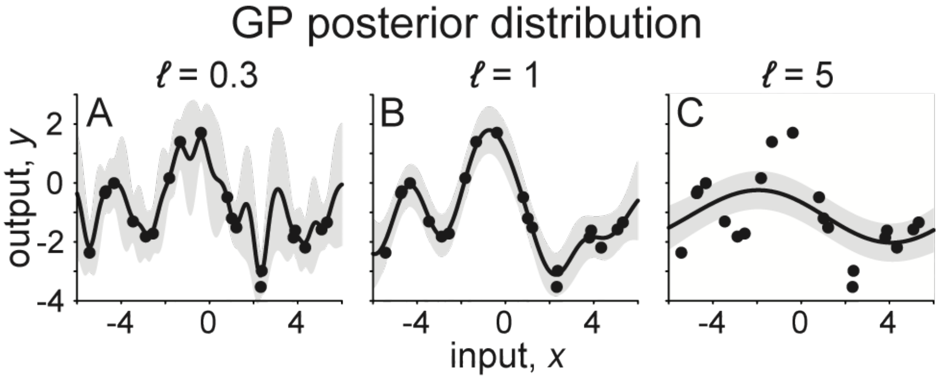

Figure 2.3: Illustration of the effects of a length scale hyperparameter. For the same set of observations (black dots), the subplots show posterior distributions for SE covariance functions with (A) ℓ = 0.3, (B) ℓ = 1, and (C) ℓ = 5. Solid lines denote the posterior mean and shaded gray areas denote 2 standard deviations about the mean.

It is sometimes sufficient to directly specify the hyperparameters $\boldsymbol{\theta}$ of the GP, assuming we have prior knowledge that compels us to do so. Often, however, we do not know the values of these hyperparameter *a priori* and must estimate them from the data. Assume we have chosen a GP prior $p\left(f \middle| \mathbf{x}, \boldsymbol{\theta}\right) = \mathcal{GP}\left(f; \mu\left(\mathbf{x}, \boldsymbol{\theta}\right), K\left(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}\right)\right)$. We can quantify the quality of fit to the observed data by computing the marginal likelihood, or the probability of observing the given data under the selected prior:

$$p\left(\mathbf{y}\middle|\mathbf{X},\boldsymbol{\theta}\right) = \int p\left(\mathbf{y}\middle|\mathbf{f},\mathbf{X},\boldsymbol{\theta}\right)p\left(\mathbf{f}\middle|\mathbf{X},\boldsymbol{\theta}\right)d\mathbf{f}, \tag{2.10}$$

where we marginalize out the unknown function values $\mathbf{f}$. For a Gaussian likelihood in a typical regression framework, this probability can be analytically computed. In the case of classification, the integral in the expression is analytically intractable due to the nonlinear sigmoidal likelihood, so we must use a Gaussian approximation for the marginal likelihood (Williams and Barber, 1998; Minka, 2001; Rasmussen and Williams, 2006).

One method of obtaining a "best fit" to the observed data is to choose a set of hyperparameters $\boldsymbol{\theta}$ that maximizes their log marginal likelihood $\log p\left(\mathbf{y}\middle|\mathbf{X},\boldsymbol{\theta}\right)$. This is done by taking the partial derivative of the log marginal likelihood with respect to each hyperparameter: $\frac{\partial}{\partial\theta_j}\log p\left(\mathbf{y}\middle|\mathbf{X},\boldsymbol{\theta}\right)$. For the case of classification, an approximation technique is used to estimate the log marginal likelihood, whose partial derivative is then computed. This best-fit hyperparameter selection process is an application of the second level in Bayesian hierarchical inference, in which the most appropriate model is chosen given the data (Jefferys and Berger, 1992; Gibbs, 1998; Xiang and Fackler, 2015). We can also choose specific hyperpriors (priors on hyperparameter distributions) given compelling prior information, but a standard choice is a uniform hyperprior, representing maximum entropy.

2.2. Active Sampling

One common requirement in any inference framework is to have a set of observed data points from which the model can be built. In the context of parametric modeling, these observations are used to select best-fitting parameters; in the context of Bayesian techniques, these observations are often used to generate a posterior distribution; and in the context machine learning, these observations are used to "train" the model for prediction of unobserved "test" cases.

Often, the exact data that are observed are not under the control of the experimenter. For instance, the data may be a set of medical records from patients in routine hospital visits or financial data from investors in the stock market, neither of which can be scheduled or decided on by the experimenter. However, in other cases, particularly in the design of a new scientific experiment or routine, the experimenter does have control over which data are to be collected.

Many standard procedures for selecting observations exist. Examples include random sampling, grid sampling (in which we select a regularly-spaced set of points in the query space), or space-filling samples such as the Halton sequence (Halton, 1964). In reality, however, time and computing power are typically limited resources, and it is advantageous to carefully choose observations that are most useful for our model. Not all potential observations are equally informative, and the concept of active sampling broadly explores the question of how, at any given time, to select the "best" sample to query.

2.2.1. Definition of Active Sampling

Broadly, active sampling (typically called "active learning" in the machine learning or "optimal experimental design" in statistics) is a system in which an algorithm is able to choose the labeled data on which it performs inference, with the goal of improving performance for small amounts of data (Olsson, 2009; Settles, 2009). Active sampling procedures are particularly useful in situations when data can be collected iteratively (rather than, say, a situation in which all data is already available), or when the resource cost of obtaining each new labeled data point is high (for instance, each observation could represent the result of a scientific experiment). Although active sampling can be utilized for both regression and classification machine learning problems, we will focus our attention on the classification case, which is relevant to this thesis.

Three common active sampling scenarios are described in the literature:

- *Stream-based sampling:* In this scenario, unlabeled points are sequentially drawn from some distribution, and the learner can select whether to query each example's label or discard it (Atlas *et al.*, 1989; Cohn *et al.*, 1994).
- *Pool-based sampling:* In this scenario, a finite set of unlabeled points are available for query, and the learner selects the most informative point or set of points whose labels should be obtained (Lewis and Gale, 1994).
- *Query synthesis:* In this scenario, the learner has precise control over the qualities or parameters of points that are queried within the input space, effectively generating each query instance *de novo* (Angluin, 1988; 2004)

Note that with a large pool of unlabeled instances that spans the input space with sufficiently high resolution, pool-based sampling can effectively imitate query synthesis. For instance, if the input space is the frequency and intensity of a pure-tone sound stimulus, we can construct a pool whose members contain each unique frequency/intensity pair from a high-resolution range for each of frequency and intensity, effectively approximating full control over stimulus parameters up to our predefined resolution. We will therefore focus the active sampling development in this section on pool-based sampling, with the assumption of a sufficiently large pool.

Figure 2.4 shows a diagram of a standard pool-based active sampling cycle (Settles, 2009). Let us first define a pool $\mathbf{X}^*$ from which new queries can be drawn. At each iteration, we use the set of currently observed data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ to learn a model using our algorithm of choice. From the pool $\mathbf{X}^*$, we then select a new sample point $\mathbf{x}$ which we have deemed optimal, and obtain its corresponding label y. (We can also choose more than one sample point at a time, but we focus here on single iterative queries.) This new observation $\{\mathbf{x}, y\}$ is then added to the set of observed data $\mathcal{D}$, and this cycle repeats until termination. Note that we have used $\mathbf{X}^*$ here to describe the pool, which is also used to define our test set for GP inference (see Section 2.1); this is deliberate because in the work presented in this thesis, our pool is defined as the test set itself, which limits the space of possible queries to the values within the test set.

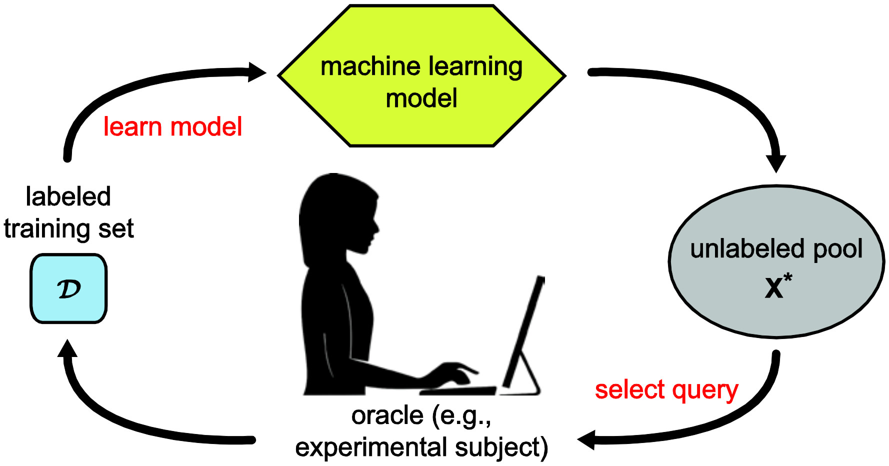

Figure 2.4: A diagram of a standard active learning cycle. In each cycle, a machine learning model is trained on some observed data, and then a new query is selected from the unlabeled pool, its label is given by the oracle, and the new observation is added to the set of labeled data. Reproduced with minor changes from (Settles, 2009).

There are many methods for selecting the optimal point on each cycle, including more local schemes such as uncertainty sampling (Lewis and Catlett, 1994; Lewis and Gale, 1994) and query-by-committee (Seung *et al.*, 1992) as well information-theoretic metrics such as expected model change (Settles *et al.*, 2008), expected error reduction (Roy and McCallum, 2001), and variance reduction (Cohn, 1994). To broadly unify these distinct strategies, we use the concept of an acquisition function, borrowing terminology from Bayesian optimization literature (Guestrin *et al.*, 2005; Osborne *et al.*, 2009; Brochu *et al.*, 2010). On any given iteration, the acquisition function $A(\mathbf{x})$ quantifies the desirability of an input point $\mathbf{x}$. To select the next query point, we select $\hat{\mathbf{x}}^*$ from among the points in the unlabeled pool $\mathbf{X}^*$ corresponding to the highest value of $A(\mathbf{x})$, i.e. $\hat{\mathbf{x}}^* = \arg\max_{\mathbf{x}\in\mathbf{X}^*} A(\mathbf{x})$.

The acquisition function $A(\mathbf{x})$ can be defined using any metric we desire, but is generally a value that can be computed based on the model learned from the current set of observed data $\mathcal{D}$. Information-theoretic metrics typically use the concept of expected utility from decision theory (Berger, 1985), quantifying a utility function U and selecting the point at each iteration with the highest expected utility $U(\mathbf{x}|\mathcal{D})$ (Chaloner and Verdinelli, 1995; Park, 2013). Such approaches can be conceptualized as performing a 1-step look-ahead, unlike greedier methods which often choose quantities computed directly from the current model (Settles, 2009). Two particular active sampling methods used in this thesis are described in Sections 2.2.2 and 2.2.3.

It is noteworthy that many of the “adaptive sampling” techniques for PFs described in Section 1.4.2 can be conceptualized within the active sampling framework. While procedural techniques such as staircase methods (Dixon and Mood, 1948; Levitt, 1971; Kingdom and Prins, 2010) do not fall into this category, other psychometric sampling techniques do iteratively choose the next sample point to minimize or maximize some objective function. To offer a few examples: the best PEST method (Pentland, 1980) can be assigned an acquisition function of the Bernoulli variance: $A\left(x\right)=p\left(x\right)\left(1-p\left(x\right)\right)$; the Ψ-method (Kontsevich and Tyler, 1999) can be assigned an acquisition function of the negative expected entropy: $A\left(x\right)=-\mathbb{E}\left[H_t\left(x\right)\right]$.

2.2.2. Uncertainty Sampling

As perhaps the most commonly employed active sampling strategy, uncertainty sampling queries the instance for which the uncertainty is the highest (Lewis and Catlett, 1994; Lewis and Gale, 1994). A logical quantity to represent uncertainty is the posterior variance. In the case of binary classification, points of highest variance correspond to points whose probability of belonging to class 1 is closest to 0.5; which can be shown using the variance of a Bernoulli distribution: $p\left(x\right)\left(1-p\left(x\right)\right)$. This uncertainty measure can also be interpreted as the expected 0/1-loss, or the model's belief that it will mislabel any point (Settles, 2009).

In the context of GP classification, y is the value that quantifies the probability of belonging to class 1. The posterior variance of y, σ_y^2, is provided by the computation of the posterior distribution $p\left(y^* = 1 \middle| \mathbf{X}, \mathbf{y}, \mathbf{x}^*\right)$ and is a logical choice for the uncertainty sampling framework. Note that the highest values of σ_y^2 also correspond to the points at which the posterior mean of y, μ_y, are closest to 0.5. Therefore, we use the following acquisition function for this method:

$$A\left(\mathbf{x}\right)=\sigma_y^2\left(\mathbf{x}\right), \tag{2.11}$$

where $\sigma_y^2\left(\mathbf{x}\right)$ refers to the posterior variance of y at the input point $\mathbf{x}$.

2.2.3. Bayesian Active Learning by Disagreement

Bayesian active learning by disagreement, or BALD (Houlsby *et al.*, 2011) is an information-theoretic approach in which we seek to reduce the number of possible hypotheses maximally fast (Cover and Thomas, 2012). We therefore seek the point that results in the maximal decrease in posterior entropy: $H\left[\boldsymbol{\theta} \middle| \mathcal{D}\right] - \mathbb{E}_{y}\left[H\left[\boldsymbol{\theta} \middle| \mathcal{D}, \mathbf{x}, y\right]\right]$, where $\boldsymbol{\theta}$ is the set of model parameters and H is Shannon's entropy (Shannon, 2001), an uncertainty measure. The first term is the current entropy and the second term is the expected entropy after having observed data point $\left\{\mathbf{x}, y\right\}$. As shown in (Houlsby *et al.*, 2011), this expression in possibly infinite-dimensional parameter space can be rewritten in low-dimensional y space: $H\left[y \middle| \mathbf{x}, \mathcal{D}\right] - \mathbb{E}_{\boldsymbol{\theta}}\left[H\left[y \middle| \mathbf{x}, \boldsymbol{\theta}\right]\right]$. This expression is maximized when the first term is high (model is marginally very uncertain about y), but the second term is low (individual settings of $\boldsymbol{\theta}$ are very confident). We can therefore interpret this expression as the degree to which parameters under the posterior disagree (Houlsby *et al.*, 2011).

In the context of GP classification, the parameter set $\boldsymbol{\theta}$ becomes the infinite-dimensional latent parameter f, e.g. $H\left[y \middle| \mathbf{x}, \mathcal{D}\right] - \mathbb{E}_{f}\left[H\left[y \middle| \mathbf{x}, f\right]\right]$. By using several approximations, we can write the following expression for the acquisition function (Houlsby *et al.*, 2011):

$$A\left(\mathbf{x}\right) = \mathrm{h}\left(\Phi\left(\frac{\mu_f(\mathbf{x})}{\sqrt{\sigma_f^2(\mathbf{x})+1}}\right)\right) - \frac{\sqrt{\frac{\pi\ln 2}{2}}\exp\left(\frac{\mu_f^2(\mathbf{x})}{2\left(\sigma_f^2(\mathbf{x})+\frac{\pi\ln 2}{2}\right)}\right)}{\sqrt{\sigma_f^2(\mathbf{x})+\frac{\pi\ln 2}{2}}}, \tag{2.12}$$

where $h\left(p\right) = -p\log_2 p - \left(1-p\right)\log_2\left(1-p\right)$ is the binary entropy function, $\mu_f(\mathbf{x})$ and $\sigma_f^2(\mathbf{x})$ are respectively the posterior mean and posterior variance of f corresponding to the input point $\mathbf{x}$, and Φ is the sigmoidal likelihood function for classification (2.6).

2.3. Concluding Remarks

The Gaussian process is a Bayesian inference framework that encodes relationships between variables rather than requiring a parametric form for the function to be estimated. Given an appropriate choice of mean and covariance functions, it can capture a diverse set of function behaviors, and can also incorporate prior constraints on function shapes given prior information. When trained on some (possibly binary) observed data, the GP posterior provides an entire probability distribution on test points, rather than point estimates.

Taken together, these qualities make the GP an attractive framework for performing inference on audiometric functions. Its nonparametric nature supports various audiogram shapes, which cannot be easily parameterized, and its estimation of entire probability distributions allows for painless integration with active sampling frameworks. Overall, the GP represents a flexible and efficient framework for performing audiometric inference.

Chapter 3: Automated Estimation of Human Threshold Audiograms Using Active Machine Learning

Note: The research presented in Chapter 3 has been published in Ear and Hearing (Song *et al.*, 2015).

3.1. Introduction

As described in Chapter 1, current methods of determining a threshold audiogram exhibit many shortcomings. The clinical Hughson-Westlake staircase method (Carhart and Jerger, 1959; Katz *et al.*, 2009), along with the numerous automated techniques that replicate the procedure (Mahomed *et al.*, 2013), provide thresholds only at a small number of (6-9) standard audiogram frequencies. Moreover, the procedure for determining threshold at any particular frequency is both inefficient and predictable; it presents multiple identical stimuli and selects tones at sound levels where the listener's response is already quite certain.

To address primarily the first shortcoming above, techniques that sweep tone stimuli through multiple frequencies have been developed, including Békésy audiometry and Audioscan® (von Békésy, 1947; Meyer-Bisch, 1996; Ishak *et al.*, 2011). While these techniques can in fact provide relatively continuous threshold curves as a function of frequencies, they show some limitations. Békésy audiometry is comparatively quick for a sweep-based method, but results in a somewhat "jagged" estimate of the threshold curve that lacks specificity along the intensity dimension. On the other hand, Audioscan® offers a smoother estimate; however, the estimate is still quantized to discrete intensity levels, and the procedure is much more time-intensive to perform.

A particularly promising set of audiometric procedures have been Bayesian methods (Özdamar *et al.*, 1990; Stadler, 2009). Unlike standard procedural methods such as HW, these methods select at every iteration the optimal stimulus frequency and intensity to present, informed by both prior data and the set of all other responses collected so far. The use of all observed responses across multiple frequencies to inform the current estimate stands in stark contrast to the HW procedure, in which all samples for a particular frequency are discarded after the corresponding threshold has been determined. Studies have shown large efficiency gains using these Bayesian techniques, but like standard clinical techniques, these methods still constrain the choice of possible frequencies to the 6-9 standard locations.

This chapter describes the development of the first version of the machine learning audiogram (MLAG), which is designed for estimating the threshold audiogram. The algorithm utilizes Gaussian process classification (GPC) and can be conceptualized as an extension of the Bayesian techniques: at every iteration, prior constraints and the set of all observations are used to form an estimate, and an optimal query point is selected. The final threshold estimate is approximately continuous along both frequency and intensity dimensions, and active selection of samples allows for efficient threshold audiogram estimation.

3.2. Methodology

3.2.1. Machine Learning Algorithm

For the machine learning (ML) audiogram algorithm, we employed Gaussian process (GP) classification (Rasmussen and Williams, 2006) to construct tone detection audiogram estimates from human listeners in real time. We also utilized active sampling procedures in order to select the most useful queries for each iteration. Sections 2.1 and 2.2 describe GPs and active sampling in more mathematical detail. The details of GP setup and algorithm functionality specific to the current technique are described below.

Variable space: For audiogram estimation, the input variable $\mathbf{x}$ is the frequency and intensity of presented pure tones, i.e. $\mathbf{x} = (\omega, \iota)$. Any observation y_i is a binary variable encoding whether or not a tone $\mathbf{x}_i$ was detected: 0 for undetected and 1 for detected. The GP was trained to predict the probability of a listener's tone detection as a function of these variables $p\left(y = 1 \middle| \mathbf{x}\right)$, which takes on continuous values between 0 and 1 over all combinations of frequency and intensity. For inference, we choose a finely spaced test grid $\mathbf{X}^*$ of samples: 0.125 to 16 kHz in semitone increments for frequency and −20 to 120 dB in 1-dB increments for intensity. This allows us to compute a posterior distribution on effectively the entire input space of interest.

Mean function: As typical for GP inference, we selected a constant mean function: $\mu\left(\mathbf{x}\right) = c$. Although this assumes that the prior central tendency of the latent function is flat, all of the variation around the mean can be captured using the covariance function $K\left(\mathbf{x}, \mathbf{x}'\right)$.

Covariance function: To pick an appropriate GP covariance function for this application, we incorporated constraints that reflected prior knowledge about psychometric functions. Most crucially, the probability of listener detecting a tone is monotonically increasing as a function of tone intensity, but need not have an explicit dependence upon frequency except that the overall function is continuous. To reflect this scenario, we used a monotonic linear kernel $K\left(\iota, \iota'\right) = s^2\left(\iota \cdot \iota'\right)$ in the intensity dimension and a more flexible squared exponential (SE) kernel $K\left(\omega,\omega'\right)=s^2\exp\left[-\left(\omega-\omega'\right)^2/2\ell^2\right]$ in the frequency dimension. The overall covariance function in $\mathbf{x}$ was therefore a sum of these individual components, plus a noise term:

$$K\left(\mathbf{x},\mathbf{x}'\right)=K\left(\left(\omega,\iota\right),\left(\omega',\iota'\right)\right)=s_1^2\exp\left[\frac{-\left(\omega-\omega'\right)^2}{2\ell^2}\right]+s_2^2\left(\iota\iota'\right)+s_{\text{noise}}^2. \tag{3.1}$$

Likelihood function: Following standard GP classification procedure and to ensure that the GP returned a probability estimate in the range $\left[0,1\right]$, we transformed the latent function using a cumulative Gaussian likelihood function: $p\left(y_i=1\middle|f_i\right)=\Phi\left(f_i\right)=\int_{-\infty}^{f_i}\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz$. This choice is consistent with standard PF inference frameworks that model PFs as monotonically increasing sigmoidal functions (Wichmann and Hill, 2001a; Kingdom and Prins, 2010).

Computation of hyperparameters: We used 5 covariance function hyperparameters for this model: a mean constant c, a scaling factor s_1 for the SE component, a characteristic length scale ℓ for the SE component, a scaling factor s_2 for the linear component, and a noise parameter s_{noise}, i.e. $\boldsymbol{\theta}=\left(c,s_1,\ell,s_2,s_{\text{noise}}\right)$. A best-fitting set of hyperparameters was chosen automatically after each response by maximizing the log marginal likelihood of these hyperparameters with respect to the sampled data, following the procedure outlined in Section 2.1.6.

Calculation of posterior distribution: Following each new tone presentation, all data sampled up to that point were used to compute the predictive posterior distribution $p\left(y^*=1\middle|\mathbf{X},\mathbf{y},\mathbf{x}^*\right)$. At each iteration, we computed the posterior mean, representing the current point estimate about the detection probability, and the posterior variance, representing uncertainty on the estimates at each input location in $\mathbf{X}^*$. An example for one run can be seen in **Figure 3.1**: Figure 3.1A shows an example of the posterior mean during data acquisition for one audiogram estimate and Figure 3.1B shows the corresponding posterior variance.

Active sampling: After initializing with a few pseudorandom samples, the ML algorithm selects only points deemed to be informative to the estimate for subsequent samples. We used an uncertainty sampling acquisition framework (see Section 2.2.2) in which at each iteration, the next chosen sample point was one whose class identity (i.e. 1 or 0) was the most uncertain (Lewis and Catlett, 1994; Lewis and Gale, 1994; Settles, 2009). Based upon the calculated posterior variance of y, σ_y^2, for the current iteration (2.11), we selected the frequency/intensity pair from the test set $\mathbf{X}^*$ corresponding to the highest value in the variance function as the next point to sample (Figure 3.1B). If multiple points were tied for maximum variance, a point was selected at random from this set. After determining the listener's response, the posterior distribution was updated for the next iteration (updated posterior mean shown in Figure 1C). This cycle of hyperparameter estimation, posterior calculation, and uncertainty sampling was repeated until convergence criteria were met, which are detailed in Section 3.2.3.

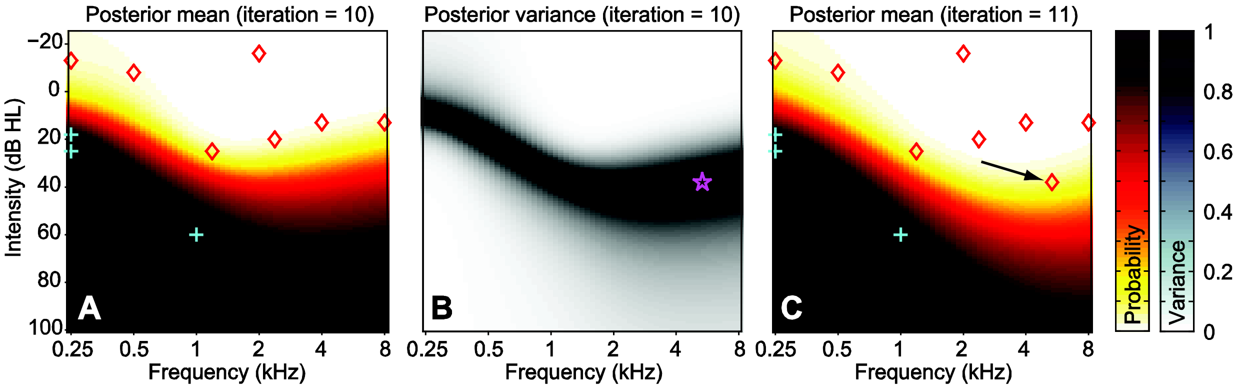

Figure 3.1: Illustration of the machine learning (ML) audiogram technique. (A) Posterior mean is computed by the GP using the sampled points. Red diamonds indicate the tone was inaudible; blue pluses, audible. (B) Posterior variance is computed by the GP using the sampled points, and the point of maximum variance is identified (purple star). (C) The point of maximal variance is queried for listener audibility (black arrow). In this example once it is determined that the listener did not hear the tone, the updated set of points is used by the GP to re-compute the posterior mean with a more elevated threshold near the corresponding frequency.

3.2.2. Participants

A total of 21 participants (8 male, 13 female) were recruited from the Department of Adult Audiology at Washington University School of Medicine Central Institute for the Deaf and the Research Participant Registry at Washington University in St. Louis. All participants were between 18 and 90 years of age (mean 47), fluent English speakers and with no history of neurological disorder. Approval for completion of the study was received from Washington University in St. Louis' Human Research Protection Office (HRPO), and all participants provided informed consent before any testing protocol began. One listener (listener number 17) fell asleep during one part of the study. This listener's data were therefore omitted from the group averages but were presented separately to demonstrate how the algorithm operates with a noncompliant listener (see Discussion).

3.2.3. Experimental Procedure

For each listener, we performed 2 repetitions of the automated ML-based audiogram and 1 repetition of a standard manual HW audiogram. An air-conduction pure-tone audiogram was obtained in each case, and each auditory stimulus consisted of a three-pulse sequence of 200-ms pure tones with inter-pulse intervals of 200 ms. Listeners were seated within a sound isolation booth, and all auditory stimuli were delivered using a Toshiba Portege R700 laptop computer running custom MatLab code and Sennheiser HD280 circumaural headphones. Computer audio output was calibrated to match the output of a GSI-61 two-channel clinical audiometer. The relative order for the ML and HW audiograms was randomized for each listener, and experimenters conducting the HW audiogram were blinded to the listeners' ML audiogram scores. Listeners were asked to remove any hearing-assist devices prior to data collection. Short periods of rest (~2 mins) were administered between each set of audiogram runs. Figure 3.2 shows a photo of the sound isolation booth used to conduct these experiments.

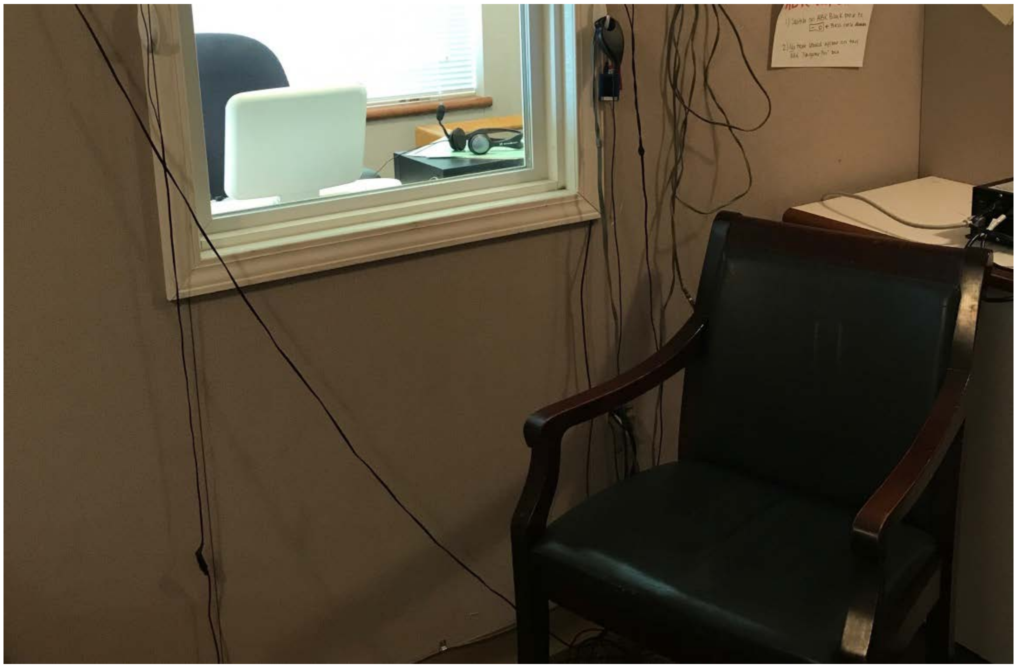

Figure 3.2: Photo of the sound isolation booth used to conduct experiments.

Manual HW audiometry: A conventional audiogram was conducted by an audiologist according to accepted standards (American National Standards Institute, 2004a; American Speech-Language-Hearing Association, 2005). Each listener was instructed to raise his or her hand upon detection of a presented pure-tone stimulus. Hearing ability was assessed at standard audiogram frequencies (0.25, 0.5, 1, 2, 4, and 8 kHz), with the possible intensity ranging from -20 to 100 dB HL in a minimum of 5-dB increments. For an individual frequency, a pure tone was first presented at an audible intensity based upon the audiologist's clinical judgment, then reduced in 10-dB increments until the listener failed to respond. Henceforth, the intensity was increased in 5-dB increments following detected tones and decreased in 10-dB increments following undetected tones. The threshold for that frequency was determined by the lowest-intensity tone to elicit a response in at least 2 of 3 ascending trials. The manual audiogram was conducted separately for left and right ears. This manual method is the modified Hughson-Westlake ascending-descending procedure and is referred to here as HW audiometry (Hughson and Westlake, 1944; Carhart and Jerger, 1959; Katz *et al.*, 2009).

Automated ML-based audiometry: The ML framework was incorporated into a user interface for real-time integration of listener responses. Listeners were instructed to click a mouse button upon detection of any stimulus. Each stimulus was separated by a randomized inter-trial interval of between 500 and 2000 ms to minimize listener prediction of stimulus presentation times. A response within 1500 ms following the onset of the tone sequence was marked as a detected sample (value of 1); no response was counted as an undetected sample (value of 0). The range of possible sample points fell within 250-8000 Hz in semitone increments centered at 1000 Hz along the frequency dimension, and −25-100 dB HL in 1-dB increments centered at 0 dB HL along the intensity dimension. Sampling was initially conducted pseudo-randomly throughout both frequency and intensity space until at least 1 sample was collected at each standard audiogram frequency (0.25, 0.5, 1, 2, 4, and 8 kHz) and at least one detected and one undetected sample had occurred. After this point, the algorithm followed the iteration cycle of hyperparameter training, posterior estimation and informative sampling of next stimulus as described in Section 3.2.1. This cycle was iterated for a minimum of 36 presentations and until two specific convergence criteria were met: 1) the average posterior variance and 2) the posterior mean change since the previous iteration were both sufficiently low. This automated ML audiometry procedure is summarized by the diagram provided in **Figure 3.3**.

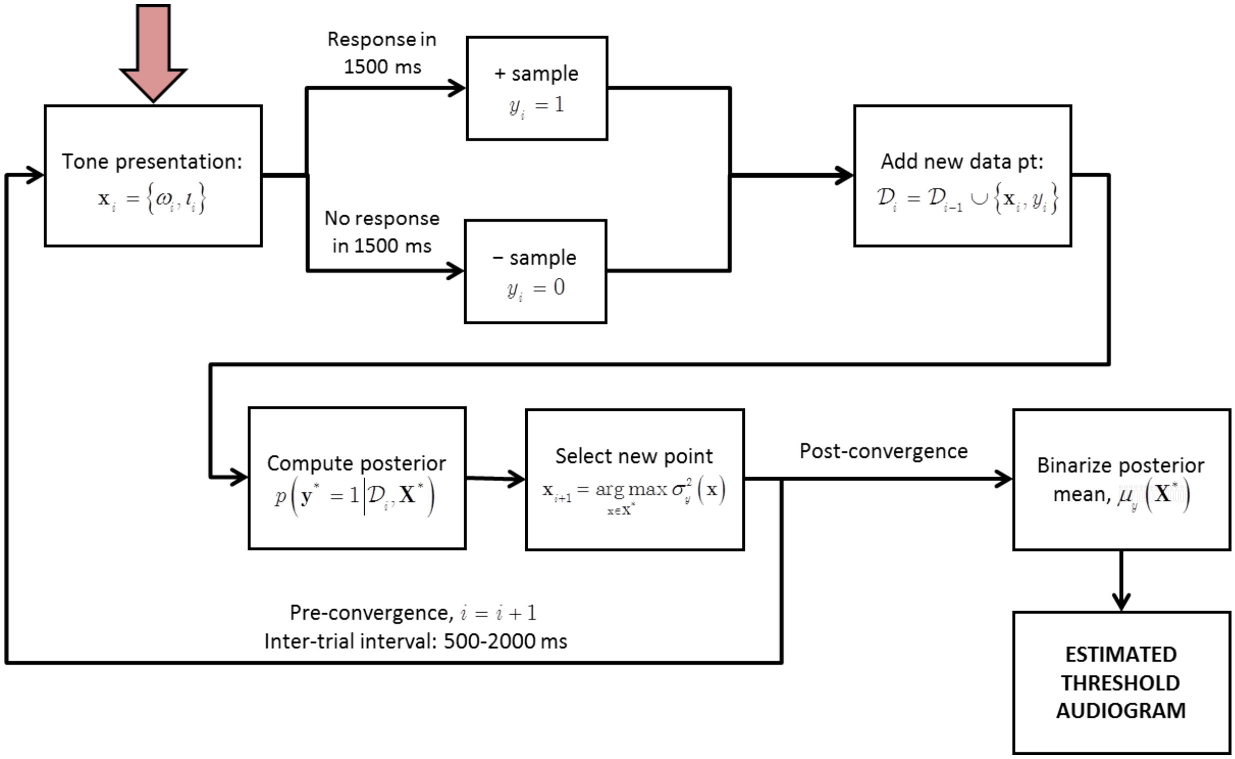
Response in 1500 ms
+ sample
$y_i = 1$
Tone presentation:
$\mathbf{x}_i = \{\omega_i, \iota_i\}$
No response in 1500 ms
− sample
$y_i = 0$
Add new data pt:
$\mathcal{D}_i = \mathcal{D}_{i-1} \cup \{\mathbf{x}_i, y_i\}$
Compute posterior
$p(\mathbf{y}^* = 1 | \mathcal{D}_i, \mathbf{X}^*)$
Select new point
$\mathbf{x}_{i+1} = \arg\max_{\mathbf{x} \in \mathbf{X}^*} \sigma_y^2(\mathbf{x})$
Post-convergence
Binarize posterior mean, $\mu_y(\mathbf{X}^*)$
Pre-convergence, $i = i + 1$
Inter-trial interval: 500-2000 ms
ESTIMATED THRESHOLD AUDIOGRAM

Figure 3.3: Diagram of the automated ML audiogram procedure. Red arrow indicates algorithm starting point.

“Heard” responses for which no tone presentations occurred within 1500 ms (i.e., false positives) were not used in evaluating the GP or in training the hyperparameters. The automated audiogram was conducted separately for left and right ears. To maximize user comfort, delivered tone intensities never exceeded 10 dB HL louder than the maximum intensity delivered up to that point in the test. Whether or not convergence criteria were met, the algorithm terminated after a maximum of 64 iterations.

3.2.4. Data Analysis

Following completion of the automated audiogram, we binarized each GP posterior mean at a detection probability of 0.707, the standard probability of a positive response at convergence for a transformed 2-up, 1-down method like the modified HW procedure (Levitt, 1971). Points for which the probability of detection was greater than or equal to 0.707 were labeled as "detected," and points for which the probability of detection was less than 0.707 were labeled as "undetected." This binary surface was then used to construct an estimate of the audiogram: for each frequency, the smallest intensity in 1-dB increments greater than the transition from "detected" to "undetected" was selected as the threshold value for that frequency. Because of the monotonic constraint enforced upon the estimator in the intensity dimension, there could be a maximum of only 1 transition point at each frequency. The threshold values at each frequency therefore become a continuous (in frequency) estimate of the listener's threshold audiogram.

The ML and HW threshold audiograms were compared at the standard audiogram frequencies. We assessed accuracy of the automated algorithm via comparison to the results of the HW audiogram by calculating 1) the mean difference and standard deviation of threshold between the ML and HW audiograms, 2) the mean absolute difference and standard deviation of threshold between the ML and HW audiograms, 3) the median absolute difference and interquartile range of threshold between the ML and HW audiograms, and 4) the percent 5-dB difference, or percentage of all ML audiogram values within 5 dB of the corresponding HW audiogram values (Swanepoel *et al.*, 2010; Mahomed *et al.*, 2013). We assessed test-retest reliability (precision) of the automated audiogram by 1) the mean difference and standard deviation of thresholds and 2) the absolute difference and standard deviation of thresholds between the audiogram estimates produced by the 2 runs of the ML algorithm (Mahomed *et al.*, 2013). Calibration correction was applied equally to both manual and automated estimates and therefore had no effect upon the comparisons between them; both methods used the same stimuli and the same hardware.

3.3. Results

The total number of stimulus presentations delivered to each listener for the manual HW and the two runs of the automated ML audiogram are shown in **Table 3.1**. This includes the samples presented to both the left and right ears. The HW procedure required an average of 97.0 ± 15.8 (mean ± standard deviation) samples to estimate the threshold audiogram, while the first and second runs of the ML procedure averaged 78.4 ± 11.0 and 78.9 ± 14.6 samples, respectively. This difference in number of samples between the HW audiogram and each run of the ML audiogram was statistically significant ($p = 0.0012$ and 1.5×10^{-4}, respectively; paired-sample t-test). Note that numerous runs of the ML audiogram terminated after 72 stimuli, the minimum number of samples after which the algorithm was allowed to terminate for each listener. Therefore, the actual mean number of stimuli required to achieve convergence criteria in the ML algorithm without this constraint is likely to be lower. All but 1 of the 40 included ML audiogram runs terminated prior to the maximum allowable number of iterations.

Listener Designation	# Samples (HW)	# Samples (ML 1)	# Samples (ML 2)
13	126	73	104
10	117	98	128
11	117	72	78
7	116	77	76
5	112	72	72
8	106	84	72
12	105	72	72
6	103	72	74
20	98	72	78
19	97	72	72
21	94	72	72
23	93	72	72
18	91	72	72
24	90	72	72
4	89	74	72
16	84	82	73
14	78	103	99
9	77	78	72
22	76	72	73
15	69	106	76
Mean	***97.0***	***78.4***	***78.9***
Standard deviation	***15.8***	***11.0***	***14.6***

Table 3.1: Summary of delivered samples for both HW and ML procedures. Total number of samples delivered by the HW and ML audiogram estimation procedures (across both ears) for each listener, presented in decreasing order of the number of HW samples required. The minimum and maximum numbers of ML audiogram samples allowed for the automated technique are 72 and 128, respectively. Listener 17's data are omitted because the listener fell asleep during part of the study.

Samples obtained during both the manual HW and automated ML methods are shown in **Figure 3.4** for one representative listener, with the final audiogram estimates shown as superimposed lines. It can be noted that the HW method searched each standard audiogram frequency across a number of intensities, with several repeat presentations of specific stimuli. The ML procedure, however, sampled across a more diverse set of frequencies with no repeats.

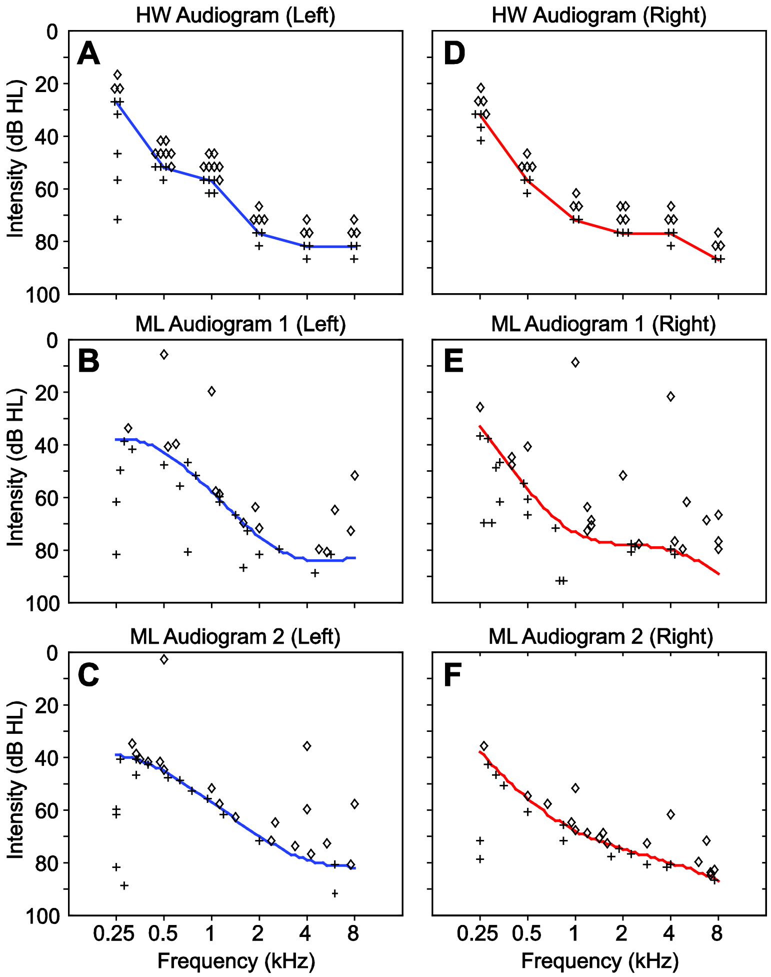

Figure 3.4: Sample plots of HW, ML1, and ML2 audiogram results. Plots of left- and right-ear audiograms obtained and samples conducted for a representative listener (Listener 4) using the manual HW technique (A, D), the first run of the ML algorithm (B, E), and the second run of the ML algorithm (C, F). Marks represent the frequencies and intensities of the stimuli that were presented, with pluses denoting listener detections and diamonds denoting misses. The superimposed curves are the final audiogram estimates produced by each technique. Note that the small displacements along the frequency axis in (A) and (B) only are to make repeat stimuli more visible and do not reflect actual deviations in the frequency of presented tones.

The degree of similarity among the different audiogram estimates for each ear is readily apparent, despite the differences in sampling procedure for each. The skilled audiologist was able to rapidly discover reversals and spent the most time probing right around threshold. A less skilled individual may have spent more time sampling points farther from threshold. These examples concisely demonstrate the utility of the HW procedure in trained hands and help explain why it is still in use many decades after its development.

Agreement between the 2 ML and 1 HW results for all ears can been seen visually in **Figure 3.5**. **Figure 3.6** compares HW and ML audiogram estimates for 3 specific ears: an ear with approximately normal hearing (Figure 3.6A), an ear with sloping high-frequency hearing loss (Figure 3.6B), and an ear with no-response at a subset of standard audiogram frequencies (Figure 3.6C). Again, the ML audiogram is able to produce a continuous audiogram estimate that compares favorably with the standard HW procedure at the standard audiogram frequencies. Moreover, while the HW procedure cannot provide a principled estimate at frequencies where no response was elicited, the ML procedure can and does, although the threshold estimate at 8 kHz in Figure 3C is not visible because of the limited range of values plotted. Hence, the similarity in estimates cannot be assessed at 8 kHz for this ear, but the ML estimate is likely closer to the actual threshold than any estimate that could be extrapolated from the HW data in this case.

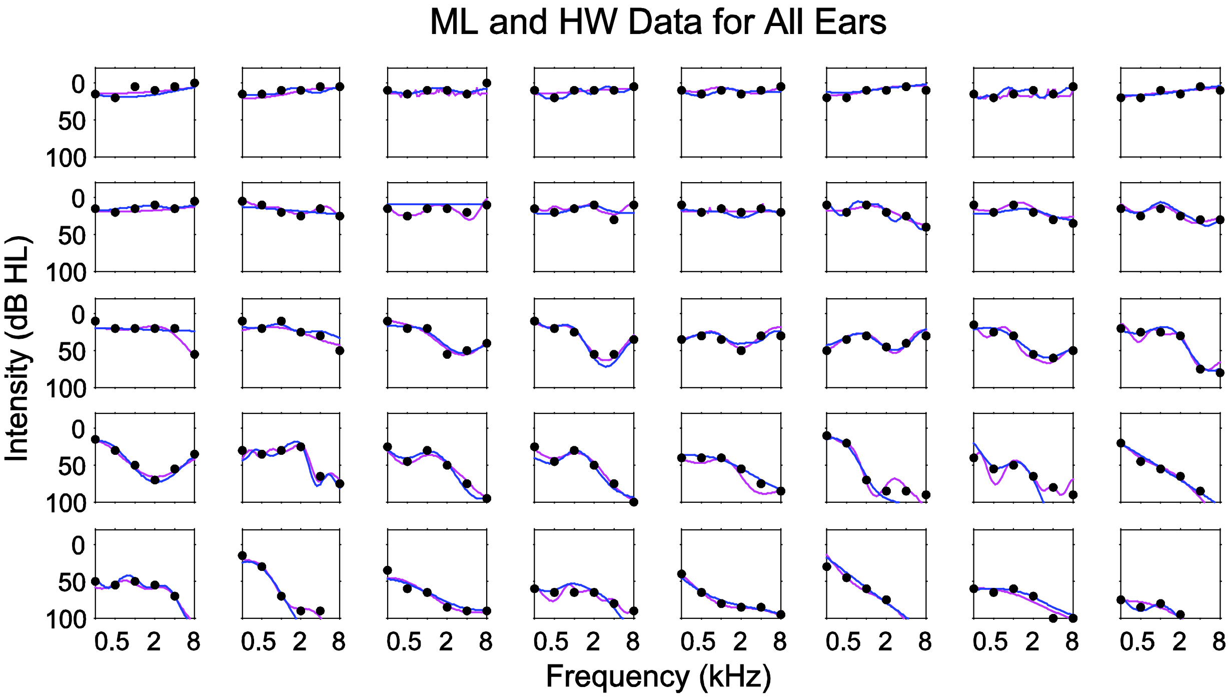

Figure 3.5: Agreement between ML and HW results for all valid ears. Magenta and blue curves show the results from the first and second runs of the ML audiogram, respectively, while black dots show the HW threshold results at the corresponding audiogram frequencies.

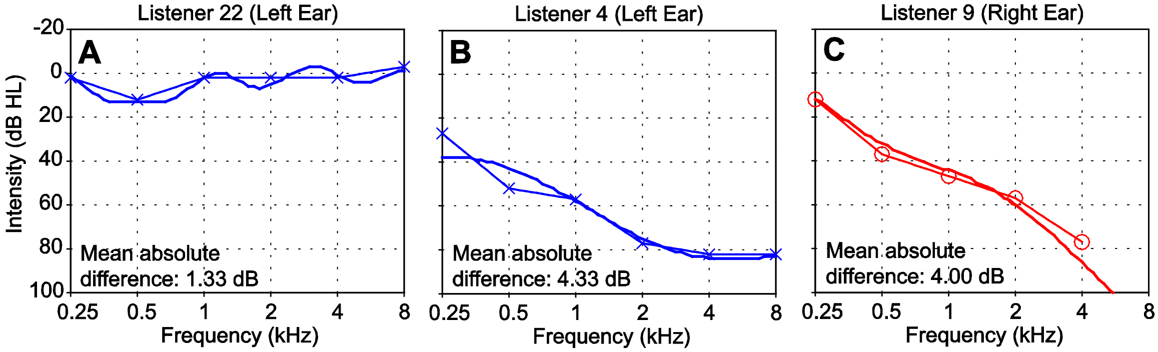

Figure 3.6: Sample HW and ML results for 3 distinct ears. Ears include (A) an ear with relatively normal hearing; (B) an ear with sloping high-frequency hearing loss; and (C) an ear with a no-response at 8000 Hz. “X” and “O” marks denote values estimated from the manual HW audiogram (connected by straight lines). The superimposed curves show the results from the automated ML audiogram.

Table 3.2 shows the results of evaluating the accuracy of the ML audiogram at standard audiogram frequencies relative to the HW audiogram, averaged across all listeners and runs. For the 6 standard audiogram frequencies, the mean estimated threshold difference was −0.011 ± 5.61 dB HL, the mean absolute estimated threshold difference was 4.16 ± 3.76 dB HL, the median absolute estimated threshold difference was 3.00 dB HL with an interquartile range of 5.00 dB HL, and the percent 5-dB difference in threshold estimates was 66.25. These values compare favorably with historical differences in audiogram estimation methodologies (Gosztonyi Jr. *et al.*, 1971; Schmuziger *et al.*, 2004; Ishak *et al.*, 2011; Mahomed *et al.*, 2013). Judging from the relatively low percent 5-dB difference yet comparable mean absolute difference, the ML procedure appears to produce somewhat more outlier estimates at individual frequencies than methods that estimate directly at those frequencies.

Frequency (kHz)	0.25	0.5	1	2	4	8	All
Mean differences and standard deviations vs. HW							
Mean difference (dB HL)	1.80	−1.43	0.138	0.244	1.14	−1.69	−0.011
Standard deviation (dB HL)	6.25	4.88	4.48	4.38	5.57	7.23	5.61
Average absolute differences and deviations vs. HW							
Mean absolute difference (dB HL)	4.80	3.75	3.44	3.53	4.48	5.17	4.16
Standard deviation (dB HL)	4.36	3.41	2.85	2.57	3.46	5.30	3.76
Median absolute difference (dB HL)	4.00	3.00	3.00	3.00	3.00	4.00	3.00
Interquartile range (dB HL)	6.00	4.00	4.00	3.00	4.25	4.25	5.00
Percent 5-dB maximum difference from HW							
Percent 5-dB max difference	61.25	82.5	80.0	78.75	61.25	48.75	68.75

Table 3.2: Differences between the ML audiogram estimate and the HW estimate.

We evaluated the ML audiogram's clinically relevant performance by classifying HW and ML audiogram results into conventional categories of hearing loss (normal, mild, moderate, severe, and profound) using the pure-tone average (Katz *et al.*, 2009; Stach, 2010). The categorical classifications produced by ML and HW audiogram estimates in our listeners were in agreement 95.0% of the time, and the disagreements in pure-tone average classification resulted in adjacent clinical categories. This result further suggests that the ML audiogram generates information that is clinically equivalent to the conventional HW audiogram using current standards.

Table 3.3 shows the results of evaluating the test-retest reliability of the automated ML audiogram at standard frequencies averaged across all listeners and estimation runs. Across all frequencies, the mean signed difference between automated audiogram runs was 0.75 ± 6.29 dB HL, the mean absolute difference between runs was 4.51 ± 4.45 dB HL, and the median absolute difference between runs was 3.00 dB HL with interquartile range 4.00 dB HL. These values are comparable to previously reported absolute test-retest differences for manual audiometry: 3.2 ± 3.9 dB HL (Fausti *et al.*, 1990; Swanepoel *et al.*, 2010; Mahomed *et al.*, 2013). This degree of similarity in final estimate between runs where the different initial randomization led to non-overlapping probe stimuli in the two cases indicates the robustness of the ML procedure.

Frequency (kHz)	**0.25**	**0.5**	**1**	**2**	**4**	**8**	**All**
Mean differences and standard deviations							
Mean difference (dB HL)	−0.15	1.55	1.63	0.26	1.03	0.032	0.75
Standard deviation (dB HL)	6.27	7.03	4.14	5.34	6.78	8.11	6.29
Average absolute differences and deviations							
Mean absolute difference (dB HL)	4.80	5.05	3.58	3.95	5.03	4.74	4.51
Standard deviation (dB HL)	3.97	5.07	2.60	3.55	4.59	6.52	4.45
Median absolute difference (dB HL)	5.00	3.00	3.00	3.00	4.00	3.00	3.00
Interquartile range (dB HL)	4.00	6.00	3.50	4.00	4.00	5.00	4.00

Table 3.3: Test-retest reliability of ML audiogram.

To further quantify the agreement between ML and HW estimates as well as ML reliability, we directly compared the threshold estimates from the first and second runs of ML and HW. This comparison is plotted in **Figure 3.7**; Figure 3.7A compares HW and the first ML run, Figure 3.7B compares HW and the second ML run, and Figure 3.7C compares the first and second runs of ML. Coefficients of determination for all three comparisons were very high and linear slope lines were close to 1, indicating good agreement between the two variables. Correlation coefficients were 0.9565, 0.9413, and 0.9351 for HW vs. ML1, HW vs ML2, and ML1 vs. ML2, respectively, indicating strong dependence between estimates for the compared variables.

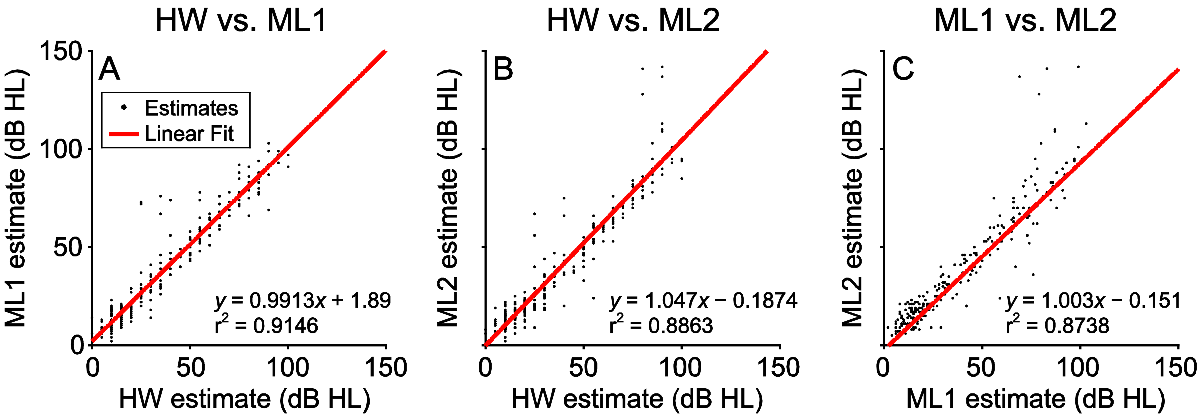

Figure 3.7: Direct comparisons of ML1, ML2, and HW threshold estimates. Black points represent individual data pairs, and the red line is a linear fit to the data. Equations describing the line and r^2 values are inset.

Figure 3.8 shows the accuracy of the ML procedure as a function of algorithm iteration (number of samples collected). This post-hoc analysis was performed by constructing an ML threshold audiogram estimate from the posterior distribution after each iteration of the GP algorithm and then evaluating the absolute difference from the final HW threshold audiogram at the six standard audiogram frequencies. Figures 3.8A, B show this trend for two representative ears, and Figure 3.8C shows the accuracy as a function of algorithm iteration averaged across all listeners and GP algorithm runs that terminated following 36 iterations. In both the individual data and the population data, the accuracy of the ML algorithm tended to improve systematically as a function of iteration. The ML estimate tends to achieve close to its final absolute difference value in only 20 samples or so. In some cases the difference function becomes shallow quickly but remains at some positive value (e.g., Figure 3.8A). Finally, note that the first 10 iterations show little systematic improvement in estimate quality, which is caused by the random sampling at this early stage before the informative sampling procedure begins.

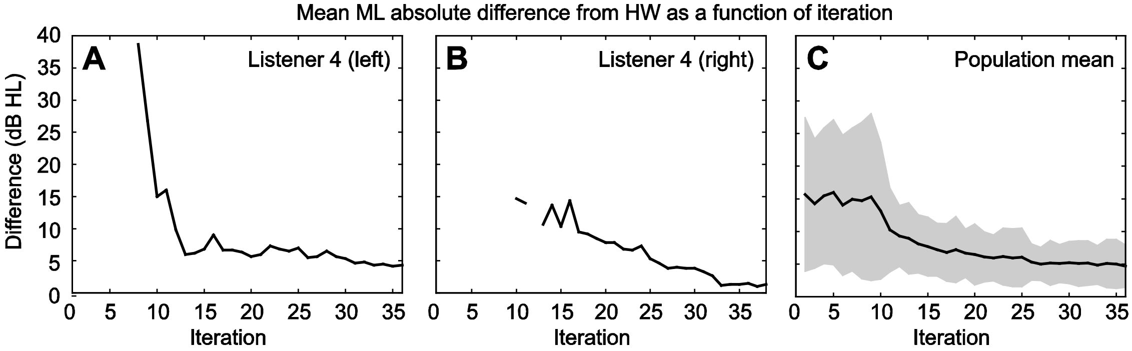

Figure 3.8: Agreement between ML and HW as a function of algorithm iteration. At each iteration, mean absolute difference was calculated by obtaining the current ML estimate of the threshold audiogram during one run, then calculating the absolute difference between that estimate and the HW threshold audiogram, averaged across all 6 audiogram frequencies. (A) and (B) show examples for two ears (Listener 4, as in Figure 3.4), and (C) shows this trend averaged across all runs where the ML audiograms terminated at 36 iterations (53 of 80 runs). Blank areas denote points at which the ML procedure did not produce a posterior mean with a clear boundary, so error could not be assessed (but in practice is very high). Gray shading on (C) indicates ±1 standard deviation from the mean.

The normalized GP posterior variance is shown as a function of ML algorithm iteration in **Figure 3.9**. At each iteration, we computed the normalized variance by summing each value in the posterior variance, which spans values $\left[0,1\right]$, and dividing by the total number of values (i.e. the number of input points on the test grid $\mathbf{X}^*$). Figures 3.9A, B show this trend for the same runs as in Figure 3.8, and Figure 3.9C shows this trend averaged across all listeners and runs. In general, the normalized posterior variance tends to decrease as a function of iteration, implying that the ML audiogram produces a less uncertain (more confident) estimate with an increasing number of samples. This function alone or in combination with other factors could therefore be used to evaluate the overall quality of an estimate.

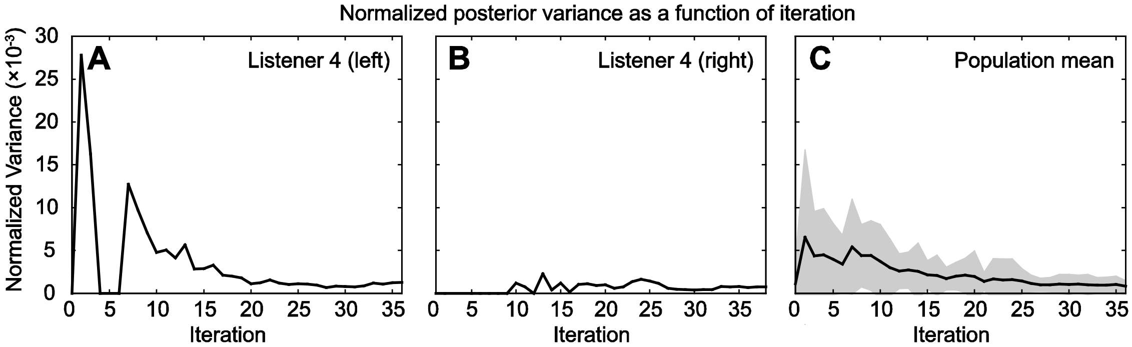

Figure 3.9: Normalized posterior variance as a function of algorithm iteration. Normalized posterior variance was calculated by dividing the sum all values in the variance function at each iteration by the total size of the variance function matrix. (A) and (B) show examples for two ears (Listener 4, as in Figure 3.4), and (C) shows this trend averaged across all listeners whose ML audiograms terminated at 36 iterations. Gray shading on (C) indicates ±1 standard deviation from the mean.

3.4. Discussion

We have described the development and verification of a novel automated technique for pure-tone audiometry, MLAG, which uses machine learning classification. This technique is able to provide a continuous estimate of a listener's pure-tone threshold audiogram across all frequencies, much like sweep-based audiometric techniques (von Békésy, 1947; Meyer-Bisch, 1996; Ishak *et al.*, 2011). However, this technique maintains efficiency by selecting only informative points, overcoming a main disadvantage of other continuous-estimate methods.

While obtaining threshold estimates at all frequencies with accuracy comparable to other algorithms, the ML audiogram required significantly fewer samples compared to conventional HW. Conventional HW approaches query frequencies individually and obey a rigorous rule for selecting tone intensities. In practice, this means that many samples collected by the conventional HW approach are not particularly informative, e.g., several relatively loud intensities in a row that the listener is very likely to hear. In contrast, the GP algorithm successively selects sample points evaluated by uncertainty sampling to be maximally informative at that point in time about the perceptual space. The rapid accumulation of information in the ML audiogram case is demonstrated in Figure 3.8, where the accuracy of the GP algorithm approaches reasonable values many iterations before the algorithm eventually terminates. Such rapid convergence cannot be accomplished with the HW approach because of its rigid sampling criteria.

Another unique property of our ML audiogram procedure is that an estimate of accuracy is automatically included with each newly computed posterior. In general, both the estimate error (in this case, correspondence with the HW estimate) and normalized posterior variance decrease as a function of algorithm iteration (Figures 3.8 and 3.9). The trend in accuracy appears more reliable than the trend in variance: additional samples will typically generate a more accurate estimate of the audiogram because there is more information about the function space. The ML procedure could possibly generate a low-variance yet inaccurate audiogram estimate with very few samples by either underfitting or overfitting, which is responsible for the dramatic drop in GP variance shown in the first 5 samples of Figure 3.9A. Multiple methods exist to deter underfitting or overfitting (Murphy, 2012); the simplest is perhaps to enforce a minimum number of iterations while ensuring that the algorithm is still sampling widely, which was deployed in the current experiment. After the first few iterations, the steady decrease in error implies that more samples lead to more accurate audiogram estimates. As Figure 3.8 suggests, however, this is likely only necessary for individuals whose GP estimates do not converge quickly.

The ML audiogram was generally robust to false positives. The covariance function allows the GP to classify unexpected responses as anomalies rather than true responses, assuming there are sufficiently many true responses to offset the false positives. If, however, the listener provides multiple false positives for very soft tones (or alternatively, misses multiple clearly audible tones), the ML audiogram may be unable to correctly reject those responses, as the evidence is no longer overwhelming in favor of rejecting them. While we did not experience this scenario with our listeners, the variance function inherent to our ML audiogram is in any case a natural quantification of estimate quality. Estimates that do not converge fully by the end of the ML audiogram can be used to signal the test operator that a poor reading resulted, thereby directing him or her to start the test over or pursue an alternate estimation strategy.

A related situation is a listener who responds inconsistently, to which the ML procedure is sensitive. **Figure 3.10** shows an example of one listener who provided inconsistent results by falling asleep during the ML audiogram procedure. The ML threshold audiogram deviated from the HW threshold audiogram obtained for the same ear (Figure 3.10A), and the inconsistency in responses that produced this result can be seen in Figure 3.10B. Note that sample points very close in intensity/frequency space elicited different responses, which is physically unrealistic. The ML procedure produced a threshold audiogram estimate that attempted to best match this inconsistent data. It can also be seen from Figure 3.10C that the ML algorithm hit the ceiling on the number of allowable iterations for that ear, 64, due to high posterior variance. Figure 3.10C further reveals that the normalized posterior variance did not generally decrease as a function of iteration; in fact, following iteration 15, the normalized variance gradually increased. The normalized variance may sometimes dramatically increase when the GP hyperparameters change substantially due to a particularly informative sample, but a gradual increase in normalized variance indicates that obtained samples may be of poor quality because additional samples are making the posterior less, rather than more, well-defined. Sufficient native quantification therefore appears to exist within the ML procedure to signal when a poor estimate is being obtained, in which case an alternate audiogram estimation strategy may be pursued.

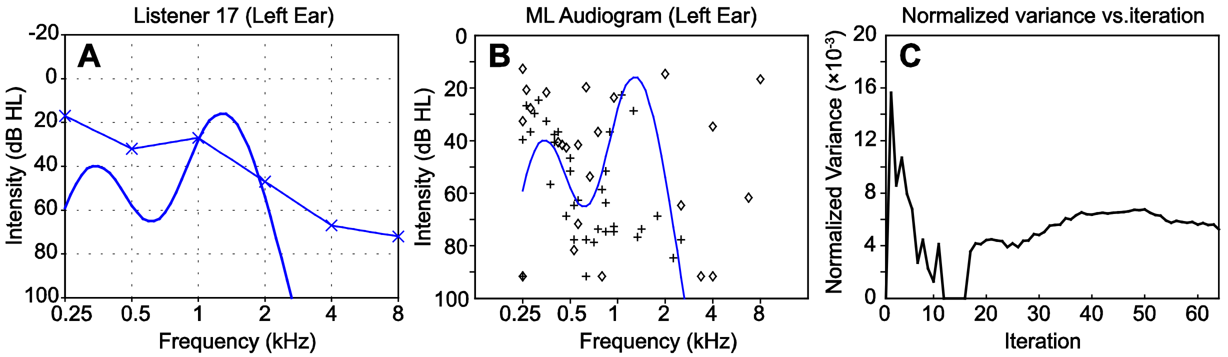

Figure 3.10: Data from Listener 17, who fell asleep during ML estimation. (A) The final ML audiogram from one ear, superimposed upon the HW audiogram obtained for the same ear (“X”). (B) Samples collected while conducting the ML audiogram. Note the inconsistency in responses, with detections and misses in very close proximity. (C) Plot of normalized posterior variance as a function of iteration for this listener. This listener reached the ceiling on the number of allowable iterations for this ear, 64. Unlike the variance trends in Figure 5, the variance in this ear actually begins to increase after approximately iteration 15 and remains high even after 64 iterations.

A key advantage of our ML audiogram technique is its ability to operate without direct human supervision. The algorithm used in this experiment necessitated experimenter intervention only upon switching ears, which was primarily as a courtesy for the listeners so that the ear switch could be announced. If this feature is removed or automated, the ML audiogram becomes a “plug-and-play” procedure that need only be initialized and will otherwise proceed on its own until termination, with no need for direct supervision by clinicians or experimenters other than to verify that the equipment is operating as desired and subjects are appropriately engaged. In other words, a technician could effectively oversee the test procedure and relay the results to a clinical audiologist for interpretation and possibly a decision to retest using a different methodology. Alternately, if it is possible to deliver only a very few stimuli, such as with very young children, the ML procedure could run decoupled from the stimulation apparatus and simply inform a clinician where to manually deliver the next sound to provide the most information about that patient’s hearing. Based upon our findings, 20 samples using this method should be enough to obtain a reasonable estimate of the threshold audiogram.

As indicated in Table 3.2, mean threshold estimates corresponded closely between the predictable, sequential HW procedure and the unconstrained, roving ML procedure. In general, both one-interval and two-interval detection and discrimination tasks have shown elevated thresholds when one or more stimulus parameters are roved (Berliner and Durlach, 1973; Mori and Ward, 1992; Amitay *et al.*, 2005; Mathias *et al.*, 2010; Bonino *et al.*, 2013). This is widely interpreted to be an attentional rather than a purely perceptual phenomenon because roving under masked conditions leads to observations best described by informational rather than energetic masking. The lack of threshold elevation with the roving ML stimulus presentations in the current study is therefore somewhat surprising. Our unmasked detection condition may have contributed to the similarity in thresholds. Other potential mitigating factors include our delivery of relatively long tones (Ward, 1991), relatively long inter-stimulus intervals (Berliner and Durlach, 1973) and, perhaps most significantly, repeated tone presentations (Kidd *et al.*, 2003; Burk and Wiley, 2004; Leibold and Bonino, 2009; Guest *et al.*, 2010).

One logical improvement to the accuracy of the algorithm is to further expand the frequency range from which the ML algorithm may sample. From Table 3.2 it is apparent that the greatest discrepancy between the HW and ML procedures occurred at 250 Hz and 8000 Hz, and the least discrepancy occurred at 1 kHz and 2 kHz. This most likely means that ML estimation at the extremes of the sampled frequencies is adversely affected by edge effects. This limitation apparently exists despite the observation that many samples are taken near the edge frequencies (e.g., Figure 3.4). Correcting this limitation would undoubtedly increase overall accuracy of the procedure as well as efficiency and could be accomplished in several ways, with one obvious solution being to sample frequencies lower than 250 Hz and higher than 8000 Hz.

A second improvement to the efficiency and precision of ML estimates would be to use explicit tone detection priors to drive initial sampling rather than learning the shape of the tone detection function completely empirically by a random priming sequence (Özdamar *et al.*, 1990). These priors can be represented in the mean and/or covariance function hyperparameters, and may be either specifically selected based upon the literature or empirically learned from real audiometric data. A further improvement may be to investigate different choices of acquisition function to inform the selection of each sample point. Our technique currently employs uncertainty sampling, but other techniques from psychophysics or Bayesian active learning may prove better-suited for this application (King-Smith *et al.*, 1994; Kontsevich and Tyler, 1999; Roy and McCallum, 2001; Settles, 2009; Houlsby *et al.*, 2011). The flexibility of the GP technique allows other active sampling methods to be swapped in relatively painlessly. Improving sampling consistency will also likely improve the accuracy of alternate classification strategies that might be developed in the future and thereby add to the overall value of the proposed procedure.

A final advantage of the ML-based algorithm is that it is more difficult for users to deliberately manipulate results than with traditional methods. The conventional HW algorithm is quite predictable, and any amount of familiarity with the procedure allows inclined individuals to manipulate their responses in order to obtain a deliberately inaccurate audiogram. On the other hand, manipulating responses to obtain a deliberately inaccurate audiogram is a much harder task using the ML estimation procedure because it does not follow the predictable structure inherent to HW. The ML audiogram samples widely across frequency and intensity from trial-to-trial, making it challenging for a listener to discern which responses would intentionally skew the test results in a particular direction. Attempts to thwart the test would also be readily discernible by the algorithm as response outliers, resulting in an inconclusive test and instruction to the operator to start over or pursue a different estimation strategy.

3.5. Concluding Remarks

This chapter describes an automated algorithm for conducting pure-tone air-conduction audiometry that selects appropriate test stimuli in real time based upon current estimate uncertainty. Our results indicate that the accuracy of this algorithm is comparable to other manual and automated methods while requiring fewer samples. At the same time, a continuous threshold audiogram is determined for all frequencies within a specified range. This algorithm also produces its own estimate of accuracy, which can be driven to high values by continuing to deliver more sample stimuli with the same criteria. The algorithm was not optimized specifically for audiogram estimation; therefore, much room for improvement remains possible for audiometry. Taken together, these advantages make this technique a compelling advance in pure-tone audiometry that can add immediate value to hearing diagnostic procedures.

Chapter 4: Uni- and Multidimensional Audiometric Function Estimation Using Gaussian Process Classification

Note: The research presented in Chapter 4 has been published in The Journal of the Acoustical Society of America (Song *et al.*, 2017).

4.1. Introduction

In Chapter 3, we presented the MLAG algorithm, which was able to provide an estimate of the threshold audiogram continuous across frequency (Song *et al.*, 2015). However, the GP framework provides not only estimates of threshold, but a probability of detection for each frequency and intensity. Therefore, the GP appears to provide sufficient information to form a continuous estimate of the entire audiometric function, or the 2-dimensional psychometric function across frequency and intensity. A full audiometric function would provide not only estimates of threshold, but also of spread, a measure of psychometric uncertainty that is at present rarely, if ever, estimated for audiograms.

Previous work has shown that PFs for both pure-tone detection (Allen and Wightman, 1994; Bargones *et al.*, 1995) and pure-tone intensity discrimination (Buss *et al.*, 2008) exhibit higher spread in young children compared to adults, consistent with a hypothesis of increased internal noise in children (Buss *et al.*, 2006). Furthermore, research in individuals with hearing loss has demonstrated shallower frequency discrimination PFs localized to hearing-loss frequencies (Nelson and Freyman, 1986; Freyman and Nelson, 1991). A detailed estimate of the audiometric

PF across frequency/intensity space could allow for better characterization of particular pathologies or cognitive states that affect psychometric spread as well as threshold.

Estimation of PFs has been a well-studied problem, but attention has been overwhelmingly focused on the unidimensional case. Except for adaptive methods seeking only thresholds (Levitt, 1971; Kollmeier *et al.*, 1988; Treutwein, 1995; Leek, 2001) and the rare nonparametric approach (Miller and Ulrich, 2001; Zychaluk and Foster, 2009), estimating a 1D PF has historically involved modeling it using an analytical equation that approximates the subject's probability of successful task performance as a function of stimulus intensity or discriminability. Multiple methods for accurately and efficiently estimating the 1D PF have been developed over the years, employing techniques such as maximum likelihood estimation, maximum *a posteriori* estimation, bootstrap resampling, and Markov chain Monte Carlo methods (e.g., Treutwein, 1995; King-Smith and Rose, 1997; Kontsevich and Tyler, 1999; Leek, 2001; Wichmann and Hill, 2001b; Shen and Richards, 2012).

The full audiometric function, however, is 2-dimensional, including one psychometric and one non-psychometric dimension. Estimation of multidimensional PFs has been considerably more limited, with most cases focusing on particular psychophysical spaces that can be easily parameterized (Bengtsson *et al.*, 1997; Lesmes *et al.*, 2006; Shen and Richards, 2013). More flexible frameworks, in which the PF of interest can belong to a particular parametric family, have also been developed (Kujala and Lukka, 2006; Vul *et al.*, 2010; DiMattina, 2015). However, the audiometric function cannot be justifiably parameterized, as it is the particular shape of the threshold curve that determines auditory dysfunction and has diagnostic value.

The Gaussian process is a nonparametric framework that does not require parameterization of the underlying function of interest; rather, it encodes relationships between nearby variable values in the form of covariances (Rasmussen and Williams, 2006; Duvenaud, 2014). Therefore, GPs have the flexibility to encode a large number of function shapes, so long as they fit the criteria of the covariance functions specified. Because of this, we chose a GP classification (GPC) framework to perform inference on these multidimensional audiometric functions, extending the capability of the MLAG algorithm to estimate entire psychometric functions.

In Chapter 4, we describe two distinct simulation experiments using the MLAG algorithm. In the first experiment, we investigate the ability of our GPC framework to estimate 1D PFs using samples obtained from the method of constant stimuli (Fechner, 1860). For comparison, we also perform inference using maximum-likelihood probit regression (PR) with the Nelder-Mead simplex method (Kingdom and Prins, 2010), as detailed in Section 1.4.1, using the same set of observed points. In the second experiment, we use the MLAG algorithm to estimate entire 2D audiometric functions using fixed sets of deterministic samples.

4.2. Methodology: 1D Psychometric Function

In Experiment 1, we evaluated performance of the GPC algorithm for estimation of a traditional 1D PF. We conducted computer simulations for a standard auditory detection task in which listeners are presented with pure tones of fixed frequency and are instructed to indicate when they detect a tone (Kingdom and Prins, 2010). While we conceptualized this as a tone detection task, the development that follows is generic and therefore relevant to a wide variety of univariate psychometric tasks.

4.2.1 Simulation Details

For each simulated auditory detection task, a stimulus of a particular intensity x_i was presented, and a binary response (indicating whether or not the stimulus was detected) was collected from a simulated participant. The probability of simulated user responses was governed by a cumulative Gaussian PF of the following form (Kingdom and Prins, 2010):

$$\psi\left(x\right)=\frac{1}{\beta\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\tfrac{1}{2}\left(\frac{z-\alpha}{\beta}\right)^{2}\right)dz\,. \tag{4.1}$$

For each stimulus x_i, we computed the corresponding detection probability $\psi\left(x_i\right)$ from the PF. To generate the binary response y_i for that trial, we simulated a single draw from a Bernoulli distribution with success probability $\psi\left(x_i\right)$, with success labeled as 1 and failure labeled as 0 (Treutwein, 1995). Following standard procedures for the method of constant stimuli (Fechner, 1860), we sampled *j* different stimulus intensities *k* times each. We then used this set of observed samples $\left\{\left(x_i, y_i\right)_{i=1}^{n}\right\}=\left(\mathbf{x},\mathbf{y}\right)$ to construct our estimate of the PF using the GPC technique described below.

4.2.2 Gaussian Process Construction

Because we observe only whether or not a subject detects the stimulus, not the detection probability itself, the output variable y is the binary value 1 if the subject detects the stimulus and 0 otherwise. We are therefore interested in obtaining $p\left(y=1\middle|f\right)$, which can be obtained using the GPC framework, allowing for inference using binary observations. We place a GP prior on the latent function, $p\left(f\right) = \mathcal{GP}\left(\mu\left(x\right), K\left(x, x'\right)\right)$, which we then transformed using an likelihood function $p\left(y = 1 \middle| f\right)$.

In the case of the 1D PF, we have substantial prior information that can be incorporated into the model. Standard techniques for 1D PF estimation parameterize the PF using a sigmoidal function that monotonically increases with stimulus intensity (Wichmann and Hill, 2001a; Kingdom and Prins, 2010). A similar procedure can be incorporated into our GPC model by combining the forms of the covariance function and the likelihood function. For the latent function f, we select a linear covariance function:

$$K\left(x, x'\right) = K\left(\iota, \iota'\right) = s^2 \iota \iota' , \tag{4.2}$$

where ι is the intensity and s^2 is a scaling factor. This covariance function constrains the latent function f to only linear functions. When combined with a cumulative Gaussian likelihood

$$p\left(y_i = 1 \middle| f_i\right) = \Phi\left(f_i\right) = \int_{-\infty}^{f_i} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz , \tag{4.3}$$

the resulting probability function takes the form of a cumulative Gaussian function that monotonically increases with intensity. For the mean function, we choose $\mu\left(x\right) = c$, where c is a constant hyperparameter, because any deviation from the mean response can be effectively captured in the covariance function.

After observing a set of training stimuli and responses $\left(\mathbf{x},\mathbf{y}\right)$, we first compute the best-fitting set of hyperparameters $\boldsymbol{\theta}=\left(s,c\right)$ by maximizing the log marginal likelihood $\log p\left(\mathbf{y}\middle|\mathbf{X},\boldsymbol{\theta}\right)$. Upon obtaining appropriate hyperparameters, we then calculate an estimate of the latent function as the posterior distribution $p\left(\mathbf{f}^*\middle|\mathbf{x},\mathbf{y},\mathbf{x}^*\right)$. Here, $\mathbf{x}^*$ is a finely spaced test set of intensities (e.g., −20 to 120 dB in 1-dB increments), and $\mathbf{f}^*$ represents the predicted latent function values at that set of intensities. We use the Laplace technique is used for approximation whenever necessary (Williams and Barber, 1998).

Estimates for the threshold α and spread β of the PF are derived directly from the *x*-intercept and inverse slope of the predictive latent mean. We obtain our point estimate for the PF by passing the predictive latent mean through the likelihood function $p\left(y_i=1\middle|f_i\right)$.

4.2.3. Evaluation

We evaluated estimation accuracy and reliability of the GPC framework for several psychometric and sampling parameters:

- *Spread value*: We used β values of 0.2, 0.5, 1, 2, 5, 10, and 20 dB/percent to construct true PFs.
- *Number of sampled intensities/number of repetitions at each intensity*: In a standard method of constant stimuli, a fixed number of intensities is selected, and each intensity is sampled some number of times (Fechner, 1860). With the maximum number of samples fixed at 200, we evaluated the following divisions individually: 200 intensities with 1 repetition per intensity, 100 intensities with 2 repetitions per intensity, 40 intensities with

5 repetitions per intensity, 20 intensities with 10 repetitions per intensity, and 10 intensities with 20 repetitions per intensity.

- *Number of observed samples*: To investigate the independent effect of total sample number on performance, sampling was conducted randomly without replacement for each set of intensities, and this process was repeated for the specified number of repetitions. For each parameter combination, then, each simulation was advanced 1 sample at a time and incrementally evaluated up to the maximum sample number of 200.
- *Simulation repetition/threshold value*: For each unique parameter combination, we conducted 4 independent simulations, resulting in 28000 simulations overall. For each simulation, an integer value for threshold α was randomly drawn from the uniform distribution $\left[30, 70\right]$.

We compared one-dimensional GPC experimental outcomes with traditional parametric 1D PF inference using maximum-likelihood probit regression (PR) (Nelder-Mead simplex method) (Nelder and Mead, 1965; Prins and Kingdom, 2009; Kingdom and Prins, 2010), a standard technique for PF inference given a set of observations. Identical stimulus and response samples were used to train both GPC and PR methods for each simulation. We evaluated estimation performance of both methods by comparing estimated values for α and β with the known values of α and β used in the simulated PFs. Accuracy was quantified by computing the mean deviation of parameter estimates from the true values, while reliability was quantified by computing the variance of GP parameter estimates across all repetitions with identical parameter values. We also derived nonparametric numerical values from the psychometric curve model to determine if such measures might differentiate the two methods in accuracy or reliability. The measures we chose were the 50% probability point and the 25-75% interquartile range (Strasburger, 2001). All statistical tests between GPC and PR were performed using the Kolmogorov-Smirnov (K-S) test, Bonferroni corrected for multiple comparisons.

We evaluated goodness of fit to the observations was evaluated for both GPC and PR using the Pearson χ^2 statistic: $\chi^2 = \sum_i \frac{N_i \left[p\left(x_i\right) - P_i \right]^2}{P_i\left(1 - P_i\right)}$. For intensity ι_i, $p\left(x_i\right)$ is the percent correct of the data, P_i is the percent correct of the model prediction, and N_i is the number of trials at that intensity (Klein, 2001; Wichmann and Hill, 2001a). The χ^2 statistic can be interpreted as a weighted sum of squared residuals, with larger statistic values representing poorer fits. The significance of the χ^2 statistic was evaluated by comparison to the chi-square distribution with *J* degrees of freedom, where *J* is the number of distinct intensities sampled.

4.3. Methodology: 2D Psychometric Function

In Experiment 2, we used the GPC framework to solve a relevant multidimensional psychometric problem. In this problem, a sequence of pure tones varying in both frequency and intensity is presented to a simulated listener, who is instructed to respond whenever a tone is detected. This task is similar to the task used for traditional pure-tone audiometry (Hughson and Westlake, 1944; Carhart and Jerger, 1959), but with two key differences: for this task, sampling does not necessarily proceed one frequency at a time, and sampling and prediction resolutions in both input feature dimensions is considerably higher. The goal of this work is to construct a general multidimensional PF estimator from recorded binary responses that can be used immediately for pure-tone audiometry and can be readily adapted to other multidimensional psychometric estimation problems.

4.3.1 Simulation Details

To simulate the 2D psychometric field, we first defined an audiogram shape for each simulated participant. For each audiogram shape, we approximated of 1 of 4 human audiometric phenotypes (Dubno *et al.*, 2013) using spline interpolation and linear extrapolation, forming a continuous threshold curve across frequency space. Figure 4.1 shows the 4 phenotypes, which were classified using both machine learning and physiology.

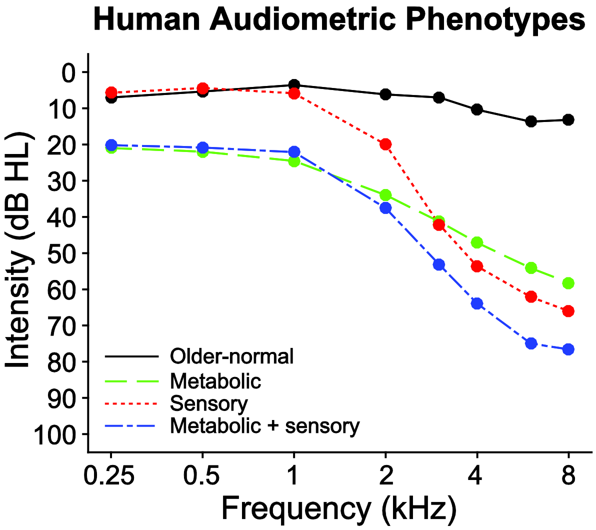

Figure 4.1: Plot of 4 human audiometric phenotypes.Reproduced from (Dubno *et al.*, 2013).

At each frequency, we used (4.1) to generate a sigmoidal psychometric curve as a function of intensity. We selected a value for spread β between 0.2 and 10 dB/percent, and we computed the value for center point α given β and the value of audiometric threshold at that frequency, which corresponded to a detection probability of 70.7% (Levitt, 1971). The overall 2D PF is therefore a combination of the audiogram shape across frequency and the sigmoidal 1D PF in intensity and provides the probability of detection $\psi\left(\mathbf{x}_i\right)$ for any input frequency/intensity pair $\mathbf{x}_i = \left(\omega_i, \iota_i\right)$.

As in the 1D case, the binary response y_i (success = 1; failure = 0) can be generated by simulating one draw from a Bernoulli distribution with success probability $\psi\left(\mathbf{x}_i\right)$. To select the set of observed frequency/intensity pairs, we use a Halton sequence (Halton, 1964), which provides a deterministic set of n well-spaced draws from the frequency/intensity domain of interest. We use these observed samples $\left\{\left(\mathbf{x}_1, y_1\right), \left(\mathbf{x}_1, y_1\right), \ldots \left(\mathbf{x}_n, y_n\right)\right\} = \left(\mathbf{X}, \mathbf{y}\right)$ as training observations for the GPC algorithm.

4.3.2. Gaussian Process Construction

For the 2D PF inference problem, we wish to estimate a subject's detection probability as a function of both the intensity and frequency of the presented stimulus. Our input variable $\mathbf{x}$ is therefore a frequency-intensity pair, or $\mathbf{x} = \left(\omega, \iota\right)$, and our output variable y is a binary response variable. We wish to infer the detection probability $\psi\left(\mathbf{x}\right) = p\left(y = 1 \middle| \mathbf{x}\right)$, and we place a GP prior on the latent function: $p\left(f\right) = \mathcal{GP}\left(\mu\left(x\right), K\left(x, x'\right)\right)$.

As in the 1D case, the dependence of detection probability on stimulus intensity is assumed to be a monotonically increasing sigmoidal function, which is captured using the linear covariance function (Equation 12) in the intensity dimension. The dependence of the detection probability on frequency is not explicit, however, and will vary across subjects based on the shape of the audiogram. A reasonable assumption is that the overall PF is continuous along the frequency dimension with some smoothness (Von Békésy, 1960; Kiang *et al.*, 1965; Green and Swets, 1966; Brant and Fozard, 1990; Leek, 2001). To reflect this behavior, we select a SE covariance function (2.9) for the frequency dimension. The full covariance function combines the linear covariance function in intensity and the SE covariance function in frequency:

$$K\left(\mathbf{x},\mathbf{x}'\right) = K\left(\left(\omega,\iota\right),\left(\omega',\iota'\right)\right)\mathbf{x} = s_1^2 \exp\left[\frac{-\left(\omega-\omega'\right)^2}{2\ell^2}\right] + s_2^2\left(\iota\iota'\right) \tag{4.4}$$

Here, s_1 and s_2 are scaling factors and ℓ is a characteristic length scale, which regulates the smoothness of the function with respect to frequency. Again, we select a constant mean function $\mu\left(\mathbf{x}\right) = c$ for this GP.

Given a set of observed samples $\left(\mathbf{X},\mathbf{y}\right)$, we again first calculate a set of best-fitting hyperparameters $\boldsymbol{\theta} = \left(c, s_1, s_2, \ell\right)$ by maximizing the log marginal likelihood $\log p\left(\mathbf{y}\,\middle|\,\mathbf{X},\boldsymbol{\theta}\right)$. We then compute the posterior distribution $p\left(\mathbf{f}^*\,\middle|\,\mathbf{X},\mathbf{y},\mathbf{X}^*\right)$ for a finely spaced grid of test samples $\mathbf{X}^*$ across frequency-intensity space: 0.125 to 16 kHz in semitone increments for frequency and −20 to 120 dB in 1-dB increments for intensity.

Unlike in the 1D case, we cannot readily specify a meaningful parametric form for the 2D PF across all frequencies and intensities. At any fixed frequency ω_i, however, we can derive an analytical expression for the PF by finding the inverse slope and *x*-intercept of the mean of the posterior latent function f at that frequency. Furthermore, the GPC method's point estimate for the full PF can be computed by passing f through the likelihood function $p\left(y_i = 1 \middle| f_i\right)$ and can be numerically compared to the true PF.

4.3.3. Evaluation

We evaluated overall performance of the GPC framework for a variety of psychometric and sampling parameters. Specifically, these parameters were manipulated:

- *Audiogram shape*: Older-normal, sensory, metabolic, and sensory + metabolic audiogram profiles (Dubno *et al.*, 2013) were used to fix simulation values of α at each frequency.
- *Spread value*: β values of 0.2, 0.5, 1, 2, 5, and 10 dB/percent, assumed isotropic across all frequencies, were used to construct the PF.
- *Number of observed samples*: 20, 50, 100, 200, 500, and 1000 pairs $\left(\mathbf{X}, \mathbf{y}\right)$ were used as observed data to condition the GP.
- *Simulation repetition*: For each unique parameter combination, we performed 40 independent repetitions of GPC inference, resulting in 5760 simulations overall.

We evaluated performance of the GPC framework by comparing the GP parameter estimates of α and β with the known values of α and β from the simulated PF. Performance was evaluated by comparing parameter values at a fine grid of frequency values (0.25 to 8 kHz in semitone increments). Edge frequencies (0.125 to 0.25 and 8 to 16 kHz) were used to train the GP but not to evaluate prediction because previous work has shown that edge effects can reduce GPC accuracy (Song *et al.*, 2015). Accuracy was evaluated by computing the mean deviation of parameter estimates from the true value, while reliability was evaluated by computing the variance of GP parameter estimates across all repetitions with the same parameter values. Once again, accuracy and reliability were verified with two nonparametric numerical values: the 50% probability point and 25-75% interquartile range.

We evaluated goodness of fit of the 2D GP posterior mean to the observations using the Pearson χ^2 statistic, consistent with the 1D case. For each frequency/intensity pair $\mathbf{x}_i = \left(\omega_i, \iota_i\right)$, the statistic $\chi^2 = \sum_i \frac{N_i \left[p\left(\mathbf{x}_i\right) - P_i \right]^2}{P_i \left(1 - P_i\right)}$ was computed and compared to the chi-squared distribution with *J* degrees of freedom, where *J* is the number of frequency/intensity pairs sampled. Because the GP framework with a Halton sampling scheme does not typically repeat observations at identical input values, *J* usually equaled the total number of observations.

4.4. Results: 1D Psychometric Function

For 1D PFs, both GPC and PR produced outlier trials for small sample numbers that disproportionately affected computation for the means and standard deviations. Following data collection from the simulations, we detected these outliers by thresholding at the 98th percentile (i.e., removing the 2% of scores farthest from the mean) across all trials and conditions for GPC and PR independently. A total of 981 and 716 outliers were detected out of 28000 total simulations each for the PR and GPC runs, respectively, primarily from trials with fewer than 20 observed samples. We excluded these outliers from the computations of means and standard deviations. We also omitted any trials using fewer than 10 observed samples from these computations because of generally poor performance at low sample numbers for both methods.

Figure 4.2 shows four representative examples of unidimensional PFs estimated by the GPC and PR techniques with four different numbers of observed samples (i.e., simulated subject responses) at fixed absolute intensity values. Identical observations were used for both methods in each panel. Qualitatively and quantitatively, GPC and PR perform very similarly, with systematically increasing estimation accuracy as the number of observed samples increases.

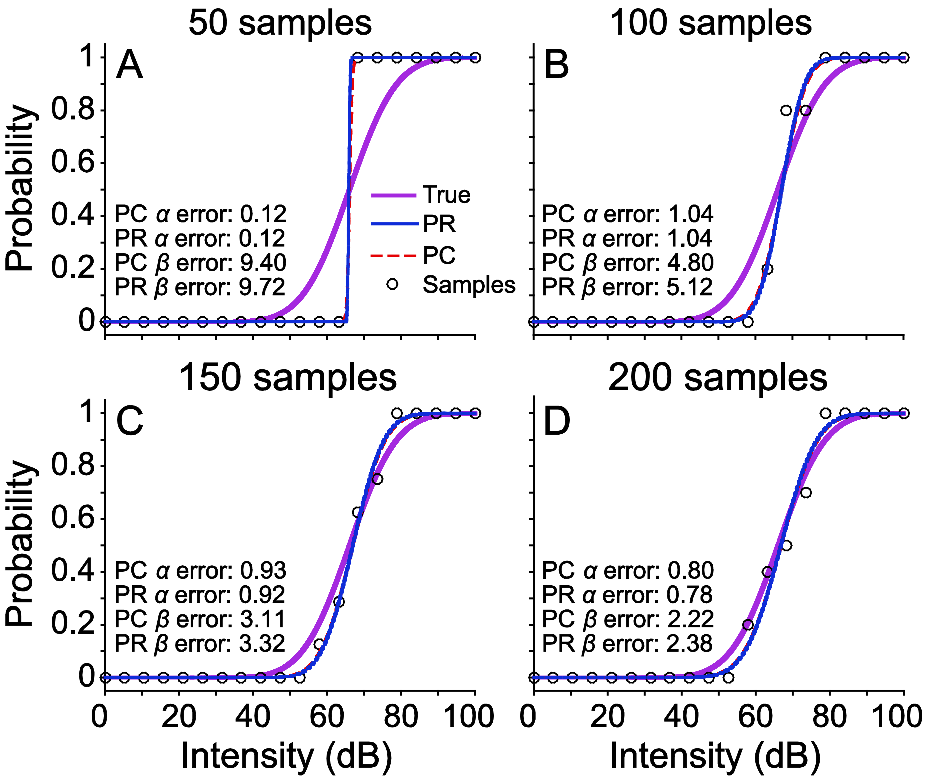

Figure 4.2: Examples of 1D PF estimation using PR and GPC. Each subplot shows model performance after (A) 20, (B) 100, (C) 150, and (D) 200 samples. True values of α and β were 66 dB and 10 dB/%, respectively, and sampling was performed at 20 distinct intensities within the interval. Units for α and β error are dB and dB/%, respectively.

Figure 4.3 shows the mean and standard deviation of absolute errors for α and β as a function of number of observed samples, averaged across all 140 conditions and trials, for both PR and GPC. As expected, the accuracy and reliability for both techniques increased with the number of samples. Estimates for α are consistent between PR and GPC. For β estimates at lower sample numbers, GPC appears to be slightly less accurate or equivalently accurate yet more reliable than

PR, although this difference disappears at higher sample numbers. Nevertheless, only 2 of 382 total comparisons showed statistically significant differences between PR and GPC (p<0.05, Kolmogorov-Smirnov test, Bonferroni corrected for multiple comparisons), consistent with the assessment that these two methods generally exhibit statistically indistinguishable performance.

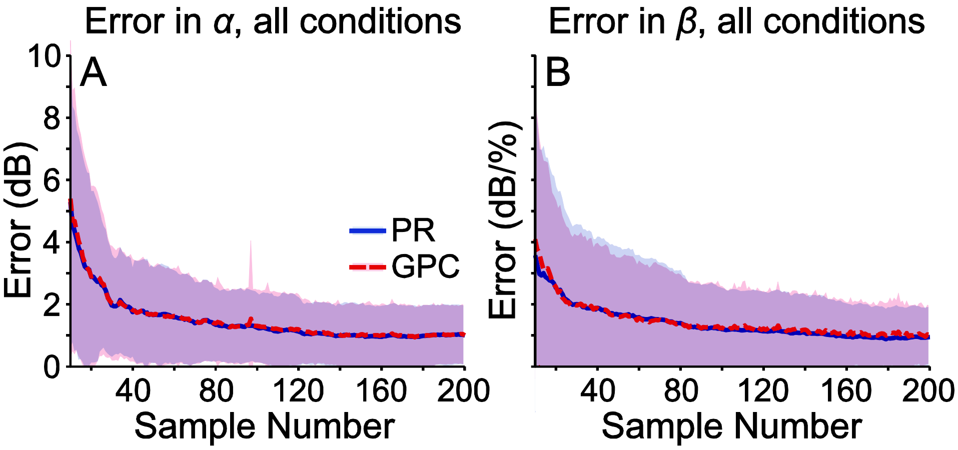

Figure 4.3: Error in 1D α and β estimates for PR and GPC across all conditions. Absolute error in estimates of (A) α and (B) β as a function of number of observed samples in the unidimensional case. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded regions designate 1 standard deviation above and below the mean.

Because of increased uncertainty in the transition zone for higher values of β, we evaluated estimator performance for larger versus smaller spreads as a function of β value. **Figure 4.4** shows the mean and standard deviation absolute errors in α and β for each β value tested. In all cases, both accuracy and reliability generally increase as a function of sample number. However, large values of β decrease the accuracy of both GPC and PR, particularly for lower numbers of observed samples. Overall, the trends for GPC and PR are generally quite similar, revealing no consistent difference in estimator quality between the two methods.

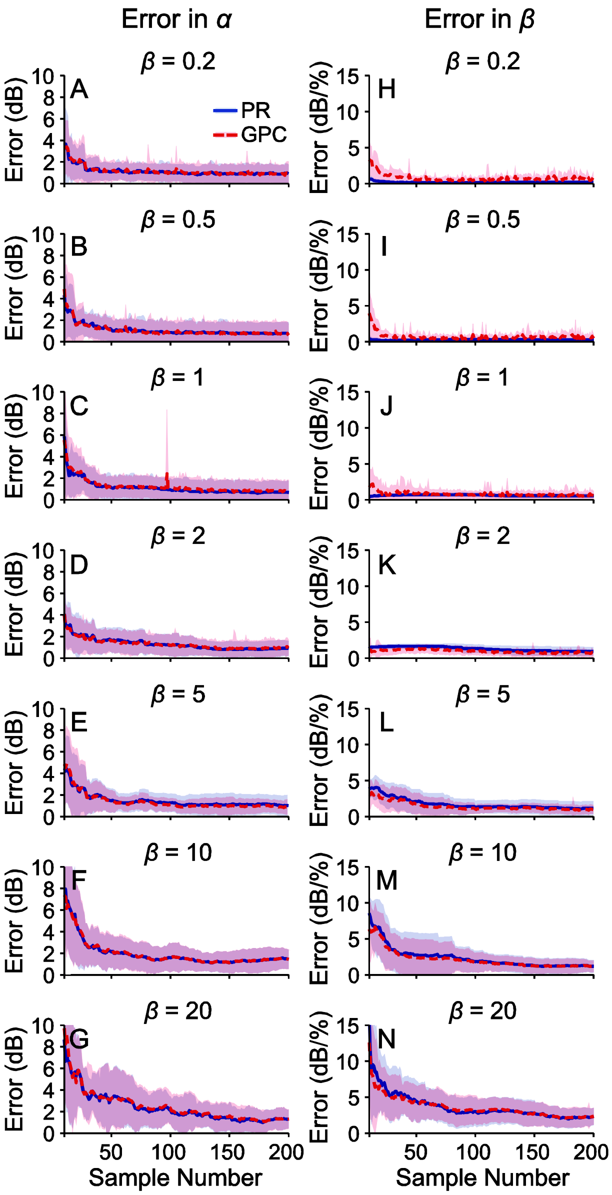

Figure 4.4: Error in 1D PR and GPC α and β estimates for different β values. Absolute error in estimates of (A-G) α and (H-N) β as a function of number of observed samples for unidimensional PFs. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded region designates 1 standard deviation above and below the mean. Each subplot corresponds to a distinct β value.

Because a fixed number of samples can be distributed across intensity in different ways, we investigated the effect that the number of distinct intensities and the number of repetitions per intensity had on the performance of both estimators. **Figure 4.5** shows the mean and standard deviation of absolute errors in α and β for each unique trial count per intensity. Both accuracy and reliability generally increase as a function of sample number. Again, overall estimator performance is quite similar between the two methods, with 0 of 1910 comparisons resulting in statistically significant differences for either α or β ($p<0.05$, K-S test, Bonferroni corrected for multiple comparisons).

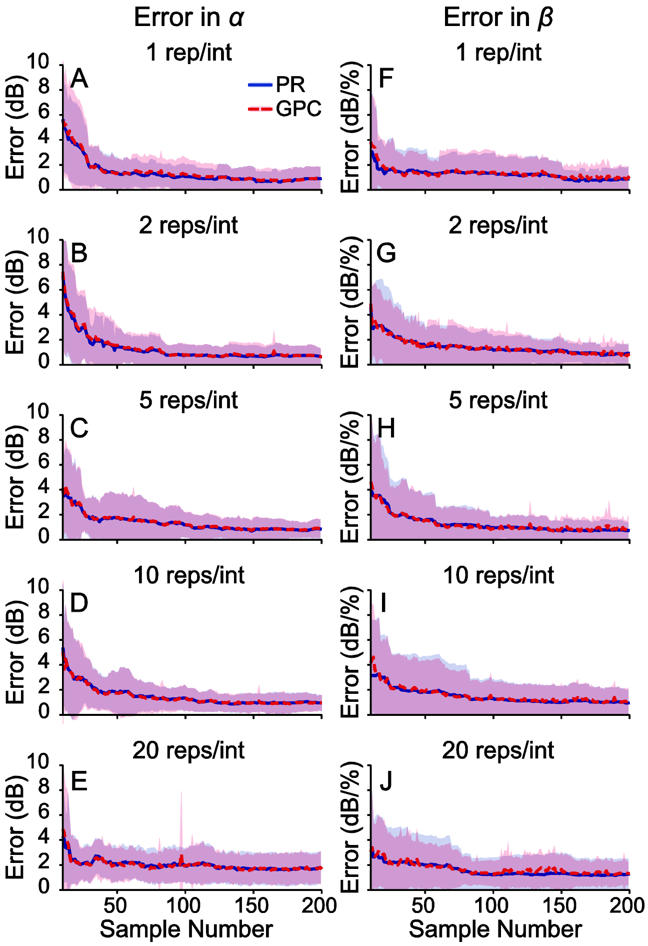

Figure 4.5: Error in 1D PR and GPC α and β estimates for different sampling distributions. Absolute error in estimates of (A-E) α and (F-J) β as a function of number of observed samples for unidimensional PFs. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded region designates 1 standard deviation above and below the mean. Each subplot corresponds to a distinct condition for number of intensities and number of repetitions per intensity.

We repeated all of the analysis described above for the numerical accuracy and reliability values of 50% point and 25-75% interquartile range. We observed identical trends for these measures, again indicating that GPC results in functionally indistinguishable estimator performance compared to parametric maximum-likelihood PR for 1D PF estimation. These results are shown in Appendix 1, **Figures A1.1–A1.3**.

To further investigate the agreement between GPC and PR estimates on 1D PFs, we directly compared the GPC and PR performance for estimates derived from the same set of data. **Figure 4.6** shows plots of GPC estimates versus PR estimates, with α and β behavior shown in Figure 4.6A and Figure4.6B, respectively. Note that 1 outlier was removed from the comparison for α. Coefficient of determination values for both α and β linear fits were very high, indicating that the linear functions were good fits for the data. For both α and β, linear slope terms were very close to 1 and linear intercept terms were near 0, indicating high PR and GPC estimate agreement. Correlation coefficients between PR and GPC estimates were 0.9992 and 0.9941 for α and β, respectively, again indicating a high degree of agreement between estimates. GPC does appear to overestimate small β values compared to PR, consistent with the data in Figure 4.4H-J.

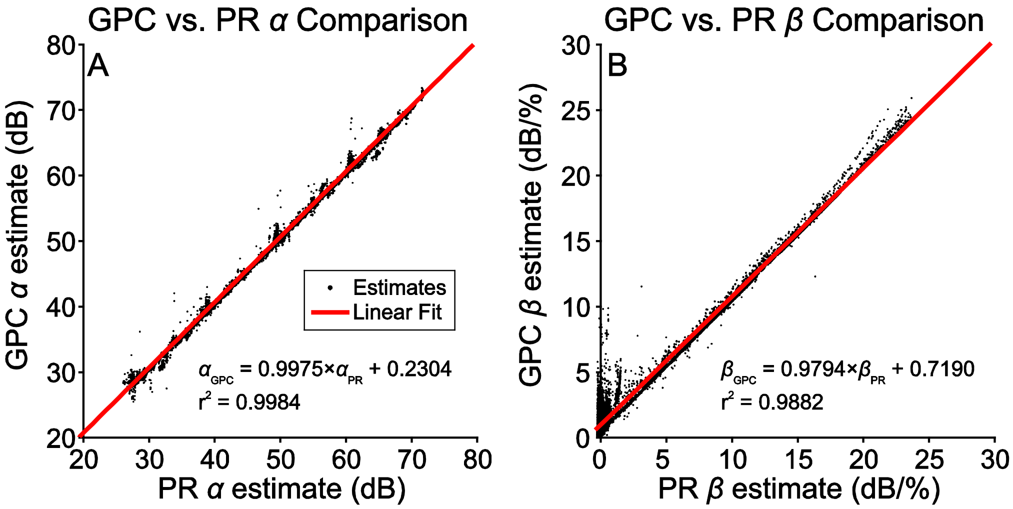

Figure 4.6: Direct comparison of 1D PR and GPC α and β estimates. Black points represent individual PR/GPC pairs for estimates derived from the same set of data, and the red line is a linear fit to the data. Equations describing the line and r^2 values are inset. 1 outlier was removed from the α comparison plot.

Across all 28000 2D trials, the median χ^2 statistic was 3.45 for GPC trials and 5.28 for PR trials. Using a significance level of $p < 0.05$, Bonferroni corrected for multiple comparisons, 165 of 28000 GPC simulations (0.59%) demonstrated statistically poor fits, while 201 of 28000 PR simulations (0.72%) demonstrated statistically poor fits.

4.5. Results: 2D Psychometric Function

Following data collection from simulations, we detected 4 trials out of 5760 as outliers by thresholding parameter values at the 99.93 percentile. These trials were omitted from the calculations for mean and standard deviation. Qualitative assessment of these outliers revealed the posterior mean surfaces were basically flat, resulting in β estimates of close to infinity.

Figure 4.7 shows a representative plot obtained by a single run of the 2D GP method. Figure 4.7A shows the samples and posterior mean obtained by the 2D GP method after observing 200 Halton samples. Figure 4.7B shows the GP prediction of α compared to the true values of α as a function of frequency, which demonstrates close agreement. Figure 4.7C shows a sample "slice" of the posterior mean at ω = 1 kHz, superimposed with the true PF at that frequency, which illustrates the agreement in β. GPC performance closely matched the simulation ground truth. In this kind of 2D PF estimation, however, no standard psychometric estimation method exists with which to compare GPC performance.

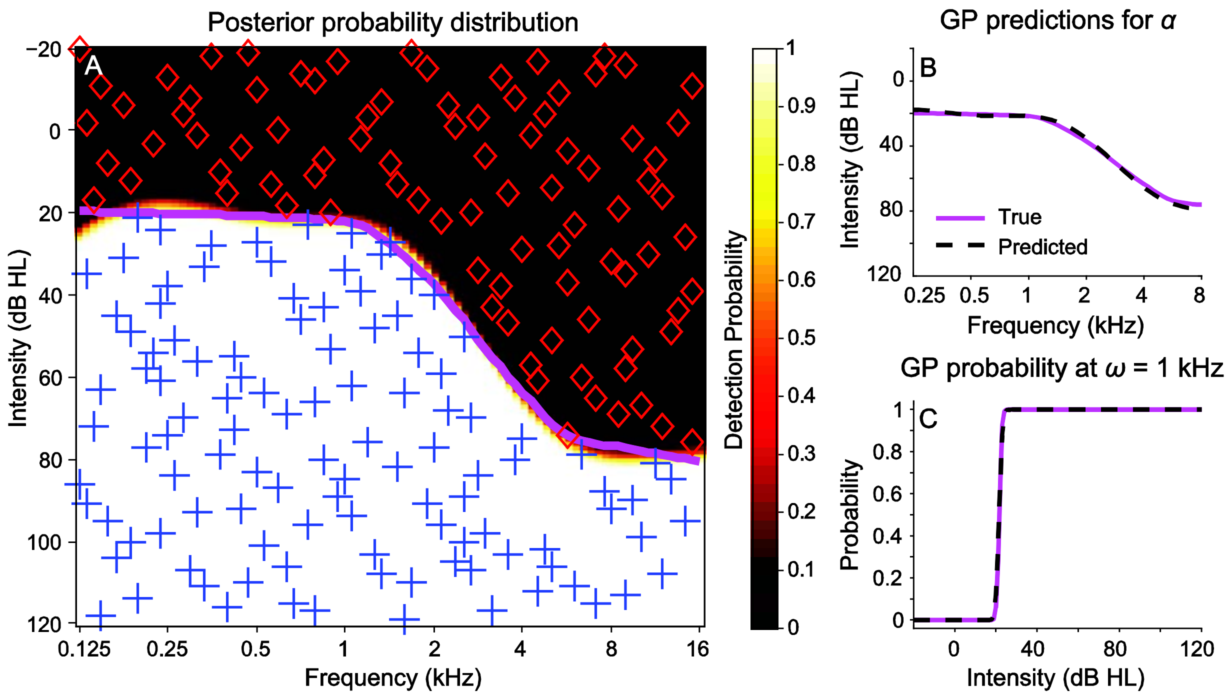

Figure 4.7: Sample posterior surface with samples and threshold/spread estimates. (A) A 200-point Halton sample set and resulting predictive posterior mean. The true psychometric function consisted of a metabolic + sensory audiogram type with a spread of 1 dB/percent. Superimposed magenta curve shows the true value of α as a function of frequency. Blue plus and red diamond symbols denote detected and missed tones, respectively. (B) The GP estimate of α (black dashed line) compared with the true values of α (magenta solid line) as a function of frequency. (C) A slice of the predictive posterior at 1 kHz (dashed line) compared with the slice of the true psychometric function (solid line). Colorbar represents detection probabilities.

Figure 4.8 shows representative GPC behavior for each of the four archetypical audiogram phenotypes (Dubno *et al.*, 2013). In each case, 200 Halton samples achieved reasonable accuracy for parameter and numerical estimates.

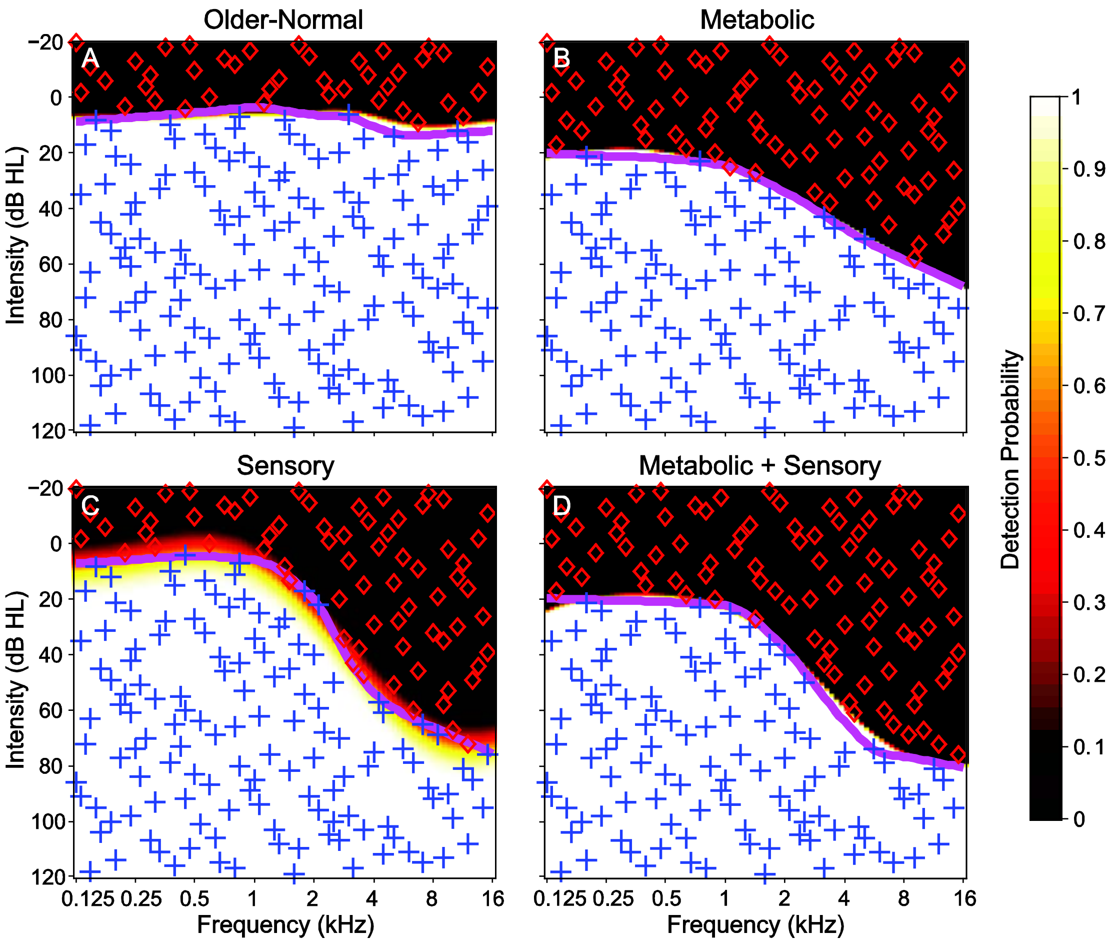

Figure 4.8: Representative posterior distributions for four audiogram phenotypes. Phenotypes are (A) older-normal, (B) metabolic, (C) sensory and (D) metabolic + sensory shapes. All plots have β = 1 dB/% as ground truth and are sampled using 200 Halton samples. Blue plus and red diamond symbols denote detected and missed tones, respectively. The magenta curve shows the true values of α across frequency for each audiogram phenotype. Colorbar represents detection probabilities.

Tables 4.1-4.3 summarize the accuracy and reliability of the 2D GP across all trials and conditions. **Table 4.1** shows the mean and standard deviation of the absolute errors in α and β values, separated by total number of samples. As expected, accuracy and test-retest reliability for both parameters increase as a function of total number of samples. **Tables 4.2 and 4.3** show the mean and standard deviation of the absolute errors in α and β values separated by β value and audiogram shape, respectively. These two tables used only data collected with 200 Halton samples, which achieved a reasonable estimate for both parameters. As the β value increased, accuracy and test-retest reliability tended to decrease for α. Results across different audiogram shapes were similar for α and β.

Number of samples	**20**	**50**	**100**	**200**	**500**	**1000**
Mean α absolute error (dB)	6.63	4.35	3.03	2.08	1.30	0.933
Std. α absolute error (dB)	5.42	7.28	2.57	1.74	1.23	0.913
Mean β absolute error (dB/%)	3.22	2.68	1.50	1.32	0.676	0.429
Std. β absolute error (dB/%)	2.63	2.62	1.64	1.18	0.576	0.345

Table 4.1: Absolute errors of 2D GP α and β estimates by number of samples. Accuracy and reliability are quantified using mean and standard deviation, respectively.

β (dB/%)	**0.2**	**0.5**	**1**	**2**	**5**	**10**
Mean α absolute error (dB)	1.53	1.60	1.70	1.86	2.47	3.32
Std. α absolute error (dB)	1.03	1.09	1.17	1.32	1.90	2.60
Mean β absolute error (dB/%)	1.04	0.843	1.11	1.49	1.77	1.70
Std. β absolute error (dB/%)	1.11	1.14	1.01	0.874	1.35	1.25

Table 4.2: Absolute errors of 2D GP α and β estimates by value of β. Accuracy and reliability are quantified using mean and standard deviation, respectively. Sample number is fixed at 200 samples.

Audiogram Phenotype	**Older**	**Metabolic**	**Sensory**	**Metabolic + Sensory**
Mean α absolute error (dB)	2.19	1.69	2.22	2.22
Std. α absolute error (dB)	1.68	1.47	1.96	1.75
Mean β absolute error (dB/%)	1.07	1.03	2.29	0.916
Std. β absolute error (dB/%)	1.14	0.888	1.14	0.993

Table 4.3: Absolute errors of 2D GP α and β estimates by audiogram phenotype. Accuracy and reliability are quantified using mean and standard deviation, respectively. Values are averaged across all β values, and sample number is fixed at 200 samples.

Numerical accuracy and reliability for the 50% point and interquartile range were similar to the trends in α and β, respectively. These results are shown in **Tables A1.1-A1.3** in Appendix 1.

Across all 5760 2D GP simulations, the median χ^2 statistic value was 1.96. After computing probability values on the chi-square distribution with the appropriate degrees of freedom for each simulation, 86 of 5760 trials, or 1.5%, demonstrated statistically poor fits at a significance level of $p < 0.05$, Bonferroni corrected for multiple comparisons.

4.6. Discussion

Throughout the history of psychophysics, traditional methods of full PF estimation have almost always employed parametric regression. We have described a novel technique for estimating PFs using nonparametric probabilistic classification with Gaussian processes and Bayesian inference. Simulations indicate that this technique is able to estimate standard 1D PFs with accuracy comparable to that of maximum-likelihood probit regression. Despite representing a new form of psychometric inference, GPC is able to achieve as accurate results in traditional applications as perhaps the most commonly used PF estimator today.

The true value of this method, however, comes with applications to scenarios more complex than 1D PF estimation. We have also shown that GPC can accurately estimate a variety of 2D PF shapes within the same framework. To our knowledge, this is the first technique for estimation of arbitrary audiometric full PFs. Unlike other existing methods for multidimensional PF estimation (Patterson, 1976; Lesmes *et al.*, 2006), we need not specify an explicit parametric form for the function, thereby imbuing this method with great flexibility to estimate arbitrarily shaped multidimensional PFs. We have evaluated this technique for 1D and 2D functions, but it is straightforward to extend GPC to PFs spanning even more dimensions.

For 1D PFs, GPC demonstrated the ability to accurately infer values for α and β across a number of distinct β values and sampling densities. Within approximately 30 samples, absolute estimation error for both α and β dropped to around 2 dB and 2 dB/%, respectively, with asymptotic performance approaching 1 dB and 1 dB/%, respectively. Crucially, GPC demonstrated equivalent performance to the Nelder-Mead simplex method for maximum-likelihood parametric probit regression, showing parity with a common current method.

The main advantages of the GP technique only become evident in estimation of the 2D audiometric function. Despite a limited number of samples across the entire frequency/intensity range, GPC is able to produce accurate estimates of α and β. Within approximately 50 samples, the error in α drops below 5 dB (see Table 4.1), matching the reported accuracy for both manual and automated audiometric techniques having much lower resolution (Fausti *et al.*, 1990; Swanepoel *et al.*, 2010; Mahomed *et al.*, 2013). That sufficient accuracy and reliability on this 2D estimation task can be achieved with a limited number of samples can be largely attributed to the covariance function, which allows for information to be shared across nearby frequencies. An advantage of the Halton sampling technique over traditional sampling techniques is that it does not repeat measurements at identical frequencies, up to the resolution limit along the frequency dimension. Instead, each observation influences the local PF estimate across frequency via the squared exponential portion of the covariance function. This structure allows an observation to influence predictions at nearby frequencies within a domain specified by the SE length scale.

Both the 1D sampling procedure and the 2D Halton sampling procedure used here are variants of the method of constant stimuli (Fechner, 1860). This sampling approach was deliberately chosen in order to make comparisons between the methods more straightforward and to separate out the effects of estimator quality and sampling methodology. Most PF estimation procedures, however, use adaptive sampling techniques that sequentially select samples maximally informative for some acquisition function, such as maximizing the decrease in expected variance or entropy (King-Smith *et al.*, 1994; Kontsevich and Tyler, 1999; Lesmes *et al.*, 2006; Shen *et al.*, 2014). These existing adaptive techniques can readily be incorporated into the GPC framework, whereby the adaptive method is used to select the samples and the GP is used to perform inference on the PF. The Bayesian nature of GPC means that informative samples can be inferred directly from the posterior itself. Previous work has demonstrated adaptive sampling schemes using uncertainty sampling, information sampling, and active model selection (Gardner *et al.*, 2015a; Gardner *et al.*, 2015b; Song *et al.*, 2015), all of which can be implemented using the GP posterior distribution.

The GPC technique as implemented here obtains point estimates for PF parameters α and β upon approximating the latent function $f\left(\mathbf{x}\right)$. A common need in Bayesian estimation, however, is to obtain a probability distribution on those parameters. PF estimation techniques typically use bootstrap and Monte Carlo resampling procedures to obtain these distribution estimates (Wichmann and Hill, 2001b; Kuss *et al.*, 2005). In the GPC framework, PF parameter distributions can be obtained numerically by sampling from the posterior distribution (e.g., by Markov chain Monte Carlo methods) or by explicitly specifying parameters in the GP latent function itself (which is typically treated as purely nonparametric) and approximating the posterior distribution over these parameters (Rasmussen and Williams, 2006).

The GPC formulation has distinct parallels with previous work in which the 1D psychometric function model was decomposed into "core" and "sigmoid" functions (Kuss *et al.*, 2005; Fründ *et al.*, 2011). In the case of GPC, the "core" function is the latent function f, and the "sigmoid" function is the likelihood function $p\left(y=1\middle|f\right)$. As in the previous work, each of these functions can be manipulated independently in GPC to reflect different psychometric properties.

4.7. Concluding Remarks

This chapter describes a nonparametric Bayesian technique for estimating unidimensional and multidimensional psychometric functions. The specific Bayesian procedure implemented made use of Gaussian process classification, a flexible yet powerful inference engine well-suited for approximating complex psychometric functions. We assessed the accuracy and reliability of this technique using both 1D and 2D simulated audiometric functions, revealing accuracy comparable to standard parametric estimation techniques wherever such comparisons could be performed. Because of its inherent flexibility, this technique can be readily extended to approximate psychometric functions outside of the auditory domain and can easily incorporate more input dimensions and more complex covariance and likelihood functions. Adopting probabilistic classification techniques will yield a host of advantages for general psychometric function approximation relative to conventional parametric regression techniques without any apparent drawbacks relative to the simpler applications in common use.

Chapter 5: Estimation of Multidimensional Audiometric Functions Using Active Gaussian Process Classification

The research presented in Chapter 5 was performed in conjunction with Kiron A. Sukesan and is currently in preparation for Attention, Perception and Psychophysics.

5.1. Introduction

In Chapter 4, we presented the results of using the MLAG algorithm to perform inference on 1D psychometric functions and 2D audiometric functions (Song *et al.*, 2017). Results showed that for 2D audiometric functions, Halton sampling resulted in reasonable estimates for α and β in approximately 100 samples (3.03 ± 2.57 dB for α, 1.50 ± 1.64 dB/% for β), compared to clinical test-retest reliability of 5 dB for thresholds (Katz *et al.*, 2009).

Halton sampling, despite being a space-filling set, is nonetheless a deterministic method that selects samples that likely contribute little information to the estimate. As can be seen in both Figure 4.7 and Figure 4.8, the majority of samples, both detected and undetected, were in regions where the probability was ultimately very certain. Because an iterative approach was not taken for 2D audiometric function estimation, the set had to be chosen *a priori*, and as a result had to be relatively naive. However, an active sampling scheme could dramatically improve the efficiency of audiometric function estimation if implemented.

As outlined in Section 1.4.2, numerous active sampling methods have been implemented over the years to efficiently estimate unidimensional psychometric functions, each of which iteratively selects optimal query points to best form the current estimate (Treutwein, 1995; Leek, 2001).

The earliest active sampling techniques were optimized for estimating threshold only; these included methods such as best PEST, which places each trial at the point closest to threshold (Pentland, 1980), and QUEST, which places each trial at the current posterior mean of threshold (Watson and Pelli, 1983). The degree to which these particular methods can estimate spread is unclear. Active sampling strategies that are known to estimate threshold as well as spread include modified ZEST, which selects the stimulus that produces the highest expected decrease in variance (King-Smith and Rose, 1997), and the Ψ-method, which selects the stimulus that minimizes the expected entropy (Kontsevich and Tyler, 1999). Variants of these 1D techniques have also been used to estimation of parametric multidimensional PFs (Kujala and Lukka, 2006; Lesmes *et al.*, 2006; Lesmes *et al.*, 2010; Vul *et al.*, 2010; DiMattina, 2015).

As mentioned in the previous chapter, however, existing models for multidimensional PFs must be parameterized in some way and cannot reliably estimate PFs that cannot be parameterized, such as the audiometric function. MLAG results from Chapter 3 showed large efficiency gains using active sampling relative to clinical methods (Song *et al.*, 2015), and results from Chapter 4 demonstrated MLAG's ability to estimate full audiometric functions using deterministic sampling. So far, active sampling has not been used with MLAG to estimate full audiometric functions. In this chapter, we combine GP classification with several active sampling strategies to investigate the accuracy and reliability of active sampling in the MLAG framework, directly comparing their results to the performance of non-active sampling methods.

5.2. Methodology

We evaluated the performance of the probabilistic classification algorithm for estimation of a tone-detection audiometric function given simulations of actual subject responses. Although these experiments are centered upon a pure-tone detection task, the technique is general and applicable to numerous uni- and multidimensional psychometric estimation tasks.

5.2.1. Simulation Details

We simulated ground truth audiometric PFs as described in Section 4.3.1 and in (Song *et al.*, 2017). Threshold curves as a function of frequency were generated by estimation of 1 of 4 human audiometric phenotypes (Dubno *et al.*, 2013) using spline interpolation and linear extrapolation. 6 different spread parameters β between 0.2 and 10 dB/percent were considered for each PF. At each frequency, we constructed a 1D sigmoidal PF using a cumulative Gaussian equation of the following form (Kingdom and Prins, 2010):

$$\psi\left(\iota\right) = \frac{1}{\beta\sqrt{2\pi}} \int_{-\infty}^{\iota} \exp\left(-\tfrac{1}{2}\left(\frac{z-\alpha}{\beta}\right)^{2}\right) dz, \tag{5.1}$$

where ι is intensity. The audiogram threshold value at that frequency corresponded to the 70.7% detection probability point along the PF (Levitt, 1971), which we used along with β in order to compute the value of α. To construct the overall 2D PF, we combined the audiogram shape across frequency with the sigmoidal 1D PF in intensity.

For a particular frequency/intensity input $\mathbf{x}_i = \left(\omega_i, \iota_i\right)$, the PF returns a detection probability $\psi\left(\mathbf{x}_i\right)$ corresponding to that point. We can generate a binary response y_i at that input by taking a single draw from a Bernoulli distribution with parameter $\psi\left(\mathbf{x}_i\right)$. Detected and non-detected responses are represented as values of 1 and 0, respectively (Treutwein, 1995).

5.2.2. Gaussian Process Construction

For this experiment, we used an identical 2-dimensional GP framework to the one described in Section 4.3.2 and (Song *et al.*, 2017). The main points are briefly summarized here.

Variable space: For this problem, the input variable $\mathbf{x}$ is a pure-tone frequency-intensity pair $\mathbf{x} = \left(\omega, \iota\right)$, and our output variable y is a binary response variable. We place a GP prior on the latent function f, describing degree of class membership: $p\left(f\right) = \mathcal{GP}\left(\mu\left(x\right), K\left(x, x'\right)\right)$.

Mean function: As before, we choose a constant mean function: $\mu\left(\mathbf{x}\right) = c$, allowing for the covariance function to capture the variation in the latent function around the mean.

Covariance function: Tone detection probability increases with increasing intensity ι, while it does not take a fixed form as a function of frequency ω but is known to be continuous and relatively smooth (Von Békésy, 1960; Kiang *et al.*, 1965; Green and Swets, 1966; Brant and Fozard, 1990; Leek, 2001). Therefore, the full covariance function across both frequency and intensity has a linear form in ι and a squared exponential form in ω:

$$K\left(\mathbf{x}, \mathbf{x}'\right) = K\left(\left(\omega, \iota\right), \left(\omega', \iota'\right)\right)\mathbf{x} = s_1^2 \exp\left[\frac{-\left(\omega - \omega'\right)^2}{2\ell^2}\right] + s_2^2\left(\iota\iota'\right), \tag{5.2}$$

where s_1 and s_2 are scaling factors and ℓ is a SE characteristic length scale.

Likelihood function: Following the GP classification procedure (Rasmussen and Williams, 2006) and for consistency with standard PF formulations (Kingdom and Prins, 2010; Fründ *et al.*, 2011), we transform the latent function using a cumulative Gaussian likelihood:

$$p\left(y_i = 1 \middle| f_i\right) = \Phi\left(f_i\right) = \int_{-\infty}^{f_i} \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \,. \tag{5.3}$$

Computation of hyperparameters: The set of mean and covariance function hyperparameters for this GP prior is $\boldsymbol{\theta} = \left(c, s_1, s_2, \ell\right)$. Upon observing a set of samples $\left(\mathbf{X}, \mathbf{y}\right)$, we calculate a set of best-fitting hyperparameters by maximizing the log marginal likelihood $\log p\left(\mathbf{y} \middle| \mathbf{X}, \boldsymbol{\theta}\right)$.

Calculation of posterior distribution: After each set of observations $\left(\mathbf{X}, \mathbf{y}\right)$, we compute the posterior distribution $p\left(\mathbf{f}^* \middle| \mathbf{X}, \mathbf{y}, \mathbf{X}^*\right)$ for a finely spaced grid of test samples $\mathbf{X}^*$ across frequency-intensity space: 0.125 to 16 kHz in semitone increments for frequency and −20 to 120 dB in 1-dB increments for intensity.

5.2.3. Sampling Methods

We evaluated the efficacy of several sampling techniques within the GP classification framework. In total, we evaluated 4 techniques: 1 random sampling scheme, 1 deterministic sampling scheme and 2 active sampling schemes. Each is described in more detail below.

- *Random sampling:* On each iteration, the sample point selected is determined at random from the set of all possible grid points. This technique reflects a useful efficiency baseline to which each other sampling technique can be compared.

- *Halton sampling:* On each iteration, the sample point chosen is drawn from the corresponding ordered value in a Halton sequence, which provides a well-spaced deterministic set of draws on some predefined interval (Halton, 1964). The work done in

Chapter 4 utilized this approach to effectively estimate multidimensional psychometric functions (Song *et al.*, 2017) and can be used as a point of comparison.

- *Uncertainty sampling:* As described in Section 2.2.2, an uncertainty sampling framework selects samples in which the model is least certain about their corresponding identities (Lewis and Catlett, 1994; Lewis and Gale, 1994; Settles, 2009). The acquisition function for uncertainty sampling is the posterior GP variance in y, σ_y^2 (2.11). Work described in Chapter 3 has demonstrated that uncertainty sampling can produce estimates of audiogram thresholds consistent with those produced by the HW approach (Song *et al.*, 2015), and we extend this technique to inference of the entire psychometric function.

- *Bayesian active learning by disagreement (BALD):* As described in Section 2.2.3, BALD seeks sample points for which model settings most disagree about the outcome (Houlsby *et al.*, 2011). The acquisition function in this case is the expected decrease in posterior entropy, which is approximated in (2.12). BALD has previously been used to infer multidimensional neural receptive field and functional magnetic resonance imaging structure (Park *et al.*, 2011; Park, 2013) as well as simulated audiogram thresholds (Gardner *et al.*, 2015b), and is used here to infer multidimensional psychometric functions using binary responses.

Heuristics: For both active sampling methods (uncertainty sampling and BALD), we added zero-mean Gaussian noise with a small variance to each point in the normalized acquisition function, i.e. $\mathcal{N}\left(0,\sigma_n^2\right)$, where $\sigma_n^2 = 0.2$. This heuristic generally improved sampling results by reducing repeated sampling within a small area of the input space yet still allowing acquisition function magnitude to dominate for sample selection.

5.2.4. Evaluation

We evaluated performance of the technique for each of several model and sampling parameters:

- *Audiogram shapes:* We used older-normal, sensory, metabolic, and sensory + metabolic audiogram phenotypes (Dubno *et al.*, 2013) to generate α values across frequency.
- *Spread value:* β values of 0.2, 0.5, 1, 2, 5, and 10 dB/percent, assumed isotropic across all frequencies, were used to determine spread.
- *Sampling method:* Random sampling, Halton sampling, uncertainty sampling, and BALD were chosen as distinct sampling strategies within the GP framework.
- *Sample number:* To identify the effect of number of observed responses on model performance, we conducted sampling iteratively up to 100 observations, and the performance of each sample count (1-100) was evaluated.
- *Simulation repetition:* For each distinct parameter set, we conducted 10 independent repetitions of GP inference, resulting in 96000 simulations overall.

We assessed model prediction in several ways. At each frequency on the fine grid (0.25 to 8 kHz in semitone increments), the α and β value of a 1D psychometric function can be derived by finding the *x*-intercept and inverse slope of the latent function f and can be compared to the α and β values of the known true PF at that frequency. Notably, we used edge frequencies (0.125 to 0.25 and 8 to 16 kHz) to train the GP but not to evaluate prediction due to known edge effects of psychometric function estimation (Song *et al.*, 2015; Song *et al.*, 2017). We evaluated accuracy using mean deviation of parameter estimates from the true value, and we evaluated reliability using the variance of model estimates across all repetitions with the same parameter values and sampling schemes. Furthermore, the model's point estimate for the overall 2D psychometric function can be computed by passing f through the likelihood function $p\left(y=1 \middle| f\right)$ and can be numerically compared to the true psychometric function.

We evaluated goodness of fit of the GP predictions to the observations using the Pearson χ^2 statistic: $\chi^2 = \sum_i \frac{N_i\left[p\left(\mathbf{x}_i\right) - P_i\right]^2}{P_i\left(1-P_i\right)}$, where $p\left(\mathbf{x}_i\right)$ is the percent correct of the data, P_i is the percent correct of the model prediction, and N_i is the number of trials at a frequency/intensity pair $\mathbf{x}_i = \left(\omega_i, \iota_i\right)$ (Klein, 2001; Wichmann and Hill, 2001a). For each input $\mathbf{x}_i$ in the test set $\mathbf{X}^*$, we computed the statistic and compared its value to the chi-squared distribution with *J* degrees of freedom, where *J* is the number of distinct frequency/intensity pairs sampled.

5.3. Results

In 174 trials out of 96000 (0.18%), a numerical issue with the algorithm prevented it from executing properly. These failed trials were therefore removed, with the remaining 95826 successful trials used in all following analysis. Any observation that would have been provided by a failed trial was replaced with a query at a random input point. Furthermore, a small number of trials (1625 of 95826) generated outliers due to poor hyperparameter convergence. Because of their disproportionate influence on the mean trends, these trials were omitted from the averaged data in Figures 5.4-5.6 by thresholding at the 98th percentile. Similarly, trials with fewer than 15 observations were not displayed in these figures due to generally poor performance, particularly for random and Halton sampling.

A representative run of GP inference after 100 BALD samples can be seen in **Figure 5.1**. In this example ground truth was a sensory phenotype with β = 2 dB/%. Figure 5.1A shows the distribution of samples collected in frequency-intensity space for this particular run. Note that the samples are largely concentrated within a band around putative threshold where they would be particularly useful at refining the PF estimate. The posterior mean after 100 iterations is shown in the background. Figure 5.1B shows the estimated α curve across frequency compared to the true α curve. After 100 samples, the BALD procedure produces a continuous estimate of threshold as a function of frequency that closely matches ground truth. Figure X2C shows a slice of the model PF (the 1D PF) at ω = 1 kHz compared to the ground truth slice at that frequency. The close match of these two curves indicates that the BALD estimate of β matches ground truth as well as α. The agreement in spread extends across frequency, as well.

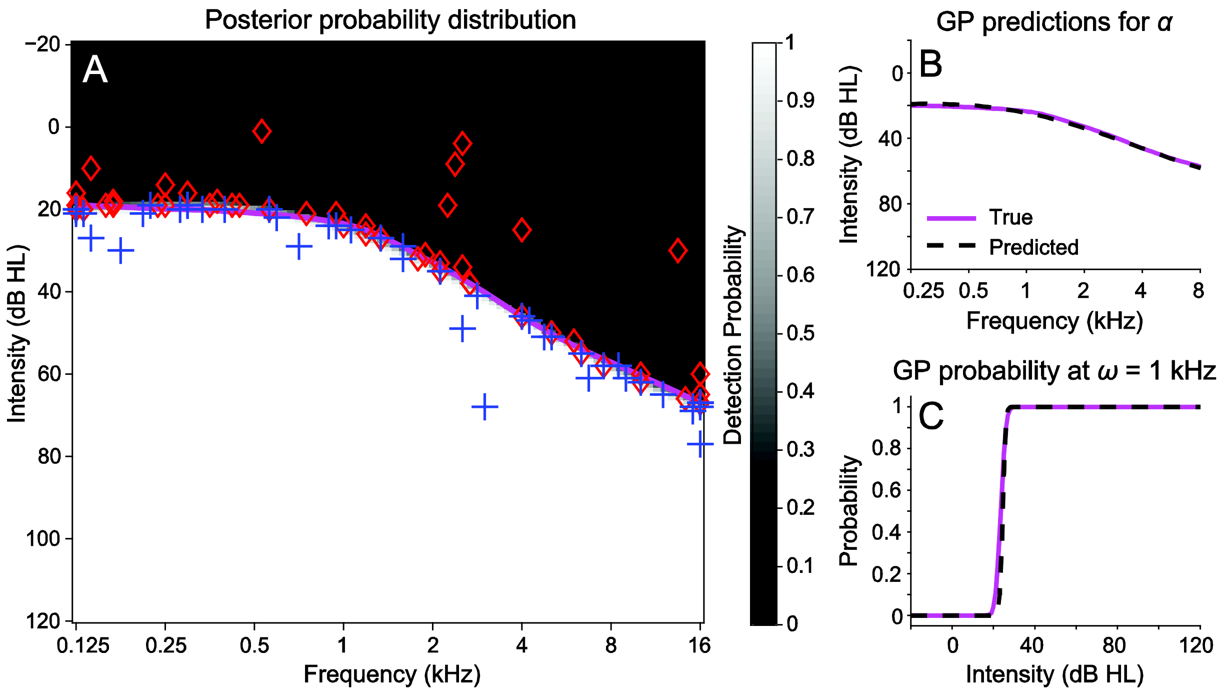

Figure 5.1: Representative example of GPC inference using active sampling. (A) Posterior mean with overlaid samples and α threshold curve for ground truth. (B) The GP estimate of α (dashed line) as compared to the true values of α (solid line). (C) A slice of the estimated psychometric function at ω = 1 kHz (dashed line), compared with a slice of the true psychometric function at that frequency (solid line). Ground truth was a metabolic phenotype with β = 1 dB/%. BALD was used as the sampling method.

The BALD sampling method produces desired estimator behavior by selecting tones near detection threshold. Its relative performance can be seen in **Figure 5.2**, which shows representative single GP estimation runs (posterior mean and selected samples) for each sampling method after 100 samples. The ground truth surface is a metabolic + sensory phenotype with β = 2 dB/%, shown in the inset figure. BALD (Figure X3A) and uncertainty (Figure X3B) sampling schemes show similar sample selection around putative threshold, with the spread measures for BALD more closely resembling ground truth. Halton sampling (Figure X3C) selects relatively well-spaced draws spanning the entire input space, but samples are not densely populated along the threshold line compared to BALD and uncertainty sampling. Random sampling (Figure X3D) shows a predicted threshold curve that is noticeably divergent from the ground truth surface for these 100 samples. These distinct sampling methods show differences in estimated PFs, with the BALD and uncertainty sampling estimates comparing most favorably to ground truth as sampling progresses.

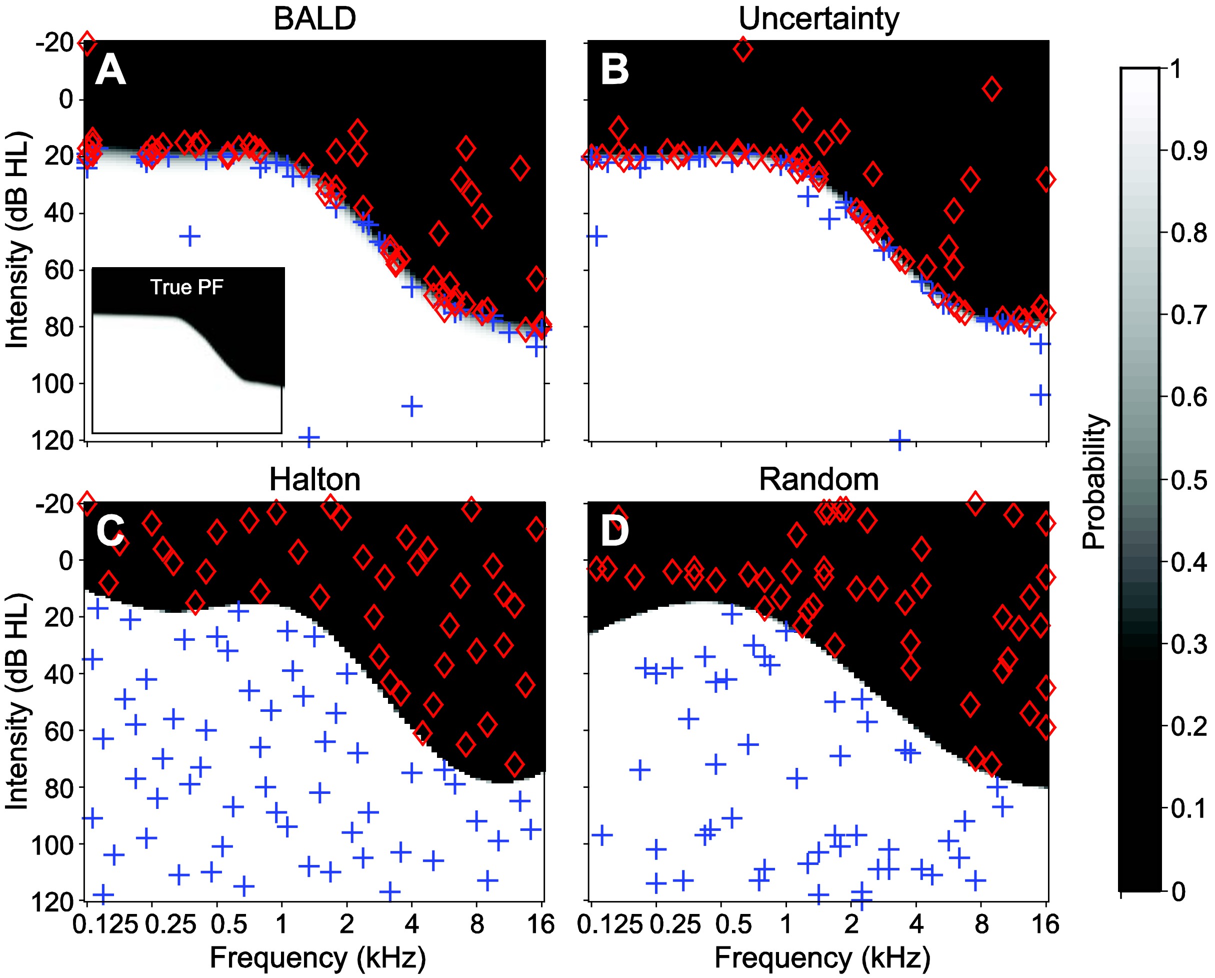

Figure 5.2: Representative posterior distributions for each sampling method. Data are shown after 100 samples, with ground truth being a metabolic + sensory phenotype with $\beta = 2$ dB/%. Each subplot corresponds to a distinct sampling method: (A) BALD sampling; (B) uncertainty sampling; (C) Halton sampling; and (D) random sampling. Inset in (A) shows the true psychometric surface used to generate all responses. Blue plus signs denote detected stimuli; red diamond signs denote undetected stimuli.

The relative tendency of any point in frequency/intensity space to be queried for a particular sampling scheme can be visualized using mean acquisition maps, which are shown in **Figure 5.3** for the same psychometric surface as in Figure 5.2. The acquisition map is constructed by averaging the acquisition function values across all trials and repetitions for each sampling method, with higher values corresponding to input locations more likely to be queried. BALD and uncertainty (Figure 5.3A and 5.3B) sampled densely around threshold, with uncertainty appearing slightly more tightly distributed near the threshold curve. Because Halton sampling (Figure 5.3C) is deterministic, only points within a small subset have high acquisition function values. Random sampling (Figure 5.3D) results in a predictably random acquisition map.

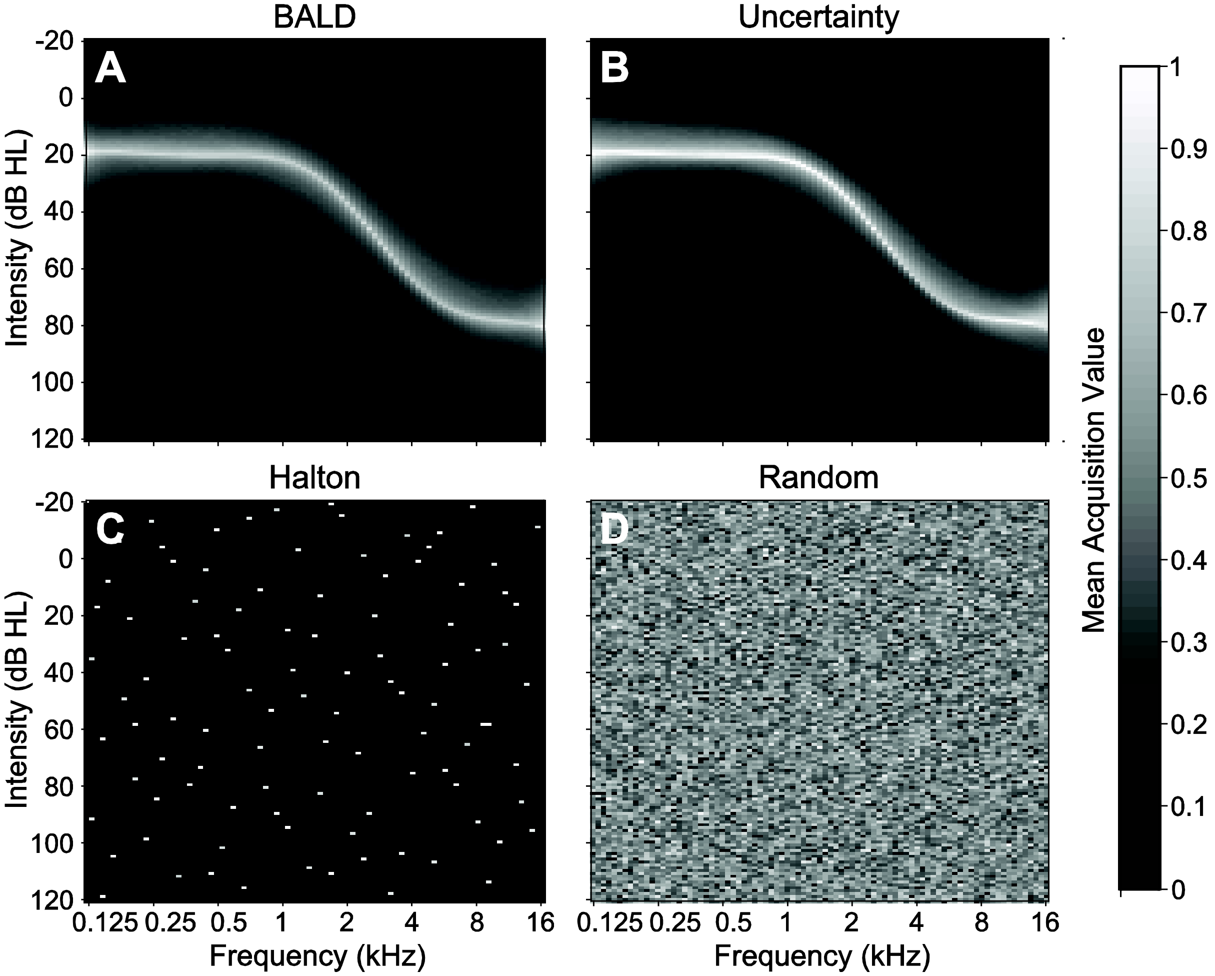

Figure 5.3: Mean acquisition maps for each sampling technique. Ground truth is a metabolic + sensory phenotype with β = 2 dB/% (Figure 5.2 inset). Each plot shows the normalized acquisition function map averaged across all iterations and repetitions for a single sampling method: (A) BALD sampling; (B) uncertainty sampling; (C) Halton sampling; and (D) random sampling.

Figure 5.4 shows the overall performance of each sampling method as a function of sample number averaged across all phenotypes, true β values and repetitions. Metrics evaluated were error in threshold α, error in spread β, and mean pointwise difference between predicted and ground truth probability surfaces, a nonparametric comparison. Active methods strongly outperformed non-active methods for α estimation and mean pointwise difference, with trends in β being more consistent between all 4 sampling types. For α prediction, active sampling methods were within clinical reliability criterion (±5 dB) across all frequencies in approximately 20 tones, with non-active methods requiring approximately 50-60 tones to achieve the same criterion. Between the non-active methods, random sampling exhibited generally poorer performance compared to Halton sampling.

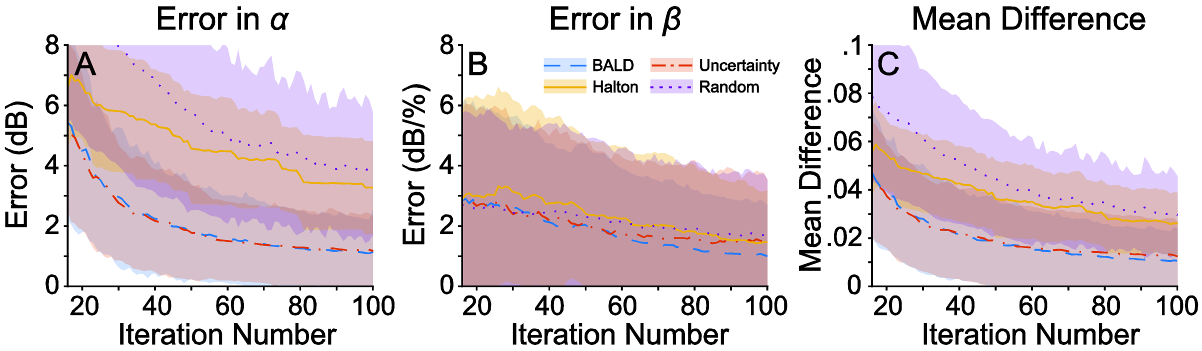

Figure 5.4: Summary of active GP performance across all conditions. Colors correspond to sampling modes; lines denote the mean and shaded areas denote 1 standard deviation above and below. As a function of iteration, (A) and (B) show differences between model predictions and ground truth for α and β, respectively, and (C) shows the mean absolute difference in probability value between the GP posterior mean and ground truth.

The influence of a particular audiogram phenotype on estimator performance is evaluated in **Figure 5.5**, which shows the prediction error of each sampling method for different audiogram profiles, averaged across true β values and repetitions. Older-normal phenotypes demonstrated the best performance due to its relatively flat shape. Performance was generally similar for the remaining 3 phenotypes, although Halton sampling appears comparatively worse for metabolic and metabolic + sensory phenotypes, particularly with small sample numbers.

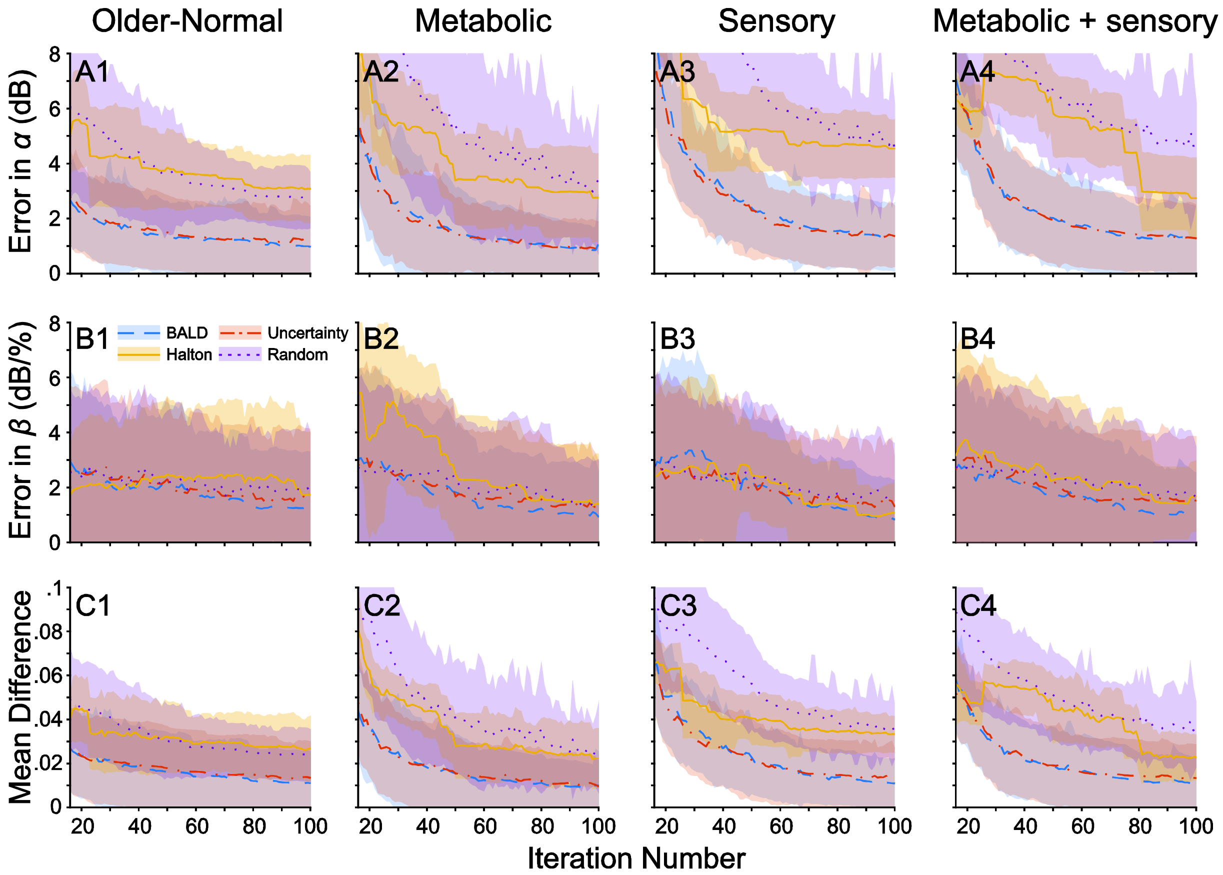

Figure 5.5: Summary of active GP performance by audiogram phenotype. Colors correspond to sampling modes; lines denote the mean and shaded areas denote 1 standard deviation above and below. Differences in α and β between model predictions and ground truth for each phenotype are shown in (A1-A4) and (B1-B4), respectively; (C1-C4) shows the mean absolute difference in probability value between the posterior mean and ground truth.

Higher values of spread β introduced increased uncertainty in the PF transition zones. The effect of β value on estimator performance was evaluated in **Figure 5.6**, which shows the prediction error for each sampling method for different β values, averaged across different phenotypes and repetitions. The results demonstrate broadly decreasing accuracy for all metrics with increasing β. Active sampling methods generally outperform non-active methods, with the exception of β inference using uncertainty sampling for the highest spread value (Figure 5.6B6).

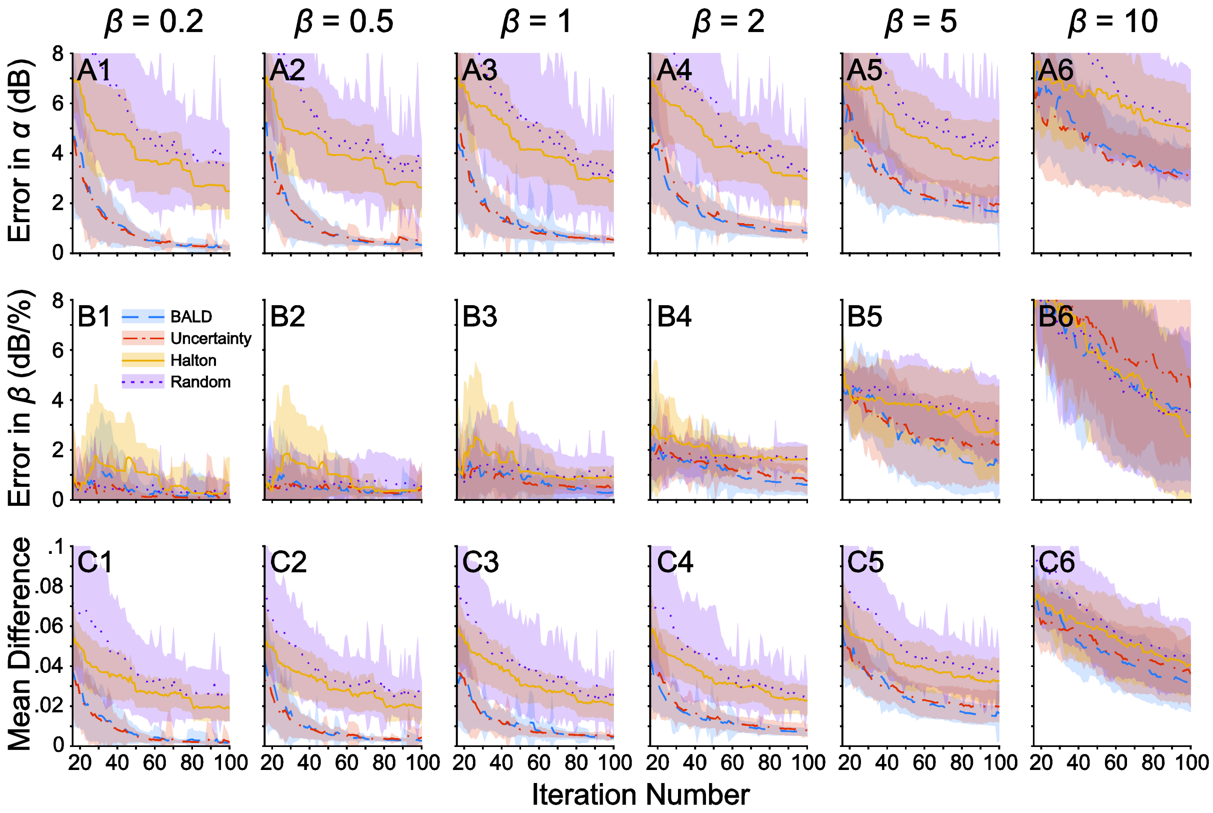

Figure 5.6: Summary of active GP performance by β (spread) value. Colors correspond to sampling modes; lines denote the mean and shaded areas denote 1 standard deviation above and below. Differences in α and β between model predictions and ground truth for each phenotype are shown in (A1-A6) and (B1-B6), respectively; (C1-C6) shows the mean absolute difference in probability value between the posterior mean and ground truth.

Across all 95826 valid simulations, the median χ^2 statistic was 5.03×10^{-5}. After using the chi-square distribution with the appropriate degrees of freedom to compute probability values for each trial, 9 of 95826 trials, or 0.0094%, were detected to be statistically poor fits at an uncorrected significance level of $p < 0.05$.

5.4. Discussion

In this chapter, we describe a method of accurately and efficiently estimating multidimensional psychometric functions by combining Gaussian process classification and active sampling. The current work extends previous efforts employing this estimation framework with deterministic sampling techniques (Song *et al.*, 2017). Active sampling produces a marked efficiency increase for the estimation of complex psychometric surfaces. Notably, active sampling methods required approximately 20-30 samples to reach the clinical reliability criterion of ±5 dB in audiometric threshold estimation (Fausti *et al.*, 1990; Swanepoel *et al.*, 2010), which is consistent with data from human subjects (Song *et al.*, 2015). This sample number improves considerably upon the approximately 50-60 samples needed to reach the ±5 dB criterion for random and Halton sampling (Song *et al.*, 2017), as well as upon the number of samples typically required for a complete Hughson-Westlake run, which often utilizes 100 or more samples per ear (Carhart and Jerger, 1959; Katz *et al.*, 2009; Mahomed *et al.*, 2013; Song *et al.*, 2015).

Overall, non-active techniques (random and Halton) performed comparatively more poorly than active techniques, with random sampling generally performing the worse of the two. The two active sampling techniques (uncertainty and BALD) exhibited similar performance, with both methods selecting samples near the putative threshold as expected. Uncertainty sampling demonstrated some difficulty with accurate β estimation for higher spread values (Figure 5.6B6) compared to the other sampling techniques, however. This result can be attributed to the fact that uncertainty sampling iteratively selects samples closest to a probability of detection of 0.5 and may not span a sufficient intensity range around threshold to adequately estimate larger β values. BALD incorporates an inherent tradeoff between exploration and exploitation (Houlsby *et al.*, 2011) which helps mitigate this issue.

Notably, the heuristic of adding zero-mean Gaussian noise $\mathcal{N}\left(0,\sigma_n^2\right)$ to the acquisition function at each input point for BALD improved the accuracy of BALD and uncertainty sampling (data not shown). Although modest improvements were observed for uncertainty sampling, performance of BALD was dramatically improved with the implementation of this heuristic.

Unmodified, BALD samples exhibited a high degree of clustering, particularly along the edges in the frequency dimension. The additive Gaussian noise allowed for selection of points that were not the absolute maximum of the acquisition function, mitigating the previously observed clustering problem. Despite this added heuristic, the modified sampling technique still produced the desired sampling behavior on average (see Figure 5.3A).

5.5. Concluding Remarks

This chapter describes a method for estimating multidimensional psychometric functions with actively selected queries, extending the work in Chapter 4 using this technique with non-active samples. The method makes use of Gaussian process classification, a nonparametric Bayesian inference framework that allows for estimating complex PFs with limited categorical observations. Results show that active sampling techniques generally outperform Halton and random sampling, reaching clinical reliability in 20-30 samples. The flexibility of this technique allows for straightforward extension to other psychophysical domains, integration with other active sampling strategies, and incorporation of more informative prior beliefs into the model. This technique therefore represents an efficient, flexible method for estimating arbitrary PFs.

Chapter 6: Summary & Future Direction

6.1. Summary of Findings

We have described the development and validation of a novel machine learning framework for estimating multidimensional audiometric functions, the machine learning audiogram (MLAG). In contrast to existing methods for multidimensional psychometric function inference, this method does not require the function of interest to be parameterized, giving it the flexibility to estimate a wide variety of PF shapes. It also integrates well with active sampling procedures, enabling an extra degree of efficiency.

In Chapter 3, we described the application of MLAG with uncertainty sampling to estimate threshold audiograms in human listeners. We found that our technique produced threshold estimates that were consistent with standard clinical Hughson-Westlake estimates and also exhibited test-retest reliability consistent with accepted techniques. However, MLAG required significantly fewer samples than HW to produce its estimates, as well as offered a principled estimate of all frequencies within the range of interest, which has not been demonstrated efficiently by other techniques for estimating threshold audiograms.

We then investigated MLAG’s ability to estimate the audiometric function (the entire PF for pure-tone detection). Because no accepted technique for estimating these multidimensional non-parameterized PFs in human subjects exists as a point of comparison, we instead evaluated the technique using simulated PFs for which ground truth was known. In Chapter 4, we described the use of our estimation framework to infer both 1- and 2-dimensional psychometric functions (thresholds and spreads). In the 1-dimensional case, our technique demonstrated consistency with maximum-likelihood probit regression using samples obtained by the method of constant stimuli. In the 2-dimensional case, MLAG with Halton sampling was able to generate estimates of a wide variety of audiometric functions with high accuracy and reliability. In Chapter 5, we extended the work from Chapter 4 on 2D audiometric functions by adding the active sampling component; we showed that active sampling produced a dramatic decrease in the number of samples required to reach an accurate estimate of the PF, with Bayesian active learning by disagreement being the most promising among our tested strategies.

In summary, we have designed and evaluated a novel technique for estimation of the audiometric function that demonstrates consistency with accepted clinical and psychophysical techniques as well as exhibits a number of other advantages. Although this thesis focuses on the particular application of the audiogram and audiometric function, this technique is general-purpose and can be easily applied to other psychophysical spaces. While the current implementation has some limitations (which are described in the next section), we believe the MLAG framework is a promising technique for audiometric and general psychometric evaluation.

6.2. Recommendations for Future Direction

In moving forward from the work presented in this thesis, there are many possible directions for future research. The simulation studies presented here should be verified with human studies wherever possible, and the robustness of the technique can be tested via application to novel scientific questions or additional psychophysical domains. The mathematical framework for the GP can be modified to account for more specific psychometric phenomena. Additional adjustments can be made to improve the efficiency of the technique, both for audiometric testing and other applications. Finally, this technique and other established methods can be widely distributed to collect additional audiometric data, helping future refinement of the technique.

6.2.1. Human Studies

In Chapter 3, we described work using the GPC method to infer audiometric thresholds in human subjects, with our method demonstrating good agreement with clinical metrics (Song *et al.*, 2015). GPC for full psychometric function estimation (Chapters 4 and 5) should similarly be verified using human subjects in future studies. Additional research using GPC to estimate audiometric PFs in humans would further establish the advantages and validity of this technique, particularly in clinical contexts, and ultimately facilitate adoption by clinicians.

Unfortunately, no techniques to efficiently estimate an entire audiometric function currently exist with which to compare the GPC technique. A logical, if time-consuming, approach is to use the method of constant stimuli (Fechner, 1860) to estimate a 1D PF at a few select frequencies $\boldsymbol{\omega}^*$ in the audiometric space. The GPC technique can then be used to generate an estimate of the 2D PF across the full frequency/intensity space, and individual 1D frequency slices corresponding to $\boldsymbol{\omega}^*$ can be compared to the PFs estimated by the method of constant stimuli. Although the number of frequencies chosen will be limited by the relative inefficiency of the method of constant stimuli, this should still act as a valid comparison against a well-established technique. Additionally, a number of adaptive schemes exist for efficient estimation of 1D PFs (Hall, 1981; Watson and Pelli, 1983; Kontsevich and Tyler, 1999; Leek, 2001), which can be used in place of the method of constant stimuli as a point of comparison if efficiency is a concern.

More exciting is the ability to use the GPC framework to answer scientific questions that have previously been difficult or prohibitive to investigate. For instance, a research area of interest is audiogram microstructure, which describes the set of often periodic fluctuations in threshold as a function of miniscule steps in frequency (Long, 1984; Baiduc *et al.*, 2014; Dewey *et al.*, 2014). This microstructure is reflective of the underlying physiology of the ear, and it is an open question how it is impacted by various disease states. A few simple changes in the Gaussian process construction, for instance a smaller and finer range for test grid $\mathbf{X}^*$ and a periodic component in the covariance function $K\left(x, x'\right)$, can enable us to effectively model audiogram microstructure. The ability to investigate this particular example and many other questions may be enabled by this flexible psychometric framework.

6.2.2. Psychometric Extensions

While the specific framework as presented in this thesis can encode a variety of psychometric function shapes, it does nevertheless make some assumptions about the form of the PF within the psychophysical space. Fortunately, GPC as a whole has large degree of inherent flexibility that allows us to encode additional properties of PFs in a straightforward manner within the same framework, with a few tweaks to how the GP is defined.

The current likelihood function $p\left(y = 1 \middle| f\right)$ is a sigmoidal function that spans the entire range $\left[0, 1\right]$ (Kuss *et al.*, 2005; Song *et al.*, 2017). This choice implicitly assumes that false positive and lapse rates take values of 0. For detection tasks, these rates are generally close to 0, but failing to account for actual lapse rates can introduce bias into the estimate when lapses truly exist (Wichmann and Hill, 2001a). The GPC framework can account for arbitrary lapse rates by modification of the likelihood function, though requiring additional hyperparameters and complexity. This generalization would also allow for ready extension of the GP framework to other psychometric tasks, such as *n*-alternative forced choice (Kingdom and Prins, 2010).

Because of the linear term in the current covariance function $s^2\iota\iota'$, the current model assumes that the value of β does not vary as a function of frequency. However, frequency-dependent changes in psychometric attributes have been observed in hearing-loss individuals (Nelson and Freyman, 1986; Freyman and Nelson, 1991), suggesting possible non-uniform psychometric uncertainty along frequency. To model a psychometric function whose β values vary with frequency, we can modify the covariance function by including a frequency-dependent multiplicative term in addition to the linear term. Note that because a SE covariance function places constraints on covariances rather than function shapes, no similar limitation exists for α: the overall function smoothness need only be constrained to be uniform across the entire domain. Other covariance functions with periodic behavior or whose length scales change with some dimension can also be chosen to encode certain function behaviors, given prior knowledge.

Finally, perhaps the most logical extension of the current work is for inference on another psychophysical domain. While this thesis has focused on using GPC for estimation of audiometric functions, the method is general-purpose and can be extended to any uni- or multidimensional psychometric domain with binary responses. One straightforward application is the visual field (Heijl and Krakau, 1975; Bengtsson *et al.*, 1997), a 3-dimensional space that can describe the integrity of a patient's central or peripheral vision. The input dimensions for this modality are $\mathbf{x} = \left(x_h, x_v, \iota\right)$: non-psychometric horizontal and vertical positions x_h, x_v and a psychometric dimension intensity ι. Responses y remain binary variables, allowing the GPC framework to proceed intact except for the added dimension and new choices of mean and/or covariance functions to reflect prior information about visual fields. A few tweaks to the GP setup would also allow our method to be used for regression problems or for classification problems with more than 2 possible categories.

6.2.3. Efficiency Improvements

The GPC algorithm as implemented in this thesis uses fairly uninformative priors, with very little information encoded specific to hearing. Its only assumptions about the shape of the full audiometric function are that it is uniformly smooth along frequency and sigmoidal along intensity. Further development of the algorithm could incorporate reasonable constraints on hyperparameters or on the final estimated function shape based upon the physics of sound transmission and known factors of cochlear function. The choice of constant mean function in the current work, although effective, is particularly naive; more informative prior mean functions could be generated by examining trends in typical human PFs. The current implementation of this algorithm also uses uniform priors on its covariance function hyperparameters, but certain hyperparameter ranges are more realistic than others. If and when sufficient data is available, reasonable hyperprior probabilities should be chosen to reflect PF shape distributions in reality.

The work in this thesis employed two active sampling methods: uncertainty sampling (Lewis and Catlett, 1994; Lewis and Gale, 1994; Settles, 2009) and to a somewhat lesser extent, Bayesian active learning by disagreement (Houlsby *et al.*, 2011). These methods are simple (uncertainty sampling, specifically, is a particularly greedy approach) and have shown to be fairly effective, but it is worth investigating other active sampling methods to assess their efficiency and estimation accuracy. Numerous other active sampling frameworks have demonstrated favorable results for PF estimation or in other domains (King-Smith *et al.*, 1994; Kontsevich and Tyler, 1999; Leek, 2001; Lesmes *et al.*, 2006; Park, 2013; Shen and Richards, 2013) and are compatible with the GP framework. Furthermore, clinical methods such as HW can be used to select samples, with the GPC framework still performing more detailed inference of the psychometric space. Using established clinical procedures for sampling and machine learning for inference may serve as a useful transition for clinical adoption.

Currently, our framework deliberately assumes that the measurements between left and right ears are independent, consistent with accepted clinical standards (American National Standards Institute, 2004a; Katz *et al.*, 2009). However, within the same individual there is often substantial correlation between measurements in one ear and the other; for instance, both ears are typically exposed to the same acoustic environments that can lead to progressive hearing loss. Qualitatively, this can be seen in Figure 6.1, which shows substantial overlap between left and right ear threshold audiograms within the same subject. We can modify the covariance function to account for this by using a discrete covariance term between left- and right-ear sample points, allowing for measurements in one ear to provide information about measurements in the other. Like other hyperparameters, a best-fitting value for this covariance can also be learned given the data. By broadly expanding this concept, we can also encode relationships between separate but related psychophysical domains. Some examples from audiometry include air conduction and bone conduction, noise-masked and unmasked pure-tone delivery, and tone-based and word-based detection tasks (Katz *et al.*, 2009; Stach, 2010; Martin and Clark, 2015).

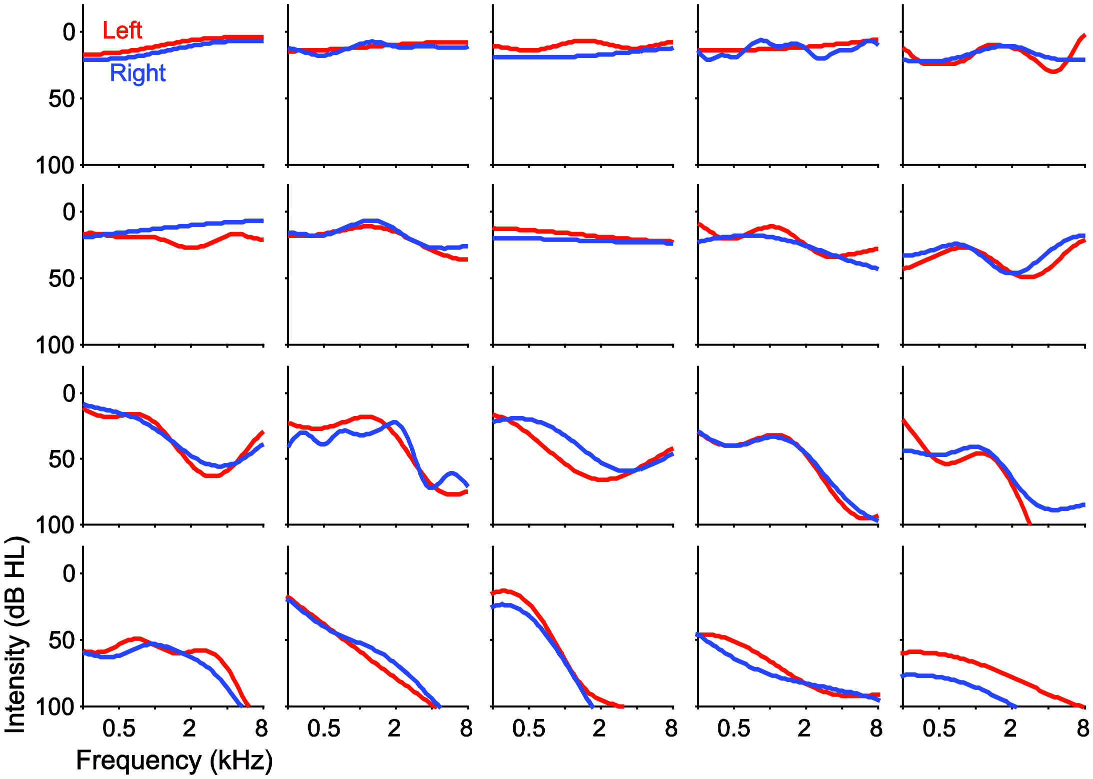

Figure 6.1: Left- and right-ear GP human audiogram estimates. Data is from the experiment performed in Chapter 3. Note the generally high correlation between left and right ears across all subjects.

The active sampling employed in this thesis queries points one at a time. However, previous work has demonstrated additional information gain when 2 tones are presented on each iteration and the subject responds when *either* stimulus is detected, leading to a corresponding efficiency gain (Gardner *et al.*, 2015b). While this technique is currently computationally expensive and difficult to use in real time, further advances that mitigate this disadvantage could lead to even more rapid estimates of audiometric threshold or PF. Additionally, this “or-channel” framework could be extended to 3 or more simultaneous stimulus deliveries, which could produce further efficiency gains but will likely face diminishing returns.

The current form of the GP mean and covariance functions is very flexible, allowing for a wide variety of audiometric shapes to be represented. However, the downside to this flexibility is a relative lack of specificity; an algorithm specific to a particular shape may generate a more detailed estimate of the space as well as reach the estimate more efficiently. In practice, there are a limited number of typical audiogram categories; while initially unknown for any particular subject, the ability to quickly deduce to which category this individual's audiogram belongs could allow the remaining samples to be tailored to estimating this particular shape, improving both accuracy and efficiency. Previous work (Gardner *et al.*, 2015a) has utilized Bayesian active model selection (Ali *et al.*, 2014) to rapidly distinguish between smooth audiogram shapes and noise-induced hearing loss (Rabinowitz, 2000; Shargorodsky *et al.*, 2010), which exhibits narrow notches that are difficult for the current algorithm to detect efficiently. This work can be extended to additional models corresponding to various hearing pathologies, and can also be weighted by their relative prevalence in reality.

A common need in both clinical audiology as well as certain hearing conservation programs is to obtain repeated threshold audiogram measures in the same individual (Sataloff and Sataloff, 2005). However, current retests require performing the entire inference procedure; this high time cost is an obstacle to more regular retests, which could be particularly useful in occupations that have high noise exposure. However, there is substantial prior information on any individual taking a retest: that individual's previous test results. If these previous data could be incorporated into the retest, we could greatly increase the efficiency of the new test. One strategy is to include all observed data from previous tests with the retest, but this will increase runtime due to the computational complexity of the GP. One promising solution is the sparse GP, which chooses a new, much smaller set of samples that induces a similar posterior distribution to that observed with a full set of points (Lawrence *et al.*, 2003; Seeger, 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006). By using a sparse representation on the retest, we can incorporate information about the previous tests' posterior distributions while minimally increasing computational cost.

6.2.4. Large-Scale Distribution

As mentioned in Section 6.2.3, we can use existing audiometric data to inform our choices for Gaussian process mean or covariance function or to assign reasonable hyperparameter prior distributions. However, most audiometric data exists in the form of pure-tone audiograms, which report thresholds at 6-9 discrete frequencies but offer relatively little information for GP construction. Although the HW approach queries numerous samples to determine the threshold audiogram, these observations are almost never retained in practice because sample points are tuned manually by audiological professionals. This is unfortunate, because these observations offer substantially more information than the thresholds themselves.

An accessible, user-friendly platform for conducting audiometric testing could enable collection of audiometric data in both larger quantities and in higher detail compared to current practice. Such a platform could be implemented through a website that directly delivers audiometric tests for use by clinicians, researchers and other interested individuals. Figure 6.2 shows a screenshot of a current website implementation, which offers GPC inference as well as an automated version of the clinical HW procedure. For any user account, all associated data including individual tests, threshold and probability surfaces, queried sample points and user responses, are saved server-side and can be reviewed at any time.

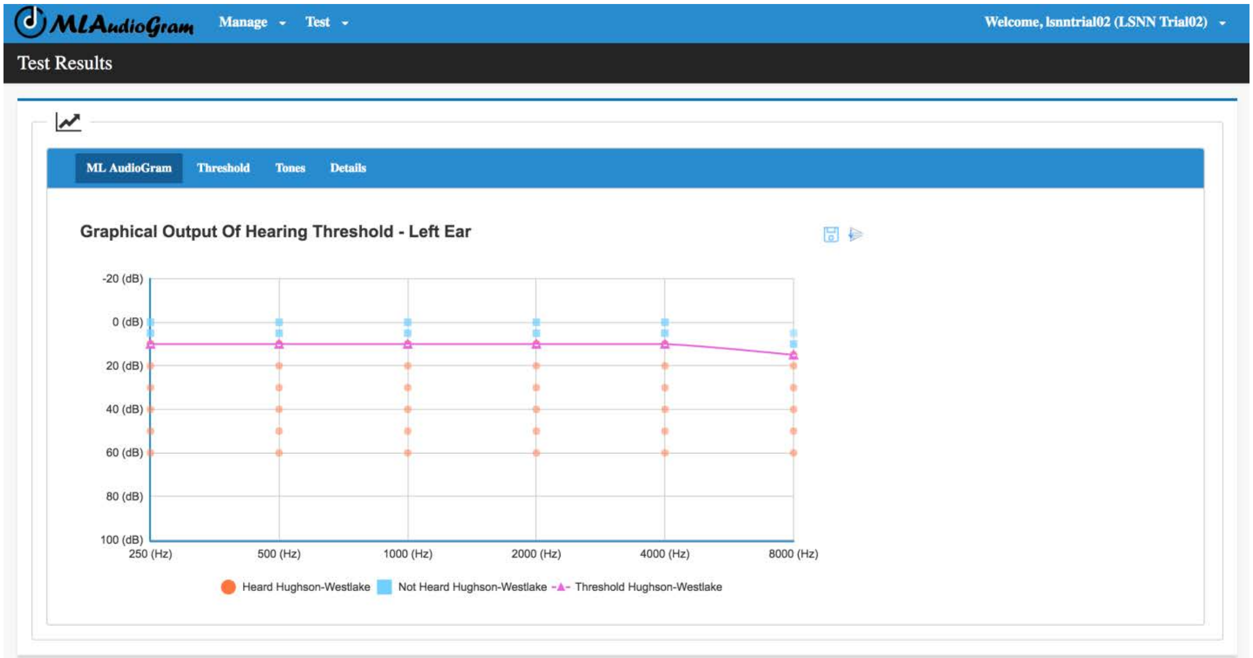

Figure 6.2: Screenshot of the current website implementation for remote audiometric testing. The current screenshot is of the remote version of Hughson-Westlake, although the GPC audiogram is available as well.

A challenge facing computer-based remote diagnosis is accounting for calibration differences between machines. Distinct computers typically use different soundcards and may deliver audio through a variety of headsets or speakers, making comparison of tests conducted on different machines difficult. A calibration scheme should be developed to facilitate cross-device comparison. One possibility is to use Gaussian process regression (Rasmussen and Williams, 2006) and some active sampling technique (Settles, 2009) to rapidly infer a transformation function between a server-side putative tone intensity and a reading from a sound level meter on the user end. Even if uncalibrated, however, it is possible to compare different tests conducted on the same machine as long as the hardware and software settings are held constant.

6.3. Concluding Remarks

This thesis has presented a method for estimating audiometric functions using machine learning classification. The technique has demonstrated the ability to estimate threshold audiograms in human subjects and full audiometric functions in simulated data, comparing favorably to existing techniques and clinical standards whenever possible. An important next step will be to verify the simulation results for audiometric function estimation in human subjects. Future work can also address the limitations of the current implementation by incorporating additional psychometric theory or by further improving the algorithm's efficiency for human testing. In summary, this work presents a promising audiometric estimation method for potential use in clinic or for research, with clear possibilities for expanding upon the current work in moving forward.

References

Ali, A., Caruana, R., and Kapoor, A. (**2014**). "Active Learning with Model Selection," in *AAAI Conference on Artificial Intelligence*, pp. 1673-1679.

Allen, P., and Wightman, F. (**1994**). "Psychometric functions for children's detection of tones in noise," J Speech Lang Hear Res **37**, 205-215.

American National Standards Institute (**2004a**). "Methods for manual pure-tone threshold audiometry," in *ANSI 3.21* (New York).

American National Standards Institute (**2004b**). "Specification for audiometers," in *ANSI 3.6* (New York).

American Speech-Language-Hearing Association (**2005**). "Guidelines for manual pure-tone threshold audiometry."

Amitay, S., Hawkey, D. J., and Moore, D. R. (**2005**). "Auditory frequency discrimination learning is affected by stimulus variability," Percept Psychophys **67**, 691-698.

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (**2003**). "An introduction to MCMC for machine learning," Mach Learn **50**, 5-43.

Angluin, D. (**1988**). "Queries and concept learning," Mach Learn **2**, 319-342.

Angluin, D. (**2004**). "Queries revisited," Theor Comput Sci **313**, 175-194.

Appollonio, I., Carabellese, C., Frattola, L., and Trabucchi, M. (**1996**). "Effects of sensory aids on the quality of life and mortality of elderly people: a multivariate analysis," Age Ageing **25**, 89-96.

Atlas, L. E., Cohn, D. A., Ladner, R. E., El-Sharkawi, M. A., Marks, R. J., Aggoune, M., and Park, D. (**1989**). "Training Connectionist Networks with Queries and Selective Sampling," in *Adv Neural Inf Process Syst* (Morgan Kaufmann), pp. 566-573.

Baiduc, R. R., Lee, J., and Dhar, S. (**2014**). "Spontaneous otoacoustic emissions, threshold microstructure, and psychophysical tuning over a wide frequency range in humans," J Acoust Soc Am **135**, 300-314.

Bargones, J. Y., Werner, L. A., and Marean, G. C. (**1995**). "Infant psychometric functions for detection: Mechanisms of immature sensitivity," J Acoust Soc Am **98**, 99-111.

Bayes, T., and Price, R. (**1763**). "An essay towards solving a problem in the doctrine of chance," Philos Trans R Soc London **53**, 370-418.

Bengtsson, B., Olsson, J., Heijl, A., and Rootzén, H. (**1997**). "A new generation of algorithms for computerized threshold perimetry, SITA," Acta ophthalmol **75**, 368-375.

Berger, J. O. (**1985**). *Statistical Decision Theory and Bayesian Analysis* (Springer, New York).

Berliner, J. E., and Durlach, N. I. (**1973**). "Intensity perception. IV. Resolution in roving-level discrimination," J Acoust Soc Am **53**, 1270-1287.

Bexelius, C., Honeth, L., Ekman, A., Eriksson, M., Sandin, S., Bagger-Sjöbäck, D., and Litton, J. E. (**2008**). "Evaluation of an internet-based hearing test—comparison with established methods for detection of hearing loss," J Med Internet Res **10**.

Bishop, C. M. (**2006**). *Pattern Recognition and Machine Learning* (Springer, New York).

Blackwell, D. L., Lucas, J. W., and Clarke, T. C. (**2014**). "Summary health statistics for US adults: national health interview survey, 2012," Vital and health statistics. Series 10, 1-161.

Blamey, P. J., Sarant, J. Z., Paatsch, L. E., Barry, J. G., Bow, C. P., Wales, R. J., Wright, M., Psarros, C., Rattigan, K., and Tooher, R. (**2001**). "Relationships among speech perception, production, language, hearing loss, and age in children with impaired hearing," J Speech Lang Hear Res **44**, 264-285.

Bogardus Jr, S. T., Yueh, B., and Shekelle, P. G. (**2003**). "Screening and management of adult hearing loss in primary care: clinical applications," JAMA **289**, 1986-1990.

Bond, M., Mealing, S., Anderson, R., Elston, J., Weiner, G., Taylor, R. S., Hoyle, M., Liu, Z., Price, A., and Stein, K. (**2009**). "The effectiveness and cost-effectiveness of cochlear implants for severe to profound deafness in children and adults: a systematic review and economic model," Health Technology Assessment **13**, 1-330.

Bonino, A. Y., Leibold, L. J., and Buss, E. (**2013**). "Effect of signal-temporal uncertainty in children and adults: tone detection in noise or a random-frequency masker," J Acoust Soc Am **134**, 4446.

Box, G. E., and Tiao, G. C. (**1992**). *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, Hoboken, NJ).

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J Acoust Soc Am **111**, 2801-2810.

Brant, L. J., and Fozard, J. L. (**1990**). "Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging," J Acoust Soc Am **88**, 813-820.

Briscoe, J., Bishop, D. V., and Norbury, C. F. (**2001**). "Phonological processing, language, and literacy: A comparison of children with mild-to-moderate sensorineural hearing loss and those with specific language impairment," J Child Psychol Psychiatry **42**, 329-340.

Brochu, E., Cora, V. M., and De Freitas, N. (**2010**). "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," arXiv preprint arXiv:1012.2599.

Burk, M. H., and Humes, L. E. (**2008**). "Effects of long-term training on aided speech-recognition performance in noise in older adults," J Speech Lang Hear Res **51**, 759-771.

Burk, M. H., and Wiley, T. L. (**2004**). "Continuous versus pulsed tones in audiometry," Am J Audiol **13**, 54-61.

Buss, E., Hall III, J. W., and Grose, J. H. (**2006**). "Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli," J Acoust Soc Am **120**, 2777-2788.

Buss, E., Hall III, J. W., and Grose, J. H. (**2008**). "Psychometric functions for pure tone intensity discrimination: Slope differences in school-aged children and adults," J Acoust Soc Am **125**, 1050-1058.

Carhart, R., and Jerger, J. (**1959**). "Preferred method for clinical determination of pure-tone thresholds," J Speech Hear Disord **24**, 330-345.

Chaloner, K., and Verdinelli, I. (**1995**). "Bayesian experimental design: A review," Stat Sci **10**, 273-304.

Chen, H. L. (**1994**). "Hearing in the elderly. Relation of hearing loss, loneliness, and self-esteem," J Gerontol Nurs **20**, 22-28.

Chien, W., and Lin, F. R. (**2012**). "Prevalence of hearing aid use among older adults in the United States," Arch Intern Med **172**, 292-293.

Chisolm, T. H., Johnson, C. E., Danhauer, J. L., Portz, L. J., Abrams, H. B., Lesner, S., McCarthy, P. A., and Newman, C. W. (**2007**). "A systematic review of health-related quality of life and hearing aids: final report of the American Academy of Audiology Task Force on the Health-Related Quality of Life Benefits of Amplification in Adults," J Am Acad Audiol **18**, 151-183.

Clark, G. M., Pyman, B. C., and Bailey, Q. R. (**1979**). "The surgery for multiple-electrode cochlear implantations," J Laryngol Otol **93**, 215-223.

Cohen, S. M., Labadie, R. F., Dietrich, M. S., and Haynes, D. S. (**2004**). "Quality of life in hearing-impaired adults: the role of cochlear implants and hearing aids," Otolaryng Head Neck Surg **131**, 413-422.

Cohn, D., Atlas, L., and Ladner, R. (**1994**). "Improving generalization with active learning," Mach Learn **15**, 201-221.

Cohn, D. A. (**1994**). "Neural network exploration using optimal experiment design," in *Adv Neural Inf Process Syst* (Denver, CO), pp. 679-679.

Collet, D. (**2003**). *Modelling binary data* (Chapman & Hall, Boca Raton).

Cover, T. M., and Thomas, J. A. (**2012**). *Elements of information theory* (John Wiley & Sons, New York).

Cressie, N. (**1990**). "The origins of kriging," Math geol **22**, 239-252.

Cruickshanks, K. J., Tweed, T. S., Wiley, T. L., Klein, B. E., Klein, R., Chappell, R., Nondahl, D. M., and Dalton, D. S. (**2003**). "The 5-year incidence and progression of hearing loss: the epidemiology of hearing loss study," Arch Otolaryngol Head Neck Surg **129**, 1041-1046.

Dai, H. (**1995**). "On measuring psychometric functions: a comparison of the constant-stimulus and adaptive up-down methods," J Acoust Soc Am **98**, 3135-3139.

Dalton, D. S., Cruickshanks, K. J., Klein, B. E., Klein, R., Wiley, T. L., and Nondahl, D. M. (**2003**). "The impact of hearing loss on quality of life in older adults," Gerontologist **43**, 661-668.

De Souza, C., and Glasscock, M. E. (**2003**). *Otosclerosis and stapedectomy: diagnosis, management, and complications* (Thieme, Stuttgart, Germany).

Dewey, J. B., Lee, J., and Dhar, S. (**2014**). "Effects of contralateral acoustic stimulation on spontaneous otoacoustic emissions and hearing threshold fine structure," J Assoc Res Otolaryngol **15**, 897-914.

DiMattina, C. (**2015**). "Fast adaptive estimation of multidimensional psychometric functions," J Vis **15**, 5-5.

Dixon, W. J., and Mood, A. M. (**1948**). "A method for obtaining and analyzing sensitivity data," J Am Stat Assoc **43**, 109-126.

Dorn, P. A., Piskorski, P., Gorga, M. P., Neely, S. T., and Keefe, D. H. (**1999**). "Predicting audiometric status from distortion product otoacoustic emissions using multivariate analyses," Ear Hear **20**, 149-163.

Dubno, J. R., Eckert, M. A., Lee, F.-S., Matthews, L. J., and Schmiedt, R. A. (**2013**). "Classifying human audiometric phenotypes of age-related hearing loss from animal models," J Assoc Res Otolaryngol **14**, 687-701.

Duvenaud, D. (**2014**). "Automatic Model Construction with Gaussian Processes," in *Department of Engineering* (University of Cambridge, Cambridge, England), pp. 1-132.

Eilers, R. E., Ozdamar, O., and Steffens, M. (**1993**). "Classification of audiograms by sequential testing: Reliability and validity of an automated behavioral hearing screening algorithm," J Am Acad Audiol **4**, 172-181.

Falmagne, J.-C. (**2002**). *Elements of psychophysical theory* (Oxford University Press, New York).

Fan, J., Heckman, N. E., and Wand, M. P. (**1995**). "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," J Am Stat Assoc **90**, 141-150.

Fausti, S. A., Frey, R., Henry, J., Knutsen, J., and Olson, D. (**1990**). "Reliability and validity of high-frequency (8–20 kHz) thresholds obtained on a computer-based audiometer as compared to a documented laboratory system," J Am Acad Audiol **1**, 162-170.

Fechner, G. T. (**1860**). *Elements of Psychophysics* (Holt, Rhinehart & Winston, New York).

Findlay, J. (**1978**). "Estimates on probability functions: A more virulent PEST," Percept Psychophys **23**, 181-185.

Finney, D. J. (**1971**). *Probit analysis* (Cambridge University Press, Cambridge, UK).

Fitzgerald, D. C. (**1996**). "Head trauma: hearing loss and dizziness," J Trauma Acute Care Surg **40**, 488-496.

Franks, J. (**2001**). "Hearing measurement," Occupational exposure to noise: evaluation, prevention and control. World Health Organization, Dortmund, 183-232.

French, N., and Steinberg, J. (**1947**). "Factors governing the intelligibility of speech sounds," J Acoust Soc Am **19**, 90-119.

Freyman, R. L., and Nelson, D. A. (**1991**). "Frequency discrimination as a function of signal frequency and level in normal-hearing and hearing-impaired listeners," J Speech Lang Hear Res **34**, 1371-1386.

Fründ, I., Haenel, N. V., and Wichmann, F. A. (**2011**). "Inference for psychometric functions in the presence of nonstationary behavior," J Vis **11**, 1-19.

Fu, Q. J., and Galvin, J. J., 3rd (**2008**). "Maximizing cochlear implant patients' performance with advanced speech training procedures," Hear Res **242**, 198-208.

García-Pérez, M. A. (**1998**). "Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties," Vision Res **38**, 1861-1881.

Gardner, J., Malkomes, G., Garnett, R., Weinberger, K. Q., Barbour, D., and Cunningham, J. P. (**2015a**). "Bayesian active model selection with an application to automated audiometry," in *Adv Neural Inf Process Syst* (Morgan Kaufmann), pp. 2377-2385.

Gardner, J. M., Song, X. D., Cunningham, J. P., Barbour, D. L., and Weinberger, K. Q. (**2015b**). "Psychophysical testing with Bayesian active learning," in *Uncertain Artif Intell* (Morgan Kaufmann Publishers Inc.), pp. 286-295.

Gescheider, G. A. (**2015**). *Psychophysics: the fundamentals* (Psychology Press, Hove, United Kingdom).

Gibbs, M. N. (**1998**). "Bayesian Gaussian Processes for Regression and Classification," in *Department of Physics* (University of Cambridge, Cambridge, England), pp. 1-123.

Goodman, A. (**1965**). "Reference zero levels for pure-tone audiometer," ASHA **7**, 1.

Gorga, M. P., Neely, S. T., Ohlrich, B., Hoover, B., Redner, J., and Peters, J. (**1997**). "From laboratory to clinic: A large scale study of distortion product otoacoustic emissions in ears with normal hearing and ears with hearing loss," Ear Hear **18**, 440-455.

Gosztonyi Jr., R. E., Vassallo, L. A., and Sataloff, J. (**1971**). "Audiometric reliability in industry," Archives of Environmental Health: An International Journal **22**, 113-118.

Green, D. M. (**1993**). "A maximum-likelihood method for estimating thresholds in a yes-no task," J Acoust Soc Am **93**, 2096-2105.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (John Wiley & Sons, Inc., New York).

Guan, N. (**2011**). "Bayesian Optimal Pure Tone Audiometry with Prior Knowledge," in *Electrical Engineering* (KTH Royal Institute of Technology, Stockholm, Sweden).

Gubner, J. A. (**2006**). *Probability and random processes for electrical and computer engineers* (Cambridge University Press, Cambridge, UK).

Guest, D., Kent, C., and Adelman, J. S. (**2010**). "Why additional presentations help identify a stimulus," J Exp Psychol Hum Percept Perform **36**, 1609-1630.

Guestrin, C., Krause, A., and Singh, A. P. (**2005**). "Near-optimal sensor placements in gaussian processes," in *Proceedings of the 22nd international conference on Machine learning* (Association for Computing Machinery), pp. 265-272.

Hall, J. L. (**1981**). "Hybrid adaptive procedure for estimation of psychometric functions," The Journal of the Acoustical Society of America **69**, 1763-1769.

Halton, J. H. (**1964**). "Algorithm 247: Radical-inverse quasi-random point sequence," Commun ACM **7**, 701-702.

Handzel, O., Ben-Ari, O., Damian, D., Priel, M. M., Cohen, J., and Himmelfarb, M. (**2013**). "Smartphone-based hearing test as an aid in the initial evaluation of unilateral sudden sensorineural hearing loss," Audiology and Neurotology **18**, 201-207.

Hastie, T., Tibshirani, R., and Friedman, J. (**2009**). *The Elements of Statistical Learning* (Springer).

He, N.-j., Dubno, J. R., and Mills, J. H. (**1998**). "Frequency and intensity discrimination measured in a maximum-likelihood procedure from young and aged normal-hearing subjects," J Acoust Soc Am **103**, 553-565.

Hecox, K., and Galambos, R. (**1974**). "Brain stem auditory evoked responses in human infants and adults," Arch Otolaryngol **99**, 30-33.

Heijl, A., and Krakau, C. (**1975**). "An automatic static perimeter, design and pilot study," Acta ophthalmol **53**, 293-310.

Helvik, A.-S., Jacobsen, G., and Hallberg, L. R. (**2006**). "Psychological well-being of adults with acquired hearing impairment," Disabil Rehabil **28**, 535-545.

Helvik, A., Krokstad, S., and Tambs, K. (**2009**). "Hearing loss and risk of early retirement," Eur J Public Health **23**, 617-622.

Henshaw, H., and Ferguson, M. A. (**2013**). "Efficacy of individual computer-based auditory training for people with hearing loss: a systematic review of the evidence," PLoS One **8**, e62836.

Ho, A. T. P., Hildreth, A. J., and Lindsey, L. (**2009**). "Computer-assisted audiometry versus manual audiometry," Otology & Neurotology **30**, 876-883.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. (**2011**). "Bayesian active learning for classification and preference learning," arXiv preprint arXiv:1112.5745.

Hughson, W., and Westlake, H. (**1944**). "Manual for program outline for rehabilitation of aural casualties both military and civilian," Trans Am Acad Ophthalmol Otolaryngol **48**, 1-15.

Ishak, W. S., Zhao, F., Stephens, D., Culling, J., Bai, Z., and Meyer-Bisch, C. (**2011**). "Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry," Audiological Medicine **9**, 40-46.

Jaynes, E. T. (**2003**). *Probability Theory: The Logic of Science* (Cambridge University Press, New York).

Jefferys, W. H., and Berger, J. O. (**1992**). "Ockham's razor and Bayesian analysis," Am Sci **80**, 64-72.

Jerger, J. (**1960**). "Bekesy audiometry in analysis of auditory disorders," Journal of Speech, Language, and Hearing Research **3**, 275-287.

Jerger, J., and Jerger, S. (**1980**). "Measurement of hearing in adults," Otolaryngology **2**, 1225-1250.

Jerlvall, L., and Arlinger, S. (**1986**). "A comparison of 2-dB and 5-dB step size in pure-tone audiometry," Scand Audiol **15**, 51-56.

Jewett, D. L., Romano, M. N., and Williston, J. S. (**1970**). "Human auditory evoked potentials: possible brain stem components detected on the scalp," Science **167**, 1517-1518.

Kaernbach, C. (**1991**). "Simple adaptive testing with the weighted up-down method," Percept Psychophys **49**, 227-229.

Kärber, G. (**1931**). "Beitrag zur kollektiven Behandlung pharmakologischer Reihenversuche [A contribution to the collective treatment of a pharmacological experimental series]," Naunyn-Schmiedeberg's Archives of Pharmacology **162**, 480-483.

Katz, J., Medwetsky, L., Burkhard, R., and Hood, L. (**2009**). *Handbook of Clinical Audiology* (Lippincott Williams & Wilkins).

Kemp, D. T. (**1978**). "Stimulated acoustic emissions from within the human auditory system," J Acoust Soc Am **64**, 1386-1391.

Kiang, N. Y. S., Watanabe, T., Thomas, E. C., and Clark, L. F. (**1965**). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (The MIT Press, Cambridge, MA).

Kidd, G., Jr., Mason, C. R., and Richards, V. M. (**2003**). "Multiple bursts, multiple looks, and stream coherence in the release from informational masking," J Acoust Soc Am **114**, 2835-2845.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., and Supowit, A. (**1994**). "Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation," Vision Res **34**, 885-912.

King-Smith, P. E., and Rose, D. (**1997**). "Principles of an adaptive method for measuring the slope of the psychometric function," Vision Res **37**, 1595-1604.

Kingdom, F. A. A., and Prins, N. (**2010**). *Psychophysics: A Practical Introduction* (Elsevier, London).

Klein, S. A. (**2001**). "Measuring, estimating, and understanding the psychometric function: a commentary," Percept Psychophys **63**, 1421-1455.

Kochkin, S. (**1993**). "MarkeTrak III: Why 20 Million In US Don't Use Hearing Aids For Their Hearing Loss-Part 1," Hear Jour **46**, 20-20.

Kochkin, S. (**2007**). "MarkeTrak VII: Obstacles to adult non-user adoption of hearing aids," Hear Jour **60**, 24-51.

Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (**1988**). "Adaptive staircase techniques in psychoacoustics: a comparison of human data and a mathematical model," J Acoust Soc Am **83**, 1852-1862.

Kontsevich, L. L., and Tyler, C. W. (**1999**). "Bayesian adaptive estimation of psychometric slope and threshold," Vision Res **39**, 2729-2737.

Krige, D. G. (**1951**). "A statistical approach to some mine valuation and allied problems on the Witwatersrand," (University of Witwatersrand, Johannesburg, South Africa).

Kujala, J. V., and Lukka, T. J. (**2006**). "Bayesian adaptive estimation: The next dimension," J Math Psychol **50**, 369-389.

Kuss, M., Jäkel, F., and Wichmann, F. A. (**2005**). "Bayesian inference for psychometric functions," J Vis **5**, 8.

Lawrence, N., Seeger, M., and Herbrich, R. (**2003**). "Fast sparse Gaussian process methods: The informative vector machine," in *Adv Neural Inf Process Syst* (MIT Press, Vancouver, B.C.), pp. 625-632.

Leek, M. R. (**2001**). "Adaptive procedures in psychophysical research," Percept Psychophys **63**, 1279-1292.

Leek, M. R., Dubno, J. R., He, N. J., and Ahlstrom, J. B. (**2000**). "Experience with a yes–no single-interval maximum-likelihood procedure," J Acoust Soc Am **107**, 2674-2684.

Leek, M. R., Hanna, T. E., and Marshall, L. (**1992**). "Estimation of psychometric functions from adaptive tracking procedures," Percept Psychophys **51**, 247-256.

Leibold, L. J., and Bonino, A. Y. (**2009**). "Release from informational masking in children: effect of multiple signal bursts," J Acoust Soc Am **125**, 2200-2208.

Lesmes, L. A., Lu, Z. L., Baek, J., and Albright, T. D. (**2010**). "Bayesian adaptive estimation of the contrast sensitivity function: the quick CSF method," Vision Res **10**, 17 11-21.

Lesmes, L. L., Jeon, S. T., Lu, Z. L., and Dosher, B. A. (**2006**). "Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method," Vision Res **46**, 3160-3176.

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J Acoust Soc Am **49**, 467-477.

Lewis, D. D., and Catlett, J. (**1994**). "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 148-156.

Lewis, D. D., and Gale, W. A. (**1994**). "A sequential algorithm for training text classifiers," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (Springer-Verlag New York, Inc.), pp. 3-12.

Linschoten, M. R., Harvey, L. O., Eller, P. M., and Jafek, B. W. (**2001**). "Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure," Percept Psychophys **63**, 1330-1347.

Long, G. R. (**1984**). "The microstructure of quiet and masked thresholds," Hear Res **15**, 73-87.

Mahomed, F., Eikelboom, R. H., and Soer, M. (**2013**). "Validity of automated threshold audiometry: A systematic review and meta-analysis," Ear Hear **34**, 745-752.

Margolis, R. H., Glasberg, B. R., Creeke, S., and Moore, B. C. (**2010**). "AMTAS: Automated method for testing auditory sensitivity: Validation studies," Int J Audiol **49**, 185-194.

Margolis, R. H., and Morgan, D. E. (**2008**). "Automated pure-tone audiometry: an analysis of capacity, need, and benefit," Am J Audiol **17**, 109-113.

Martin, F. N., and Clark, J. G. (**2015**). *Introduction to audiology* (Pearson Education, Boston).

Mathers, C., Smith, A., and Concha, M. (**2000**). "Global burden of hearing loss in the year 2000," in *Global Burden of Disease* (World Health Organization, Geneva, Switzerland).

Mathias, S. R., Micheyl, C., and Bailey, P. J. (**2010**). "Stimulus uncertainty and insensitivity to pitch-change direction," J Acoust Soc Am **127**, 3026-3037.

McCullagh, P., and Nelder, J. A. (**1989**). *Generalized Linear Models* (CRC press).

Ménière, P., and Atkinson, M. (**1961**). "Ménière's original papers: reprinted with an English translation together with commentaries and biographical sketch," Acta Otolaryngol.

Meyer-Bisch, C. (**1996**). "Audioscan: a high-definition audiometry technique based on constant-level frequency sweeps - A new method with new hearing indicators," Int J Audiol **35**, 63-72.

Miller, J., and Ulrich, R. (**2001**). "On the analysis of psychometric functions: the Spearman-Karber method," Percept Psychophys **63**, 1399-1420.

Minka, T. P. (**2001**). "Expectation propagation for approximate Bayesian inference," in *Uncertain Artif Intell 17* (Morgan Kaufmann Publishers Inc., Seattle, WA), pp. 362-369.

Mohr, P. E., Feldman, J. J., Dunbar, J. L., McConkey-Robbins, A., Niparko, J. K., Rittenhouse, R. K., and Skinner, M. W. (**2000**). "The societal costs of severe to profound hearing loss in the United States," Int J Technol Assess Health Care **16**, 1120-1135.

Molander, P., Nordqvist, P., Öberg, M., Lunner, T., Lyxell, B., and Andersson, G. (**2013**). "Internet-based hearing screening using speech-in-noise: Validation and comparisons of self-reported hearing problems, quality of life and phonological representation," BMJ Open **3**, e003223.

Morgan, B. J. (**1992**). *Analysis of quantal response data* (Springer, New York).

Mori, S., and Ward, L. M. (**1992**). "Intensity and frequency resolution: masking of absolute identification and fixed and roving discrimination," J Acoust Soc Am **91**, 246-255.

Mulrow, C. D., Aguilar, C., Endicott, J. E., Velez, R., Tuley, M. R., Charlip, W. S., and Hill, J. A. (**1990**). "Association between hearing impairment and the quality of life of elderly individuals," J Am Geriatr Soc **38**, 45-50.

Murphy, K. P. (**2012**). *Machine Learning: A Probabilistic Perspective* (MIT press).

Neal, R. M. (**1993**). "Probabilistic inference using Markov chain Monte Carlo methods," (University of Toronto, Toronto, Ontario).

Nelder, J. A., and Mead, R. (**1965**). "A simplex method for function minimization," Comput J **7**, 308-313.

Nelson, D. A., and Freyman, R. L. (**1986**). "Psychometric functions for frequency discrimination from listeners with sensorineural hearing loss," J Acoust Soc Am **79**, 799-805.

NIDCD (**2014**). "Quick Statistics About Hearing," pp. https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing.

Olsson, F. (**2009**). "A literature survey of active machine learning in the context of natural language processing," (Swedish Institute of Computer Science).

Olusanya, B. O., Neumann, K. J., and Saunders, J. E. (**2014**). "The global burden of disabling hearing impairment: a call to action," Bull World Health Organ **92**, 367-373.

Osborne, M. A., Garnett, R., and Roberts, S. J. (**2009**). "Gaussian processes for global optimization," in *3rd international conference on learning and intelligent optimization (LION3)* (Trento, Italy), pp. 1-15.

Özdamar, Ö., Eilers, R. E., Miskiel, E., and Widen, J. (**1990**). "Classification of audiograms by sequential testing using a dynamic Bayesian procedure," The Journal of the Acoustical Society of America **88**, 2171-2179.

Park, M., Horwitz, G., and Pillow, J. W. (**2011**). "Active learning of neural response functions with Gaussian processes," in *Adv Neural Inf Process Syst* (Curran Associates), pp. 2043-2051.

Park, M. J. (**2013**). "Bayesian learning methods for neural coding," in *Electrical and Computer Engineering* (The University of Texas at Austin, Austin, Texas), p. 157.

Pascolini, D., and Smith, A. (**2009**). "Hearing Impairment in 2008: a compilation of available epidemiological studies," Int J Audiolog **48**, 473-485.

Patterson, R. D. (**1974**). "Auditory filter shape," The Journal of the Acoustical Society of America **55**, 802-809.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," J Acoust Soc Am **59**, 640-654.

Pentland, A. (**1980**). "Maximum likelihood estimation: The best PEST," Perception & Psychophysics **28**, 377-379.

Popelka, M. M., Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E., and Klein, R. (**1998**). "Low prevalence of hearing aid use among older adults with hearing loss: the Epidemiology of Hearing Loss Study," J Am Geriatr Soc **46**, 1075-1078.

Prins, N., and Kingdom, F. (**2009**). "Palamedes: Matlab routines for analyzing psychophysical data."

Probst, R., Lonsbury-Martin, B., Martin, G., and Coats, A. (**1987**). "Otoacoustic emissions in ears with hearing loss," Am J Otolaryngol **8**, 73-81.

Probst, R., Lonsbury-Martin, B. L., and Martin, G. K. (**1991**). "A review of otoacoustic emissions," J Acoust Soc Am **89**, 2027-2067.

Quiñonero-Candela, J., and Rasmussen, C. E. (**2005**). "A unifying view of sparse approximate Gaussian process regression," J Mach Learn Res **6**, 1939-1959.

Rabinowitz, P. M. (**2000**). "Noise-induced hearing loss," Am Fam Phys **61**, 2759-2760.

Rasmussen, C. E. (**1996**). "Evaluation of Gaussian processes and other methods for non-linear regression," in *Department of Computer Science* (University of Toronto).

Rasmussen, C. E., and Williams, C. K. I. (**2006**). *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA).

Robinson, D., and Sutton, G. (**1979**). "Age effect in hearing-a comparative analysis of published threshold data," Audiology **18**, 320-334.

Roy, N., and McCallum, A. (**2001**). "Toward optimal active learning through monte carlo estimation of error reduction," in *International Conference on Machine Learning* (Williamstown).

Saliba, J., Al-Reefi, M., Carriere, J. S., Verma, N., Provencal, C., and Rappaport, J. M. (**2016**). "Accuracy of Mobile-Based Audiometry in the Evaluation of Hearing Loss in Quiet and Noisy Environments," Otolaryngol Head Neck Surg, 0194599816683663.

Sataloff, R. T., and Sataloff, J. (**2005**). *Hearing loss* (CRC Press, Boca Raton, FL).

Schacht, J., and Hawkins, J. E. (**2006**). "Sketches of Otohistory Part 11: Ototoxicity: Drug-Induced Hearing Loss," Audiol Neurotol.

Schmuziger, N., Probst, R., and Smurzynski, J. (**2004**). "Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones," Ear Hear **25**, 127-132.

Seeger, M. (**2003**). "Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations," in *Division of Informatics* (University of Edinburgh, Edinburgh, UK).

Selters, W. A., and Brackmann, D. E. (**1977**). "Acoustic tumor detection with brain stem electric response audiometry," Arch Otolaryngol **103**, 181-187.

Settles, B. (**2009**). "Active learning literature survey," in *Computer Sciences Technical Report 1648* (University of Wisconsin, Madison).

Settles, B., Craven, M., and Ray, S. (**2008**). "Multiple-instance active learning," in *Adv Neural Inf Process Syst* (Morgan Kaufmann, Vancouver, B.C.), pp. 1289-1296.

Seung, H. S., Opper, M., and Sompolinsky, H. (**1992**). "Query by committee," in *Proceedings of the fifth annual workshop on computational learning theory* (ACM, Pittsburgh, PA), pp. 287-294.

Shannon, C. E. (**2001**). "A mathematical theory of communication," Bell System Technical Journal **5**, 3-55.

Shargorodsky, J., Curhan, S. G., Curhan, G. C., and Eavey, R. (**2010**). "Change in prevalence of hearing loss in US adolescents," JAMA - J Am Med Assoc **304**, 772-778.

Sharma, A., Dorman, M. F., and Spahr, A. J. (**2002**). "A sensitive period for the development of the central auditory system in children with cochlear implants: implications for age of implantation," Ear Hear **23**, 532-539.

Shen, Y., Dai, W., and Richards, V. M. (**2015**). "A MATLAB toolbox for the efficient estimation of the psychometric function using the updated maximum-likelihood adaptive procedure," Behav Res Meth **47**, 13-26.

Shen, Y., and Richards, V. M. (**2012**). "A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention," J Acoust Soc Am **132**, 957-967.

Shen, Y., and Richards, V. M. (**2013**). "Bayesian adaptive estimation of the auditory filter," J Acoust Soc Am **134**, 1134-1145.

Shen, Y., Sivakumar, R., and Richards, V. M. (**2014**). "Rapid estimation of high-parameter auditory-filter shapes," J Acoust Soc Am **136**, 1857-1868.

Snelson, E., and Ghahramani, Z. (**2006**). "Sparse Gaussian processes using pseudo-inputs," in *Adv Neural Inf Process Syst* (Morgan Kaufmann, Vancouver, B.C.).

Song, X. D., Garnett, R., and Barbour, D. L. (**2017**). "Psychometric function estimation by probabilistic classification," J Acoust Soc Am **141**, 2513-2525.

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., and Barbour, D. L. (**2015**). "Fast, continuous audiogram estimation using machine learning," Ear Hear **36** e326-e335.

Spearman, C. (**1908**). "The method of ‘right and wrong cases’(‘constant stimuli’) without Gauss's formulae," Br J Psychol **2**, 227-242.

Stach, B. A. (**2010**). *Clinical audiology: an introduction* (Delmar, Cengage Learning, Clifton Park, NY).

Stadler, S. (**2009**). "Probabilistic modelling of hearing: Speech recognition and optimal audiometry," in *Electrical Engineering* (KTH Royal Institute of Technology, Stockholm, Sweden).

Stapells, D. R., and Oates, P. (**1997**). "Estimation of the pure-tone audiogram by the auditory brainstem response: a review," Audiol Neurotol **2**, 257-280.

Strasburger, H. (**2001**). "Converting between measures of slope of the psychometric function," Percept Psychophys **63**, 1348-1355.

Strawbridge, W. J., Wallhagen, M. I., Shema, S. J., and Kaplan, G. A. (**2000**). "Negative consequences of hearing impairment in old age: a longitudinal analysis," Gerontologist **40**, 320-326.

Stuart, A., Stenstromb, R., Tompkins, C., and Vandenhoff, S. (**1991**). "Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults," Int J Audiol **30**, 82-90.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., and Miyamoto, R. T. (**2000**). "Language development in profoundly deaf children with cochlear implants," Psychol Sci **11**, 153-158.

Swanepoel, D. W., Mngemane, S., Molemong, S., Mkwanazi, H., and Tutshini, S. (**2010**). "Hearing assessment—reliability, accuracy, and efficiency of automated audiometry," Telemed J E Health **16**, 557-563.

Swanepoel, D. W., Myburgh, H. C., Howe, D. M., Mahomed, F., and Eikelboom, R. H. (**2014**). "Smartphone hearing screening with integrated quality control and data management," Int J Audiol **53**, 841-849.

Sweetow, R., and Palmer, C. V. (**2005**). "Efficacy of individual auditory training in adults: a systematic review of the evidence," J Am Acad Audiol **16**, 494-504.

Szudek, J., Ostevik, A., Dziegielewski, P., Robinson-Anagor, J., Gomaa, N., Hodgetts, B., and Ho, A. (**2012**). "Can Uhear me now? Validation of an iPod-based hearing loss screening test," Otolaryngol Head Neck Surg **41**.

Taylor, M. M., and Creelman, C. D. (**1967**). "PEST: Efficient estimates on probability functions," The Journal of the Acoustical Society of America **41**, 782-787.

Thomas, P. D., Hunt, W. C., Garry, P. J., Hood, R. B., Goodwin, J. M., and Goodwin, J. S. (**1983**). "Hearing acuity in a healthy elderly population: Effects on emotional, cognitive, and social status," J Gerontol **38**, 321-325.

Treutwein, B. (**1995**). "Adaptive psychophysical procedures," Vision Res **35**, 2503-2522.

Treutwein, B., and Strasburger, H. (**1999**). "Fitting the psychometric function," Percept Psychophys **61**, 87-106.

Tye-Murray, N. (**2014**). *Foundations of aural rehabilitation: Children, adults, and their family members* (Nelson Education, Scarborough, ON).

Vermeire, K., Brokx, J. P., Wuyts, F. L., Cochet, E., Hofkens, A., and Van de Heyning, P. H. (**2005**). "Quality-of-life benefit from cochlear implantation in the elderly," Otol Neurotol **26**, 188-195.

Vogel, D. A., McCarthy, P. A., Bratt, G., and Brewer, C. (**2007**). "The clinical audiogram: its history and current use," Commun Disord Rev **1**, 81-94.

Von Békésy, G. (**1960**). *Experiments in Hearing* (McGraw-Hill, New York).

von Békésy, G. V. (**1947**). "A new audiometer," Acta Oto-Laryngologica **35**, 411-422.

Vul, E., Bergsma, J., and MacLeod, D. I. (**2010**). "Functional adaptive sequential testing," Seeing Perceiving **23**, 483-515.

Wallhagen, M., Strawbridge, W., and Kaplan, G. (**1996**). "6-year impact of hearing impairment on psychosocial and physiologic functioning," Nurse Pract **21**, 11-14.

Ward, L. M. (**1991**). "Informational and neural adaptation curves are asynchronous," Percept Psychophys **50**, 117-128.

Watson, A. B. (**2017**). "QUEST+: A general multidimensional Bayesian adaptive psychometric methodWatson," J Vis **17**, 1-27.

Watson, A. B., and Pelli, D. G. (**1983**). "QUEST: A Bayesian adaptive psychometric method," Perception & Psychophysics **33**, 113-120.

Watson, C. S., Kidd, G. R., Miller, J. D., Smits, C., and Humes, L. E. (**2012**). "Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version," J Am Acad Audiol **23**, 757-767.

Watt, R., and Andrews, D. (**1981**). "APE: Adaptive probit estimation of psychometric functions," Current Psychological Reviews **1**, 205-213.

Weinstein, B. E., and Ventry, I. M. (**1982**). "Hearing impairment and social isolation in the elderly," J Speech Lang Hear Res **25**, 593-599.

Wetherill, G., and Levitt, H. (**1965**). "Sequential estimation of points on a psychometric function," Br J Math Stat Psychol **18**, 1-10.

Wichmann, F. A., and Hill, N. J. (**2001a**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept Psychophys **63**, 1293-1313.

Wichmann, F. A., and Hill, N. J. (**2001b**). "The psychometric function: II. Bootstrap-based confidence intervals and sampling," Percept Psychophys **63**, 1314-1329.

Williams-Sanchez, V., McArdle, R. A., Wilson, R. H., Kidd, G. R., Watson, C. S., and Bourne, A. L. (**2014**). "Validation of a Screening Test of Auditory Function Using the Telephone," J Am Acad Audiol **25**, 937-951.

Williams, C. K., and Barber, D. (**1998**). "Bayesian classification with Gaussian processes," IEEE Trans Pattern Anal Mach Intell **20**, 1342-1351.

Williams, C. K., and Rasmussen, C. E. (**1996**). "Gaussian processes for regression," in *Adv Neural Inf Process Syst*, edited by D. Touretzky, M. Mozer, and M. Hasselmo (MIT Press, Cambridge, MA).

Williams, C. K. I. (**1998**). "Prediction with Gaussian processes: From linear regression to linear prediction and beyond," in *Learning in Graphical Models*, edited by M. I. Jordan (Springer Netherlands, Dordrecht), pp. 599-621.

World Health Organization (**2012**). "WHO global estimates on prevalence of hearing loss."

Xiang, N., and Fackler, C. (**2015**). "Objective Bayesian analysis in acoustics," Acoust Today **11**, 54-61.

Yoshinaga-Itano, C. (**2003**). "From screening to early identification and intervention: Discovering predictors to successful outcomes for children with significant hearing loss," J Deaf Stud Deaf Educ **8**, 11-30.

Yoshinaga-Itano, C., Sedey, A. L., Coulter, D. K., and Mehl, A. L. (**1998**). "Language of early- and later-identified children with hearing loss," Pediatr **102**, 1161-1171.

Yueh, B., Shapiro, N., MacLean, C. H., and Shekelle, P. G. (**2003**). "Screening and management of adult hearing loss in primary care: scientific review," JAMA **289**, 1976-1985.

Zhao, F., Stephens, D., and Meyer-Bisch, C. (**2002**). "The Audioscan: a high frequency resolution audiometric technique and its clinical applications," Clinical Otolaryngology & Allied Sciences **27**, 4-10.

Zhao, F., Stephens, S. D. G., Ishak, W. S., and Meyer-Bisch, C. (**2014**). "The characteristics of Audioscan and DPOAE measures in tinnitus patients with normal hearing thresholds," Int J Audiol **53**, 309-317.

Zychaluk, K., and Foster, D. H. (**2009**). "Model-free estimation of the psychometric function," Atten Percept Psychophys **71**, 1414-1425.

Appendix 1: Supplemental Figures

A1.1. Numerical Estimates for 1D Psychometric Functions

Figures A1.1-A1.3 show plots of error in 50% point and 25-75% interquartile range (IQR) for unidimensional PFs estimated using probit regression and probabilistic classification. 50% point is the numerical equivalent to α, while the IQR is a numerical analog for β. **Figure A1.1** shows 50% point and IQR error averaged across all conditions; **Figure A1.2** shows 50% point and IQR error for various true values of β, and **Figure A1.3** shows 50% point and IQR error for various distributions of number of intensities sampled/number of samples per intensity. These figures correspond to the plots for parametric estimates in Figures 4.3, 4.4, and 4.5, respectively.

Trends in each plot for numerical estimates are identical to those in the corresponding plot for parametric estimates from Section 4.3.1. As with the parametric estimates for α and β, there is no appreciable difference in performance between PR and GPC.

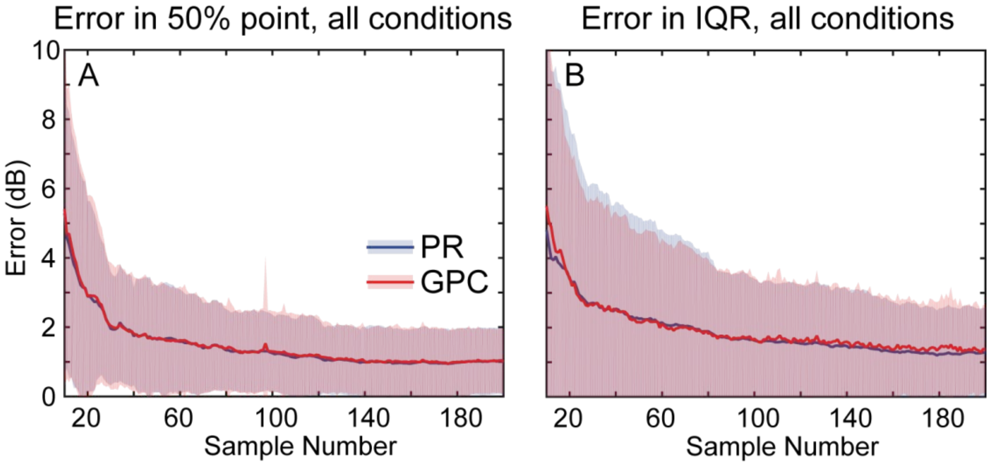

Figure A1.1: Error in 1D 50% point and IQR estimates for PR and GPC across all conditions. Absolute error in estimates of (A) 50% point and (B) IQR as a function of number of observed samples in the unidimensional case. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded regions designate 1 standard deviation above and below the mean.

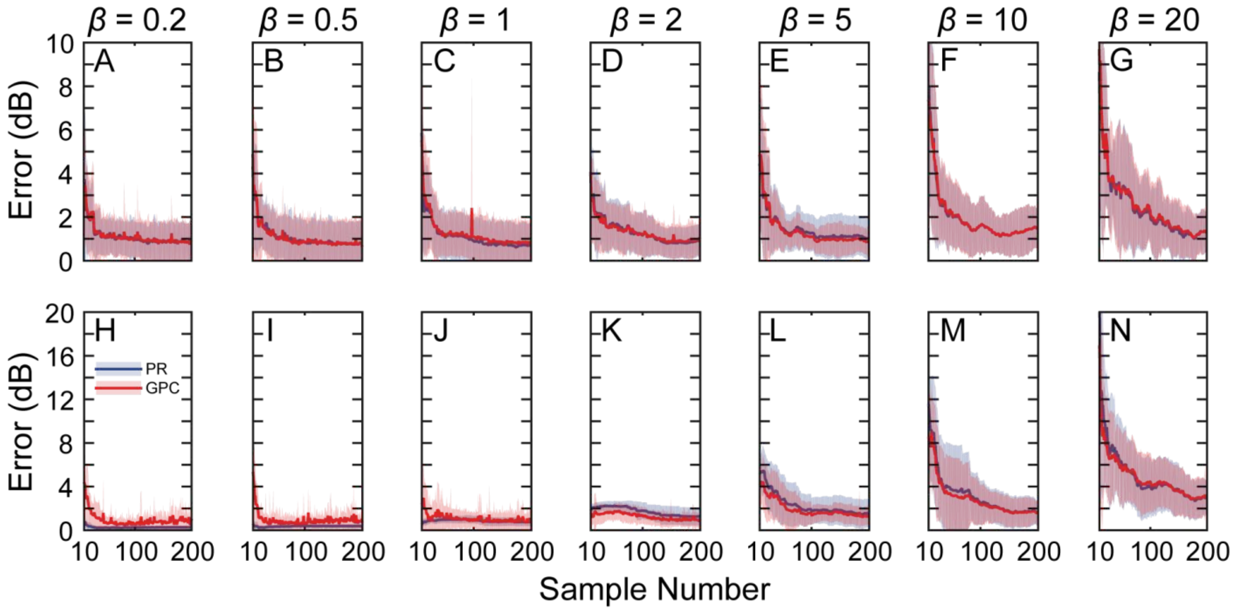

Figure A1.2: Error in 1D PR and GPC 50% point and IQR estimates for different *β* values. Absolute error in estimates of (A-G) 50% point and (H-N) IQR as a function of number of observed samples for unidimensional PFs. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded region designates 1 standard deviation above and below the mean. Each subplot corresponds to a distinct *β* value.

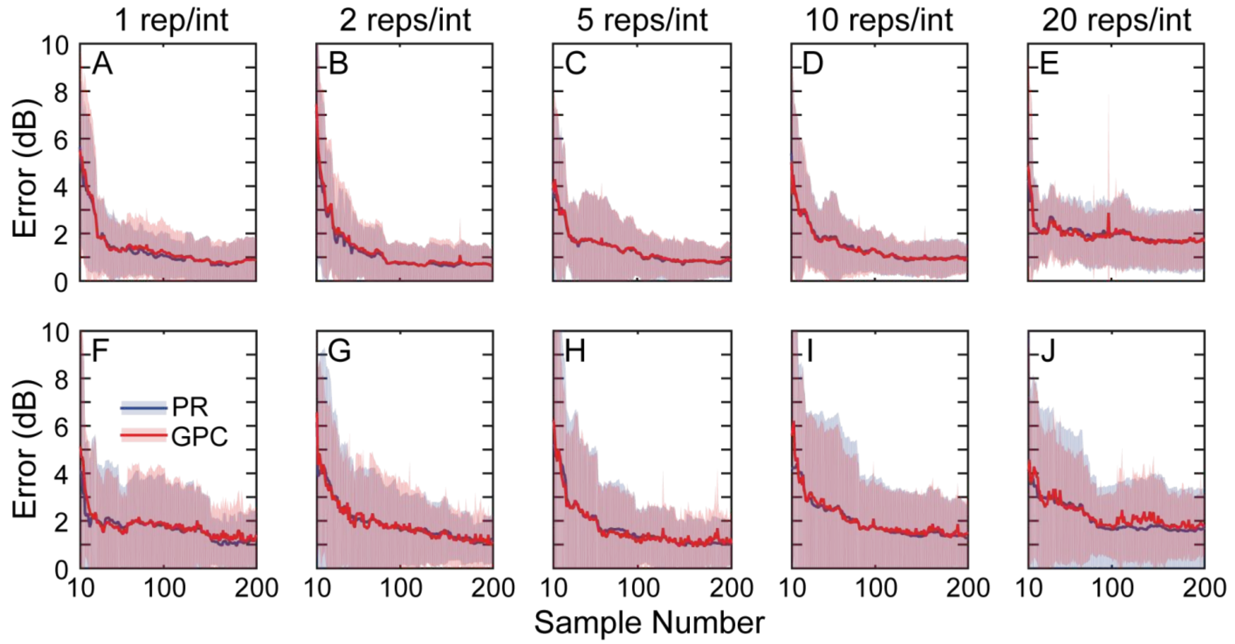

Figure A1.3: Error in 1D PR and GPC 50% point and IQR estimates by sample split. Absolute error in estimates of (A-E) 50% point and (F-J) IQR as a function of number of observed samples for unidimensional PFs. Blue solid and red dashed lines denote mean absolute errors of the PR and GPC estimates, respectively, and the matching shaded region designates 1 standard deviation above and below the mean. Each subplot corresponds to a distinct condition for number of intensities and number of repetitions per intensity.

A2.1. Numerical Estimates for 2D Psychometric Functions

Tables A1.1-A1.3 show tables of error in 50% point (equivalent to α) and 25-75% IQR (analog for β) for multidimensional PFs estimated using Gaussian process classification. **Figure A1.1** shows 50% point and IQR error for various total sample numbers; **Table A1.2** shows 50% point and IQR error for various true values of β, and **Table A1.3** shows 50% point and IQR error for various audiogram phenotypes (Dubno *et al.*, 2013). These figures correspond to the tables for parametric estimates in Tables 4.1, 4.2, and 4.3, respectively.

Trends in each table for numerical estimates are identical to those in the corresponding table for parametric estimates from Section 4.3.2.

Number of samples	**20**	**50**	**100**	**200**	**500**	**1000**
Mean 50% point abs. error (dB)	6.71	4.29	3.04	2.08	1.30	0.937
Std. 50% point abs. error (dB)	6.31	5.32	2.82	1.74	1.23	0.915
Mean IQR absolute error (dB)	4.29	3.52	1.97	1.74	0.867	0.548
Std. IQR absolute error (dB)	3.55	3.20	2.22	1.59	0.763	0.461

Table A1.1: Absolute errors of 2D GP 50% point and IQR estimates by number of samples. Accuracy and reliability are quantified using mean and standard deviation, respectively.

β (dB/%)	**0.2**	**0.5**	**1**	**2**	**5**	**10**
Mean 50% point abs. error (dB)	1.54	1.60	1.70	1.86	2.47	3.32
Std. 50% point abs. error (dB)	2.03	1.09	1.17	1.32	1.90	2.60
Mean IQR absolute error (dB)	1.26	1.10	1.44	1.97	2.38	2.29
Std. IQR absolute error (dB)	1.43	1.52	1.39	1.18	1.80	1.69

Table A1.2: Absolute errors of 2D GP 50% point and IQR estimates by value of β. Accuracy and reliability are quantified using mean and standard deviation, respectively. Sample number is fixed at 200 samples.

Audiogram Phenotype	**Older**	**Metabolic**	**Sensory**	**Metabolic + Sensory**
Mean 50% point abs. error (dB)	2.19	1.69	2.23	2.22
Std. 50% point abs. error (dB)	1.68	1.47	1.96	1.74
Mean IQR absolute error (dB)	1.40	1.33	3.03	1.20
Std. IQR absolute error (dB)	1.53	1.19	1.52	1.35

Table A1.3: Absolute errors of 2D GP 50% point and IQR estimates by phenotype. Accuracy and reliability are quantified using mean and standard deviation, respectively. Values are averaged across all β values, and sample number is fixed at 200 samples.

Curriculum Vitae

Education

Washington University in St. Louis, St. Louis, MO 2011-2017

PhD, Biomedical Engineering, *May 2017*

M.S. Biomedical Engineering, *May 2014*

Duke University, Durham, NC 2007-2011

B.S. Biomedical Engineering, *May 2011*

B.A. Music, *May 2011*

Research Experience

Laboratory of Sensory Neuroscience and Neuroengineering, Washington Univv. 2012-2017

Doctoral Student, Biomedical Engineering; Mentor: Dennis Barbour

- Developed novel machine learning technique for more efficient auditory testing and tested this technique in human subjects, finding that this technique can significantly reduce the number of trials needed to estimate a standard pure-tone threshold audiogram
- Demonstrated the ability of machine-learning algorithm to accurately and efficiently estimate nonparametric multidimensional psychometric functions
- Designed, developed, and programmed video games for remote diagnosis of auditory disorders as well as subsequent training to improve function
- Maintained collaborations with other researchers in the Departments of Computer Science, Psychology and Audiology & Communication Sciences
- Pioneered new research direction in laboratory and secured over $100,000 in funding

Laboratory of Neural Computation and Motor Behavior, Washington Univ. 2011-2012

Rotation Student, Biomedical Engineering; Mentor: Kurt Thoroughman

- Investigated electrocorticography (ECoG) signals as an indicator of motor adaptation
- Obtained and performed spectral analysis of ECoG recordings from primary motor cortex of behaving rhesus macaque monkeys and kinematic data from a robot manipulandum

Vision and Image Processing Laboratory, Duke Univ. 2009-2011

Pratt Research Fellow, Biomedical Engineering; Mentor: Sina Farsiu

- Developed algorithms for automatic alignment of retinal optical coherence tomography (OCT) images by combining registration and segmentation techniques
- Developed algorithms for reconstruction of a three-dimensional model of retina shape

from collection of OCT slices for better diagnosis of disorders such as glaucoma

- Programmed a graphical user interface for manual segmentation of optic disc and cup, combined with automatic area calculations, for use by doctors and researchers

Duke Interactive Studio, Duke Univ. 2009-2010

Undergraduate Researcher, Music Technology; Mentor: Scott Lindroth

- Composed, designed, and implemented interactive music exhibit for Museum of Life and Science (Durham, North Carolina) Halloween program
- Wrote code for dynamic image processing and audio synthesis in Supercollider that translated user movement captured by mounted cameras to real-time changes in music

Teaching Experience

Biological Neural Computation Minicourse Instructor, Washington Univ. 2015-2017

- Developed and taught intensive 1-week MATLAB courses for new programmers

Quantitative Physiology Teaching Assistant, Washington Univ. 2013-2014

- Managed laboratory of biomedical engineering class; held office hours for assignments

Cognitive, Computational & Systems Neuroscience Outreach Program 2012-2013

- Presented auditory neuroscience demonstrations to the public at St. Louis Science Center

Peer Tutoring Program, Duke Univ. 2009-2011

- Held individual tutoring sessions for Duke University students in undergraduate physics

Awards & Honors

Center for Integration of Medicine and Innovative Technology Primary Healthcare Prize 2014-2015

- Annual national award granted to proposed collaborative research projects that demonstrate high potential for improving patient care in the primary healthcare setting, with emphasis on novel ideas and technologies that may benefit several disciplines
- Second place, 2015: **$110,000** awarded
- Finalist (top 10), 2014: **$10,000** awarded

Cognitive, Computational & Systems Neuroscience Fellowship 2012-2014

- Full tuition and fellowship awarded to doctoral candidates in the brain-related sciences who demonstrate the capacity to develop a project that uniquely approaches a particular research question utilizing tools and concepts from across neuro-related disciplines
- Completed required curricular pathway including coursework from disciplines along with a project course catered to the crafting of an interdisciplinary research project

Pratt Undergraduate Research Fellowship 2009-2011

- Competitive stipend fellowship awarded to high-achieving Duke University undergraduate students to perform intensive research in their engineering major in collaboration with a faculty member

Publications

Song X.D., Sukesan K.A., D.L. Barbour. "Active probabilistic classification for psychometric field estimation." *Attention, Perception & Psychophysics* (in submission).

Song X.D., Garnett R., D.L. Barbour. "Psychometric function estimation by probabilistic classification." *The Journal of the Acoustical Society of America* 141(4), 2513-2525, 2017.

Gardner J.R., **Song X.**, Weinberger K.Q., Barbour, D., J.P. Cunningham. "Psychophysical detection testing with Bayesian active learning." *Uncertainty in Artificial Intelligence*, 2015.

Song X.D., Wallace B.M., Gardner J.R., Ledbetter N.M., Weinberger K.Q., D.L. Barbour. "Fast, continuous audiogram estimation using machine learning." *Ear and Hearing* 36(6), e326-335, 2015.

Conference Presentations/Abstracts

Howard, R.T., **Song, X.D.**, Metzger, N.M., DiLorenzo, J.C., Snyder, B.R.D., Sukesan, K.A., D.L. Barbour. "Web-Based Audiometric Threshold Estimation." *American Auditory Society*, Poster No. 18, Scottsdale, AZ, 2017.

Barbour D.L., **X.D. Song**. "The Development of Machine Learning Audiometry." *Association for Research in Otolaryngology*, Poster No. PS-320, Baltimore, MD, 2017.

Song X.D., Sun W., D.L. Barbour. "Rapid estimation of neuronal frequency response area using Gaussian process regression." *Society for Neuroscience*, Poster No. 231.20, Chicago, IL, 2015.

Barbour D.L., Wallace B.M., Gardner J.R., Ledbetter N.M., Weinberger K.Q., **X.D. Song**. "Optimizing pure-tone audiometry using Gaussian processes." *Association for Research in Otolaryngology*, Poster No. PS-171, Baltimore, MD, 2015.

Song X.D., Ledbetter N.M., D.L. Barbour. "A platform for automated diagnosis of speech and hearing disorders." *IEEE Engineering in Medicine and Biology Society*, Chicago, IL, 2014.

Song X.D., Ledbetter N.M., Sommers M.S., Tye-Murray N, D.L. Barbour. "Auditory games as a novel tool for aural rehabilitation." *Association for Research in Otolaryngology*, Poster No. PS-516, San Diego, CA, 2014.

Song X.D., Ledbetter N.M., Sommers M.S., Tye-Murray N, D.L. Barbour. "Training the brain with auditory games." *Entertainment Software and Cognitive Neurotherapeutics Society*, Poster No. A19, Los Angeles, CA, 2013.

Song X., Estrada R., Chiu S.J., Dhalla A.H., Toth C.A., Izatt J.A., S. Farsiu. "Segmentation-based registration of retinal optical coherence tomography images with pathology." *Association for Research in Vision and Ophthalmology*, Program No. 1309, Fort Lauderdale, FL, 2011.