Spring 5-15-2020

Joint Estimation of Perceptual, Cognitive, and Neural Processes

Katherine Heisey
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Neurosciences

Dissertation Examination Committee:
Dennis L. Barbour, Chair
Todd Braver
Roman Garnett
Camillo Padoa-Schioppa
Jonathan Peelle

Joint Estimation of Perceptual, Cognitive, and Neural Processes

by
Katherine L. Heisey

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

May 2020
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

My path to a PhD has been far from normal and as it comes to a close, I would like to offer my gratitude to those who have helped me reach this finish line.

First and foremost, my advisor, Dr. Barbour. I came to you three years into my program, unsure if completing a PhD was something I even desired to do anymore. You graciously created a space for me to explore research while still keeping one foot out the door. Your patience and guidance was evident from our first meeting, and you continued to mentor me once I did, officially, join the lab. You have always supported my post-graduate goals and have advocated for me accordingly. I am exceedingly thankful to have had the opportunity to learn from you. Your passion and tenacity has profoundly impacted the way I view science and medicine.

I have had the joy of working with very talented and motivated lab members throughout my years in the Barbour lab. My research would not have been possible without the assistance and early contributions James, David, Kiron, and other lab members who pioneer this work and laid a solid foundation from which my research grew. To Jenna and Alex, who helped with audiogram data collection and to Kevin, who help developed and test the code used in Chapter 5: I greatly appreciate each of your contributions and many hours you committed to making this work better.

Outside of the Barbour lab, this thesis would not have been possible without the support from Dr. Peelle and his lab. You welcomed me into your lab as if I were one of your own, and it has been a delight to spend weekly imaging meetings with you all. Dr. Peelle, your patience, thorough explanations of imaging methods, and thoughtful suggestions have been invaluable to me. Chapter 6 would not have been possible without assistance from many members of the Peelle lab, but I would like to specifically thank Mike: for your willingness to conduct imaging analysis in a short time frame, and Sarah: for all your help with coordinating our research studies and always being willing to track down the information I need.

I am thankful for my committee members, Dr. Braver, Dr. Garnett, and Dr. Padoa-Schioppa. You have each made my thesis better, and I am grateful for your expertise and advice. Dr.

Padoa-Schioppa: thank you for trekking over to the Danforth campus for committee meetings and serving as my committee chair for all these years.

I also want to acknowledge my funding source: the National Science Foundation's Graduate Research Fellowship Program. This work would not have been possible without the financial support you provided.

I am particularly grateful for friends in St. Louis and elsewhere who have walked with me through this journey. Whether it was evening dinners, morning runs, or long talks, I would not have made it to this point without you.

Finally, to my parents: you have always seen the best things in me and have taught me, from a young age, to pursue big dreams. Thank you and I love you.

Katherine L. Heisey

Washington University in St. Louis
May 2020

ABSTRACT OF THE DISSERTATION

Joint Estimation of Perceptual, Cognitive, and Neural Processes

by

Katherine L. Heisey

Doctor of Philosophy in Biology and Biomedical Sciences

Neurosciences

Washington University in St. Louis, 2020

Dr. Dennis L. Barbour, Chair

Humans are remarkable in their ability to perform highly complicated behaviors with ease and little conscious thought. Successful speech comprehension, for example, requires the collaboration of multiple sensory, perceptual, and cognitive processes to focus attention on the speaker, disregard competing cues, correctly process incoming audio stimuli, and attach meaning and context to what is heard. Investigating these phenomena can help unravel crucial aspects of human behavior as well as how the brain works in health and disease. However, traditional methods typically involve isolating individual variables and evaluating their decontextualized contribution to an outcome variable of interest. While rigorous and more straightforward to interpret, these reductionist methods forfeit multidimensional inference and waste data resources by collecting identical data in every participant without considering what is the most relevant for any given participant. Methods that can optimize the exact data collected for each participant would be useful for constructing more complex models and for optimizing expensive data collection. Modern tools, such as mobile hardware and large databases, have been implemented

to improve upon traditional methods but are still limited in the amount of inference they can provide about an individual. To circumvent these obstacles, a novel machine learning framework capable of quantifying behavioral functions of multiple variables with practical amounts of data has been developed and validated. This framework is capable of linking even loosely related input domains and measuring shared information in one comprehensive assessment.

The work described in this thesis first evaluates this framework for active machine learning audiogram (AMLAG) applications. AMLAG customizes the generalized framework to efficiently, accurately, and reliably estimate audiogram functions. Audiograms provide a measure of hearing ability for each ear in the inherently two-dimensional domain of frequency and intensity. Where clinical methods rely on reducing audiogram acquisition to a one-dimensional assessment, AMLAG has been previously verified to provide a continuous, two-dimensional estimate of hearing ability in one ear.

Modeling two ears that are physiologically distinct but are defined in the same frequency-intensity input domain, AMLAG was extended to bilateral audiogram acquisition. Left and right ears are traditionally evaluated completely unilaterally. To realize potential gains, AMLAG was generalized from two unilateral tests to a single bilateral test. The active bilateral audiogram allows observations in one ear to simultaneously update the model fit over both ears. This thesis shows that in a cohort of normal-hearing and hearing-impaired listeners, the bilateral audiogram converges to its final estimates significantly faster than sequential active unilateral audiograms.

The flexibility of a framework capable of informative individual inference was then evaluated for dynamically masked audiograms. When one ear of an individual can hear significantly better than the other ear, assessing the worse ear with loud probe tones may require delivering masking

noise to the better ear in order to prevent the probe tones from inadvertently being heard by the better ear. Current masking protocols are confusing, laborious and time consuming. Adding a standardized masking protocol to the AMLAG procedure alleviates all of these drawbacks by dynamically adapting the masking to an individual's specific needs. Dynamically masked audiograms are shown to achieve accurate threshold estimates and reduce test time compared to current clinical masking procedures used to evaluate individuals with highly asymmetric hearing, yet can also be used effectively and efficiently for anyone.

Finally, the active machine learning framework was evaluated for estimating cognitive and perceptual variables in one joint assessment. Combining a verbal N-back and speech-in-noise assessment, a joint estimator links two disjoint assessments defined by two unique input domains and, for the first time, offers a direct measurement of the interactions between two of the most predictive measures of cognitive decline. Young and older healthy adults were assessed to investigate age-related adaptations in behavior and the inter-subject variability that is often seen in low-dimensional speech and memory tests. The joint cognitive and perceptual test accurately predicted standalone N-back but not speech-in-noise performance. This first implementation did not reveal significant interactions between speech and memory. However, the joint task framework did provide an estimate of participant performance over the entire two-dimensional domain without any experimenter-observed scoring and may better mirror the challenges of real-world tasks. While significant age-related differences were apparent, substantial within group variance led to evaluating joint test performance in predicting individual differences in neural activity.

Speech-in-noise tests may activate non-auditory specific networks of the brain as age and task difficulty increase. Some of these regions are domain-general networks that are also active

during verbal working memory tests. Functional brain images were collected during an in-scanner speech-in-noise test for a portion of the joint test participants. Individual brain activity at regions of interest in the frontoparietal, cingulo-opercular, and speech networks was correlated to performance on the joint speech and memory test. No significant correlations were found, but the joint estimation of neural, cognitive, and perceptual behaviors through this framework may be possible with further test adaptations. Generally, the lack of significant findings does not detract from the feasibility and utility of a generalized framework that can accurately model complex cognitive, perceptual, and neural processes in individuals. As demonstrated in this thesis, high-dimensional, individual testing procedures facilitate the direct assessment of complicated human behaviors empowering equitable, informative, and effective test methods.

Chapter 1: Introduction and Motivation

1.1 Motivation: The Need for Individual-Specific Models

Fast and accurate individual assessments of cognition and perception have the potential to change the way clinical medicine and scientific research are conducted. Previous research shows that individual differences in cognition and perception can indicate functional changes in brain activity, occasionally disputing long-held assumptions (for example, Lafer-Sousa, Hermann, & Conway, 2015; Wallisch, 2017). However, current methods are formulated to interpret potentially informative variance as noise or error when computing population-based analysis. Often, data deemed to be too distant from the majority trend is removed from analysis completely. If not, all data are averaged together, potentially obscuring the predictive power of any one point. But, on occasion, highly variable data may result from informative individual differences in a participant's specific life context that could inform the results. In these cases, removing participants from a data set or averaging their results together with the majority group will prevent potentially important research conclusions from being considered.

This is demonstrated in a series of recent studies identifying significant individual differences in the activation of whole brain functional magnetic resonance imaging (fMRI) networks (Gordon, Laumann, Adeyemo, & Petersen, 2017; Gordon, Laumann, Gilmore, et al., 2017; Laumann et al., 2015; Marek et al., 2018; Mueller et al., 2013). Earlier studies were vital in revealing the central tendencies of specific neural networks, like the default mode network, but individual specificity could not be researched due to the perceived cost of data collection, be it in scan time or resource scarcity. However, technical advancements and the commitment of the research community have

paved the way for extensive data collection to be done in individual brains and has empowered the examination of individual differences in neural networks. Consequently, replicable, individual-specific features of many brain systems have been identified. Group averaged results smooth away disparate imaging data, but focusing on individual brain systems provides the foundation to explore individual-specific neural function or connectivity and its relation to personality, aging, and disease. These studies demonstrate that group averaged results do not always fully represent an individual and may be concealing revelatory facets of neural systems. Relying solely on group level analysis could limit the applicability of fMRI data-driven conclusions, and there is merit in considering individual differences in neural organization.

Besides brain networks, the last 20 years witnessed an emerging body of research across disciplines implicating socioeconomic class, cultural upbringing, race, gender, or even musical training in impacting basic cognitive and perceptual behaviors (for example, Aneshensel, Ko, Chodosh, & Wight, 2012; Kagan, 2018; Krecic-Shepard et al., 2000; Magee, Blum, Lates, & Jusko, 2001; McFarland, 2017). Most concerningly, when researchers and clinicians fail to incorporate individual life context into their studies, critical medical decisions can be affected. A recent study by Obermeyer et al (2019) revealed implicit racial bias in the proprietary algorithms used to determine health care needs and the distribution of additional care support programs to patients. This outcome resulted from the algorithm's intentional exclusion of race as a predictive factor combined with the manufacturer's decision to predict health care costs above other metrics (such as avoidable future costs). A more equitable algorithm estimating a multidimensional variable that combines a prediction of patient health and avoidable future costs was suggested. Simply omitting predictor variables can perpetuate model bias. However, without the ability to

determine the most relevant predictive factors and the flexibility to incorporate multidimensional estimates into the model, this particular health disparity would have continued indefinitely.

Decontextualizing human behavior based on a small vector of pre-selected features is proving to be problematic in real-world applications. It is becoming apparent that approaches considering individual context in addition to group-level analysis are necessary to successfully link the results of basic science research to relevant and practical applications. This is even more critical in view of the current push for reproducible science. As a result of numerous pervasive shortcomings in experimental design, analysis, and publication culture, many previously published results have failed to be reliably reproduced (Ioannidis, 2005; Open Science Collaboration, 2015). In 2016, the NIH recognized that one gap in reproducible research is the historically inadequate consideration of a basic individual factor, sex, as a predictive variable. Study proposals are now mandated to explicitly address sex as a biological variable (National Institute of Health, 2016). Sex and race are just two examples of individual-specific variables that have been ignored to the detriment of the research community and the advancement of useful, scientifically sound outcomes. The reality is probably much bleaker, as it is challenging to determine *a priori* which individual-specific variables should be included in current models. Researchers try to incorporate diversity into their subject pool and perform statistical measures to ensure the predictive power of their study, but the uniqueness of life experience, genetics, and other identifying features means any one person cannot be fully represented by a cohort (Rose, Rouhani, & Fischer, 2013). Narrowing the scope of investigation to ignore the complexity of the human experience not only makes it difficult to reproduce results in a new cohort, but it limits our understanding of basic science, affects patient outcomes, and may introduce unexpected bias.

1.2 Motivation: The Limitations of Current Multidimensional Models

Most complex human behaviors, including speech comprehension and working memory tasks, are inherently multidimensional (Cacace & McFarland, 2013; McFarland, 2017). Unfortunately, high-dimensional models cannot often be constructed for individuals because of the immense data requirements—psychophysical and psychometric tests require a large amount of data in order to draw robust conclusions for individual participants.

Despite technological advancements, researchers are limited in how many queries can be made in one experiment. Multidimensional measures very quickly encounter what is referred to as the 'curse of dimensionality' or 'big p, small N' bottlenecks. Namely, increasing the dimension of the feature space being assessed necessitates an exponential increase in the number of observations needed to make substantial claims and avoid overfitting the data (Alyass, Turcotte, & Meyre, 2015; Barbour, 2019; Johnstone & Titterington, 2009). Necessarily, current methods are largely constrained to delivering a series of unidimensional behavioral tests, meaning, testing one domain at a time without context. This procedure ensures subjects do not fatigue, which would lead to excessive errors or lapses. Accumulated data can fit a model to relate the observed behavior directly to the single dimension of the domain being assessed. Studies aiming to explore the possible interactions between unidimensional measures are limited to deploying numerical methods after data are collected in order to determine correlations between stimulus features.

Parametric models, such as generalized linear models, and advanced machine learning methods are commonly implemented to determine the relationships between multiple input domains.

These models can capture interactions but may require predictor variables to be determined empirically based on large amounts of data already collected. Even innovative, high-dimensional machine learning methods require extensive computing resources, time, and large data sets. Still, they essentially reduce participants to a set of pre-selected features. Additionally, most models are inflexible in that once they are trained under an assumed function, all predictions on unseen data are restricted to the specific model definition. There is little room for individuals or subsets of a cohort to adapt the model in real time.

Although favored because data collection is more feasible and highly controlled experiments are easier to interpret, unidimensional methods waste data collection resources by collecting identical data in every participant, without considering what is the most relevant data for any given participant. Methods to investigate the interactions of multiple stimulus dimensions are severely underpowered, and many complex behaviors, such as working memory and speech comprehension, are not adequately modeled by current methods (Paivio, 2014; Read, 2015). Relying on correlation measures to hypothesize about the profoundly intricate aspects of complex human behavior is often reductionist, and verifying that any given correlation is accurate or meaningful is burdensome (Varoquaux & Poldrack, 2019). Effectively modeling multidimensional behaviors would make efficient use of data collecting resources by reducing redundant queries probing overlapping domains and would provide more informative estimates of complex, individual behaviors.

1.3 Concluding Remarks

Adapting algorithms and developing models capable of multidimensional, individual inference requires time and tools that many research and clinical teams do not have. The burden must be

on researchers and clinicians to discover methods capable of incorporating individual differences in a reasonable time and with enough detail to inform clinical decisions. A framework that can assess only the most informative features for any given participant, while allowing those exact features to vary from participant to participant, would allow for equitable, informative, and effective testing procedures.

To that end, a novel machine learning framework has been developed. This framework employs a Gaussian process (GP) Bayesian inference method along with active learning techniques to model multidimensional input domains with practical amounts of data. Collecting more informative data in less time, this machine learning framework can address some of the shortcomings of conventional methods. The GP framework can flexibly encode relationships between domain spaces in real time rather than estimating the relevant parametric form after data collection. Prior beliefs can be incorporated into the framework, but, given the appropriate definitions, the GP can adapt to observed data and is not restricted to experimenter assumptions about the underlying structure of the data. Active learning techniques can optimize data collection for each participant by choosing the most informative next point to probe given all of the previously collected data in that participant. Individual test sessions can vary in what data are observed to best model the domain of interest. Making efficient use of data and exploiting the advantages of an iterative, Bayesian inference algorithm, multidimensional behaviors can be estimated in individuals.

The thesis work presented here leverages this active machine learning GP framework to model complex, individual behaviors in perception and cognition. Concepts relevant to this thesis will be introduced in Chapter 2. Chapter 3 will extend a previously validated application of the GP

framework in audiogram acquisition in one ear to estimate bilateral audiogram functions. Bilateral audiogram estimation represents a four-dimensional space and efficiently performs simultaneous assessments of two ears in one test. Chapter 4 will evaluate the flexibility of a framework capable of informative individual inference by extending machine learning audiogram acquisition to include dynamically masked audiograms, which typically requires long test times for a small set of patients. Chapters 3 and 4 will have demonstrated the flexibility and efficiency of a framework capable of multidimensional, individual inference for perceptual behaviors that are physiologically distinct but are similarly defined. Chapter 5 will extend the framework from modeling multidimensional, individual behaviors in perception (hearing ability) to estimate perceptual and cognitive variables in one joint assessment. Combining a verbal N-back and speech-in-noise assessment, a joint estimator links two disjoint assessments defined by two unique input domains and, for the first time, offers a direct measurement of the interactions between two of the most predictive measures of cognitive decline. Chapter 6 will evaluate joint test performance in predicting individual differences in neural activity.

Chapter 2: Background

2.1 Inference Background

Human discovery is predicated on the principle that underlying systems govern most experienced phenomena. An understanding of these systems is developed and tested based on observations about the world. Using these observations, inferences can be made that predict the consequences of future actions. Mathematical models have been developed to formalize these inquiries and approximate the properties of latent systems based on a set of observations.

Psychophysics studies the relationship between measurable physical properties and their behavioral response. Typically, data are collected by systematically adjusting a feature of a universally understood stimulus and recording the corresponding behavioral response (Fechner, 1860). Data are often fitted to a psychometric function. Psychometric functions help decipher how sensory information is encoded and how perception is affected by varying stimulus features (Read, 2015). In its simplest form, a psychometric function is a unidimensional sigmoid modeling the probability of participant detection or discrimination of stimuli across the input domain (**Figure 2.1**).

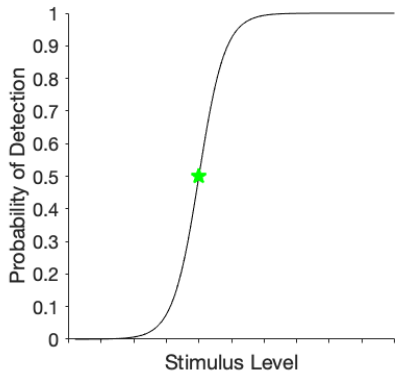

Figure 2. 1: Example of a psychometric function. The 50% threshold is indicated by a green star. Stimuli presented at levels above threshold have a higher probability of being detected.

Psychometrics attempt to quantitatively measure cognitive processes by means of behavioral assessment. Many psychometric tests assign a participant a score that, when considered with respect to a larger population, quantifies an aspect of an individual's cognitive ability. Standardized educational testing and intelligence tests are demonstrative of psychometric assessments in the real world. In research, psychometric tests are often used to ascertain the properties of a population in order to describe the underlying systems of cognition and to categorize 'normal' function.

Both traditional psychophysical and psychometric assessments demand large amounts of data to be collected. Useful perceptual models often necessitate individual estimates of the interactions between variables being studied, and robustly estimating unidimensional psychometric functions requires substantial data in individual participants. This is almost never done in cognitive models. While both could leverage distributions across populations, cognitive models have focused almost exclusively on that. Population-based models are sensitive to the properties of the larger group. Differences within and between populations may alter the reliability of a measure and could lead to inaccurate inferences (Cooper, Gonthier, Barch, & Braver, 2017). To address this concern, researchers must ensure that their study pools are sufficiently large and representative of the population to power reproducible conclusions. Thus, intersubject inference often requires data to be collected in large or fairly homogeneous populations. Informative intrasubject inference is therefore challenging in many cognitive tests because cohort-level analysis is often a prerequisite to meaningful individual inference.

Intersubject variability inference can always be peformed with intrasubject variability models, but the converse is not true. Therefore, improving the process of forming intrasubject models would have broad impact.

2.2 Machine Learning Background

2.2.1 General Machine Learning Background

Due to their ability to deduce meaning from large, complex sets of data, machine learning methods have become popular in a wide variety of applications from finance to health care to neural networks. Machine learning has become an umbrella term that encompasses a variety of methods and, for clarity, has been subdivided into supervised and unsupervised learning.

In supervised machine learning methods, the previously observed relationships between features of data and the resulting output measurements are used to train models to predict unobserved measurements. A wide variety of supervised machine learning algorithms have been developed to model complicated datasets. Models can be parametric or non-parametric. Parametric models make assumptions about the shape and characteristics of the underlying function, f, simplifying the prediction process. Necessarily, parametric models constrain the form of the underlying function, which can limit the fit of the model to the data. Non-parametric models offer more complex modeling and usually require substantially more computational effort and observed data to train. An advantage of non-parametric models is that f can be deduced from the features observed and can still accurately fit the data even if confident prior assumptions of f cannot be made.

Further classifying machine learning methods, supervised learning can be subdivided into regression and classification techniques. Regression endeavors to define the function, f, that relates a set of feature variables to the dependent measurements extracted from a finite set of observations. The value of unobserved measurements can then be predicted based on f. Classification, on the other hand, defines the function, f, that separates the observed data into a proper grouping scheme with respect to the selected features. Delineating cats from fish based on the number of legs is a simple example. The probability of future observations belonging to either of the groups can be calculated from f. The work in this thesis employs classification and regression of a non-parametric, supervised learning model: a Gaussian process (GP) model.

2.2.2 Gaussian Process Framework

A GP is a set of random variables such that any subset sampling exhibits a multivariate Gaussian distribution (Rasmussen & Williams, 2006). Being Gaussian in nature, any function drawn from a GP is fully explained by its mean and covariance function. The mean function describes the central tendency of the underlying function while the covariance function accounts for its structure. Any parameters of the mean and covariance functions are referred to as hyperparameters. Hyperparameters can be learned or fixed and can encode information about the domain or retain an uninformative distribution. Hyperparameters that are learned as the algorithm iterates allow the shape of the estimated function to change in global structure as more data are observed. The flexibility of a GP is evidenced in the freedom to represent the covariance and mean functions in any functional form that best reflects the assumptions over the latent function being modeled. GPs can be used for regression and, with modification, for classification.

GPs are capable of capturing nonlinear relationships between the input and output data, and application-specific prior beliefs of the underlying function can be incorporated through prior distributions (Rasmussen & Williams, 2006). In the framework used in this thesis, observed data condition a GP prior using Bayesian inference.

Bayesian Inference

Bayesian inference techniques have become increasingly popular in building models of perception, cognition, and neural processes (for reviews see Chater, Oaksford, Hahn, & Heit, 2010; Parr, Rees, & Friston, 2018). Many well-designed models attempt to isolate the phenomenon they are observing, aiming to maximize the confidence in which they postulate an underlying system's properties. However, the complex systems that govern cognitive and perceptual behaviors are often dynamic and must contend with epistemological variance in any observed data. Bayesian inference, unlike other methods, is equipped to incorporate noisy observations directly into model design. Fundamental beliefs about the latent system can be encapsulated into a prior distribution, and a carefully chosen likelihood function can model how observations are generated (including assumptions of variance). As data are observed, Bayesian inference applies Bayes' Theorem to derive a posterior distribution that describes the updated beliefs about the underlying system (Bayes & Price, 1763; Jaynes, 2003). The posterior distribution takes into account the observational model and the prior assumptions and returns a prediction of uncertainty. Bayesian inference performs well even with relatively small data sets. Additionally, it is straightforward to iteratively update the model as new data are collected. In this case, the posterior distribution is reassigned as the prior distribution and updated using the new likelihood function that incorporates a new observation. Being a composite measure, neither the prior distribution nor the likelihood function enforces complete control over the shape of the

posterior distribution. Iterative implementations allow the model to adjust and deviate as new data dictate.

In the GP framework, the prior distribution, which is a GP, and the likelihood function are updated via Bayesian inference as data are observed. The resulting posterior distribution is also a GP, which becomes the new prior in the next iteration of data collection. Unlike many machine learning methods, GPs specify a posterior probability distribution of the underlying function for every point in the input domain. The posterior probability provides a confidence estimation of the model. The mean of the posterior distribution denotes the best prediction of the underlying function given all inputs. The uncertainty of the estimation can be represented by the variance of the posterior. Besides being non-parametric, the choice of an iterative Bayesian inference GP framework is advantageous in that it pairs well with active sampling techniques. Active sampling optimizes model performance by selecting the most informative next point at which to query. An acquisition function encapsulates the specific active sampling technique and defines what qualifies as the 'most informative' data to be sampled. In this thesis, the acquisition function is based on the posterior distribution's variance when used in regression. For classification, new points are queried according to Bayesian active learning by disagreement, which minimizes the entropy of the posterior distribution (Garnett, Osborne, & Hennig, 2013; Houlsby, Huszar, Ghahramani, & Lengyel, 2011). In this way, each new query embodies the point at which the model is most uncertain given the previous stimulus and response pairs.

This new framework can model a single psychometric function, improving on previous models by employing GPs and active sampling. This is referred to as disjoint estimation. Disjoint estimation resembles traditional psychometric models in that it samples and estimates within the

same input domain. Disjoint estimation is incredibly flexible in that it can estimate a standard, low-dimensional psychometric function over a continuous domain. Or, disjoint estimation can be used on complex, high-dimensional domains previously unexplored (combining speech-in-noise and working memory, for example).

The GP framework is an iterative, Bayesian inference framework that can model non-parametric relationships between input data and observed measurements. Prior information can be encoded in the mean and covariance functions, and new data can be efficiently queried based on active learning techniques. The GP framework is designed to be flexible and efficient, even with small amounts of data. This enables intrasubject variability models that can scale to high-dimensional input domains while maintaining an efficiency and accuracy comparable to low-dimensional assessments.

2.3 Audiology Background

The initial applications of the machine learning framework have focused on the pure-tone audiogram. Pure-tone audiograms are the most commonly used assessment of hearing ability and represent a complex, yet well understood input domain. Accordingly, they are an ideal choice to validate new testing methodology.

2.3.1 Hughson-Westlake Audiograms

Pure-tone audiograms are inherently two-dimensional as each tone is defined by its frequency and intensity. Current clinical methods reduce audiogram estimation to a series of unidimensional tests, discretizing the frequency dimension. Pure-tone audiograms measure the lowest intensity at which an individual can detect a pure tone for a given set of frequencies. Typically, frequencies are selected at octave or half-octave intervals from 250 Hz to 8000 Hz

(American Speech-Language-Hearing Association, 2005). Tone intensities are often measured in units of hearing level (HL), which are calculated relative to their offset from a population-representative 'normal' hearing curve. Audiograms deliver tones with intensities from −20 dB HL to 100 – 120 dB HL, depending on the frequency, in 5 dB increments (American Speech-Language-Hearing Association, 2005). Clinically, a threshold at a fixed frequency is estimated by systematically varying the intensity of the tone delivered based on the individual's reported response following the Hughson-Westlake audiogram (HWAG) procedure (**Figure 2.2a**) (Carhart & Jerger, 1959; Hughson & Westlake, 1944).

At each frequency, the initial tone is presented at a level that is expected to be easily detected. Subsequent tones are delivered at lower and lower intensities until the tone is no longer reported as audible. At this point the intensity level is increased until it again reaches a level detected by the listener. When an individual's response switches from 'heard' to 'not heard' it is considered a reversal. Adaptive up-down staircase methods are commonly used to estimate models of perception and cognition. In audiometry, this modified up-down method determines the 70.7% threshold of detection based on the averaged intensity of the reversals (Carhart & Jerger, 1959; Hughson & Westlake, 1944). This procedure is repeated for each frequency and for each ear.

HWAG cannot provide a continuous threshold estimate of hearing ability across the frequency domain, but must linearly interpolate between discrete frequency estimates (**Figure 2.2b**). It follows that HWAG must determine the threshold estimate of one frequency before proceeding with subsequent frequencies, and incomplete frequency estimates cannot be exploited to improve the final threshold estimate. On average, pure-tone HWAG administered in the clinic require ~100 tone presentations to obtain a six-octave audiogram threshold estimation for both ears

(Song et al., 2015). Details of an individual's pathology (for example: narrow notched-shaped loss) can be missed if it occurs solely between the frequencies estimated (Kwak & Kwak, 2007). Simply adding more frequencies to the audiogram is often not practical as test time increases linearly with the number of frequencies estimated. Additionally, each frequency estimation often begins with highly uninformative stimuli, delivering tones well above threshold (see Figure 2.2a).

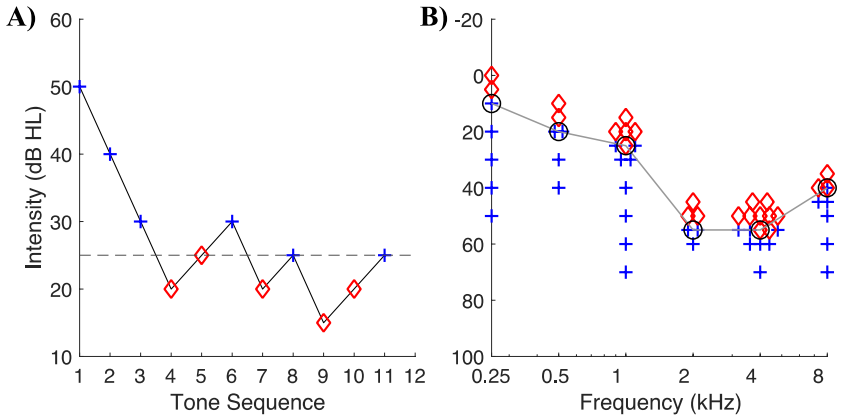

Figure 2. 2: Hughson-Westlake Audiogram (HWAG) procedure. Red diamonds denote 'not heard' responses, blue pluses denote 'heard' responses. A) Example of tones and responses for one frequency. Reversals between the 'heard' and 'not heard' responses determine the threshold. This procedure is repeated for each frequency. B) The final audiogram is a linear interpolation between discrete frequency threshold estimates.

Automated audiometry methods present the opportunity for standardization and uniformity of hearing assessments regardless of patient hearing status. While manual HWAG is considered the clinical standard for threshold estimation, automated and adaptive techniques have demonstrated similar accuracy and reliability to manual audiometry (Ho, Hildreth, & Lindsey, 2009; Mahomed, Swanepoel, Eikelboom, & Soer, 2013; Shojaeemend & Ayatollahi, 2018; Swanepoel, Mngemane, Molemong, Mkwanazi, & Tutshini, 2010). These methods have yet to see widespread adoption.

2.3.2 Hearing Loss Categorization

Hearing ability can be categorized as normal hearing, symmetric hearing loss, and asymmetric hearing loss (**Figure 2.3**). Normal hearing is defined as threshold estimates between -20 dB and 15 dB at all frequencies. As the name implies, symmetric hearing loss refers to individuals with left and right ear thresholds within 10 dB of one another, matched at all frequencies. Asymmetric hearing loss individuals present with a minimum difference between left and right thresholds of 10 dB at three contiguous frequencies or a 15 dB difference at any two or more frequencies (Margolis & Saly, 2008).

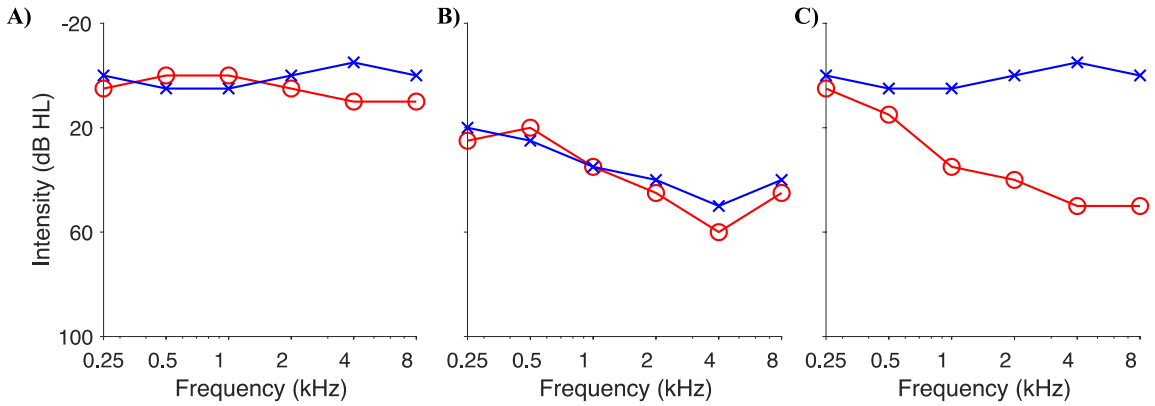

Figure 2. 3: Hearing ability categorization. Red circles denote right ear thresholds; blue X's denote left ear thresholds. A) Normal hearing. All thresholds are between −20 and 20 dB HL. B) Symmetric hearing loss. Left and right ear thresholds matched for frequency are within 10 dB of each other. C) Asymmetric hearing loss. Left and right ear thresholds matched for frequency are greater than 10 dB different for at least three contiguous frequencies. Two or more non-contiguous frequencies greater than 15 dB difference is also considered asymmetric hearing loss.

Sound can be transmitted through air or through bone vibrations. Bone-conducted sound is heard directly by the inner ear and bypasses the outer and middle ear components. Sound waves transmitted through the air, on the other hand, pass through the eardrum before traveling to the cochlea and auditory nerve. Air- and bone- conduction audiograms assess the functionality of each pathway. Hearing loss can also be subdivided according to which part of the ear is

damaged. Damage to the outer or middle ear is considered conductive loss. Individuals with conductive loss are identified by normal bone-conduction thresholds despite air-conduction hearing loss, also known as an air-bone gap (**Figure 2.4b**). Inner ear damage to hair cells in the cochlea or the auditory nerve is classified as sensorineural hearing loss. Most symmetric hearing loss individuals have sensorineural loss (Dubno, Eckert, Lee, Matthews, & Schmiedt, 2013; Ho et al., 2009). Since damage to the inner ear obstructs sound transmission through bone and air pathways, sensorineural hearing loss is defined by bone-conduction thresholds that are similar to air-conduction thresholds, or a lack of an air-bone gap (**Figure 2.4a**). Mixed hearing loss refers to individuals with both conductive and sensorineural loss in the same ear. Individuals with mixed hearing loss are identified by an air-bone gap in which bone-conduction thresholds are not normal (**Figure 2.4c**).

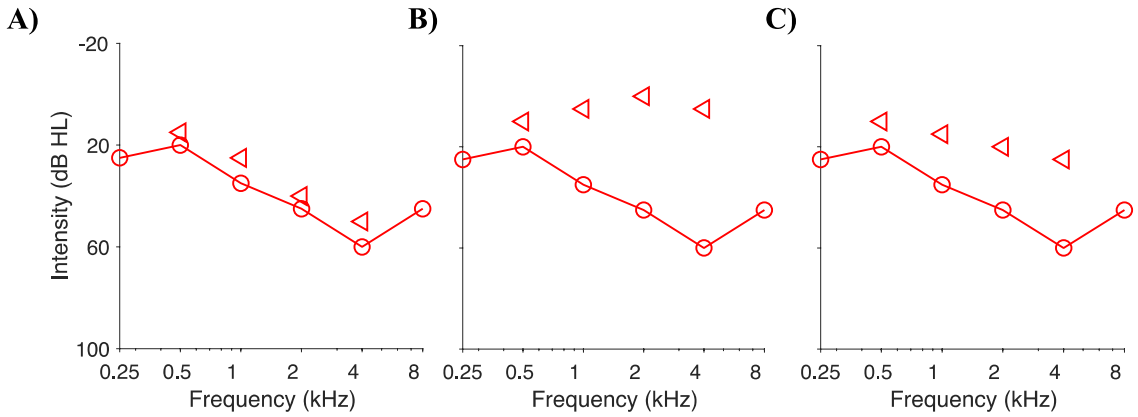

Figure 2. 4: Types of hearing loss. Red circles denote right ear air-conduction thresholds, red triangles denote right ear bone-conduction threshold. Bone-conduction thresholds are typically only tested at frequencies between 500 and 4000 Hz. A) Sensorineural loss. Bone-conduction thresholds are within 10 dB of air-conduction thresholds at all frequencies. B) Conductive hearing loss. Bone-conduction thresholds are near normal despite air-conduction loss. This is called an air-bone gap. C) Mixed hearing loss. Bone-conduction thresholds are not normal but there is still an air-bone gap.

2.3.3 Cross Hearing

Useful audiograms depend on confident threshold estimates for each ear. One challenge to this procedure is when the cross-hearing of tones occurs. Cross hearing arises when loud tones

presented to the test ear cross over and are heard by the non-test ear via bone conduction through the skull (Martin & Blosser, 1970). If a tone presented to the test ear is actually heard by the non-test ear, what was intended to be an independent assessment of the test ear's hearing ability is now confounded by the contralateral ear's response. In such cases, the estimated threshold of the test ear is artifactually lower than the true threshold.

A sound delivered to the test ear will lose some intensity as it travels to the contralateral ear, and it arrives at the non-test ear at a reduced sound level compared to its starting intensity. Interaural attenuation reflects the amount of sound energy that dissipates as the tone travels from the ipsilateral test ear to the contralateral non-test ear. Because interaural attenuation varies for each individual based on the dimensions of their skull, transducers used, frequency of the sound, and other testing factors, current compensatory testing methods rely on a conservative estimate of interaural attenuation for each transducer. For supra-aural and circumaural headphones, 40 dB is used across all frequencies (Brännström & Lantz, 2010; C. R. Smith, 1968). Having less contact with the skull, insert headphones have a higher estimated interaural attenuation of 50 dB – 75 dB depending on the frequency tested (M C Killion, Wilber, & Gudmundsen, 1985; Munro & Contractor, 2010; Sklare & Denenberg, 1987). Tones are conventionally considered at risk of cross-hearing only if their intensities are greater than the interaural attenuation estimate plus the hearing threshold of the non-test ear.

To offset the effects of cross-hearing, narrowband noise is introduced to the non-test ear to mask any potential cross tone detection in that ear (Denes & Naunton, 1951; Hood, 1960; Studebaker, 1964). Current methods do not assess the need for masking until after initial unmasked threshold estimates are determined. Only frequencies with significantly asymmetric left and right ear

thresholds after this testing are suspected to be incorrect due to contralateral ear responses. A masking procedure is then employed to re-estimate the thresholds at the identified frequencies. Because measuring the actual interaural attenuation at each frequency is impractical with conventional testing, a complex yet generic protocol must be used to re-evaluate the threshold estimates and achieve effective masking levels without overmasking (i.e., allowing the masker to cross over and affect tone detection in the test ear). One method to administer masking requires multiple iterations of re-establishing the threshold in the test ear while systematically adjusting the amount of masking in the non-test ear (Hood, 1960). Optimized methods requiring fewer iterations have been proposed (C. R. Smith, 1968; Turner, 2004a, 2004b) but are similarly constrained by the need to perform masking after initial unmasked audiograms are completed, thus substantially increasing true threshold estimation time.

Only individuals with severely asymmetric hearing or air-bone gaps are at risk for cross-hearing. In individuals with a small air-bone gap due to low asymmetry or symmetric hearing, any cross-tone is below the bone-conduction thresholds of the non-test ear and does not affect the test-ear audiogram. As a rule of thumb, masking is required when either:

1) $Air\ Conduction_{TestEar} - Bone\ Conduction_{NonTestEar} \geq interaural\ attenuation$
2) $Air\ Conduction_{TestEar} - Air\ Conduction_{NonTestEar} \geq interaural\ attenuation$

2.3.4 Active Machine Learning Audiogram

The Barbour lab has developed and validated a Gaussian process machine learning framework for audiogram acquisition, the active machine learning audiogram (AMLAG). AMLAG, as shown in **Figure 2.5**, delivers continuous threshold estimates over the entire frequency domain in fewer tone presentations than conventional methods (Song et al., 2015). Analyzing over one

million audiograms in the National Institute for Occupational Safety and Health database, strong concordance was found between neighboring frequencies (**Figure 2.6**) (Barbour, DiLorenzo, et al., 2019). AMLAG is capable of exploiting the shared information between adjacent frequencies where HWAG cannot, significantly reducing test time.

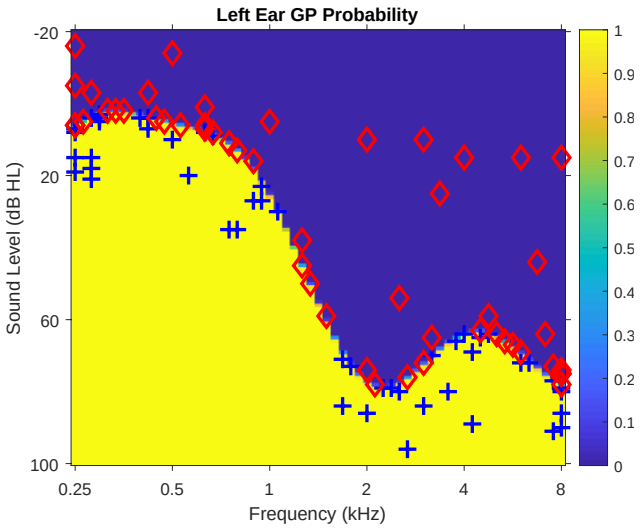

Figure 2. 5: Final audiogram acquired with AMLAG. Red diamonds denote 'not heard' responses and blue pluses denote heard responses. Threshold estimate is continuous across the frequency domain. Tones are optimally selected near threshold.

AMLAG deploys active machine learning to estimate an individual's threshold audiogram. The current implementation of AMLAG is an iterative, Bayesian inference GP classification method. An uninformative prior distribution allows the model to adjust according to the observed data without the constraint of any threshold assumptions. Future implementations could employ a more informative prior such as a previous audiogram or a population or sub-population average. An uninformative prior is used in this work with the intent of demonstrating the flexibility of the GP model to accurately assess hearing ability in individuals with no prior knowledge and to serve as a worst-case limit on model efficiency.

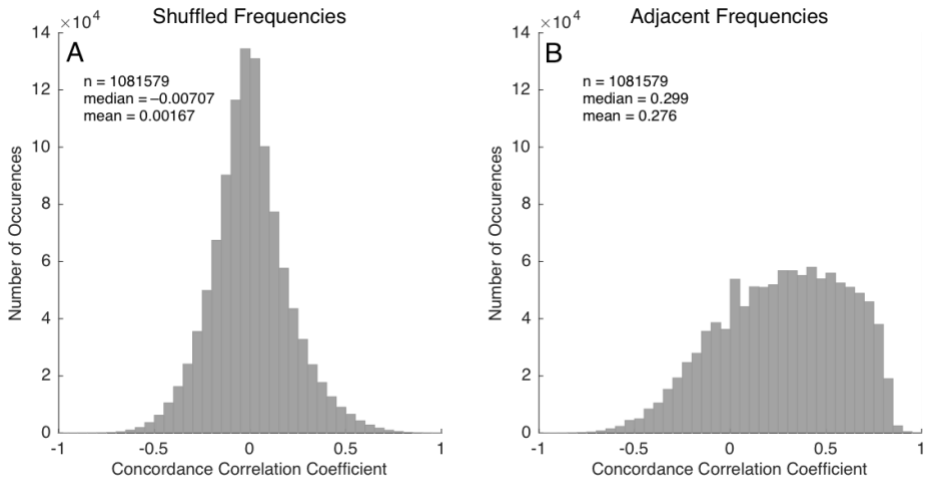

Figure 2. 6: Concordance correlation of 1,000,000+ NIOSH audiogram thresholds between A) arbitrary frequencies and B) adjacent frequencies. Threshold at a particular frequency could be used to speed up estimation of thresholds at adjacent frequencies. AMLAG does this, but conventional methods do not.

The observed data is a binary variable encoding an individual's response (or lack of response) to a detection task. Similar to HWAG, a listener is tasked to response when a tone is detected. AMLAG tones are selected from the frequency and intensity domain defined by semitone octave frequencies from 250 Hz to 8000 Hz and 1 dB intensity increments ranging from –20 to 100 dB HL. After each observation of tone response, the posterior distribution is updated to reflect all responses that have been observed up to this point. The posterior distribution reflects the probability that a tone will belong to the "heard" group. The mean of the posterior distribution signifies the psychometric function in that it produces the probability of detecting a tone specified by a given frequency and intensity (**Figure 2.7a**). The class boundary between "heard" and "unheard" responses is calculated at the 0.707 detection probability and corresponds to the HWAG threshold estimation. Variance of the posterior distribution suggests the model's uncertainty (**Figure 2.7b**). New frequency-intensity pairs are selected according to Bayesian

active learning by disagreement such that each successive stimulus is optimally chosen to best inform the model (**Figure 2.7c**). Because new points are always selected to be most informative, AMLAG very quickly focuses its sampling on the frequency and intensity pairs where the probability of tone detection is close to 0.5.

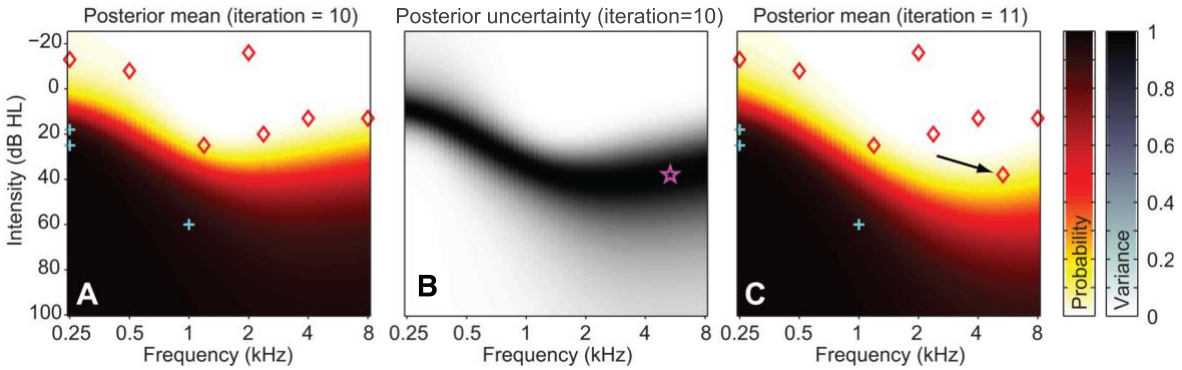

Figure 2. 7: Illustration of the sampling algorithm used by the Gaussian process (GP) for AMLAG. A) Posterior mean is computed by the GP using the sampled points. Red diamonds indicate the tone was inaudible; blue pluses, audible. B) Posterior uncertainty is computed by the GP using the sampled points, and the point of maximum uncertainty is identified (purple star). C) The point of maximal uncertainty is queried for listener audibility (black arrow). Once it is determined that the listener did not hear this tone, the updated set of points is used by the GP to re-compute the posterior mean with a more elevated threshold near the frequency of that tone.

AMLAG uses a constant mean function ($\mu(x) = c$) and a composite covariance function ($K(x, x') = K_\omega(x, x') + K_\iota(x, x')$) that integrates audiology-specific assumptions of the frequency (ω) and intensity (ι) domains. As a tone increases in intensity, its probability of being heard also increases. Thus, a monotonically increasing linear covariance function was placed in the intensity dimension: $K_I(x, x') = s_1 \cdot (I \cdot I')$. To ensure a sigmoidal probability function, K_I is transformed with a cumulative Gaussian likelihood. The frequency domain is assumed to be smooth and continuous, dictating the choice of a squared exponential function: $K_\omega(x, x') = s_2^2 \cdot$

$\exp\left(-\frac{(\omega-\omega')^2}{2\ell^2}\right)$. The hyperparameters of the mean function (c) and the covariance functions (scalar factors s_1 and s_2 and length constant ℓ), are learned by gradient descent.

By iteratively querying a listener at optimized points in the frequency-intensity domain, AMLAG's posterior distribution always represents the model's best prediction of a listener's hearing ability and an audiogram threshold estimate can be made at after any iteration.

2.4 Speech-in-Noise Background

Speech-in-noise tests are a psychophysical measure of speech comprehension ability in the presence of background noise. While pure-tone audiograms measure audibility thresholds to define hearing type and configuration, speech-in-noise tests more accurately represent auditory challenges encountered outside of experimentally controlled environments (Taylor, 2003). Pure-tone audiometry is often not very predictive of a subject's speech comprehension in noise (M C Killion & Niquette, 2000; Moore et al., 2014). Instead, specific assessments have been developed to test this perceptual ability directly.

2.4.1 Speech-in-Noise Assessments

Comprehension of speech that is acoustically degraded by noise is quantified using a signal to noise ratio (SNR). The lower the SNR, the more prominent the competing noise and the more difficult the perceptual test. While the exact parameters may vary, speech-in-noise tests require a listener to repeat back a stimulus presented at systematically adjusted or fixed SNRs (Egan, 1948; Fletcher, 1929). If enough data are collected, the relationship between the successful repetition of the stimulus and the SNR can be modeled with a psychometric function. Similar to pure-tone audiogram acquisition, adaptive speech-in-noise assessments often utilized staircase

methods to determine the SNR at which a specific percentage of the stimuli are correctly repeated. SNRs in adaptive speech-in-noise tests typically range from –15 dB SNR to 15 dB SNR with normal hearing listeners performing at the 50% threshold near 0 dB SNR. While the 50% threshold is a commonly selected threshold level, successful comprehension in noisy environments requires greater than 50% understanding, motivating a higher threshold level to be considered (Robinson & Casali, 2003). Alternatively, if observed data are fitted to a psychometric function, multiple performance levels can be estimated. Fixed speech-in-noise assessments present stimuli at predetermined SNRs and record the percentage of stimuli correctly repeated back by the listener. The advantage of fixed speech-in-noise tests is their ability to directly assess the listener's performance in SNRs commonly encountered outside of the laboratory (Le Prell & Clavier, 2017). Most speech-in-noise assessments must be scored by a human observer either during the assessment or at a later time if the responses were recorded.

There is no standardized speech-in-noise assessment. Frequently used stimuli are single words, sentences, or even phonemes. Word-based stimuli have a variety of manipulatable defining features. Among others, they can vary in frequency, familiarity, syllables, or phonological neighborhoods. Similarly, background noise can vary from test to test. Common choices of background noise are white noise, noise filtered to match the speaker's speech spectrum (referred to as speech-shaped noise), or speech babble. Stimulus choice and noise type alters the perceptual challenge (Brungart, Sheffield, & Kubli, 2014; Le Prell & Clavier, 2017). It is likely that a listener's performance will differ with varying speech-in-noise parameters; although their threshold SNR between tests would be highly correlated (Spyridakou & Bamiou, 2015).

2.4.2 Neural Components of Speech-in-Noise Assessments

Speech processing is a complex task that requires multiple, hierarchical stages to successfully be performed (Davis & Johnsrude, 2003; Okada et al., 2010; Peelle, Johnsrude, & Davis, 2010). Primary auditory cortex processes auditory aspects of speech and are sensitive to the acoustic structure of incoming stimuli. Bilateral superior temporal gyrus, left inferior frontal gyrus, and other temporal lobe regions near the auditory cortex allow access to the mental lexicon and maps speech to stored semantic representations that attach meaning to incoming sound (Davis & Johnsrude, 2003; Narain et al., 2003; Scott Blank, Catrin, Rosen, Stuart, and Wise, Richard J.S., 2000). The exact nature of linguistic property processing and the degree of acoustic sensitivity is still debated and is an area of active research.

By modifying the test parameters, speech-in-noise measures can offer a cognitively demanding perceptual test (Heinrich, Schneider, & Craik, 2008; Rudner, Foo, Rönnberg, & Lunner, 2007). As the task difficulty escalates, cognitive resources beyond purely perceptual systems are thought to contribute to behavior (**Figure 2.8**) (Peelle, 2018). Attentional control and executive functions assist performance as test challenge increases and the listener must concentrate their focus on the target stimulus (Wingfield, Tun, & Mccoy, 2005). Proposed theories on speech comprehension suggests that working memory resources inevitably engage when listening to noisy speech stimuli, even if comprehension is ultimately successful (Rönnberg, 2003; Rönnberg, Rudner, Foo, & Lunner, 2008). In these models, speech unencumbered by background noise is quickly processed through the auditory-perceptual network and meaning is attached by accessing long-term memory storage. When speech is degraded, listeners need to store and process incoming signals for an extended time compared to clear speech. Working memory

resources can then offset the ambiguity caused by degraded stimuli by inferring meaning and context from surrounding signals.

Brain imaging during speech-in-noise tests at high levels of perceived difficulty show boosted activity in the frontoparietal network (Davis & Johnsrude, 2003; Peelle, 2018; Wingfield et al., 2005). The frontoparietal network consists of regions in the frontal and parietal lobes and is active in a myriad of domain-general functions and executive functions of some working memory models (Marek & Dosenbach, 2018). Also active is the cingulo-opercular network (Erb, Henry, Eisner, & Obleser, 2013; Vaden et al., 2016, 2013). This network is commonly associated with performance or error monitoring (Vaden, Kuchinsky, Ahlstrom, Dubno, & Eckert, 2015; Vaden, Teubner-Rhodes, Ahlstrom, Dubno, & Eckert, 2017). Notably, increased activity is observed even before the listener's performance begins to suffer, and activity in this network may predict future successful speech comprehension (Vaden et al., 2013). Generally, there is evidence that domain-general resources lend cognitive support to aid in performance maintenance during speech-in-noise tests as they increase in task challenge. The nature of that support has yet to be fully determined.

One manipulation of test challenge used in this thesis is in the intentional selection of a stimuli's phonological neighborhood. Phonological neighborhoods are defined as groups of words that differ by only one phoneme (Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980). Studies contend that words stemming from dense phonological neighborhoods create more demand on cognitive and perceptual resources compared to words with few phonological neighbors (Chen, Vaid, Boas, & Bortfeld, 2011). One cause of increased cognitive demand might be the extra

inhibition required to select the correct word from similar competing words in the mental lexicon.

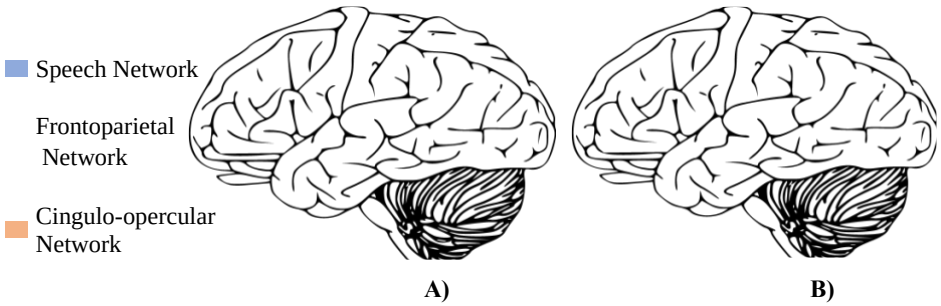

Figure 2. 8: Brain regions active during successful speech comprehension. A) Speech networks active during all speech comprehension. B) domain-general regions thought to support challenging speech comprehension that are also active during working memory tests.

2.4.3 Speech-in-Noise Assessments and Age

As people age, speech comprehension can be relatively well preserved, even in the presence of age-related cognitive and perceptual decline (Peelle, Troiani, Wingfield, & Grossman, 2010; Wingfield & Grossman, 2006; Wingfield, Mccoy, Peelle, Tun, & Cox, 2006). One hypothesis is that additional cognitive resources are recruited to support auditory processing (Lin et al., 2011; Pichora-Fuller, Schneider, & Daneman, 1995; Tun, Mccoy, & Wingfield, 2009). Age-related compensation from domain-general cognitive operations is similar to the neuronal recruitment observed in complex speech-in-noise tests perceived as challenging. In older adults, SNRs that result in correct responses may already be recruiting neural resources that are not necessary in younger adults exhibiting similar task performance. As task difficulty increases, neural resources are more quickly exhausted and performance more readily deteriorates compared to younger subjects (Harris, Dubno, Keren, Ahlstrom, & Eckert, 2009; Moore et al., 2014; Peelle, Troiani, et al., 2010; Pichora-fuller, Schneider, & Daneman, 2006). Designing a test that can explore the diverse cognitive demands of speech comprehension in the presence of competing background

noise can help determine how the brain prioritizes and supports task performance in healthy aging.

2.5 Working Memory Assessments

Multiple theories on the exact construct of working memory are debated in the current literature (Baddeley, 1986; Cowan, 1999; Nairne, 1990; Postle, 2006). One of the most cited paradigms is Baddeley's component model. Under this system for working memory, a central executive mechanism directs the limited attentional resources and dictates access to long-term memory storage. Stimulus specific sub-systems process visual information through the visuospatial sketchpad or auditory information through the phonological loop. An episodic buffer has been proposed to assist with grouping incoming information into related chunks, mediate between the visuospatial sketchpad and phonological store, and access long-term memory support (Baddeley, 2000).

Verbal working memory, which is implicated in speech-in-noise tasks, primarily stores and manipulates incoming verbal information via the phonological loop, central executive processes, and the episodic buffer. The phonological loop maintains a verbal trace of the, stimulus and, through silent articulation, keeps it active in memory (Baddeley, 2003).

Many behavioral methods have attempted to assess working memory. A defining feature of such tests is the temporary storage and manipulation of incoming information. Working memory can retain a limited number of incoming stimuli (Miller, 1956), referred to as working memory capacity. As a test nears a subject's working memory capacity, response time and accuracy begin to suffer (B. M. J. Kane & Engle, 2002).

2.5.1 N-back Assessments

One commonly implemented measure for working memory is the N-back test (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kirchner, 1958). An N-back test presents a series of stimuli, and the participant must recall if the current stimulus had been presented N presentations ago. Like speech-in-noise assessments, there is no standardized N-back test. The load, or the N of the N-back test, typically varies from a 0- to a 4-back. Higher loads are possible but less frequently implemented. Participants may be asked to response only when they identify a positive match, or they may be asked provide a binary (yes/no) response after every stimulus presentation. Many N-back assessments include foils in their test design. Foils are repetitions of stimuli previously presented that do not match the current N-back target. For example, during a 3-back test a foil could be two stimuli presented back-to-back (a 1-back presentation in a 3-back test). Foils help deter participants from simply matching previously presented stimuli based on recognition. The number of N-back loads, the number of targets and foils, as well as the stimuli used vary from study to study. Visuospatial, visual, or verbal N-backs are commonly used and assess different aspects of working memory. Purely auditory-verbal N-backs, with no visual component, appear to be more rare, however (Hancock, LaPointe, Stierwalt, Bourgeois, & Zwaan, 2007; Monk, Jackson, Nielsen, Jefferies, & Olivier, 2011).

N-back tests have strong face validity as a working memory measure evidenced by the need to maintain, update, match, and encode the set of N previous stimuli (Jonides et al., 1997). A main appeal of the N-back as a working memory test is that it is straightforward to increase working memory load by increasing the N of the N-back. Load manipulation produces robust increases in reaction time and errors (Jaeggi et al., 2010).

A series of recent studies have focused on the validity of the N-back test as a psychometric measure of working memory. Weak correlations between N-back and complex span tests have extrapolated that different aspects of working memory are engaged by these two tests (M. J. Kane, Conway, Miura, & Colflesh, 2007). However, N-back accuracy and reaction time correlate with validated measures of task switching and updating, interference control, attention, and processing speed (Gajewski, Hanisch, Falkenstein, Thönes, & Wascher, 2018) substantiating it as a useful measure of a hard-to-define construct.

2.5.2 Neural Components of N-back Assessments

Brain imaging studies have widely used the N-back to examine activation associated with verbal working memory. Activity in frontoparietal and the cingulo-opercular networks is often found, regardless of N-back modality (Chein & Fiez, 2010; Honey et al., 2002; B. M. J. Kane & Engle, 2002; Owen, McMillan, Laird, & Bullmore, 2005), and activity is highly sensitive to the manipulation of memory load (Braver et al., 1997; Jonides et al., 1997).

The dorsolateral regions of the prefrontal cortex are suspected to participate in a wide array of working memory processing including monitoring and maintaining incoming stimuli (Owen, 1997; Wang et al., 2018). Activity in this region has been particularly implicated for being a key contributor to N-back performance (Barbey, Koenigs, & Grafman, 2013; Braver et al., 1997; Rodriguez-Jimenez et al., 2009). As previously mentioned, activation in the cingulo-opercular network is usually reflected in performance monitoring, attention, and increased test effort (Barch et al., 2001; Vaden et al., 2017). Working memory assessments, generally, and N-back tests, specifically, often see additional activation in other brain regions and deactivation in default mode networks. The extent of neural activation seems to vary with different N-back

parameters, and elucidating how each region contributes to overall performance is a persistent goal of the field.

Overall, the N-back test is a good candidate for probing working memory as demonstrated by its proven validity as a psychometric measure and the correlated activation of brain regions governing specific aspects working memory.

2.5.3 N-back Assessments and Age

Verbal working memory capacity is highly variable across individuals (DeCaro, Peelle, Grossman, & Wingfield, 2016; B. M. J. Kane & Engle, 2002). As one ages, an individual's working memory capability declines, but variability within age cohorts remain (DeCaro et al., 2016). This variability and the wide range of contributing brain regions make it difficult to assert generalizations beyond an overall shift in performance. Individual differences in cognitive decline, life experience, and neural connectivity compound the challenge of teasing out exactly which mechanisms underlie age-related shifts in working memory. However, similar to the resource strain in complex speech-in-noise tasks, it is hypothesized that additional domain-general resources are recruited to assist working memory tasks at the onset of age-related decline (Grady, 2013; Kirova, Bays, & Lagalwar, 2015; Peelle, Troiani, et al., 2010; Wingfield & Grossman, 2006). Specifically, the N-back test has been shown to closely measure age-related shifts in executive and attentional control. Older adults consistently have longer reaction times as well as lower working memory capacities (Braver & West, 2008; Gajewski et al., 2018; Mattay et al., 2006). Neuroimaging of young and older adults has revealed differences in brain activation during N-back tests. Older adults frequently display bilateral brain activation compared to young adults who depict more specialized, left-lateralized activation (Mattay et al., 2006; Nyberg,

Dahlin, Stigsdotter Neely, & BÄckman, 2009; Reuter-Lorenz et al., 2000). The additional neural recruitment is thought to help older adults maintain performance as they age.

2.6 Working Memory and Speech-in-Noise

It has been suggested that age-related changes in cognition are predicated on deficits in sensory processing (Humes, Busey, Craig, & Kewley-port, 2013; Humes, Kidd, & Lentz, 2013). In the case of verbal working memory tests, one must consider the possibility that a decline in auditory processing is contributing to shifts in both brain function and behavior. Age-related decline in hearing ability has been identified as a direct predictor of future cognitive function and Alzheimer's Disease progression (G. a Gates, Anderson, Feeney, Susan, & Larson, 2008). Given that deficits in memory are also a reliable predictor of cognitive decline, examining the relationship between memory and hearing ability may provide further insight to age-related changes in health and disease.

The current methods of measuring speech-in-noise and verbal working memory treat these behaviors as two completely separate constructs. Advancements made with neuroimaging indicate a much more intricate theory where domain-general resources entwine these two measures to support function throughout the lifespan. A test that evaluates both of these abilities together has potential value as a more sensitive behavioral test of brain function than separate tests.

2.7 Concluding Remarks

The work presented in this thesis integrates concepts from machine learning, psychometric and psychophysical model design, and individual differences in neural activity to evaluate a machine

learning framework that can perform joint estimation of perceptual, cognitive, and neural processes.

Chapter 3: Bilateral Audiogram

Note: The research presented in this chapter has been published in *Acta Acustica* (Heisey, Buchbinder, & Barbour, 2018).

3.1 Introduction

Hearing naturally involves two ears, though clinicians and researchers typically evaluate one ear at a time, resulting in two independent, unilateral audiograms. As described in Chapter 2, AMLAG provides a compelling method to evaluate hearing ability in the two-dimensional domain of frequency and intensity. AMLAG can be used to optimize data acquisition for each ear independently, leading to substantial efficiency gains (Song et al., 2015). Proceeding sequentially by ear, AMLAG efficiently and accurately estimates the hearing thresholds across each ear's stimulus domain separately. Although human sound transduction is not physiologically linked between the ears, the ears do share many features in common, including genetics, physical proximity, lifetime sound exposure, blood supply, downstream neural processes, etc. Therefore, one might expect thresholds between most individuals' two ears to be similar. This indeed is the case, with 50% concordance between left and right ear thresholds in over 1 million working-age adults (**Figure 3.1**) (Barbour, DiLorenzo, et al., 2019; Masterson et al., 2013). The similarity between two ears could represent additional information usable to speed model estimation concurrently in both ears with a conjoint estimator.

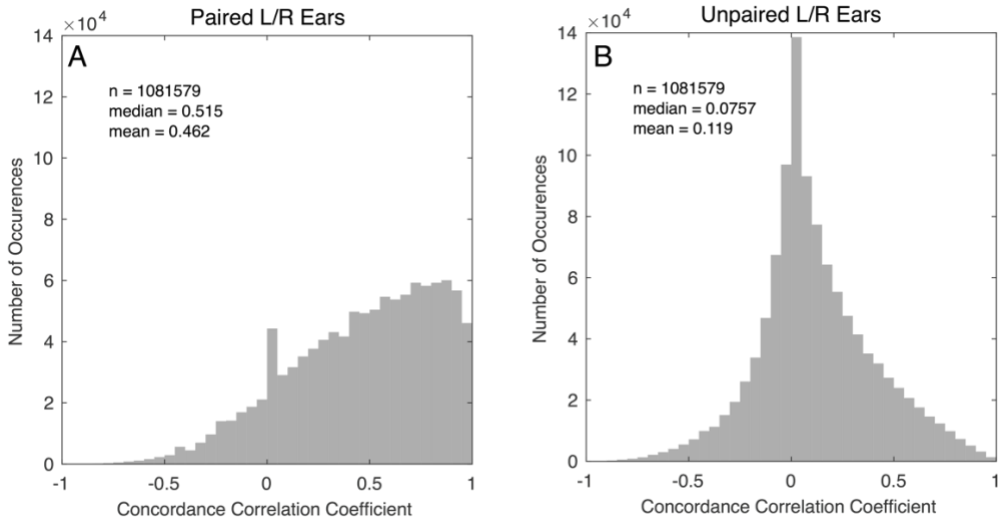

Figure 3. 1: A) Pairs of audiogram thresholds derived from ears in the same individual ("paired") yield high positive concordance. B) Pairs of audiogram thresholds deriving from ears in different individuals ("unpaired") yield concordance much closer to 0, though still with a tendency toward positive values.

A conjoint psychometric estimator is defined as one that updates the model fit of two or more psychometric functions from observations over the stimulus domain of one of them. In this way, even loosely related input domains can be linked together and shared covariance between input domains can be exploited. Conjoint estimators are distinct from disjoint estimators, which are traditional estimators that observe and model in the same input domain; unilateral AMLAG uses disjoint estimation. The extension to conjoint estimation is possible because the method used to implement the probabilistic classifier is a Bayesian kernel method capable of learning nonlinear relationships between variables of interest. As long as the input domains share some interrelationship, a method that can learn and exploit this information could produce accurate multidimensional estimates in less time.

The first logical psychoacoustic conjoint estimator to develop is the bilateral audiogram, where observations from one ear mutually reinforce hearing estimates of the contralateral ear.

Effectiveness of such an estimator is predicted from high average similarity between paired ears. This Chapter shows that bilateral audiogram estimation delivers accurate hearing thresholds in significantly less time than serial unilateral estimation for a variety of hearing loss and hearing asymmetry profiles.

3.2 Methods

3.2.1 Participants

Twenty subjects were recruited for this study, 6 with normal hearing and 14 with known sensorineural hearing loss ranging from moderate to profound loss. All participants provided informed consent prior to testing. The experimental protocol was approved by the Human Research Protection Office of Washington University.

Two of the 20 subjects failed to complete any AMLAG test due to a hardware misconfiguration. For three subjects an algorithmic error in tone delivery prevented one or more tests from executing correctly. Incomplete tests were removed from analysis. Two subjects had profound hearing loss in their right ears with hearing thresholds above the highest sound level delivered. Of the 40 ears that entered the study, 30 were included for analysis, 15 left ears and 15 right ears.

Participants with normal hearing and symmetric and asymmetric hearing loss were recruited; however, care was taken to ensure that all participants were not at risk for cross hearing.

3.2.2 Procedure

Three air-conduction AMLAGs were administered to each subject: one disjoint unilateral right ear, one disjoint unilateral left ear, and one conjoint bilateral. Test order was randomized. Listeners were seated within a sound isolation booth, and all auditory stimuli were delivered

using a Dell XPS laptop computer running custom MatLab code and Etymotic Research 3A insert earphones paired with a DragonFly Black 32-bit DAC (AudioQuest, Irvine, CA). Stimulus delivery and data acquisition were performed with the Bonauria online audiometry platform (Barbour, Howard, et al., 2019). Listeners were asked to remove any hearing-assist devices prior to data collection and an otoscopic observation was performed to confirm that there was no concerning ear canal occlusion in either ear.

Each stimulus consisted of a three-pulse sequence of 200-ms pure tones with silent inter-pulse intervals of 200 ms. Listeners were instructed to press a button whenever they detected a tone presentation. Each tone had a frequency between 250 and 8000 Hz in semitone increments and a level between –20 and 100 dB HL in 1 dB increments. Right, left, and bilateral audiograms delivered a total of 50, 50, and 100 tones, respectively. To prevent listeners from anticipating stimulus presentations, tone deliveries were separated by a randomized silent interval between 3 and 8 seconds. Each response was logged as "Heard" if occurring within 2000 ms of stimulus onset or "False Positive" otherwise. If no response was recorded within 2000 ms following stimulus onset, a "Not Heard" response was logged.

3.2.3 Bilateral AMLAG

Bilateral AMLAG adapts the GP classification model defined in unilateral AMLAG. The ith stimulus $\mathbf{x}_i$ for the bilateral audiogram is augmented from unilateral tone frequency and intensity to include a third discrete "ear" dimension: $x_i = (\omega_i, L_i, e_i)$. The GP kernel function was derived from prior knowledge about the behavior of audiograms. Like unilateral AMLAG, bilateral AMLAG uses a constant mean function: $\mu(x) = c$. The composite covariance kernel incorporates the bilateral "ear" dimension: $K(x, x') = K_e(x, x')(K_\omega(x, x') + K_L(x, x'))$.

Bilateral AMLAG estimates hearing ability in two ears that are not physiologically the same but share the same two-dimensional input domain of frequency and intensity. Logically, prior beliefs that determined the frequency and intensity kernels, $(K_\omega(x,x')\ and\ K_L(x,x'))$, in unilateral AMLAG dictated the use of the same frequency and intensity covariance functions in bilateral AMLAG. Namely, a linear covariance function in the intensity dimension and a squared exponential in frequency. Covariation between all pairs of inputs in ear 1, all pairs in ear 2, and all pairs between the ears is reflected in a discrete conjoint kernel: $K_e\left(\mathbf{x},\mathbf{x}'\right)=\begin{cases} s_{11}\ if\ x,x'\in e_1 \\ s_{12}\ if\ e\neq e' \\ s_{22}\ if\ x,x'\in e_2 \end{cases}$. Hyperparameter s_{12} is referred to as the conjoint correlation and quantifies the psychometric function similarity between the ears. Fixing $s_{12} = 0$ creates a disjoint kernel for each ear, which leads to two independent model fits and is identical to unilateral AMLAG.

In querying a participant's audiogram, the conjoint estimator determines in which ear to deliver the tone as well as the frequency and intensity of tone delivered. Each next stimulus is selected by Bayesian active learning to elicit the subject's response that will maximize the information gain given all previous data and hyperparameters. Hyperparameter learning occurs by gradient descent and is initiated after one heard and one not-heard response has been recorded for each ear being tested. The posterior mean function of the GP is calculated after each probe tone and represents point estimates of detection probability as a function of tone frequency and sound intensity (i.e., the psychometric function). Detection probability at 0.5 was used as an estimation

of detection threshold. Lapses, guesses, and other nonstationarities, such as criterion drift, were not modeled in this implementation of AMLAG.

3.2.4 Data Analysis

Analysis was performed utilizing an offline version of AMLAG written in custom Matlab code. At every query, bilateral AMLAG updates the model of both left and right ears. To compare bilateral AMLAG to unilateral AMLAG tests, all threshold estimates were analyzed as paired ears, each pair was probed with 100 tones (50 from each ear's unilateral test combined, 100 tones for the bilateral test). Paired sample t-test were used to analyze statistical significance in the differences between unilateral and bilateral AMLAG convergence.

3.3 Results

One individual participant's intermediate threshold estimates for each AMLAG type are shown in **Figure 3.2**. Both disjoint and conjoint tests converged to similar threshold estimates at the final tone count. For this subject, conjoint AMLAG learned that both ears share similar hearing functions and used that information to more quickly construct an accurate model.

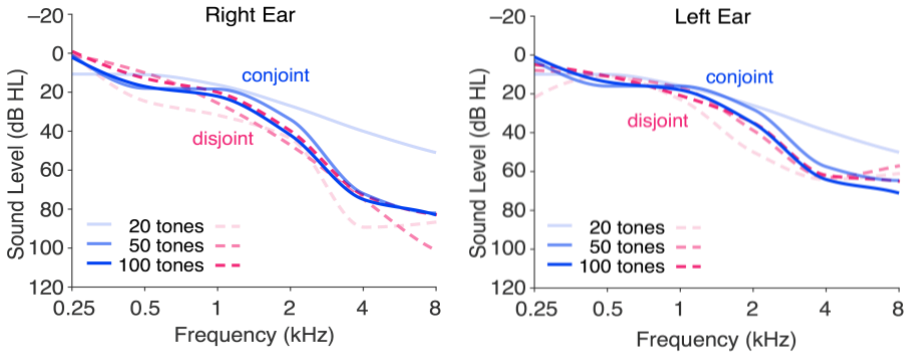

Figure 3. 2: Hearing thresholds for both ears of Subject 1 measured two ways: serial disjoint unilaterally (red dashed) and concurrent conjoint bilaterally (blue solid). Final estimates are similar for the two methods. Tone counts are for both ears combined

To quantify the relative performance of the unilateral and bilateral estimators, the threshold functions estimated after 100 tones were averaged for each AMLAG type. The mean absolute difference between that function and a threshold function estimated by each AMLAG following every tone increment could then be used to determine the convergence rate of the two models. **Figure 3.3** shows these results for one subject. Both disjoint and conjoint methods achieve thresholds near their final estimates within a relatively small number of tones, consistent with previous studies. Under the conditions tested, disjoint estimates tend to vary more with early tone counts.

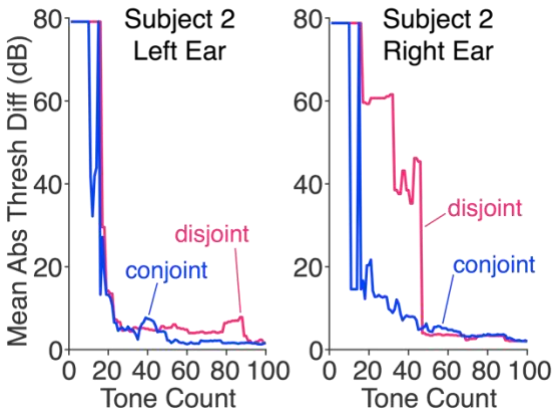

Figure 3. 3: Average absolute difference in thresholds between the final estimate at 100 tones and estimates with each incremental tone. Variation can be seen between convergence of the two methods, but in general, the conjoint estimator tends to achieve its final threshold estimate with fewer tones. Tone counts are for both ears combined.

The relative performance of the two AMLAG types for this population was evaluated by averaging the threshold difference curves for all ears, as shown in **Figure 3.4**. As predicted, conjoint estimation considering both ears concurrently approaches its final threshold estimate values significantly more quickly than disjoint estimation ($p = 2.5\times10^{-12}$, paired-sample t test). The transition to final estimate is also smoother for the conjoint estimator. The two methods tend

to deliver similar estimates, as evidenced by the small mean absolute differences at high tone counts. AMLAG has previously been shown to reliably deliver threshold estimates similar to those of Hughson-Westlake audiometry in fewer tone deliveries (Song, 2015). Following 100 tones for both ears combined, the mean absolute threshold difference between conjoint and disjoint estimates was 5.0 dB.

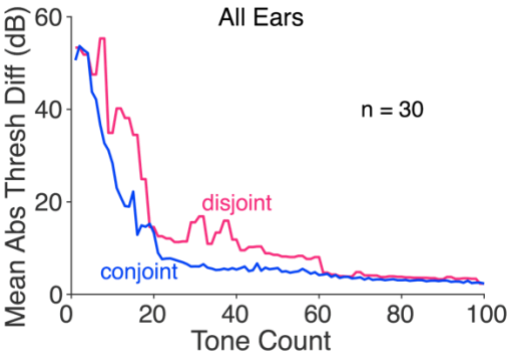

Figure 3. 4: Average absolute population difference in threshold functions between the final estimate at 100 tones and estimates with each incremental tone. The conjoint estimator achieves its final threshold estimate with significantly fewer tones. Tone counts are for both ears combined. Disjoint estimation achieves near-final threshold estimates after about 60 tones (i.e., 30 tones per ear) while conjoint estimation converges after about 30 tones (i.e., 15 tones/ear).

Implementing the conjoint estimator improves the relative convergence rate for all subjects, regardless of hearing type. **Figure 3.5** compares the performance of AMLAG for subjects with asymmetric and symmetric hearing. In both cases, conjoint estimation requires significantly fewer tones (asymmetric: $p = 3.1\times10^{-11}$; symmetric: $p = 3.1\times10^{-8}$; paired-sample t test). The mean concordance between all subjects' ear pairs in this cohort was 0.50, which is similar to the population mean of 0.46 (Figure 3.1).

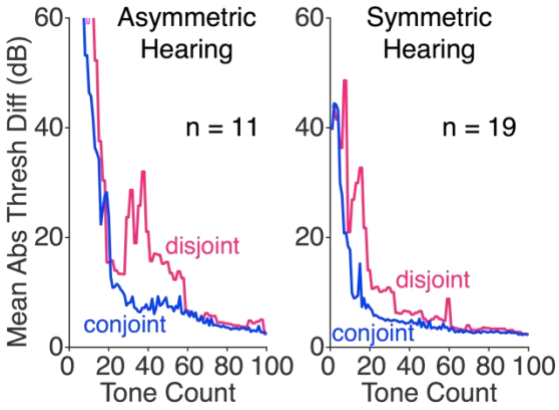

Figure 3. 5: Average absolute subpopulation difference in threshold functions between the final estimate at 100 tones and estimates with each additional tone. Asymmetric hearing subjects take longer to converge, on average. However, the conjoint estimator outperforms disjoint estimation. Tone counts are for both ears combined.

Similar analysis when the population is separated into normal-hearing (n=8) and hearing loss (n = 22) ears also reveals that conjoint estimation requires significantly fewer tones than disjoint estimation in both cases (normal: $p = 2.9\times10_{-6}$; hearing loss: $p = 1.7\times10_{-11}$; paired-sample t test) (**Figure 3.6**).

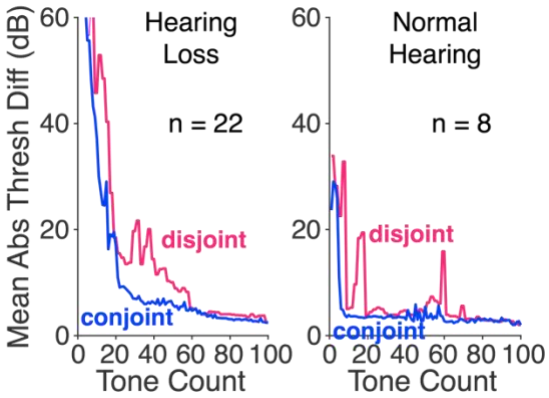

Figure 3. 6: Average absolute subpopulation difference in threshold functions between the final estimate at 100 estimates with each additional tone. Conjoint estimation converges for subjects with normal hearing and with hearing loss. Tone counts are for both ears combined.

Further, ears were arbitrarily paired from different heads across six additional participants, giving each participant one insert earphone and one response button with instructions to respond whenever they heard a tone. With only 70% of the tones, the conjoint threshold estimates for different-head ear pairs matched the disjoint threshold estimates of those ears similarly to the results observed in same-head ear pairs (**Figure 3.7**).

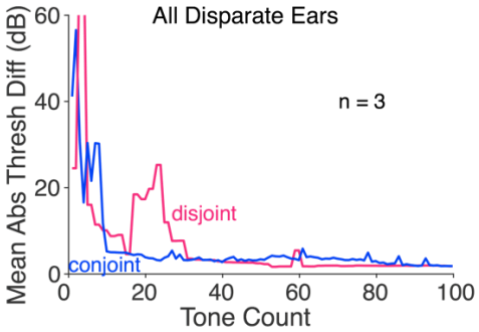

Figure 3. 7 Average absolute subpopulation difference in threshold functions between the final estimate at 100 tones and estimates with each incremental tone. In this case, each ear of the ear pair belonged to different participants ("disparate"). Conjoint estimation converged to similar estimates as disjoint estimation in these cases where the individual domains were completely independent. Shared variation between these ears results from the laws of physics and human biology.

3.4 Discussion

Bilateral audiometry differs from conventional audiometry by considering both ears simultaneously in real time as the test is being conducted. Tones are delivered to either ear as directed by the algorithm while the subject is instructed to respond whenever he or she hears a tone. While stimuli are delivered to each ear independently, inference is drawn for both ears simultaneously. Strong or weak concordance between the hearing functions of an individual's two ears may exist depending on multiple factors. Because bilateral audiometry learns the shared variation between ears for each subject, it uses this information to speed the test, even under discordant conditions (c.f., Figure 3.5). Simulations indicate that bilateral audiometry should

achieve accuracy consistent with serial unilateral audiometry with only about 60% of the tone count (Barbour, DiLorenzo, et al., 2019). The current results in humans show similar gains.

Conjoint AMLAG convergence is nearly an order of magnitude faster than traditional HWAG testing time (Song et al., 2015). Additionally, optimization for traditional methods is limited due to the constraint of testing one ear and one frequency at a time. Subject anticipation of tone delivery in such cases due to rhythmic testing patterns can lead to false positives. “Roving ear” tone presentations in the bilateral audiogram did not lead to systematically different threshold estimates than the “fixed ear” tone presentations in the unilateral audiogram (mean signed difference of conjoint minus disjoint thresholds was 1.1 dB).

Often, collecting audiograms can be time consuming and tiring for the patient, which can lead to erroneously missed tones from nonstationarities such as attention lapses or criterion drift. Traditional audiometry testing time is further extended when patients have complicated hearing loss in one or both ears. The conjoint estimator accurately estimates hearing loss generally and asymmetric hearing loss specifically using few tones, allowing less time for nonstationarities to affect estimates.

It is important to note that similarity between the ears is not required for an effective bilateral audiogram procedure. If a conjoining hyperparameter value of 0 best accounts for a subject’s data, the result will mathematically be two unilateral audiograms. Even under these conditions, however, conjoint estimation could be faster than disjoint estimation over finite input domains because the learned dissimilarity may be useful to induce more appropriate sampling.

The vast majority of clinical and experimental tests begin by formally ignoring previous knowledge and proceeding to collect data with no prior assumptions. This situation exists either because incorporating priors into existing tests is impossible due to test construction, philosophically undesirable for fear of bias, or both. The audiogram is no different. As a result, considerable audiometric information is discarded that could be used to improve audiogram test accuracy and speed.

The large amount of paired and unpaired ear data in the NIOSH database indicate that information exists from the contralateral ear that could be quite useful for incorporating into measures of the ipsilateral ear. HWAG provides a limited mechanism to do so, as the only real flexibility in the test available to the clinician or experimenter is the starting sound level. Machine learning audiometry, on the other hand, can exploit prior information by design. Active GP estimation is able to determine correlations between variables in real time as data are accumulated. While a person's two ears are not themselves physiologically linked, they do share many things in common. GP inference therefore represents an excellent method for exploiting these correlations for improving test accuracy and efficiency.

3.5 Concluding Remarks

Bilateral audiometry has demonstrated the value of the conjoint estimation approach for improving hearing threshold estimation efficiency. A multidimensional, bilateral test was able to model hearing ability more efficiently than unilateral testing, without sacrificing accuracy, for all participants in this study, regardless of hearing ability. Because any kernelized psychometric function model can be conjoined to any other kernelized model with this formulation, potential benefits can be extended well beyond hearing.

Chapter 4: Dynamically Masked Audiograms

Note: The research presented in this chapter has been accepted for publication in *Ear and Hearing*.

4.1 Introduction

In most audiology practices today a clinician manually obtains pure-tone hearing thresholds following a procedure that was recommended as the standard for audiometric testing 60 years ago (Carhart & Jerger, 1959; Hughson & Westlake, 1944): HWAG (for details, see Chapter 2). This adaptive up-down staircase method continues to be emphasized in the most recent clinical guidelines (American Speech-Language-Hearing Association, 2005) and is valued for being fast and reliable for many patients. A particular case where manual HWAG is inadequate, however, is for individuals with asymmetric hearing where cross hearing is likely to occur (see Chapter 2). While only a subset of the general population, for them contralateral masking is an essential component of hearing assessment, aiding in differential diagnosis and hearing loss management decisions. Unfortunately, masking is a time-consuming process and is often cited as one of the most challenging procedures for audiologists to learn (Gumus, Gumus, Unsal, Yuksel, & Gunduz, 2016; Hamil, 2016; Ho et al., 2009; Sanders & Rintelmann, 1964; Valente, 2009; Yacullo, 2015). No universally accepted masking standard or guideline exists. In the most recent surveys of audiologic practices conducted by the American Academy of Audiology (Martin, Armstrong, & Champlin, 1994; Martin, Champlin, & Chambers, 1998), researchers noted that audiologists were using a broad range of masking methods and further determined that over half of the respondents were using inappropriate masking procedures.

AMLAG has been shown to be as accurate as and more efficient than manual HWAG methods for normal and hearing loss populations (Barbour, Howard, et al., 2019; Song et al., 2015). In Chapter 3, bilateral AMLAG successfully learned the hearing thresholds of left and right ears in one test regardless of hearing ability. No test adaptations were necessary to efficiently estimate hearing thresholds for all participants in the study. Participants with diverse hearing abilities were recruited for that study; however, care was taken to ensure that no participant was at risk for cross hearing. To truly develop useful individualized measures of perception, assessments must be able to accommodate not just those who are outside of the 'normal' range, but those for whom traditional testing protocols are insufficient with currents methods.

To that end, a dynamic masking protocol has been integrated into AMLAG to create the masked AMLAG. Dynamic masking adds additional complexity to the existing GP framework and demonstrates the utility in leveraging an algorithm that adapts to each participant in real time. Unlike masking during manual audiometry, masked AMLAG presents suitable masking noise to the non-test ear throughout the entire audiogram test procedure. Every tone presented to the test ear is paired with masking noise in the non-test ear. Masking noise levels are derived from a combination of the interaural attenuation estimate and the intensity of the test ear tone. Every audiogram becomes a masked audiogram, and accurate thresholds are estimated directly because cross hearing is dynamically eliminated. AMLAG so rapidly homes in on hearing thresholds (Heisey et al., 2018) that individuals with fairly symmetric hearing should almost never be presented a suprathreshold masking noise, making masked and unmasked AMLAG procedurally equivalent for this large population. The work presented here shows that an active machine

learning framework provides multidimensional, individualized assessments without sacrificing accuracy, efficiency, or complexity of the behavior being modeled.

4.2 Methods

4.2.1 Participants

This study was approved by the Human Research Protection Office at Washington University School of Medicine. A total of 29 participants (20 females, 9 males) were recruited using the Research Participant Registry at Washington University in St. Louis. Participants were required to be at least 18 years of age and proficient English speakers. The 28 participants who reported their age were between 21 and 83 years of age (mean 43, SD 20). Informed consent and a voluntary demographic form were obtained from each individual prior to beginning the study. Two participant's right ears were excluded from analysis due to a temporary equipment malfunction.

4.2.2 Equipment

All testing was performed in a sound-treated booth. The unmasked and masked AMLAG tests were administered using a Dell XPS laptop computer. Tones were delivered through TDH-50P Telephonic supra-aural headphones connected to an AudioQuest Dragonfly Red USB digital-to-analog converter. An external mouse was connected through a USB port and functioned as the response button. Manual HWAG was performed by a student audiologist using a Grason Stadler GSI AudioStar Pro two-channel clinical audiometer. Thresholds were obtained using TDH-50P Telephonic supra-aural headphones, a bone oscillator, and a response button. The computer audio output was calibrated to match the output of the audiometer.

4.2.3 Experimental Procedure

The 29 participants were split into two experimental groups according to their reported hearing ability. The first group consisted of nine participants with self-reported normal hearing, designated as No Loss (NL). The remaining 20 participants reported some degree of hearing deficit and were designated as Hearing Loss (HL). The cohort with hearing loss exhibited a variety of etiologies based upon the relationships between their air-conduction and bone-conduction audiograms, including sensorineural loss, conductive loss and mixed losses (see Chapter 2 for details concerning hearing loss categorization).

NL participants completed a left and right unmasked AMLAG and a left and right masked AMLAG. HL participants were first given a manual left and right HWAG with appropriate masking protocol, if needed, to determine their hearing loss profiles. Then they were given left and right masked AMLAGs.

Unmasked and Masked AMLAG Protocol

Unmasked and masked AMLAG tests were implemented directly on the computer using custom Matlab code. The unmasked AMLAG procedure has previously been described in detail (see Chapter 2 and Song et al., 2015). Three-pulse sequences of 200 ms pure tones, with frequencies in semitone increments between 250 and 8000 Hz and sound levels from –20 to 100 dB HL, were presented with interpulse intervals of 200 ms. Inter-sequence intervals were randomized and ranged from 0.5 to 3 seconds in order to prevent predictability.

Participants were instructed to click the left mouse button whenever they heard a tone, even if it was very soft. They were informed that the frequency, or pitch, would change between each tone sequence and that there could be relatively long periods of silence. Participants were instructed

to ignore any wind or white noise they heard and were reminded to only click the mouse when they heard a pure tone. All participants were asked after each ear's test if they had heard any wind or white noise. Their responses to this question were recorded.

Any normally worn hearing devices were removed, and headphones were placed after instructions were given. Participants were seated so that they could not see the computer screen, and the order of the ears tested was randomized by the experimenter. Each AMLAG test consisted of a total of 100 tone sequences per ear and began with seven tones randomly selected from the median threshold values for normal hearing: 10 dB HL at 500 Hz, 5 dB HL at 1000 Hz, 10 dB HL at 2000 Hz, 10 dB HL at 3000 Hz, 15 dB HL at 4000 Hz, 15 dB HL at 6000 Hz, and 15 dB HL at 8000 Hz. Median normal hearing thresholds were obtained from a dataset of 1.1 million individuals developed by the NIOSH Occupational Hearing Loss Surveillance Project, Division of Surveillance, Hazard Evaluations and Field Studies (Masterson et al., 2013). If none of the seven population median threshold tones were heard, the algorithm employed Halton sampling until a heard tone response was recorded. Halton sampling ensures broad sampling across all frequencies and intensities (Song, Garnett, & Barbour, 2017). Following the first heard tone response, active sampling was initiated and the remaining tones were queried according to Bayesian active learning by disagreement (Song et al., 2017). For ears where no heard tone was ever indicated, all of the remaining tones were ultimately selected by Halton sampling.

Masked AMLAG presented 1/3 octave narrowband noise to the contralateral non-test ear while simultaneously presenting a three-pulse sequence of tones to the test ear. This procedure was performed for every tone presentation, even if a participant would not typically require masking. Masking noise began randomly in the 250 – 1500 ms interval before the onset of the pure-tone

sequence and remained on for a total of 3.0 – 5.5 sec. The noise ramped on for 100 ms at the beginning of the intersequence interval and ramped off during the final 100 ms. All masking noise presentations began seamlessly at the conclusion of the preceding noise presentation, centered at the frequency of the test-ear tone and presented at 40 dB below the tone's presentation level. This masking presentation level is based on a conservative interaural attenuation level of 40 dB for supra-aural headphones (Yacullo, 2015).

Manual HWAG Protocol

HWAGs were conducted manually by a student audiologist. During manual HWAG, participants heard pulsed pure tones through headphones and were instructed to press a button whenever they heard a tone, even if it was very soft. Air conduction thresholds were obtained for each ear at the standard octave frequencies (250, 500, 1000, 2000, 4000, and 8000 Hz) using the modified HW procedure. Bone conduction thresholds were obtained at 250, 500, 1000, 2000, and 4000 Hz using the same protocol as air conduction thresholds described above.

Masking for air conduction was performed when the air conduction threshold of the test ear was worse than the bone or air conduction threshold of the non-test ear by greater than or equal to 40 dB. To ensure the non-test ear was not responding to the tone, narrowband noise was presented at a suprathreshold level. Specifically, 10 dB was added to the air conduction threshold of the non-test ear and presented as narrowband noise. The true air conduction threshold of the test ear was then found using the plateau method (Hood, 1960; Martin et al., 1998; Yacullo, 2015). A true threshold was determined when a participant responded to a tone after the noise was raised by 5 dB three times. In other words, when the participant heard the tone even after the noise was increased by a total of 15 dB.

Masking for bone conduction was performed when there was a difference of greater than or equal to 15 dB between the air and bone conduction thresholds of the test ear. In addition to the bone oscillator, a supra-aural headphone was placed such that it covered the non-test ear but the test-ear remained unobstructed. Similar to masking for air conduction, 10 dB was added to the air conduction threshold of the non-test ear and presented as narrowband noise. The occlusion effect must be considered when testing masked bone conduction at 250, 500, and 1000 Hz, however (Edgerton & Klood, 1977; Valente, 2009). To compensate for the occlusion effect, an additional 20 dB of narrowband noise was added to the initial masking level at 250 Hz. An additional 15 dB of noise was added at 500 Hz and 10 dB was added at 1000 Hz. The true bone conduction threshold of the test ear was then found using the plateau method.

Extended details on the masking procedure and other experimental details can be found at https://osf.io/64qd7/

4.2.4 Data Analysis

AMLAG returns a continuous estimate of the probability of hearing any frequency-intensity pair in the stimulus domain. Hearing thresholds at octave frequencies were determined at the 0.707 detection probability to match the standard probability of detection for HWAG estimates. Any threshold estimate that was greater than 100 dB HL was designated as a “no response” at that frequency.

The unmasked and masked AMLAG thresholds were compared at the standard audiogram frequencies for Group NL. Efficiency and accuracy of the masked AMLAG were assessed via comparison to the unmasked AMLAG. Individual ears were evaluated by comparing the mean signed difference, mean absolute difference and root mean square difference between unmasked

and masked AMLAG. To assess the efficiency of masked AMLAG, the mean tone counts and testing times required for left and right ear threshold estimation were determined and compared to unmasked AMLAG. The effects of dynamic contralateral masking on a participant's test experience were determined by calculating the number, percentage and maximum sound level of masking noise presentations delivered above the non-test ear threshold, as well as post-test interviews.

All Group HL analysis compared masked AMLAG and manual HWAG thresholds at the standard audiogram frequencies. Accuracy and efficiency of masked AMLAG was assessed via comparison to manual HWAG. Analysis of Group HL was identical to that of Group NL but compared masked AMLAG threshold estimates to manual HWAG estimates.

To better analyze the effects of dynamic masking, Group HL analysis was subdivided according to masking needs. Eight participants with highly asymmetric hearing loss between the two ears required masking by conventional guidelines (see Chapter 2) and were separated into subgroup HL-HA. The 12 other Group HL participants had a low asymmetric hearing loss not requiring masking and were separated into subgroup HL-LA. This subdivision enabled determination of the impact of dynamic masking on audiogram acquisition for participants who would not otherwise require masking. It further allowed the analysis of dynamic masking effects for the participant subgroup that would benefit most from a more effective and standardized masking implementation.

4.3 Results

4.3.1 Group NL Analysis

Masked AMLAG thresholds estimated in Group NL listeners were consistent with thresholds estimated by unmasked AMLAG, which has been previously validated as equivalent in accuracy to HWAG (Barbour, Howard, et al., 2019; Heisey et al., 2018; Song et al., 2015). The similarity of unmasked AMLAG and masked AMLAG threshold estimates at the standard audiogram frequencies across all tests within Group NL is depicted in Bland-Altman plots in **Figure 4.1** (Bland & Altman, 1999). Differences do not appear to be a function of threshold magnitude. Means and 90% limits of agreement (1.645 × standard deviations) are depicted. Mean signed differences are close to 0, as would be expected if the two tests were evaluating the same underlying physiological process.

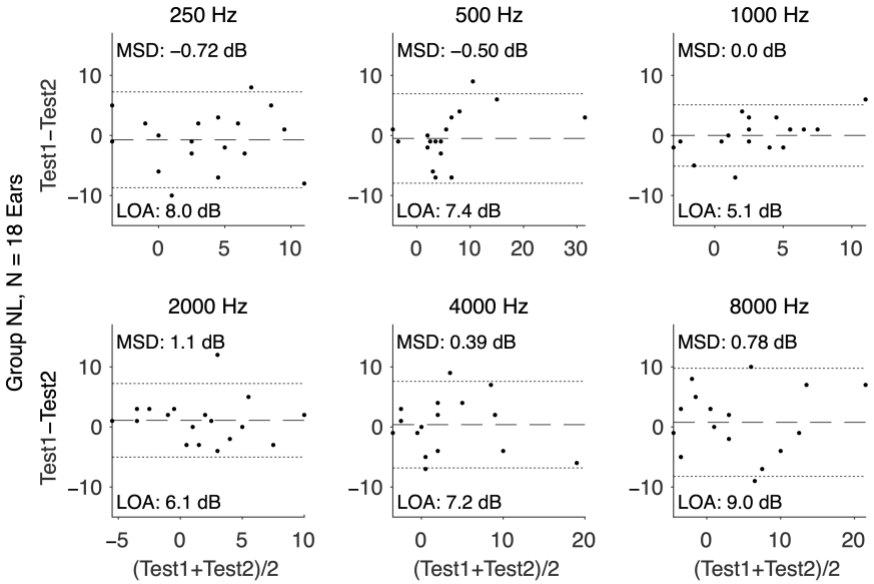

Figure 4. 1: Bland-Altman plots at the 6 frequencies of threshold comparison for unmasked AMLAG ("Test1") versus masked AMLAG ("Test2") in Group NL. Mean signed difference (MSD) in dB is indicated numerically and by a horizontal dashed line in each plot. Limit of agreement (LOA) in dB is indicated numerically and by 2 horizontal dotted lines in each plot. LOA is computed as 1.645 × the standard deviation of the signed differences, reflecting the central 90% of the estimated distribution.

Additional numerical summaries are given for Group NL in the Chapter 4 Supplemental at https://osf.io/64qd7/. Previously published studies have demonstrated that variability in pure-tone manual HWAG thresholds obtained with supra-aural transducers in the age range studied here are considered clinically relevant only when exceeding 10 dB, and mean deviations within 5 dB are commonly cited as clinically acceptable (Landry & Green, 1999; Mello, Silva, Gil, & Ram, 2015; Stuart, Stenstromb, Tompkins, & Vandenhoff, 1991). The mean absolute difference between masked and unmasked AMLAG was under 5 dB at all frequencies with an overall mean of 3.4 ± 2.7 dB. Collectively, these results indicate that masked AMLAG yields threshold estimates comparable in value to unmasked AMLAG in normal hearing individuals.

Table 4. 1: Average number of tones and minutes required to achieve threshold estimates for each participant, Group NL (N = 9 participants

	Mean ± SD Tone Count	Mean ± SD Number of Minutes
Group NL participants		
Unmasked AMLAG	37 ± 15	4.0 ± 1.6
Masked AMLAG	34 ± 12	3.7 ± 1.3

All AMLAG tests in this study were designed to deliver 100 tone presentations per ear in order to ensure confident final threshold estimates. Previous research has demonstrated that unmasked AMLAG often converges to a threshold estimate within 5 dB of the final threshold estimate in considerably fewer than 100 tone presentations per ear (Heisey et al., 2018; Song et al., 2015). For each participant in Group NL, the total number of tone presentations and average time for unmasked and masked AMLAG to converge to a threshold estimate within 5 dB of the final

estimation in both ears were calculated (**Table 4.1**). **Figure 4.2** shows the mean absolute difference between the threshold estimate at each tone presentation and the final estimate after 100 tones averaged across all Group NL ears. It demonstrates a very similar convergence profile for both tests in this group.

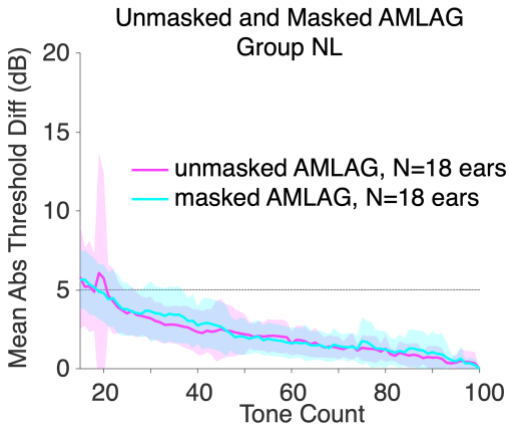

Figure 4. 2: Average ± standard deviation absolute difference in threshold estimates between the final estimate at 100 tones and estimates with each incremental tone presentation for unmasked and masked AMLAG (Group NL). Values are for each ear.

At each tone presentation, masked AMLAG presented narrow band noise in the ear contralateral to the ear being tested. Because AMLAG so rapidly identifies putative thresholds and spends most of its sampling effort at nearby intensities, the paired masking noise level was almost always subthreshold and therefore expected to be undetectable by the non-test ear (**Table 4.2 and Figure 4.3**). To determine if dynamic masking subjectively altered the test experience, participants were asked following each AMLAG test (unmasked and masked) if they had heard any white noise and if so, in which ear they had heard it. Of the 36 automated audiogram assessments for Group NL, five tests were identified by participants as having presented detectable white noise in the non-test ear. Three of those five were actually unmasked AMLAG

tests with no noise delivery at all, and the reported perception most likely was due to occlusion effects. The two masked AMLAG tests during which the participants noted hearing white noise were both tests in which suprathreshold masking levels were presented to the non-test ear. The participants commented that the masking noise was not distracting and described the noise as "soft."

Table 4. 2: Masking noise above non-test ear threshold, Group NL (N = 18 ears)

Total number of masks	1800
Masks above non-test ear threshold	3
Percent of masks above non-test ear threshold	0.17
Maximum level above non-test ear threshold (dB)	12.0

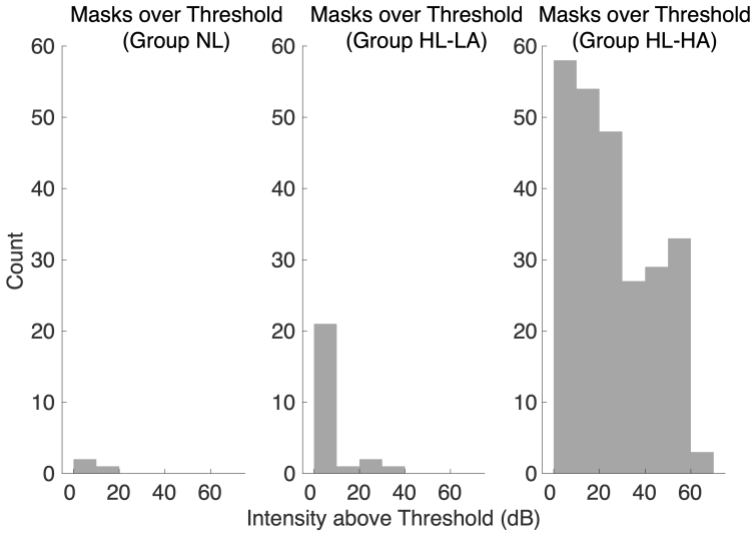

Figure 4. 3: Intensities of masking noise delivered over non-test ear threshold for all three experimental groups.

4.3.2 Group HL Analysis

The accuracy of masked AMLAG was evaluated at standard audiogram frequencies relative to manual HWAG and averaged across all tests for Group HL. The similarity of masked AMLAG and HWAG threshold estimates at the standard audiogram frequencies across all tests is depicted in Bland-Altman plots in **Figure 4.4** for group HL-LA and **Figure 4.5** for Group HL-HA. Means and 90% limits of agreement are again depicted. Differences generally do not appear to be a function of threshold magnitude, though the variability in differences appears to be higher with higher thresholds for 4 kHz, Group HL-LA. Given that this trend was not found at adjacent frequencies or for 4 kHz in other groups, it seems likely to reflect participant sampling. The large outlier at the highest threshold for 1 kHz, Group HL-LA, may be attributable to this participant's self-reported tinnitus, and is a scenario worthy of further investigation.

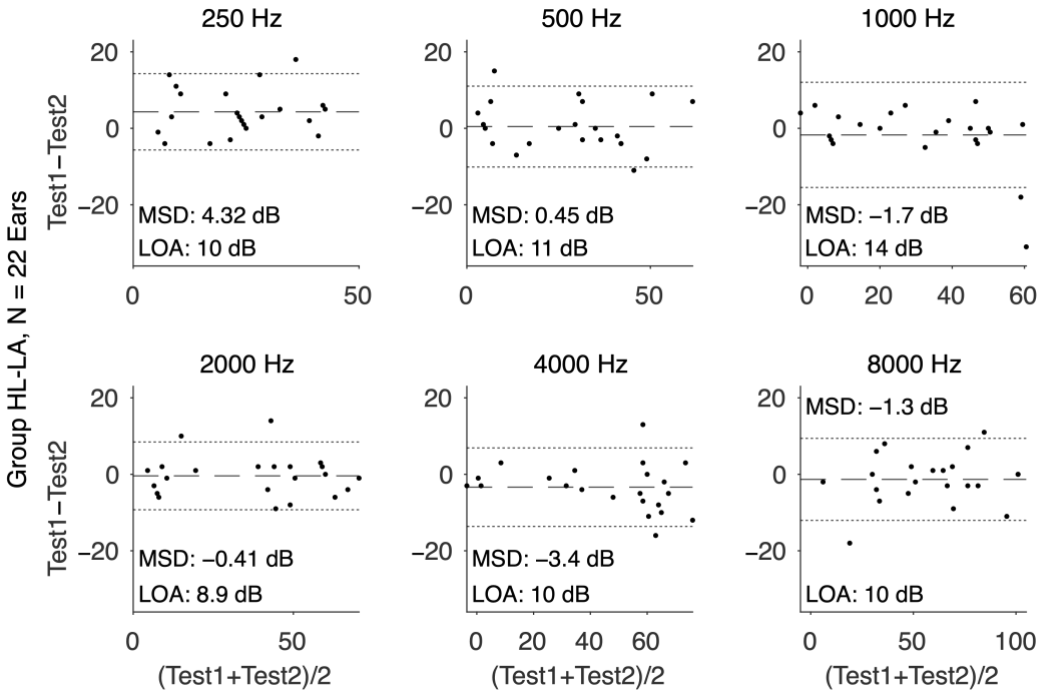

Figure 4. 4: Bland-Altman plots at the 6 frequencies of threshold comparison for HWAG ("Test1") versus masked AMLAG ("Test2") in Group HL-LA. Plot details are identical to Figure 1.

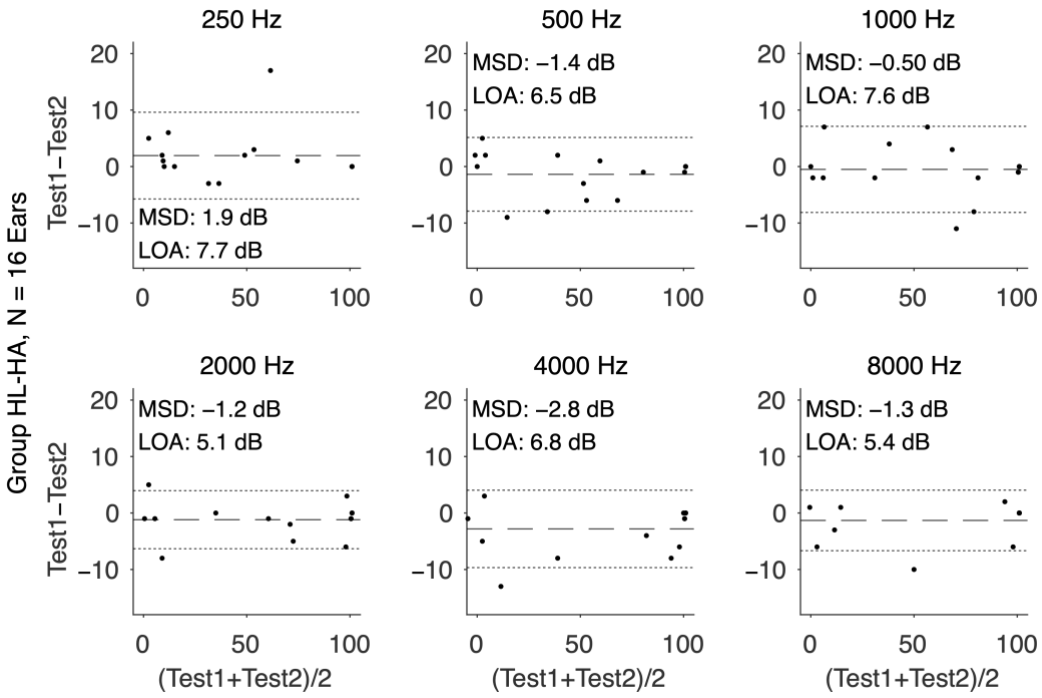

Figure 4. 5: Bland-Altman plots at the 6 frequencies of threshold comparison for HWAG ("Test1") versus masked AMLAG ("Test2") in Group HL-HA. Plot details are identical to Figure 4.1.

Group HL-LA and Group HL-HA numerical summaries are given at https://osf.io/64qd7/. Once again, mean signed differences near 0 imply that one test is not biased in its threshold estimates relative to the other. The small mean absolute differences between masked AMLAG and HWAG convey that the tests consistently deliver similar estimates. Group HL-LA had a mean absolute difference of 4.9 ± 4.5 dB and Group HL-HA had a 2.6 ± 3.1 dB difference. These results are within the published variability of 5-10 dB shown between traditional and other automated audiometry assessments (Shojaeemend & Ayatollahi, 2018).

In addition to estimating accurate pure-tone thresholds, masked AMLAG was able to generate these thresholds with significantly fewer tone presentations ($p = 3.92 \times 10^{-3}$ for Group HL-LA, $p = 2.95 \times 10^{-4}$ for Group HL-HA, paired t-tests) and significantly more quickly than manual

HWAG ($p = 5.48 \times 10^{-2}$ for Group HL-LA, $p = 5.66 \times 10^{-4}$ for Group HL-HA, paired t-tests). The efficiency of masked AMLAG was evaluated through a comparison of the average number of tone presentations and the average test time required to estimate thresholds within 5 dB of the final threshold estimates relative to manual HWAG's final threshold determinations. Because left and right threshold estimates are necessary to determine masking needs for HWAG, ears were analyzed as left and right pairs, giving overall results for each participant. The two Group HL-LA participants with a single excluded ear were removed from this analysis. Overall results are shown in **Table 4.3**. Masked AMLAG estimated thresholds for both ears with, on average, 64 fewer tones per Group HL-LA participant and 136 fewer tones per Group HL-HA participant. For hearing losses where no masking was required during manual HWAG (Group HL-LA), the average masked AMLAG test time to estimate both ears for a single participant was 3.8 minutes faster than the average manual HWAG time. For hearing losses requiring masking during manual HWAG (Group HL-HA), the difference was much greater, with masked AMLAG estimating thresholds an average of 13.1 minutes faster than manual HWAG. Clinically, bone conduction is needed to determine a participant's masking needs in the presence of an air-bone gap. Accordingly, both air conduction and bone conduction tone counts were included in the total manual HWAG convergence analysis. No Group HL participants presented an air-bone gap that required additional air conduction masking. Therefore, the mean number of tone presentations and minutes required for all HWAGs with bone-conduction assessment removed from analysis are also summarized in Table 4.3.

Table 4. 3: Average number of tones and minutes required to achieve threshold estimates for each participant, Group HL

Group	Test	N	Mean ± SD Tone Count	Mean ± SD Minutes	Mean ± SD Tone Count: Air Conduction Only	Mean ± SD Minutes: Air Conduction Only
HL-LA	Manual HWAG	10	127 ± 36	10.7 ± 4.2	93 ± 17	6.9 ± 1.7
HL-LA	Masked AMLAG	10	63 ± 31	6.9 ± 3.3	63 ± 31	6.9 ± 3.3
HL-HA	Manual HWAG	8	186 ± 61	18.5 ± 6.6	114 ± 32	9.9 ± 3.4
HL-HA	Masked AMLAG	8	50 ± 16	5.4 ± 1.7	50 ± 16	5.4 ± 1.7

Similar to the analysis for Group NL, each masking noise presentation was assessed to determine the effect of dynamic masking on Group HL tests. Group HL-LA participants did not clinically require masking, and it was anticipated that much like Group NL, most masking levels would be presented at levels below the non-test ear threshold. On the other hand, Group HL-HA participants did require masking, and it was expected that masking noise would be heard in the test ear at suprathreshold levels. These results are shown in **Table 4.4** and Figure 4.3. After each masked AMLAG test, participants were asked if they had heard any white noise and in which ear it had been heard. Listeners from Group HL-LA noted hearing masking noise in eight out of 22 masked AMLAG tests. Six of the eight were tests in which a fraction of the tones were paired with masking noise levels that would have been above the contralateral ear threshold for the Group HL-LA participants. One participant identified masking noise during left and right masked AMLAG, yet analysis shows that no suprathreshold masking noise was delivered during either test. It is suspected that occlusion effects or tinnitus might account for the perceived noise

heard during both tests. Nevertheless, the noise perception did not appear to interfere with testing procedures or results. Two AMLAG tests in Group HL-LA had tone presentations paired with suprathreshold masking noise delivered to a non-test ear that were not identified by the participant. In these tests, masking noise levels may have been infrequent or quiet enough to be unremarkable. All eight Group HL-HA participants reported hearing masking noise in their better-hearing ear during the worse-hearing test ear assessment. No masking noise was discerned in the worse-hearing ear. No participant in Group HL reported the onset of the masking noise to be distracting or to inhibit their ability to perform the task.

Table 4. 4: Masking noise above threshold of the non-test ear, Group HL (N = 38 ears)

	HL-LA	HL-HA
Total number of masks	2200	1600
Masks above non-test ear threshold	25	252
Percentage of masks above non-test ear threshold	1.14	15.8
Maximum level above non-test ear threshold (dB)	31.5	70.0

Figure 4.6 shows the mean absolute difference between the final threshold estimate at 100 tone presentations and each increment iteration of masked AMLAG averaged across all of Group HL, Group HL-LA, and Group HL-HA participants. Test results converged faster for individuals with normal hearing, most likely because the initial seven fixed frequency/intensity combinations were particularly informative for this group and enabled active learning to select tone queries that rapidly reduced errors. Highly asymmetric hearing thresholds can also be estimated relatively rapidly, presumably for the complementary reason that extremely high thresholds near or beyond the maximum stimulus can also be identified relatively quickly in an active testing

scenario. It is not surprising given this consideration that individuals with thresholds in both ears near the middle of the testing range would require the most test tones to achieve comparable accuracy. Incidentally, these individuals are exactly the patient population for whom bilateral audiometry can most speed up testing (Heisey et al., 2018).

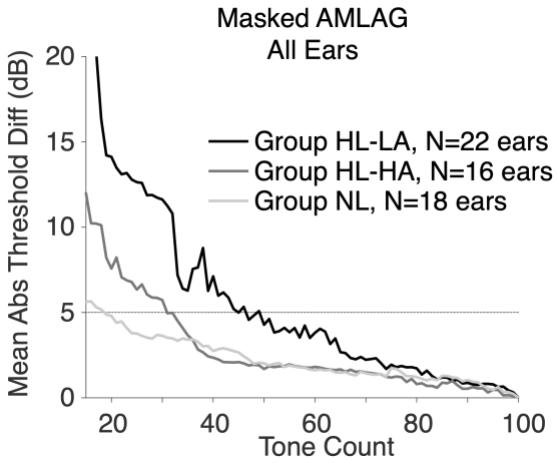

Figure 4. 6: Average absolute threshold differences (dB) between the final estimate at 100 tones and estimates at each incremental tone presentation for Group NL, Group HL-LA, and Group HL-HA masked AMLAG.

Figure 4.7 visually depicts the thresholds estimated for all ears with all air conduction tests for this study. Participants are sorted by group (NL, HL-LA and HL-HA), and within each group by pure tone average of the better hearing ear. This visualization demonstrates the variety of hearing profiles for the participants in this study, as well as the agreement between testing procedures. Most agreement is high, with occasional disparities at individual frequencies. Asymmetry alone is not associated with the disparities because Group HL-LA exhibited the least overall agreement between threshold estimates and not Group HL-HA.

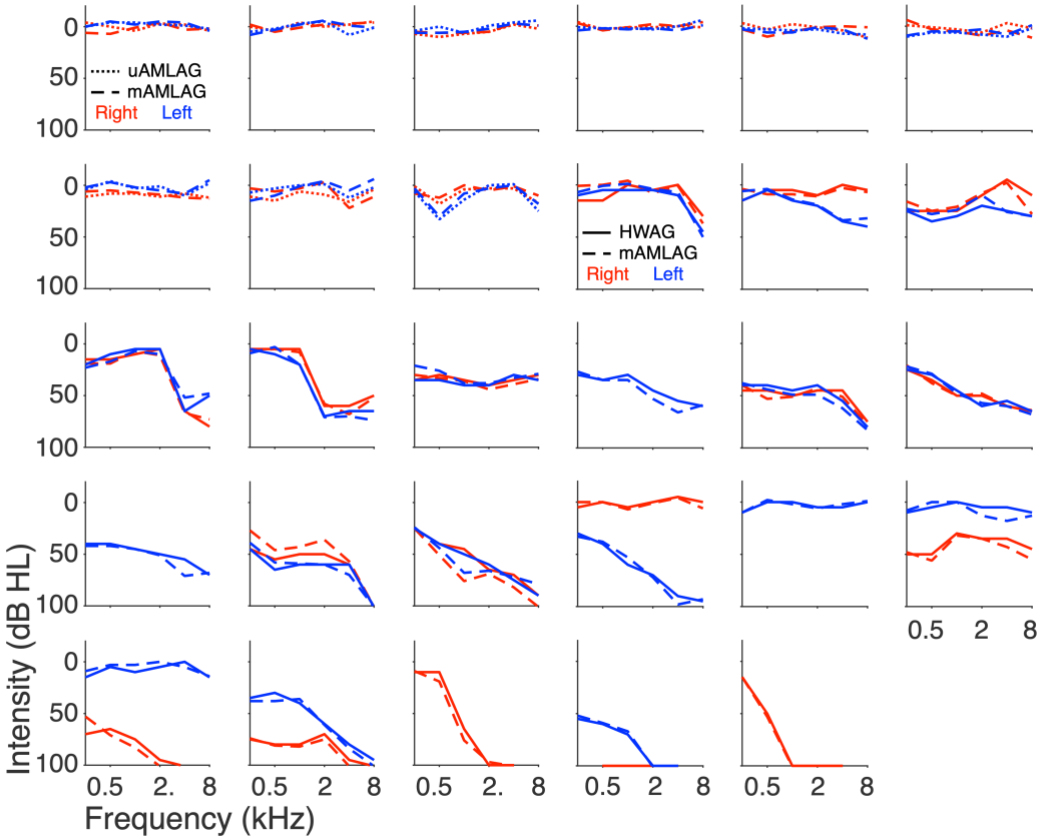

Figure 4. 7: Air conduction threshold audiograms derived from the HWAG and AMLAG procedures for every participant in this study, sorted first by group and then by pure-tone average of the better ear. Dotted lines indicate unmasked AMLAG, dashed lines indicate masked AMLAG, and solid lines indicate HWAG. Red lines denote right ears and blue lines denote left ears.

4.4 Discussion

Masked AMLAG demonstrated similar accuracy and improved efficiency when compared to unmasked AMLAG and manual HWAG. These results were observed in normal hearing, symmetric loss, and asymmetric loss participants. This finding is particularly important as it indicates that masked machine learning audiometry delivers accurate true threshold estimates even for patients with highly asymmetric hearing where substantial masking is required. Exploiting the relationships between interaural attenuation, intensity and frequency, dynamically

masked AMLAG achieves its test time reduction for patients with asymmetric hearing by eliminating the need for a separate masking step. Notably, adding contralateral masking to every tone does not significantly increase test time for listeners with normal or symmetric hearing. For most of these participants, masking levels remained below hearing thresholds and were undetected throughout the test. A dynamically masked audiogram therefore allows for individual differences in masking needs to be addressed in real time without increasing test time.

It is important to consider that masked AMLAG was set to deliver 100 tone presentations per ear even if it was confident in the estimated thresholds at earlier tone counts in order to ensure the acquisition of complete audiogram models. Therefore, tone counts and test times were calculated at the point when masked AMLAG's estimation fell within 5 dB of its final estimation. Test stopping criteria, such as were used previously (Song et al., 2015), are the subject of ongoing research. A notable difference between HWAG and AMLAG is that the former must reach the end of its testing procedure before a complete threshold estimate is available, while the latter delivers a complete estimate for any length of test, though it converges closer to a more accurate model as more tones are delivered (Figure 4.1). AMLAG is therefore very flexible in its test length and can deliver useful results even in extremely short testing scenarios, such as with pediatric patients.

Manual HWAG test time and tone counts included the collection of both air conduction, bone conduction, and any masked thresholds. This procedure likely increased both measures significantly. This comparison is reasonable, however, because the manual masking protocol used in this study requires bone conduction thresholds to determine if air conduction masking was needed due to an air-bone gap. None of the HL study participants had an air-bone gap

requiring additional air conduction masking, so masked AMLAG has yet to be tested under those conditions. While masked AMLAG is currently limited to testing air conduction, it dynamically masks all tone presentations and, therefore, does not require bone conduction thresholds to effectually mask air conduction thresholds, making it more efficient than manual HWAG. To evaluate a more direct comparison, however, Table 4.3 is appended with tallied manual HWAG tone counts for this study population that only include air-conduction threshold and masking presentations, thereby excluding bone conduction counts from analysis. Excluding bone conduction tone counts highlights that masked AMLAG is already more efficient than manual HWAG without masking and substantially outperforms manual HWAG when masking is required. For Group HL-LA, masked AMLAG estimated air conduction thresholds with fewer tone counts but in the same number of minutes as manual HWAG. For these participants, all of whom did not require contralateral air conduction masking, manual HWAG benefited from the proficiency and adaptability of an individual clinically trained to perform audiograms. The current implementation of masked AMLAG has a static response window of 1.5 seconds, regardless of when the participant responded to the tone. Future implementations could, for example, commence the inter-sequence wait time immediately after recording a heard response to more closely mimic the actions of skilled audiologists.

Three Group HL-HA participants had unilateral cochlear implants with no residual hearing in the implanted ear. For these participants, masked AMLAG for the implanted ear executed only Halton sampling because no heard tone was ever detected in that ear. **Figure 4.8** shows the final left and right ear thresholds and each tone presented for one unilaterally deaf participant. While the tones presented in the better-hearing ear are almost all focused near the threshold estimate,

the ‘dead’ ear samples canvas the entire frequency/intensity domain. Because no tone was heard and there is no initial threshold estimate, masked AMLAG is declared to be converged to “no response” at every frequency if there was no threshold estimate after 15 tones. This value accounts for a reasonable number of tones to adequately sample the domain and deduce a complete lack of hearing. In an eventual clinical version of AMLAG, Halton sampling will not be used, and a dead ear would be determinable rapidly by active sampling. The purpose of the extensive sampling in the current study was to determine if dynamic masking ever failed to properly mask a test tone. No examples of such failure were noted in 6200 tone deliveries. Ears with no residual hearing almost always elicit cross-hearing and require extensive masking when tested, as shown in the rightmost histogram of Figure 4.3. Masked AMLAG is able to effectively sample throughout the domain and cancel out all cross tones without requiring any additional procedure.

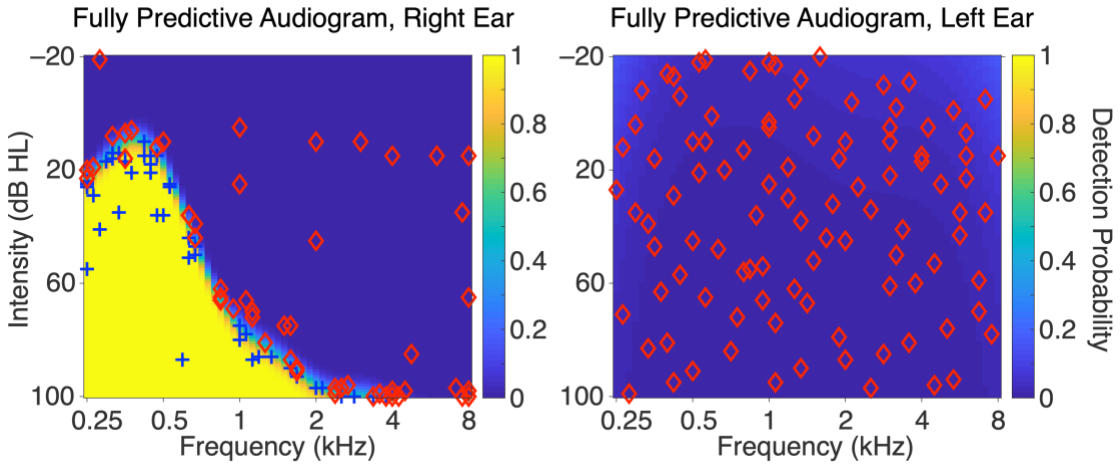

Figure 4. 8: Final masked AMLAG results for one participant (127) with a left cochlear implant and no residual hearing. Red diamonds denote unheard tones and blue pluses denote heard tones.The most intense tones at lower frequencies in the left ear were effectively masked.

Because AMLAG frequency and intensity levels are selected by an active learning algorithm, subsequent masking levels rove across the entire frequency and intensity spectrum. It is possible for long periods without suprathreshold masking to be followed by a tone presentation paired with an audible contralateral masking level. While listener performance has been shown to be unaffected by tones roving between frequency, intensity, and ears (Barbour, Howard, et al., 2019; Heisey et al., 2018; Song et al., 2015), masking is unique in that tones in a test ear are paired with masking noise in a non-test ear. For any test stimulus, both, either, or neither sound might be heard by the listener. The onset of sound, be it tone or noise, requires the listener to discern if a response is appropriate or should be inhibited. The consistent threshold estimation results in all groups demonstrate that masking noise did not disorient listeners or induce false positives. It was anticipated that participants requiring masking would have had more masking protocol exposure as a part of routine manual HWAG assessments, whereas participants to whom masking noise is a novel experience might have struggled to ignore masking noise. Eight of the 22 Group HL-LA tests, however, were presented with audible masking during masked AMLAG. Presumably, these participants had not previously experienced audiograms that included any masking protocol. These unfamiliar listeners successfully completed the assessment and had similar results as those without any audible masking.

Clinically, masked AMLAG offers several potential benefits compared to manual HWAG. As this study showed, masked AMLAG provides an opportunity for the standardization of masking, a challenging procedure with multiple variations that are frequently implemented incorrectly (Gumus et al., 2016; Hamil, 2016; Sanders & Rintelmann, 1964; Valente, 2009; Yacullo, 2015). Uniformity of clinical procedures is imperative in order to reduce inter-clinician variability and

ensure that best practices are being achieved. Automation of these methods would also allow technicians to perform some routine testing, providing audiologists with more time for complex cases and to perform other clinical duties.

Delivering contralateral masking levels fixed relative to the ipsilateral tone yields two potential disadvantages. First, undermasking and overmasking are a theoretical possibility because direct confirmation of the proper range of masking levels is not obtained in each participant. No evidence of either type of masking error was apparent in this cohort's data because thresholds were not systematically biased for either the better or worse ear in the HL-HA population. In particular, extreme asymmetry showed no evidence of systematic bias. It therefore seems unlikely that this simple method of fixed maskers would lead to undermasking or overmasking. Insert earphones with a much larger interaural attenuation would be a method to address this possibility directly.

Second, previous research using AMLAG has included discussion of the unpredictable nature of the constituent tone sequences and corresponding difficulty for malingering patients to thwart the test (Song et al., 2015). The use of consistent relative contralateral noise levels that begin prior to tone delivery reinjects some predictability into the tone sequences for individuals with asymmetric hearing loss who definitely detect the maskers. The ultimate solution in this case may be to allow masker level to vary just as tone frequency and level do and explicitly estimate the frequency-dependent interaural attenuation at every test. This next-generation automated masking audiogram would then no longer rely on rules of thumb adopted from evaluating interaural attention in small numbers of individuals in the distant past. Additional bone conduction data may be required to properly select dynamic masking levels, but total maskers

delivered would decrease over the method presented here, and the unpredictability of the masker presentations would cement a difficult if not impossible test procedure to thwart.

Bone conduction has been presented in this study as the source of a phenomenon that can confound accurate air conduction threshold estimation under some conditions. Bone conduction thresholds are useful in their own right, however, to aid in the differential diagnosis between sensorineural and conductive hearing loss (see Chapter 2). Adding pure-tone bone conduction threshold estimation to AMLAG would represent another advance toward efficient standardization. Given the demonstrated flexibility of AMLAG, doing so would be straightforward. It is possible that new transducer configurations may be needed in this case, though AMLAG may prove able to compensate for hardware limitations with an advanced software implementation.

Unmasked AMLAG includes the ability to estimate the hearing thresholds of both ears simultaneously through bilateral testing (Barbour, DiLorenzo, et al., 2019; Heisey et al., 2018). Adding air conduction masking to this procedure is straightforward. In fact, three normal hearing participants were recruited in this study to demonstrate the feasibility of masked bilateral AMLAG. All three participants were given an unmasked and a masked bilateral AMLAG. The average mean signed difference of 0.10 ± 3.7 dB and average mean absolute difference of 3.0 ± 2.0 dB between unmasked and masked bilateral AMLAG are similar to the differences seen between unmasked and masked unilateral AMLAG (Heisey et al., 2018). Additionally, the mean tone count and time to reach threshold estimates within 5 dB of the final estimate for both ears were 19 ± 25 tones and 2.1 ± 2.7 minutes for unmasked bilateral AMLAG, and 23 ± 20 tones and 2.5 ± 2.2 minutes for masked bilateral AMLAG (details of bilateral masked AMLAG analysis

can be found at https://osf.io/64qd7/). Accurate and efficient bilateral estimation of air conduction thresholds for normal hearing individuals under conditions of dynamic masking suggests the successful extension of masked bilateral AMLAG to participants with symmetric or asymmetric hearing loss.

4.5 Concluding Remarks

The incorporation of automatic dynamic masking into AMLAG demonstrates the versatility of active machine learning diagnostic procedures. AMLAG finds hearing thresholds so rapidly, most patients will never know they are taking a masked test because all the masking noise will fall below their detection thresholds. For the patients with asymmetric hearing, however, masked AMLAG delivers true thresholds much more quickly than conventional techniques and in about the same time as unmasked AMLAG would require to estimate thresholds potentially contaminated with cross hearing. Machine learning audiometry therefore has great potential to enhance patient care by simultaneously standardizing a challenging clinical procedure and optimizing both clinician and patient time. More generally, the work presented here and in Chapter 3 demonstrates that the GP framework uniquely enables complex, multidimensional assessments capable of individual inference with practical amount of data. Of particular import, these methods significantly improve perceptual testing for the most vulnerable of the population for whom standard methods are often the most time consuming and costly.

Chapter 5: Joint Estimation of Speech-in-Noise and Verbal Working Memory

5.1 Introduction

Previous studies have elucidated that certain speech-in-noise measures correlate to verbal working memory tests (Akeroyd, 2008; Daneman & Merikle, 1996; McCoy et al., 2005; S. L. Smith & Pichora-Fuller, 2015). There are a wide variety of assessments frequently used to probe working memory ability and speech comprehension. Studies reporting significant correlations are those that require the active processing, manipulating, and storing of incoming signals during working memory tests and often evaluate speech comprehension in challenging environments or hearing-impaired cohorts (reviews in Akeroyd, 2008; Daneman & Merikle, 1996). It is theorized that extra cognitive effort is needed for successful speech perception in challenging environments, such as the extended storage of incoming signals until sufficient context or an individual's mental lexicon can resolve gaps in understanding due to degraded signals. Yet, the exact contribution of working memory during challenging speech comprehension remains highly debated. Individual factors such as age, pure-tone and speech reception thresholds, and cognitive ability together with the lack of a widely accepted standard for working memory tests has hindered any universally accepted interpretation of how these constructs collaborate in individuals.

To determine the interaction between noisy speech comprehension and verbal working memory, serial test batteries are administered to collect multiple measures in individuals with the aim of correlating distinct behavioral results in a meaningful way. Test batteries take time and often make inefficient use of data collection resources by redundantly querying participants in

overlapping domains. Balancing the demands of data collection with sufficiently powered conclusions, many assessments are designed to attempt to control for all features of the environment that are not being actively measured. However, specific interactions between speech perception and working memory are unique to individuals based on, but not limited to, their life experiences, genetics, cognitive abilities, and neural encoding (Samira Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013; DeCaro et al., 2016; Millman & Mattys, 2017; Peelle, Troiani, Grossman, & Wingfield, 2011; Peelle & Wingfield, 2016). Further, these interactions morph with age as the onset of perceptual and/or cognitive decline reallocates neural resources to compensate for changes in connectivity, brain structure, dedifferentiation, or dopamine levels (Grady, 2013). Age-related shifts are not consistent across individuals since cognitive ability and optimized neural strategies are longitudinal adaptations that do not fluctuate at a constant rate but reflect the ever-changing context of an individual's life. Consequently, even in highly controlled experimental settings, it is impossible to homogenize the sample population, and low-dimensional test batteries show a high degree of inter-subject variability regardless of age (DeCaro et al., 2016; Killion & Niquette, 2000; Plomp & Mimpen, 1979). As a result, incremental and occasionally competing conclusions have complicated any understanding of how speech comprehension and working memory interact (Füllgrabe & Rosen, 2016). Definitively quantifying the interactions between these two measures and how it transforms with age is still incomplete.

Directly measuring verbal working memory and speech-in-noise ability in one test might begin to disentangle the distinctive demands placed on shared resources during complex comprehension tasks. Until now, joint cognitive and perceptual assessments have not been

feasible due to extensive data collection requirements or the inability to efficiently model such a complex domain. Utilizing the GP inference framework, noisy speech comprehension and verbal working memory can be assessed in one comprehensive test. Employing an active machine learning model that can simultaneously measure across multiple domains allows, for the first time, a direct measurement of the intra-individual interactions between speech-in-noise and verbal working memory.

Similar to AMLAG exploiting shared information in audiometry, an active machine learning method was implemented to explore and exploit shared information between noisy speech comprehension and verbal working memory. A joint active machine learning perceptual and cognitive test, or AMLPACT, directly models the interactions between speech comprehension and verbal working memory by estimating a participant's performance over a complex two-dimensional input domain. In this way, AMLPACT estimates behavior on the standalone low dimensional tests as well as the interactions between them.

AMLPACT was utilized to conduct individualized analysis of the interplay between speech-in-noise and verbal working memory and how it adapts with age. While there are many variations of speech-in-noise and working memory assessments, as a first implementation of a joint speech and memory test, AMLPACT models an auditory naming speech-in-noise assessment and a verbal working memory N-back. The N-back assessment was chosen because of its extensive use in functional magnetic resonance imaging, which will be relevant in Chapter 6, and its validation as a robust measure of verbal working memory (Gajewski et al., 2018; Jaeggi et al., 2010). Similarly, the auditory naming test was selected because it pairs well with fMRI data collected in Chapter 6, and its widespread use as a measure of noisy speech comprehension.

In this chapter, AMLPACT was administered to young and older adults to investigate both intra-individual and age-related variations in speech and memory behavior. Individual differences in how verbal working memory engages during speech-in-noise tests may help explain variances in speech comprehension or N-back accuracy that are not delineated by purely low-dimensional measures. Regardless, AMLPACT will demonstrate that high-dimensional, active machine learning methods are an innovative, practical option for combining cognitive and perceptual tests.

5.2 Methods

5.2.1 Participants

Forty-four participants (18 male, 26 female) were recruited for this study. Participant were divided into age-based cohorts of 17 young adults whose ages were between 21 and 30 (mean 25) and 27 older adults whose ages were between 65 to 77 (mean 72). All participants were native English speakers with self-reported normal hearing. Informed consent and a voluntary demographic form were obtained prior to the beginning of the study. Participants were recruited using the Research Participant Registry at Washington University in St. Louis or by referral from an ongoing speech-in-noise fMRI study. This study and the study used to refer participants were both approved by the Human Research Protection Office at Washington University.

One young and one older participant were excluded from all analysis due to a failure to correctly complete the assessments. One participant was outside of the age ranges included in this study and was excluded from all analysis. Additionally, one older adult was recently diagnosed with early stages of Alzheimer’s disease. This participant's data was omitted from all healthy cohort analysis but is presented separately in a discussion of the possible utility of joint tasks to assess

performance in populations experiencing cognitive decline. Consequently, 16 young adults and 24 older adults were included for all healthy analysis.

It is noted that one young participant was a main contributor to the development of the N-back and joint memory and speech perception software. Their extensive task exposure likely impacted their task performance, and their data has been noted in all subsequent analysis.

5.2.2 Procedure Overview

Each participant was administered four to five tests, three different speech-in-noise tests (a standard QuickSIN and one or two auditory naming tests), an N-back working memory test, and the joint AMLPACT. Test order was pseudo-randomized such that the N-back and AMLPACT were not delivered back to back in order to relieve some of the cognitive strain from the most cognitively demanding tasks. Instructions were given before each assessment. Practice trials were provided for the N-back, QuickSIN, and AMLPACT to ensure that participants understood the basic mechanics of each assessment.

Assessments were administered in a sound-treated booth. Auditory stimuli were delivered via circumaural headphones paired with a Dragonfly Red 32-bit DAC (AudioQuest, Irvine, CA) connected to a Dell XPS laptop computer. A headphone splitter allowed the experimenter to monitor the delivery of the auditory stimuli. The N-back working memory test required a mouse button response click, and an external mouse was connected via USB for the entirety of that test. The two auditory naming speech-in-noise assessments and AMLPACT were written in custom Matlab code, the standalone N-back test was written in C#, and QuickSIN was administered with the official QuickSIN CD (Etymotic Research, 2001). The laptop volume was set such that all

stimuli were delivered at or near 70 dB SPL. Before beginning the tests, participants were asked to determine if a set of words presented in quiet was audible. If not, the volume was increased to a comfortable listening level.

All three speech-in-noise assessments asked the participant to listen and repeat a word or sentence to the experimenter. The experimenter recorded whether the response was correct or incorrect. The N-back and AMLPACT tests presented a running list of words and required the participant to respond only when identifying a positive N-back match. No feedback was provided during the test session, unless requested by the participant.

5.2.3 Procedure for Speech-in-Noise Assessments

The three speech-in-noise tests administered were a QuickSIN (Etymotic Research, 2001, Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004); a dense phonological neighborhood, auditory naming staircase task; and a dense phonological neighborhood, auditory naming test at three discrete SNRs. Thirteen participants (one young and 12 older) did not complete the discrete level speech-in-noise test because it was introduced to the test battery after initial recruitment had begun.

A QuickSIN speech-in-noise test gives an assessment of a participant's speech perception in the presence of babbled background noise (Killion et al., 2004). This test is often used clinically due to its short test time (1-3 minutes). During a QuickSIN test, a female talker speaks six short sentences, each with five target words, in the presence of four-person background babble. The participant repeats back the sentence, and it is scored according to how many of the target words are correctly repeated. The first sentence starts at an SNR of 25 dB and each following sentence

reduces the SNR by 5 dB, with the final sentence presented at an SNR of 0 dB. An individual's SNR offset is calculated by totaling the correct target words repeated from each sentence and subtracting from 25.5. A practice list and two test lists were randomly selected at the beginning of each session. The score from the two test lists were averaged to determine the SNR loss of each participant. SNR loss is defined as the increase in SNR required for a participant to successfully identify 50% of the target words compared to normal listeners ((Etymotic Research, 2001; Mead C. Killion et al., 2004). Two QuickSIN lists achieve test-retest accuracy of ± 1.9 dB at a 95% confidence interval level (Killion et al., 2004).

The two auditory naming tasks utilized dense phonological neighborhoods to increase cognitive demand beyond simple SNR manipulation while keeping cognitive demand constant between participants. For both tests, subjects were expected to repeat back words in the presence of speech-shaped noise, and the test administrator recorded if a correct or incorrect response was given. Stimuli were chosen with equal likelihood from a set of 400 monosyllabic words matched for word frequency, number of phonemes, familiarity (Balota et al., 2007), and correctness (Brysbaert, Warriner, & Kuperman, 2014). Dense neighborhoods were defined as words with many neighbors (greater than 20) that differed by only one phoneme.

In the staircase assessment, 40 words were presented, and the SNR level increased or decreased by 1 dB increments depending on the incorrect or correct repetition of the word. The SNR at which the words were correctly repeated back 50% of the time was determined by averaging the SNRs at which the level reversed from decreasing to increasing (the SNRs where a response reversed from correct to incorrect).

The discrete levels auditory naming task presented 10 words at +5, 0, and −5 dB SNR. Delivering words at SNRs that can be matched across the age groups as well as matched for difficulty of task (for example, perceived difficulty of +5 SNR for older adults and 0 SNR for young adults) allows for diverse analysis of how cognition contributes to accuracy in age and task difficulty. Response accuracy at each SNR level was calculated as the ratio of the number of correctly repeated words to the total number of words in the set.

5.2.4 Procedure for N-back Assessment

The verbal N-back was designed to closely match the parameters of AMLPACT. Blocks of N-back tasks were presented with 16 audio signals, each with four positive N-back targets and four foils. A foil was considered any stimulus matching one presented less than N stimuli previously (for example, a 1- or 2-back match presented during a 3-back block). No foils were presented during the 1-back blocks.

Auditory stimuli were monosyllabic words presented in quiet, all randomly chosen from the same word set used in the auditory naming tasks. Consecutive N-back blocks linearly increased memory load from a 1-back to a 7-back. Each word presentation was followed by a 2 second response window during which participants were instructed to press a button if a word matched the target word presented N previously. If a response was recorded, the next word presentation would begin 0.5 seconds after the recorded response. If the presented word was not a match to the target, they were instructed to do nothing and the next word would begin at the conclusion of the 2 second response window. At the end of each N-back block, participants were notified of the increase in memory load and asked to press a button to confirm and continue. Response accuracy

and response time were recorded. Response accuracy for each block was calculated as the ratio of the number of correct responses to the number of total possible responses in that block.

5.2.5 Procedure for AMLPACT

A joint speech perception and working memory assessment was developed to better reflect complex, real-world environments that demand simultaneous perceptual and cognitive processing and so that it might help to disentangle the interactions of speech-in-noise and verbal working memory in individuals.

The test consisted of 20 blocks of distinct verbal N-back test items. Participants were instructed to press a keyboard button anytime an incoming audio signal matched the stimulus presented N trials previously and to ignore any non-matching signals. If a button was pressed during the 2.25 second response window following the audio stimulus, the next stimulus was delivered 0.5 seconds after the recorded button press. If no response was recorded, the next stimulus began at the conclusion of the 2.25 second response window.

Each N-back block contained 16 audio signals with four positive N-back targets and four foils. No foils were presented during 1-back blocks. Audio signals were monosyllabic words chosen randomly from the word set used in the auditory naming tests. Each N-back block presented all 16 words at the selected SNR. SNRs between −10 dB and +10 dB were achieved by introducing the stimulus word in the presence of steady, speech-shaped noise that matched the frequency spectrum of the talker. The noise began 0.5 seconds before the word was presented and remained on for 0.5 seconds after. Memory load ranged between 1-back and 7-back. An example of the auditory stimuli presented in three different blocks of an AMLPACT is depicted in **Figure 5.1**.

Response times and accuracies of button presses were recorded. The subject's mean response accuracy for each block was calculated as the ratio of the number of correct responses to the number of total possible responses in that block.

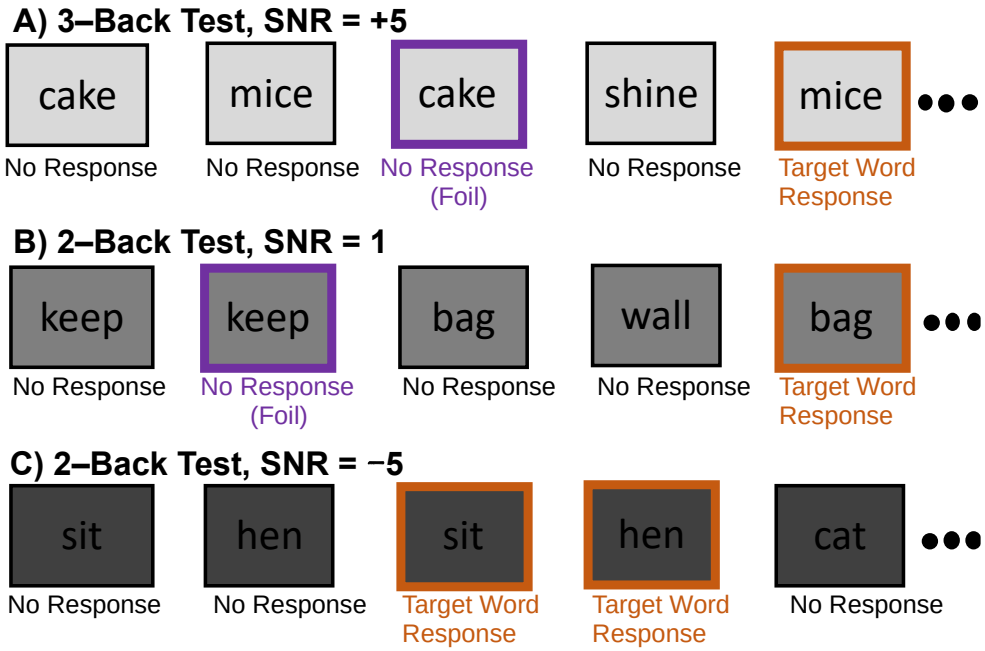

Figure 5. 1: An example of three blocks of AMLPACT stimuli. Each block of AMLPACT is an N-back paired with speech-shaped background noise presented at a set SNR. Memory load (N) and SNR are chosen by the GP framework to optimally sample the domain space. Each AMLPACT consists of 16 monosyllabic words from dense phonological neighborhoods. Each block has four **target** words (positive N-back matches) and four **foils** (non-target matches).

AMLPACT is an iterative, Bayesian inference, GP regression model capable of optimally selecting the SNR and memory load to best simultaneously explore the input domains as the test progresses. Every AMLPACT began with a 1-back at +10 dB SNR. Subsequent blocks adjusted memory load and SNR based on the uncertainty of the GP's posterior probability and a custom heuristic to penalize introducing 7-back blocks too early in the assessment.

AMLPACT's GP utilizes a constant mean function, $\mu(x) = c$, and a composite covariance function, $K(x, x') = K_\omega(x, x') + K_\eta(x, x')$ that integrates our assumptions about the memory load (ω) and SNR (η) dimensions. The mean function hyperparameter, c, was set to 0.75. This is the threshold above which participants respond to more correct than incorrect N-back targets and was selected to reflect a prior belief about average participant performance. Independently, N-back and speech-in-noise tests have a monotonically negative relationship between task difficulty (increasing N or decreasing SNR) and accuracy. However, this relationship may not always be strictly linear, and potential participant lapses dissuaded the use of linear covariance functions. Instead, a squared exponential covariance function was selected for both memory load and SNR dimensions: $K_\omega(x, x') = s_1^2 \cdot \exp\left(-\frac{(\omega-\omega')^2}{2\ell_1^2}\right)$ and $K_\eta(x, x') = s_2^2 \cdot \exp\left(-\frac{(\eta-\eta')^2}{2\ell_2^2}\right)$. The hyperparameters for the scalar factors, s_1 and s_2, and the length constants, ℓ_1 and ℓ_2, were learned by gradient descent on simulated data and refined on a 50 data point set of prior data collected on a lab member. AMLPACT set the informative hyperparameters before beginning the assessment of experimental participants, and subsequent hyperparameter learning was turned off.

Each block of the AMLPACT updated a posterior probability over the domain defined by the total set of SNR and memory load values. The posterior probability represented the model's prediction of a participant's accuracy at every SNR and N combination given the data observed. **Figure 5.2** shows the posterior probability of a participant after 1, 5, and 20 blocks are observed. As more data are collected, the GP is able to further refine its prediction of a participant's accuracy.

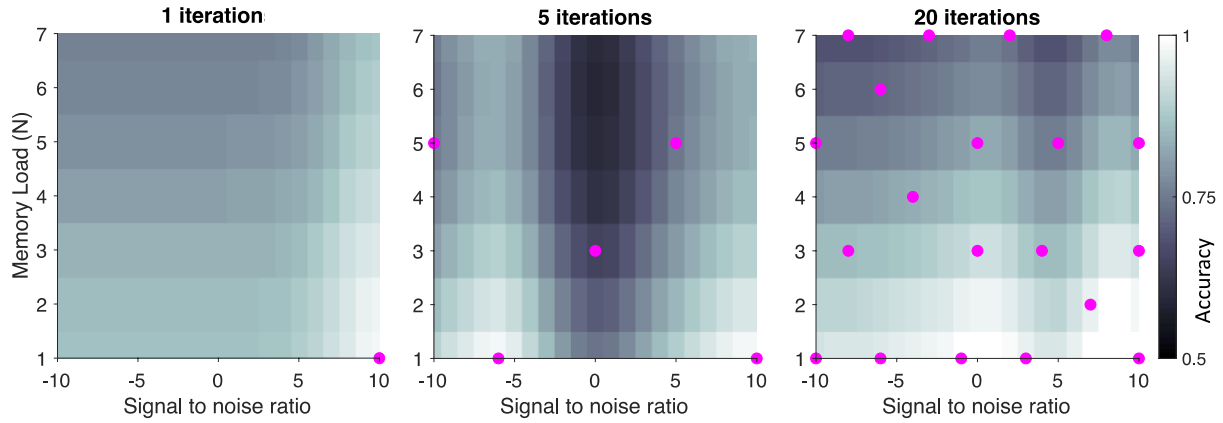

Figure 5. 2: The posterior probability mean of the GP after 1, 5, and 20 iterations for one participant. The posterior probability reflects the prior assumption of the relationships between SNR, N, and performance accuracy and the observed data. The posterior probability is updated as more data are observed. Queried points are denoted by magenta dots.

5.2.6 Data Analysis

The discrete speech-in-noise and N-back assessments were evaluated for differences between young and older adults using permutation tests. Subsets of the two-dimensional model at selected N/SNR levels were compared to the discrete working memory and speech-in-noise tests to validate the joint AMLPACT. Repeated-measures ANOVA was used to evaluate the effect of increasing memory load on AMLPACT and the N-back standalone tests. A GP and linear model were derived from the observed AMLPACT data. Parameters that informed the fit of the linear model were used to investigate the independent contributions of SNR and memory load. Summary measures of the GP and linear model were used to assess age-related shifts in performance and individual differences.

Being a novel perceptual and cognitive task, AMLPACT's test-retest reliability has yet to be determined. To that end, 10 participants (9 young and 1 older) completed two AMLPACT tests in two separate test sessions on different days.

All measures were tested for normality using the Shapiro-Wilke test as recommended for small sample sizes (Ghasemi & Zahediasl, 2012; Yap & Sim, 2011), and nonparametric statistical tests were used if the equivalence to a normal distribution could not be established for any measure being assessed.

Models of AMLPACT Performance

In this study, AMLPACT queried 20 N/SNR combinations in the two-dimensional input domain. Using these observations, two models were constructed to predict performance over the entirety of the domain. The first model is the posterior probability from the GP regression produced after every new query. The selection of nonlinear covariance functions allows the GP model to capture interactions between memory load and SNR levels within individual participants. The second model is a two-dimensional linear regression based on the observed points. This model enforces strictly linear relationships between memory load and SNR. An example of the models generated from one participant's data is depicted in **Figure 5.3.** Both models were used to analyze performance to

1) determine if a strictly linear model based on observed data is sufficient to predict performance and capture individual differences.
2) explore any added benefit of a model capable of capturing nonlinearities in investigating the interactions of speech-in-noise and verbal N-back assessments.
3) determine which model, if any, best predicts neural activity (see Chapter 6)

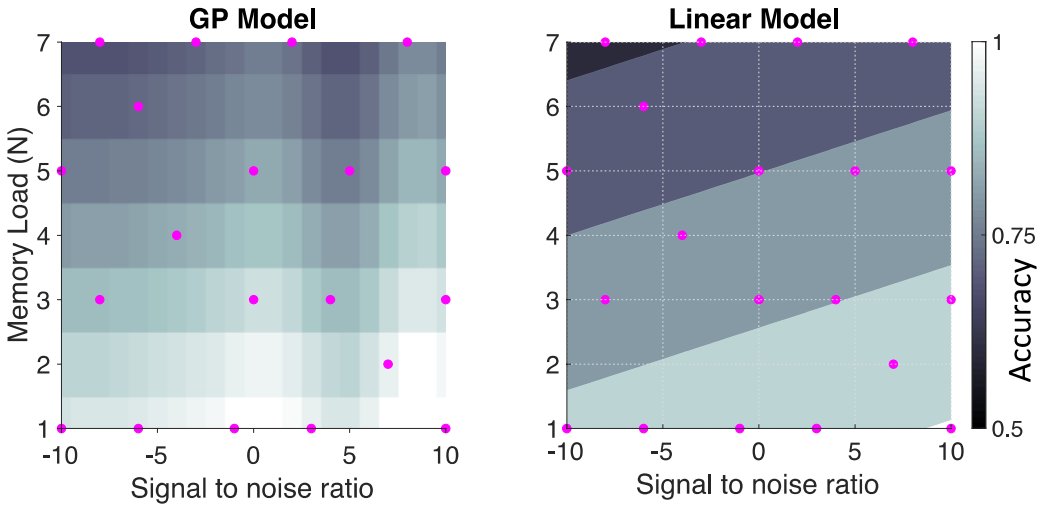

Figure 5. 3: The final GP and linear model of one participant after 20 observations. The observed blocks are denoted by magenta dots.

Summary Measures

To facilitate statistical analysis and interpret behavioral results, a summary measure of the two-dimensional surface was calculated. Reducing the two-dimensional surface to a single summary measure, performance in young and older adults was examined to extract cohort-level analysis of changes in task performance due to healthy aging. Additionally, summary measures were used to assess individual differences and the interactions of verbal working memory and speech comprehension regardless of age.

AMLPACT is a nonparametric model, so common statistical measures (such as an overall maximum or mean value) are not necessarily descriptive of the model's structure. A nonparametric statistic was developed to succinctly summarize AMLPACT performance. QuickSIN, a coarse measure of each participant's speech perception in noise, was used to determine the most predictive summary measure of the GP model. AMLPACT incorporates a speech-in-noise assessment and any summary measure chosen should be correlated to the

independent QuickSIN scores. A Pearson's correlation between QuickSIN scores and a variety of potential summary measures including weighted and unweighted sums, weighted and unweighted means, the gradient and direction of maximum slope, and the area under the volume was calculated (see https://osf.io/64qd7/ for addition details). Overall, the mean accuracy of all points where performance was above 0.75 was the best predictor of QuickSIN and a reasonable choice to represent the GP's and linear model's predictions of individual AMLPACT performance. Performance accuracy greater than 0.75 is only possible when more correct than incorrect responses are recorded, dictating its use as a threshold statistic. The mean accuracy above 0.75 threshold was used to summarize the individual shape of each participant's predicted accuracy as modeled by the GP and linear fit and to assess individual differences and predictive capabilities.

5.3 Results

5.3.1 Independent Assessments

As expected, increasing memory load in the standalone N-back resulted in increased reaction time and decreased accuracy for both age groups (**Figure 5.4**). A repeated-measures ANOVA shows effects of task load on accuracy for young adults: $F_{6,90} = 22.1, p = 7.9 \times 10^{-16}$; and older adults: $F_{6,138} = 20.0, p = 9.5 \times 10^{-17}$; as well as effects of load on reaction time for young adults: $F_{6,114} = 4.1, p = 7.6 \times 10^{-3}$; and older adults: $F_{6,90} = 3.0, p = 9.9 \times 10^{-4}$. Overall, young adults performed at a higher accuracy compared to the older adults (permutation test, p = 5.9 × 10–4) but there was no significant difference in average reaction time between young and older participants (permutation test, p = 0.89).

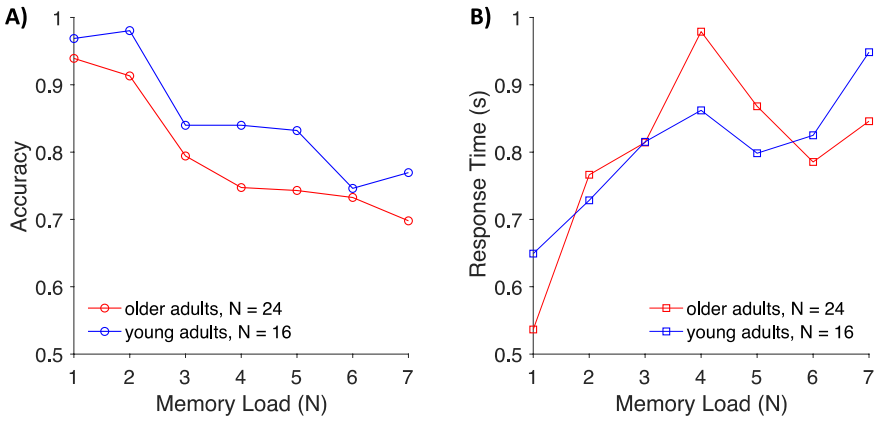

Figure 5. 4: Standalone N-back accuracy and reaction time for young and older adults. Accuracy declines and reaction time generally increases with increasing memory load.

The speech-in-noise 50% thresholds for young adults ranged from −7 to −1 dB and older adults ranged from −4 to 8 dB. There were age-related differences in the mean speech-in-noise threshold (permutation test, $p = 9.9 \times 10^{-5}$) but considerable variability within each cohort (**Figure 5.5**). Similarly, performance on the auditory naming speech-in-noise levels assessment revealed significant age-related differences (permutation test, $p = 6.0 \times 10^{-4}$) (**Figure 5.6**). The average accuracy of the young adults at SNR = 0 (0.76 ± 0.07) was similar to the average accuracy of the older adults for words presented at SNR = +5 (0.75 ± 0.09). These results are consistent with the published literature on N-back and speech-in-noise assessments.

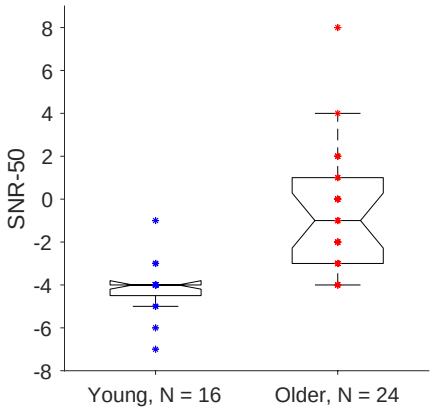

Figure 5. 5: Box plot of speech-in-noise 50% thresholds from standalone auditory naming test.

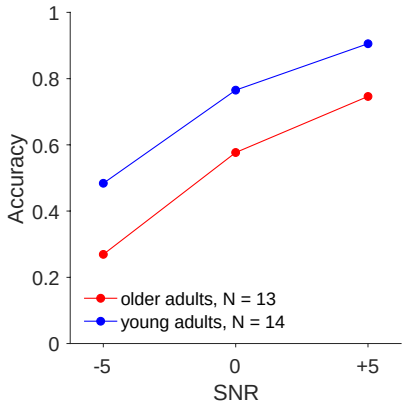

Figure 5. 6: Accuracy on auditory naming test at three SNRs. Substantial differences between young and older adults are evident.

5.3.2 Analysis of GP and Linear Model Fit and Predictive Capabilities

A 'leave-one-out' cross validation evaluated the effectiveness of the GP and linear models at predicting the mean performance accuracy of the entire N/SNR domain. Each of the 20 observations was left out once and the GP and linear models were trained with the remaining 19 observations. The trained model was then used to predict the mean accuracy of the model that included all 20 observations. The mean absolute difference between the predicted mean accuracy and the realized mean accuracy was calculated across all 20 predictions and was recorded for each participant (**Figure 5.7**). Both models accurately reflect the mean accuracy; the GP model is

slightly less accurate with a mean absolute difference of 0.0043 ± 0.001 compared to the linear model mean absolute difference of 0.0038 ± 0.001. However, neither model prediction differed significantly from the realized mean performance (Mann-Whitney U-test, p = 0.54 and p = 0.20 for GP and linear models, respectively).

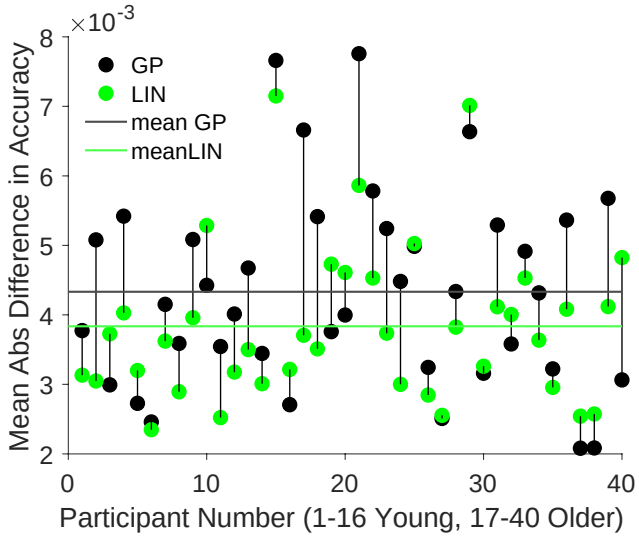

Figure 5. 7: Leave one out cross validation. Mean absolute difference between the mean accuracy and the predicted mean accuracy after 'leave one out' cross-validation of GP and linear model for each participant. On average, both models predict average participant performance with low error, but the linear model is more predictive compared to the GP model. The one young participant who had some AMLPACT training prior to recruitment is participant #1.

In addition to cross validation, the negative log likelihood of the GP model and the R^2 of the linear model were examined to measure how well the models fit the observed data. The negative log likelihood of the GP assesses the convergence rate of the model to the final posterior probability calculated with all 20 observations. It is a measure of in-sample error reduction. Assuming samples are representative of the entire input domain, once the negative log likelihood becomes asymptotic, additional observations are unlikely to significantly improve the model. At 20 observations, the GP model is not quite asymptotic and additional observations may result in

greater predictive ability (**Figure 5.8.A**). Examining a GP model from a 55 point data set collected on a lab member over multiple test sessions shows that, indeed, incremental model improvements level out as more data are collected (**Figure 5.8.B**). The R_2 of the linear models were evenly dispersed between 0.084 to 0.83 (mean: 0.49 ± 0.19) implying a lack of consistency in the model's ability to fully capture the observed data regardless of age (**Figure 5.9**). Age-related differences between R_2 values were not significant (2-sample t-test, $p = 0.19$).

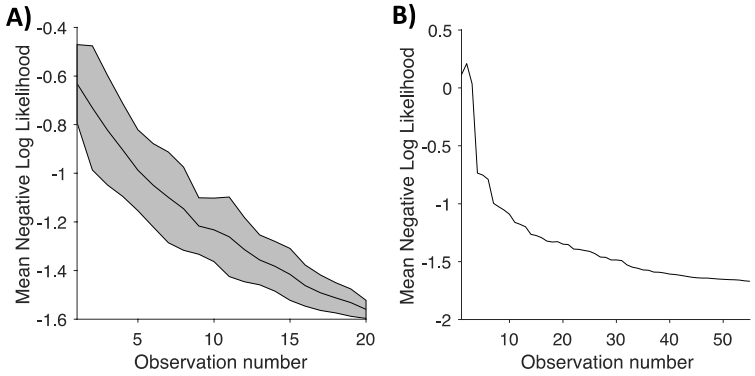

Figure 5. 8: A) Mean negative log likelihood of the posterior probability of the GP across participants. Negative log likelihood is a measure of the reduction of error as more blocks of AMLPACT are observed and is the likelihood that models constructed with fewer than 20 observations would predict the performance modeled in the final posterior probability constructed from 20 observations. B) Negative log likelihood calculated from a 55 point data set collected on one participant over multiple sessions. As more blocks are observed, subsequent models improve by smaller increments.

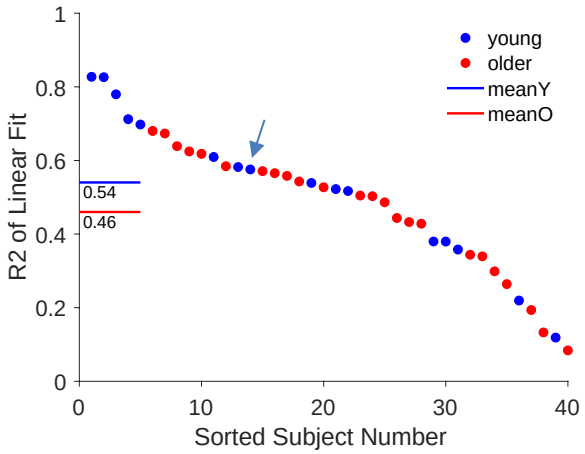

Figure 5. 9: R_2 of AMLPACT linear model for each participant. Young participants are represented by blue dots. Older participants are denoted by red dots. The mean R_2 for each age group is indicated by the long dashes along the y-axis. Age does not appear to determine the goodness of fit of the linear model to the observed data (2-sample t-test, p = 0.188). A blue arrow indicates the one young participant who had some AMLPACT training prior to recruitment.

5.3.3 Effect of Memory Load and SNR on AMLPACT Models

The regression coefficients of the linear AMLPACT model measure the approximate contribution of each independent variable included in the model construct. Inspecting the regression coefficients assigned to memory load (N) and SNR, it is apparent that the memory load dominates participant performance (**Figure 5.10**). The mean regression coefficients assigned to the memory load dependent variable was –0.035 ± 0.015 (mean t-statistic, p-value = 0.023) and to SNR was 0.0024 ± 0.0032 (mean t-statistic, p-value = 0.36). The high p-value assigned to the SNR variable indicates that the SNR coefficient does not contribute to participant performance in a statistically significant manner given the other terms in the model. The signs of the mean regression coefficients indicate that increasing memory load decreases performance accuracy (negative coefficient) while increasing SNR increases accuracy (positive coefficient). It is worth noting that while the positive mean coefficient of the SNR variable does indicate that increasing SNR results in increasing accuracy generally, 7 of the 40 participants (1 young, 6 older) had SNR

coefficients less than 0. It is counterintuitive that a participant would perform better on a N-back block with more competing noise as opposed to less when memory load is kept constant. It is more probable that the negative weights are a result of participant lapses or a shortcoming in task design. The current implementation of AMLPACT does not have the built-in capability to accommodate lapses, and this is an area for future research.

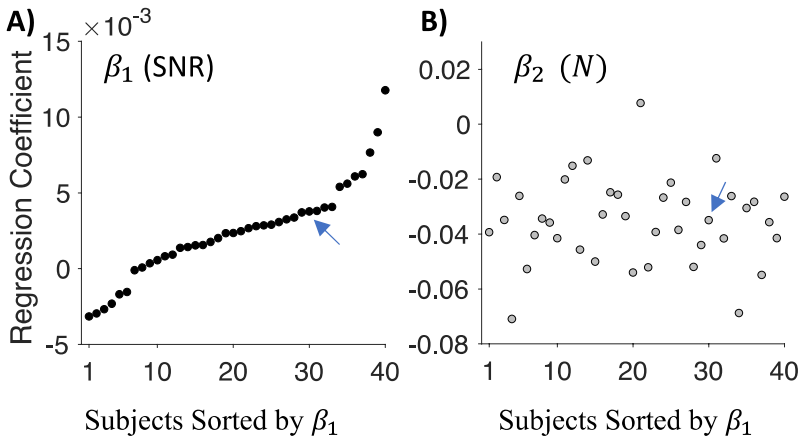

Figure 5. 10: Regression coefficients of the linear regression model fit for:

$$\text{Accuracy} \sim \beta_0 + \beta_1 * SNR + \beta_2 * N$$

Where A) shows β_1 weights and B) shows β_2. Subjects are sorted by β_1. The mostly positive β_1 reflects that increasing SNR increases performance accuracy while the mostly negative weight of β_2 reflects that increasing memory load decreases performance accuracy. Blue arrows indicate the one young participant who had some AMLPACT training prior to recruitment.

To investigate the influence of memory load and SNR on the GP model, the two-dimensional posterior probability was collapsed down to the respective memory load and SNR dimensions (**Figure 5.11**). The average change in predicted accuracy with respect to each dimension could then be evaluated. Performance accuracy declined as a function of increasing memory load from 0.93 ± 0.040 at 1-back blocks to an accuracy of 0.73 ± 0.070 at 7-back blocks (repeated-measures ANOVA: $F_{6,234} = 159.1, p = 1.2 \times 10^{-79}$). Unlike the linear model, predicted

accuracy as modeled by the GP was affected by decreasing SNR. Accuracy at the most favorable SNR = +10 dB was 0.82 ± 0.072 and accuracy at SNR = −10 dB was 0.77 ± 0.067 (repeated-measures ANOVA: $F_{20,780} = 8.6, p = 2.2 \times 10^{-23}$). The implications of these findings will be discussed further in Discussion.

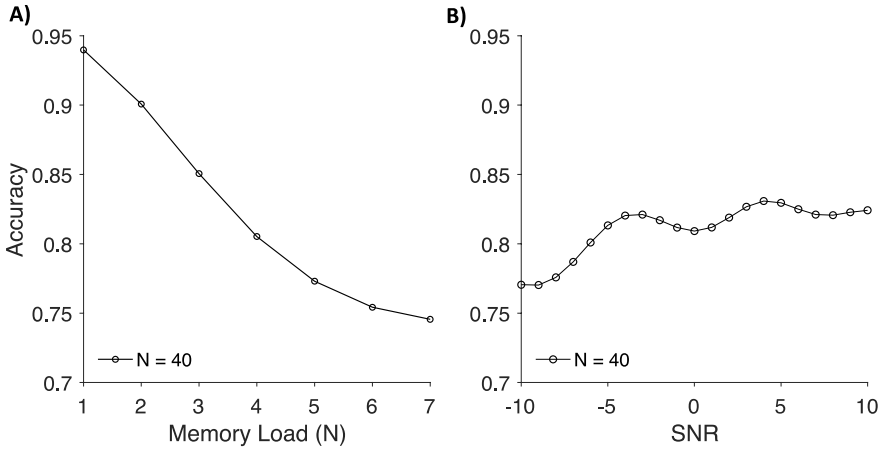

Figure 5. 11: Predicted performance accuracy collapsed down to A) memory load and B) SNR dimensions across all participants. Performance changes with respect to memory load (A) but not SNR (B).

5.3.4 Effect of Speech-in-Noise Thresholds on AMLPACT Performance

AMLPACT evaluated speech-in-noise ability by presenting words in varying levels of background noise. Participants varied in their ability to successfully comprehend speech at challenging SNRs as seen in Figure 5.5 and Figure 5.6. As a result, some blocks delivered words that were acoustically degraded to such a degree as to be unintelligible for participants. To examine the effect of intelligibility on individual AMLPACT performance, each AMLPACT model was divided with respect to the 50% signal-to-noise threshold, as determined by the staircase auditory naming task. Each participant's predicted performance could then be analyzed according to blocks paired with SNRs above or below individual signal-to-noise thresholds.

For many participants, accuracy increased *slightly* when blocks were paired with background noise above the participant's 50% threshold (**Figure 5.12**). A greater improvement was seen for the young cohort with most participants improving when words were presented above their threshold for both models (mean improvement in accuracy of 0.037 and 0.037 for the GP and linear models, respectively). Older participants had more mixed results with generally less improvement when words were presented at favorable SNRs (mean improvement in accuracy of 0.022 and = 0.020 for the GP and linear models, respectively). Overall, consistent with section 5.2.1, words presented at favorable SNRs seemed to have a small but significant effect on AMLPACT performance accuracy. A two-sample t-test confirms that performance accuracy is significantly different during blocks paired with noise above participant's individual speech-in-noise threshold compared to blocks paired with noise below a participant's threshold (GP: $p = 0.033$ and linear: $p = 0.036$).

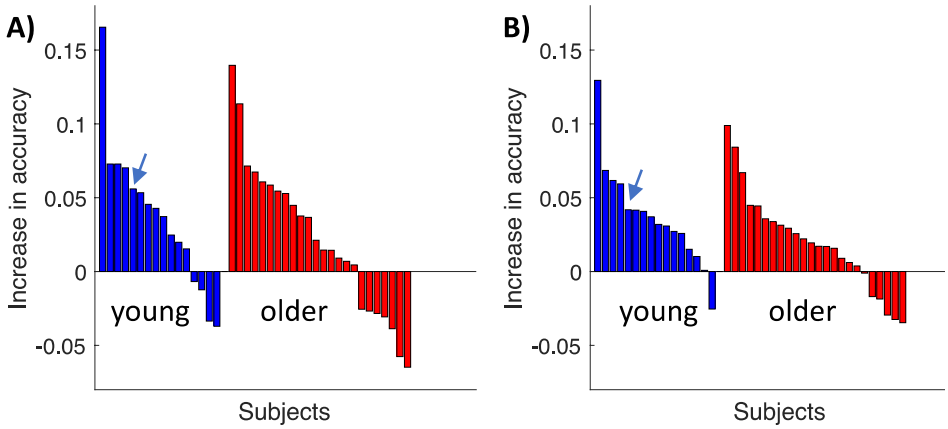

Figure 5. 12: Increase in accuracy, per participant, during blocks of AMLPACT paired with background noise above each participant's speech-in-noise threshold compared to blocks of AMLPACT paired with noise below each participant's threshold, as determined by the speech-in-noise standalone assessment, for **A)** the GP model (young participants improved, on average, 0.037 while older participants improved 0.022) **B)** the linear model of performance. (young participants improved, on average, 0.037 while older participants improved 0.020). GP and linear models are very similar. Blue arrows indicate the one young participant who had some AMLPACT training prior to recruitment.

5.3.5 Correlation of AMLPACT Summary Measures with Independent Assessments

To justify the use of AMLPACT to examine both cognitive and perceptual behaviors, it must be confirmed that AMLPACT offers a measure of a participant's speech-in-noise and working memory ability. This was assessed using a Pearson's correlation coefficient between AMLPACT's summary measures for the GP and linear model and the independent speech-in-noise and N-back assessments. Both GP and linear summary measures showed a significant correlation between mean standalone N-back accuracy (GP $r = 0.54$, $p = 3.1 \times 10^{-4}$; linear $r = 0.60$, $p = 4.3 \times 10^{-5}$), a participant's signal-to-noise 50% threshold (GP $r = -0.46$, $p = 0.003$; linear $r = -0.49$, $p = 0.001$), and the mean accuracy on the discrete levels speech-in-noise assessment (GP $r = 0.32$, $p = 0.01$; linear $r = 0.49$, $p = 0.009$).

5.3.6 AMLPACT Slices Predict Independent N-back Performance

Before examining the individual differences and the interactions of speech-in-noise and working memory, AMLPACT was first validated with respect to the N-back assessment. Predicted participant performance at slices of the two-dimensional GP and linear model were compared to participant performance on the independent N-back test. Setting the SNR to the most favorable level (SNR = +10 dB) and memory load to span the domain (N = 1 to 7), a slice of the two-dimensional predictive surface was extracted for every participant. While not in quiet, this slice most represents the discrete, words-in-quiet N-back test by minimizing the competing background noise.

Performance on the AMLPACT slice matched performance on the standalone N-back test for most memory loads (**Figure 5.13**). The mean difference between performance accuracy on the independent N-back and the predicted performance accuracy at the highest SNR slice of the GP

model was −0.0054 ± 0.031 while the linear model mean difference was −0.016 ± 0.029. **Table 5.1** shows the mean signed difference ± standard deviations for all memory loads. A two one-sided t-test evaluated the equivalency of the predicted accuracy of each model to the observed accuracy on the standalone test at each N-back load. The predicted accuracy of the GP model was statistically equivalent to the accuracy of the stand-alone N-back at all memory loads except for N = 2 and N = 3. The linear model was statistically equivalent at all memory loads except for N = 2, N = 3 and N = 4. See https://osf.io/64qd7/ for the two one-sided t-test p-values for all memory loads for each model evaluated.

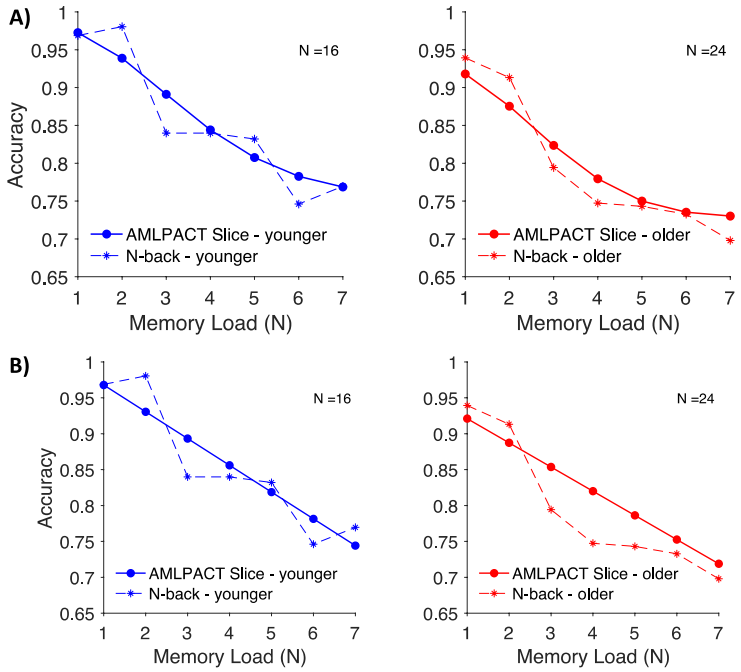

Figure 5. 13: Comparison of accuracy on standalone N-back and a slice of the AMLPACT model at the most favorable SNR (+10) for young and older adults. A) The GP model and B) linear model. The GP model better predicts standalone N-back accuracy for young and older adults.

The GP model more accurately predicted the performance of the older cohort compared to the linear model and both models were comparable in their prediction of young adult performance

(See Table 5.1 and https://osf.io/64qd7/ for details). Similar to the discrete N-back results (section 5.3.1), both models predicted statistically significant higher performance accuracy in young adults compared to older adults (2-sample t-test, $p = 9.67 \times 10_{-6}$ and $p = 0.005$ for the GP and linear slices, respectively).

Table 5. 1: Mean Signed Difference Between Predicted Accuracy on AMLPACT Slice and Actual Accuracy on Standalone N-Back for GP and Linear Models.

Memory Load (N)	1	2	3	4	5	6	7
All Participants (40)							
N-back – AMLPACT GP	0.011	0.039	−0.038	−0.021	0.006	−0.016	−0.019
	± 0.09	± 0.14	± 0.11	± 0.11	± 0.12	± 0.16	± 0.12
N-back – AMLPACT Linear	0.011	0.035	−0.057	−0.050	−0.021	−0.026	−0.002
	± 0.08	± 0.09	± 0.11	± 0.12	± 0.11	± 0.15	± 0.12
Young Participants (16)							
N-back – AMLPACT GP	−0.004	0.042	−0.051	−0.004	0.025	−0.027	0.0008
	± 0.10	± 0.09	± 0.11	± 0.13	± 0.13	± 0.16	± 0.12
N-back – AMLPACT Linear	0.0009	0.050	−0.053	−0.016	0.013	−0.035	0.025
	± 0.10	± 0.09	± 0.11	± 0.14	± 0.13	± 0.16	± 0.11
Older Participants (24)							
N-back – AMLPACT GP	0.021	0.038	−0.029	−0.032	−0.007	−0.003	−0.032
	± 0.05	± 0.09	± 0.11	± 0.09	± 0.08	± 0.15	± 0.12
N-back – AMLPACT Linear	0.018	0.026	−0.059	−0.073	−0.043	−0.020	−0.21
	± 0.06	± 0.09	± 0.10	± 0.07	± 0.07	± 0.15	± 0.13

5.3.7 AMLPACT Does Not Predict Independent Speech-in-Noise Performance

The discrete levels auditory naming task evaluates participant accuracy in correctly repeating back words at three set SNRs, –5, 0, and +5 dB. While AMLPACT performance appears to be dictated by memory load (see section 5.3.3), word intelligibility seems to have some contribution as well (section 5.3.3 and 5.34). To determine if AMLPACT can predict the discrete speech-in-noise test, AMLPACT predicted accuracy was extracted at points in the N/SNR domain that most match the discrete levels speech-in-noise test. The standalone test does not have any explicit memory component, so only AMLPACT points at N=1 were assessed to most reduce the cognitive load due to working memory. Therefore, predicted accuracy at [N, SNR] combinations of [1, –5], [1, 0], and [1, +5] were compared to accuracy on the speech-in-noise test (**Figure 5.14**). Neither the GP model nor the linear model were able to accurately predict word-repetition accuracy and predicted accuracy was significantly different than the standalone test accuracy at each SNR except for the GP prediction at [1, +5] (see **Table 5.2** for details). Error increased with decreasing SNR for both models. Given the limited effect of SNR on AMLPACT accuracy, this result is not surprising.

Table 5. 2: Mean absolute difference and Mann-Whitney U-test for significant differences between standalone speech-in-noise accuracy and predicted accuracy at AMLPACT points matched for SNR with memory load set at N=1. Significant differences indicate lack of agreement between joint and standalone.

SNR	−5	0	5
All Participants = 40			
Standalone Accuracy v. GP Predicted Accuracy	0.41 $p = 0.33 \times 10^{-9}$	0.14 $p = 1.3 \times 10^{-4}$	0.11 $p = 0.33$
Standalone Accuracy v. Linear Predicted Accuracy	0.53 2.8×10^{-10}	0.42 $p = 1.6 \times 10^{-9}$	0.13 $p = 3.0 \times 10^{-4}$

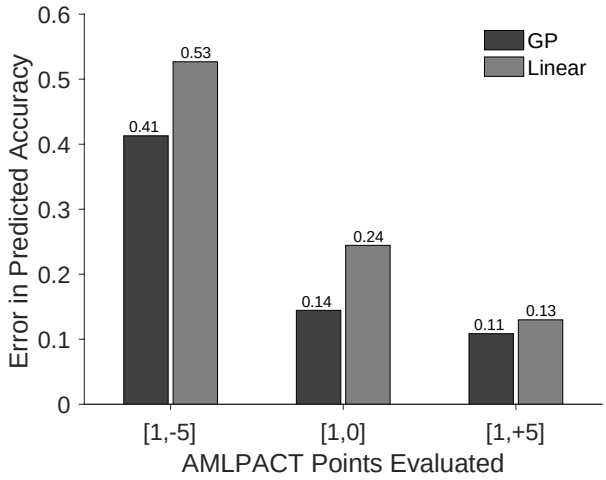

Figure 5. 14: AMLPACT error in predicting standalone speech-in-noise, auditory naming task accuracy at [N, SNR] combinations for GP and linear models. Error decreases as SNR increases. GP model has consistently less error compared to the linear model, but both models do not accurately predict the standalone assessment.

5.3.8 Test-Retest Reliability of AMLPACT

Test-retest of AMLPACT was examined to begin assessing the reliability of this novel test. Ten participants (9 young, 1 older) completed a second AMLPACT on a different day than the first assessment. Bland-Altman plots are used to show the lack of bias and the overall similarity between the summary measures of tests taken in two separate sessions (**Figure 5.15)** (Bland & Altman, 1999). Mean signed differences are near zero indicating good agreement between the two test sessions.

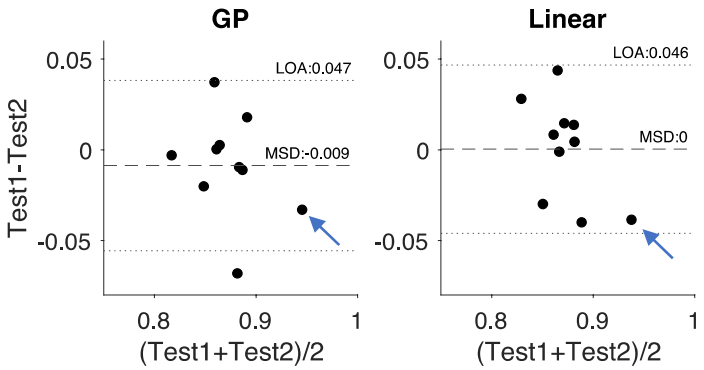

Figure 5. 15: Bland Altman plots depicting the differences between two test sessions of AMLPACT in 10 participants. The mean signed difference near zero for both models indicates good agreement between tests. Limits of agreement for the 90% CI are denoted with dotted lines. Blue arrows indicate the one young participant who had some AMLPACT training prior to recruitment.

5.3.9 Analysis of AMLPACT Performance

Older adults have been shown to have increased reaction times and produce more errors in challenging N-back and speech-in-noise tests compared to young adults (Gajewski et al., 2018; Moore et al., 2014; Wingfield et al., 2005). This shift in performance is reflected in AMLPACT false positive percentages. Controlling for the number of responses, older adults were 1.2 times as likely to incorrectly responded to non-target N-backs compared to young adults when matched for memory load and SNR. Most studies deliver N-backs at memory loads ranging between 1–3. AMLPACT was designed to tax working memory resources and delivers blocks with memory load up to 7. Therefore, it is expected that there would be very little age-related differences in AMLPACT blocks with high memory load. By analyzing the number of errors made during low (N = 1–3) memory load separate from high (N = 4–7) memory load blocks, this was apparent. During low load blocks, older participants were 1.6 times as likely to exhibit a false positive compared to young participants. However, participants were equally likely to incorrectly respond during high load blocks, and both young and older adults had false positive rates of 50%.

There were four positive N-back targets in each of the 20 blocks of AMLPACT (80 total, per test). Examining the average number of missed targets, young and older adults were matched with young adults missing 29 targets and older adults missing 33, on average. Young and older adults did not have significant differences in reaction times when matched for SNR and memory load (Mann-Whitney U-test, $p = 0.72$), and mean reaction time across young participants was only 0.05 seconds faster than older participants.

Additionally, there were significant differences between the overall performance on the AMLPACT as estimated by the chosen summary measure (**Figure 5.16**). Young adults performed at a higher accuracy (mean $= 0.87 \pm 0.02$, GP and linear) compared to the older adults (GP mean $= 0.85 \pm 0.03$ and linear mean $= 0.85 \pm 0.02$). Differences were significant with $p = 0.004$ and $p = 1.4 \times 10^{-4}$ (2-sample t-test) for the GP and linear model measures, respectively. Despite statistically different means, the heterogeneity of each age cohort can be seen in **Figure 5.17**.

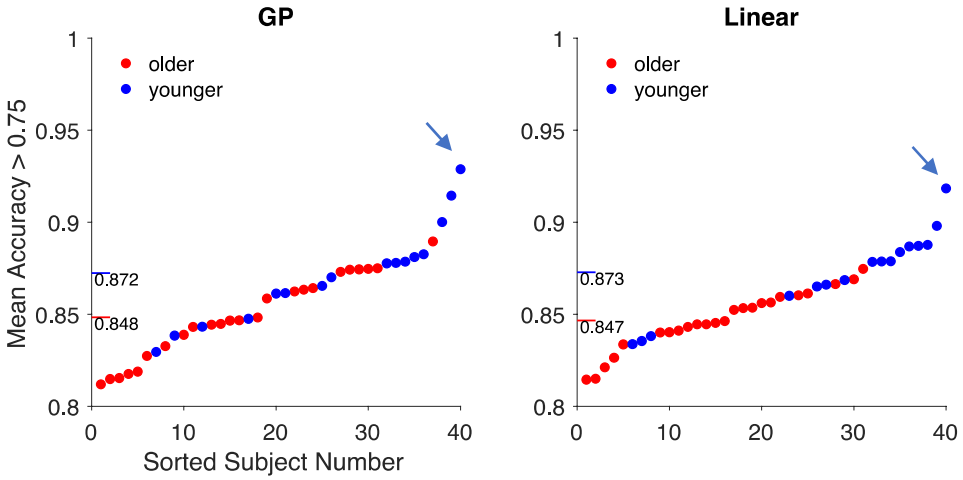

Figure 5. 16: AMLPACT summary measure of participant performance for GP and linear models. Cohort means are indicated by long dashes on the y-axis. Age-related differences are evident in the mean summary measure across cohorts. GP and linear models yield similar results. Blue arrows indicate the one young participant who had some AMLPACT training prior to recruitment.

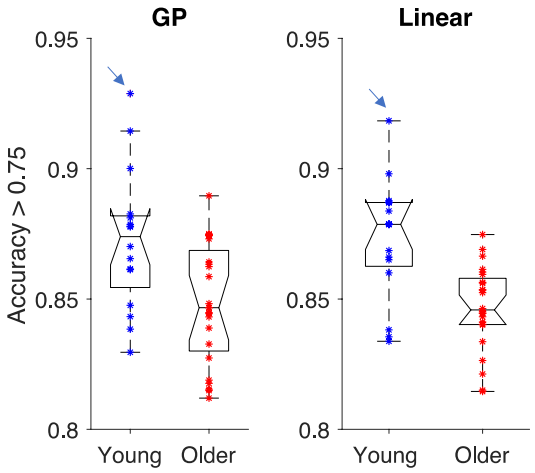

Figure 5. 17: Box plot of summary measures for young and older participants for GP and linear AMLPACT models. The long whiskers indicate high variability within age groups. Young adults are positively skewed towards higher accuracy and older adults are negatively skewed towards lower accuracy. Blue arrows indicate the one young participant who had some AMLPACT training prior to recruitment.

5.4 Discussion

One goal of this study was to demonstrate that the Bayesian inference GP framework could estimate cognitive and perceptual tasks in a single, multidimensional assessment. AMLPACT was able to sufficiently model older and younger participant's working memory ability. While it was correlated to the speech-in-noise assessments, it did not accurately predict the standalone comparison test. This is likely due to AMLPACT test design and not a flaw of the framework itself. AMLPACT did provide an estimate of performance accuracy that varied with SNR and memory load; and, similar to other working memory assessments, between-group differences and within group variances in performance were detected.

By design, AMLPACT is a continuous estimate of a subject's response accuracy over the entire SNR and memory load domain. It offers a direct measure of the interactions of memory load and SNR with relatively few observations. By incorporating nonlinear covariance functions, AMLPACT is able to model interactions that a strictly linear regression model could not. In this

study, the GP model proved to be as predictive as the linear model in estimating unobserved data. Many of the analyses showed almost no differences between the GP and linear model results.

And yet, the GP model did *slightly* outperform the linear model in accurately predicting the standalone speech-in-noise and working memory assessments (sections 5.3.6 and 5.3.7). While the linear model appeared to be unaffected by the SNR regression coefficient, GP model performance exhibited significant load effects as a result of changing SNR (section 5.3.3). Varying SNR seemed to contribute less to AMLPACT performance when compared to memory load, but its impact might be reflected in the slight improvement the GP shows in predicting low-dimensional test performance. Nevertheless, there is little evidence that the GP is capturing interactions unaccounted for by the linear model. This result demonstrates that when interactions do not appear to be present, the GP model exhibits similar accuracy to standard linear regression. A persistent advantage of the GP is that no accuracy is sacrificed by enabling a more complex model, even when a simpler model might suffice. More basic models, such as standard or generalized linear models, must wait until most data are collected before constructing an informative model, and each additional predictive variable must be evaluated and added to model design systematically. The GP, on the other hand, can flexibly estimate an infinite number of predictive functions if the appropriate covariance and mean functions are incorporated into its definition.

AMLPACT is highly correlated with the independent tests of noisy speech perception and verbal working memory used in this study. However, analyzing the regression coefficients of the linear model (section 5.3.3) suggests that there is only a small direct linear relationship between memory load and background noise. Similarly, the effect size of varying SNR on the GP estimate

is considerably smaller than that of varying memory load. This suggests that fully comprehending the word being presented is not critical to task performance, but doing so might lessen the cognitive load and improve accuracy slightly.

In inspecting Figure 5.11, the possibility remains that SNR has a small nonlinear effect on performance during joint speech and memory behaviors. If true, it could explain the inconsistent fit of the linear model to the observed data across participants, as indicated by the large spread of R_2 values. Both the GP and linear predictions would benefit from better model fits. The GP model might only require additional observations to be collected during test sessions. In the 55-point set of AMLPACT data, model error continued to diminish substantially as addition data were added into the model. Nearing 50 observations, the negative log likelihood began to become asymptotic, indicating that additional data would not improve the model fit much further. If there is a nonlinear effect of SNR on performance during the joint task, it is anticipated that additional observations would not improve the fit of the linear model for many participants. Fitting two linear models, one to the first 20 observations and another to the entire set of observations in the 55-point dataset, the R_2 value did not change as a result of additional observations (20 observation model, $R_2 = 0.67$ and 55 observation model, $R_2 = 0.66$). While GP and linear models appear to find similar results in the analyses presented here, the possibility that both models are performing at equally suboptimal levels should be considered. For the GP model, this is might be easily rectified by additional data whereas the linear model might be more limited in its potential for improvement.

It appears that this specific combination of an auditory N-back with variable speech-shaped background noise may not interact in the ways anticipated at the onset of this study. It would be

too hasty to conclude that verbal working memory and challenging speech comprehension utilize wholly independent cognitive resources, generally. More likely, AMLPACT test design is not a balanced representation of joint speech-in-noise and memory ability and adaptations are needed to better measure speech comprehension.

The previous studies that have asserted correlations between speech-in-noise and working memory are employing assessments that require a confirmation of the word or sentence intelligibility. Reading span tasks, for example, are often cited as having high correlations with speech-in-noise tasks (Akeroyd, 2008; Foo, Rudner, Rönnberg, & Lunner, 2007; Lunner, 2003; Rudner, Rönnberg, & Lunner, 2011). Reading span tasks are a working memory test in which the participant repeats back the final words from a series of previously presented sentences and are scored according to the number of words correctly repeated. N-backs differ from reading spans in that participants could perform with high accuracy on a N-back while not comprehending any of the target words. Instead, participants could be matching non-word or partial word targets and still maintain high performance. In AMLPACT, where stimuli are intentionally distorted by competing background noise, it is possible that participants were doing just that. This would have reduced the perceptual demand of AMLPACT and may have subsequently reduced any interactive effect between memory load and noisy speech comprehension. If AMLPACT does not require comprehension to accurately perform, the lower effect of SNR on the GP and linear model is also not surprising. This would also explain the inability of either model to predict accuracy on the standalone comprehension test (c. f., Figure 5.14).

Further, many of the widely cited studies that purport working memory and speech-in-noise correlations are primarily conducted in populations with hearing loss (Füllgrabe & Rosen, 2016),

whereas the participants in this study have exclusively near-normal hearing. It is suspected that older adults with hearing loss more quickly exhaust purely perceptual resources (possibly due to age-related changes in brain structure combined with declining perceptual ability (Grady, 2013) and rely more on compensatory support from cognitive or domain general functions compared to older adults with normal hearing or young adults. Recruitment of non-auditory resources to support perception leaves fewer resources to be allocated to any competing demands of concurrent working memory functions. Adults with normal hearing are more able to solely dedicate specialized resources to perform perceptual behaviors, leaving working memory resources available for simultaneous cognitive functions. Thus, older adults with hearing loss may experience more interaction between cognitive and perceptual functions during complex tasks than adults with normal hearing. Restricting this study to self-reported normal hearing listeners may have inadvertently limited the interactive effects of noisy speech comprehension and verbal working memory that could be measured.

Future studies could directly address these concerns by recruiting participants with more diverse hearing abilities and by adjusting the AMLPACT assessment to better balance speech-in-noise and memory load contributions. One of the main advantages originally envisioned of the current AMLPACT was being able to offer a measure of speech-in-noise and working memory without needing to be scored by an observer. Adjusting AMLPACT to include an intelligibility component might require observer scoring, but it could also be possible to develop a more complex version that can check for intelligibility without needing to be scored. For example, AMLPACT could have the participant select the target word from a set of multiple choice options carefully designed to included words from the same phonological neighborhood as well

as a ‘no match’ option. AMLPACT could also be modified to model a difference working memory assessment all together. Further, words consist of multiple phonemes that, when combined, might influence the ease with which they are understood in the presence of speech-shaped background noise (Billings, Grush, & Maamor, 2017; Meyer, Dentel, & Meunier, 2013). This would be compounded by age-related shifts in hearing loss which affect high-frequency sounds first. As a result, some words, regardless of phonological neighborhood, may be easier to pick out throughout speech-in-noise assessments. AMLPACT could be extended to incorporate word structure and learn which stimulus words should be delivered to each participant to better assess noisy speech comprehension. This might result in a more stable assessment of speech-in-noise ability and could provide additional information about individual hearing ability. An appeal of the GP framework is its readiness to extend to different input domains with minimal adjustments to the defining parameters.

Differences in reaction time between young and older participants were expected, but not found in either the standalone N-back or AMLPACT results. Both assessments were auditory verbal working memory tasks that require input from the participant in the form of a button press. It is possible that the processing time on these specific tasks differ too significantly from other studies where age-related reaction time differences have been found. Comparing reaction times across varying test modalities is cautioned against (Hancock et al., 2007), and the lack of findings should not be considered a flaw in either test.

Decline in hearing ability has been identified as a predictor of future cognitive function and Alzheimer’s Disease progression (Gates, Anderson, McCurry, Feeney, & Larson, 2011; Gates et al., 2008; Liu & Lee, 2019). Given that deficits in memory are a clear harbinger of cognitive

decline, AMLPACT could be useful in assessing performance in early disease. One participant in the current study was recently diagnosed with early stage cognitive decline that was likely to progress to Alzheimer's Disease as identified through a structural MRI. This participant, klh306, had a higher speech-in-noise threshold and was on the low end of N-back performance. Both results are within the inner fences of the data (1.5 x interquartile range) and neither result would be considered an outlier compared to the healthy older adults (**Figure 5.18**). In fact, klh306 does not have the highest speech-in-noise threshold or the lowest N-back performance in the older adult cohort. Similarly, AMLPACT performance as estimated by the linear model summary measure of participant klh306, is equal to the lowest healthy adult's summary measure (**Figure 5.19a**). The AMLPACT linear model of klh306 is, therefore, essentially indistinguishable from the healthy older adult data. However, their AMLPACT performance, as modeled by the GP, was noticeably worse than all healthy participants (**Figure 5.19b**). While not quite outside the inner fence of the older adult data, the distance between the GP summary measure of participant klh306 and the nearest healthy adult is one order of magnitude greater than the distance between all other healthy adults. It is possible that, in states of cognitive decline, there are nonlinear interactions between memory load and SNR that the flexible AMLPACT framework can uniquely model. Only one participant recruited for this study was confirmed to have any cognitive decline, so future research would be needed to make any substantial claim.

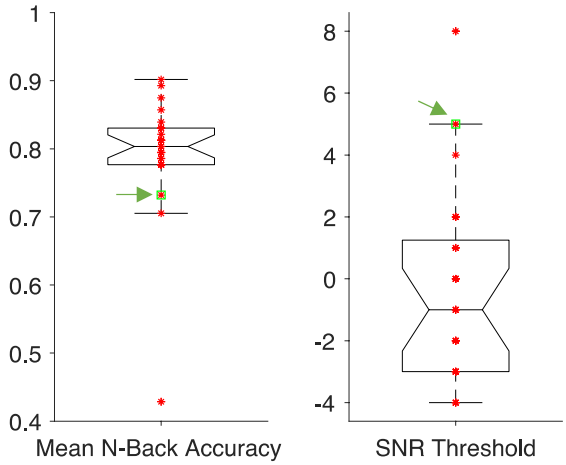

Figure 5. 18: Box plots of mean N-back accuracy and speech-in-noise 50% threshold for all healthy older adults plus cognitive decline participant klh306. Klh306 is denoted by green arrows. Note that klh306 is not an outlier in the distribution nor is he or she the worst scoring participant.

Individual variations in the noisy speech comprehension of young and older adults, despite having near-normal pure-tone hearing thresholds, is reflected in the variable speech-in-noise 50% thresholds (see Figure 5.5). It has been suspected that such variations might be explained by individual differences in verbal working memory (Akeroyd, 2008). AMLPACT, being dominated by a participant's working memory ability while still offering a measure of speech-in-noise assessment, offers a multidimensional evaluation that might better explain the variability of low dimensional assessments. However, the lack of interactions between the two AMLPACT domains limits the extra information that can be gleaned from this novel test. AMLPACT performance correlates significantly with speech-in-noise and working memory tests, which is consistent with the literature on individual differences in speech-in-noise assessments. At the very least, AMLPACT has replicated the previous research in a more direct manner by testing speech-in-noise and working memory concurrently.

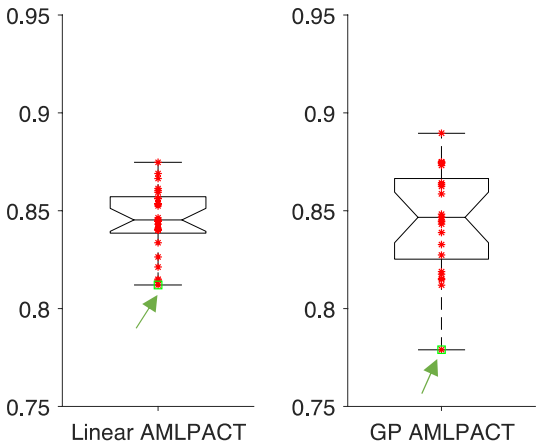

Figure 5. 19: Box plots of linear and GP AMLPACT summary measure for all healthy older adults and cognitive decline participant, klh306. Klh306 is denoted by green arrows. While klh306 has the lowest summary measure in both models, there is a much bigger gap between healthy older adults and klh306 in the GP model compared to the linear model.

Several participants in this study had the same speech-in-noise 50% threshold and mean N-back accuracy (**Figure 5.20**). Despite performing similarly on the standalone tests, their AMLPACT summary measures differed. Perhaps the concurrent demands of the joint speech and memory test affect individuals differently. Because of the possible shortcomings of the implemented AMLPACT test design and the small sample of participants with similar scores on both standalone assessments, future research is needed to investigate the implications of this finding. For example, collecting fMRI data while these participants perform the standalone and AMLPACT assessments might discover individual differences in neural strategy relevant to performance.

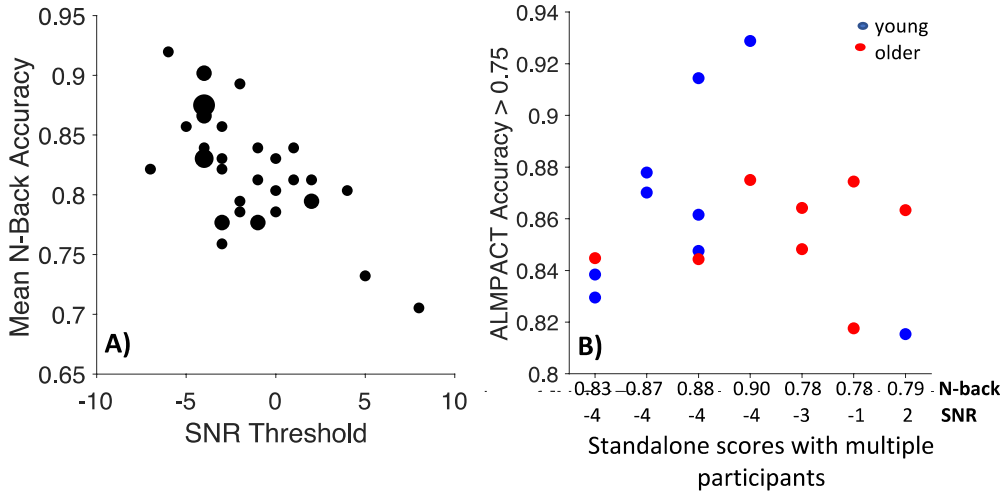

Figure 5. 20: A) Participant accuracy on standalone N-back and 50% speech-in-noise threshold. Larger circles indicate more participants with the same mean N-back accuracy and SNR threshold. Individuals with the same N-back accuracy and SNR threshold might have individual differences in the interactions between working memory and speech comprehension that AMLPACT can detect. One participant is excluded from this figure due to their substantially lower N-back accuracy (0.41) and no other participant had similar results. B) AMLPACT summary measure (modeled by the GP) for participants with the same scores on the standalone tests (the larger circles in A). Similar N-back and SNR thresholds do not necessarily lead to similar AMLPACT results.

AMLPACT has good test-retest reliability in healthy adults. Therefore, changes in individual performance from one test session to another might give significant insight into a patient's changing health. Using an initial test result or an average healthy result as a prior could allow AMLPACT to quickly assess if a patient has deviated from past behavior or from a 'normal' result.

5.5 Concluding Remarks

AMLPACT demonstrates that a joint perceptual and cognitive task is feasible with the new active machine learning framework. While substantial interactions were not found, the GP model was as accurate as the linear model in estimating participant performance in relatively few observations. The initial implementation of AMLPACT might not reflect noisy speech

comprehension to the extent necessary to fully model standard speech-in-noise assessments. Amending AMLPACT to better model the domains of interest would be a straightforward adaptation to the GP definition. The advantages of a flexible joint assessment is demonstrated in AMLPACT's ability to vary each test delivery according to a participant's previous performance and its real time evaluation of the domains being modeled. The framework underpinning AMLPACT is not specific to speech-in-noise or working memory. Adapting this framework to other cognitive test paradigms could advance the understanding of individual differences and interactions in a variety of domains.

Chapter 6: Estimating Neural Activity From Individual Differences in a Joint Speech and Memory Test

6.1 Introduction

Successful speech comprehension requires the collaboration of multiple sensory, perceptual, and cognitive processes. In noisy environments, listeners must focus attention on the speaker, disregard competing cues, correctly process incoming audio stimuli, and attach meaning and context to what is heard. It stands to reason that speech comprehension in noisy environments engages more than simple auditory processing.

Domain-general networks associated with attention, task-switching, and verbal and visual representation have been implicated in supporting complex auditory demands (S. Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013; Cacace & McFarland, 2013; Davis & Johnsrude, 2003; Peelle, 2018; Pichora-Fuller et al., 1995; Rönnberg et al., 2013). These same domain-general networks are also engaged during verbal working memory tasks (Braver et al., 1997; McCoy et al., 2005). In fact, verbal working memory may be directly recruited during speech comprehension (Lunner, 2003; Rudner et al., 2011; Ward, Rogers, Engen, & Peelle, 2016) and working memory training can improve speech comprehension ability (Wayne, Hamilton, Huyck, & Johnsrude, 2016). Additionally, age-related hearing loss and cognitive decline (which can be measured by working memory ability, among other cognitive measures) appear to be interdependent (for review, Wayne & Johnsrude, 2015). However, current methods have limited

power to investigate the interactions between these two systems (McFarland, 2017; Wayne & Johnsrude, 2015).

In-scanner tasks often deliver the same test items to all subject for ease of comparison. For example, recording the successful repetition of words in levels of background noise that are constant between cohort participants. Individual performance is then correlated to neural activity at group and individual levels to identify which brain regions contribute to task performance. Because tasks completed inside an MRI scanner must be optimized to reduce head movements, make efficient use of costly scan time, and contend with demanding acoustic noise conditions, they are often low-dimensional tests incapable of delivering complex stimuli that place multiple demands on limited neural resources at once. As a result, non-task specific neural resources are readily available to support performance.

A multidimensional behavioral test that can optimize data collection for participants within one test session and is capable of identifying individual neural strategies for successful noisy speech comprehension would begin to unravel how non-auditory brain networks contribute to individual behavior during real-world tasks. The machine learning framework that models verbal working memory and a measure of speech-in-noise ability may provide this utility. As an initial experiment to extend this framework to estimate neural activity, the joint speech and memory test (AMLPACT) piloted in Chapter 5 was evaluated as an out-of-scanner behavioral test to predict individual neural activity during an in-scanner speech-in-noise test.

AMLPACT is a multidimensional assessment that was designed to jointly require speech comprehension and verbal working memory resources; therefore, the increased cognitive

demand (compared to a low-dimensional test) might more effectually elucidate how speech, domain-general, and verbal working memory networks interact. Additionally, AMLPACT might more directly assess the individual neural strategies employed to maintain performance. Within age group cohorts, AMLPACT performance had high variance, particularly in older adults. Previous research suggests that older adults employ cognitive strategies to maintain performance at low levels of task challenge, even before comprehension accuracy declines (Pichora-Fuller et al., 1995; Tun et al., 2009; Wild et al., 2012; Wingfield & Grossman, 2006). Individual differences in AMLPACT performance might be indicative of individual cognitive strategies and differences in neural recruitment in older adults.

The frontoparietal network is one domain-general network active during many cognitive tasks, including challenging speech-in-noise tests (Peelle, 2018). It is noted to contribute to a variety of executive functions and is generally thought to coordinate and modulate cognitive control (Cole, Yarkoni, Repovš, Anticevic, & Braver, 2012; Marek & Dosenbach, 2018). The in-scanner assessment used in this study is designed to be challenging, especially for older adults, without sacrificing accuracy, with the aim of recruiting non-auditory resources that would be otherwise unnecessary for simpler auditory tasks. Within the frontoparietal network, the dorsolateral prefrontal cortex has been identified as a core contributing region to working memory ability (Cole et al., 2012; D'Esposito et al., 1998; Duncan & Owen, 2000; Fedorenko, Duncan, & Kanwisher, 2012; Rottschy et al., 2012; Wallis, Baker, Meese, & Georgeson, 2013), making it a good candidate for non-auditory activation. Additionally, the dorsolateral prefrontal cortex robustly exhibits load-dependent changes in activation during N-back assessments (Braver et al., 1997; Mencarelli et al., 2019), which are predominantly featured in AMLPACT design. For this

reason, a region of interest in the dorsolateral prefrontal cortex has been extracted for comparison to AMLPACT performance.

In addition to the frontoparietal network, three regions within the cingulo-opercular network were used: regions in the left and right frontal operculum/anterior insula and a region in the dorsal anterior cingulate. Like the frontoparietal network, the cingulo-opercular network is thought to be a domain-general network active during many cognitive tasks. Primarily contributing to the executive functions of error monitoring and attentional salience, activity in this network, increases during missed targets and may precede an increase in accuracy on the next target during speech-in-noise tests (Harris et al., 2009; Kuchinsky et al., 2013; Vaden et al., 2017). Besides contributing during challenging speech comprehension, the cingulo-opercular network is active during verbal working memory (Owen et al., 2005; Sadaghiani & D'Esposito, 2015), prompting its inclusion in the following correlation analysis.

Performance on AMLPACT was paired with neural activity collected during an in-scanner speech-in-noise test to analyze variance in individual performance and neural recruitment. Differences in brain function during the in-scanner test might be related to differences in working memory and the interaction between it and speech comprehension as modeled by AMLPACT. The correlation between AMLPACT performance and neural activity in areas associated with verbal working memory may quantify the variations in signal due to age, working memory ability, and individual neural recruitment strategy.

6.2 Methods

6.2.1 Participants

Of the forty participants who completed the joint noisy speech and working memory test, twenty participants (5 young and 15 older) also completed a speech-in-noise test paired with fMRI scanning as part of a different study. All participants who participated in the fMRI study were right-handed, matched for education level, used no hearing assist devices, and exhibited no evidence of neurological disease. To date, fMRI data from 11 of the older adults have been pre-processed and are analyzed in this chapter. In addition to fMRI data, audiograms were collected in each of the 11 participants. All but two participants presented with better-ear pure-tone averages in the normal range (mean hearing threshold across 500, 1000, 2000, 4000 Hz is less than 20 dB). The mean pure-tone average of all 11 participants was 17 ± 6 dB HL. The two participants with better-ear pure-tone averages greater than 20 dB were classified as having mild hearing loss with pure-tone averages of 25 dB HL and 28 dB HL.

6.2.2 Procedure

The speech-in-noise test used in the scanner is similar to the discrete levels, auditory naming test delivered in Chapter 5. Auditory stimuli were monosyllabic words matched for word frequency, number of phonemes, familiarity (Balota et al., 2007) and correctness (Brysbaert et al., 2014). Words were grouped according to phonological neighborhood density. Words with many neighbors (greater than 20) that differed by only one phoneme were categorized into dense neighborhoods, while words with fewer than six single phoneme neighbors were considered to be in sparse neighborhoods. Participants completed an auditory naming test under two experimental conditions: words from sparse and dense neighborhoods presented at a set level of acoustic clarity. Young and older participants were presented words at an SNR of +3 dB. Each

test consisted of 40 words. In addition to the two auditory naming tests, 40 trials of single channel noise-vocoded words were presented to each participant in +3 dB SNR speech-shaped background noise as a control condition.

6.2.3 MRI Acquisition and Processing

MRI data were acquired using a Siemens Prisma scanner (Siemens Medical Systems) at 3 T equipped with a 32-channel head coil. Scan sequences began with a T1-weighted structural volume using an MPRAGE sequence (repetition time (TR) = 2.4s, echo time (TE) = 2.2 ms, flip angle = 8°, 300 × 320 matrix, voxel size = 0.8 mm isotropic). Blood oxygenation level-dependent (BOLD) functional MRI images were acquired using a multiband echo planar imaging sequence (Feinberg et al., 2010) [TR =3.07 s, TA = 0.770 s, TE = 37 ms, flip angle = 90°, voxel size = 2 mm isotropic, multiband factor = 8). To mediate the challenge of outstanding acoustic noise during standard MRI collection, a sparse imaging design in which there was a 2.3 second delay between scanning acquisitions and the TR was longer than the acquisition time to allow for minimal scanning noise during stimulus presentation and audio recording of participant responses (Edmister, Talavage, Ledden, & Weisskoff, 1999; Hall et al., 1999; Wong et al., 2009) was used. This method inserts a brief pause at every scan and allows short stimuli to be delivered in relative quiet. During the auditory-naming task, participants were asked to repeat back the words heard in the scanner during the pause in scanning to minimize head motion that would degrade the scan quality. Results were scored for accuracy at a later date.

Analysis of the MRI data was performed using Automatic Analysis (Cusack et al., 2015) which scripted a combination of SPM12 (Wellcome Trust Centre for Neuroimaging) and FSL (FMRIB Analysis Group; Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). Data were realigned

using rigid-body image registration, and functional data were co-registered with the bias-corrected T1-weighted structural image. Spatial and functional images were normalized to MNI space using a unified segmentation approach (Ashburner & Friston, 2005), and resampled to 2 mm. Finally, the functional data were smoothed using an 8 mm FWHM Gaussian kernel.

For the listening-only condition, there were no measures of accuracy, so all trials were analyzed. For the auditory naming conditions, only trials associated with correct responses were analyzed. For both, the noise condition was modeled in addition to words.

Motion effects were of particular importance given that participants were speaking during the auditory naming condition. To mitigate the effects of motion, a thresholding approach in which high motion frames were individually modeled for each subject using a delta function in the GLM was used (see e.g. Siegel et al., 2014). Motion was quantified using framewise displacement (FD), calculated from the 6 motion parameters estimated during realignment assuming the head is a sphere having a radius of 50 mm (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012).

A threshold that resulted in 10% data exclusion across all participants was selected with the rationale that some participants move more, and thus produce worse data; therefore, a single threshold for all participants was used, resulting in more data exclusion from high-motion participants. For each frame exceeding this threshold, a column was added to that participant's design matrix consisting of a delta function at the time point in question, which effectively excludes the variance of that frame from the model.

Activity during all word presentations that were correctly repeated was collapsed down across brain densities for each experimental condition. Activity was then assessed with respect to the control noise condition. For each voxel, only intensities that were significantly greater ($p < 0.05$) than activity during the passive listening noise condition were included. Contrast images from single subject analyses were analyzed at the second level using permutation testing (FSL randomize; 5000 permutations) with a cluster-forming threshold of $p < 0.001$ (uncorrected) and results corrected for multiple comparisons based on cluster extent ($p < 0.05$) (Gorgolewski et al., 2015). This resulted in whole brain maps of activity during correct trials that was greater than activity during the noise control for each participant. These maps could then be averaged across participants to get the average activity of the cohort (see F**igure 6.1**).

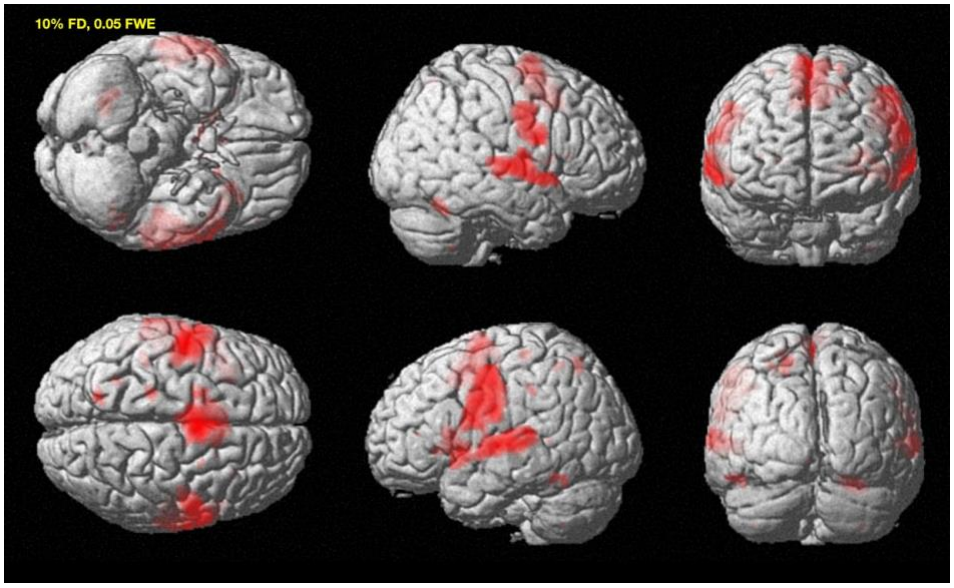

Figure 6. 1: Whole brain thresholded t-map for older adults during dense phonological test. Activity is thresholded at $p < 0.05$ for dense activity greater than activity during the noise-only trials

6.2.4 Regions of Interest

MNI coordinates that were central in regions of interest for the frontoparietal attention network and the cingulo-opercular network were used to select distinct parcels from a 400 parcellation of

the human brain (Schaefer et al., 2018; Thomas Yeo et al., 2011). To assess frontoparietal attention network activity during the in-scanner test, a parcel in the left dorsolateral prefrontal cortex (MNI coordinates: [–44, 14, 29]) was selected. Regions that include these coordinates are specifically active during N-back assessments and have been repeatedly identified as characteristic of working memory activity (Cole et al., 2012; Lamichhane, Westbrook, Cole, & Braver, 2020; Rottschy et al., 2012). Three parcels in the cingulo-opercular network that have demonstrated activity during word recognition tests (Vaden et al., 2013) were chosen: one in the dorsal anterior cingulate (MNI coordinates [5, 35, 34]), one in the left anterior insula/frontal operculum (MNI coordinates [−45, 21, −8]), and one in the right anterior insula/frontal operculum (MNI coordinates [32, 27, −9]). For region of interest analysis of primary auditory cortex, probabilistic maps based on postmortem human histological staining were used (Morosan et al., 2001). These are available in the SPM Anatomy toolbox (Eickhoff et al., 2005). All chosen regions are depicted in **Figure 6.2**.

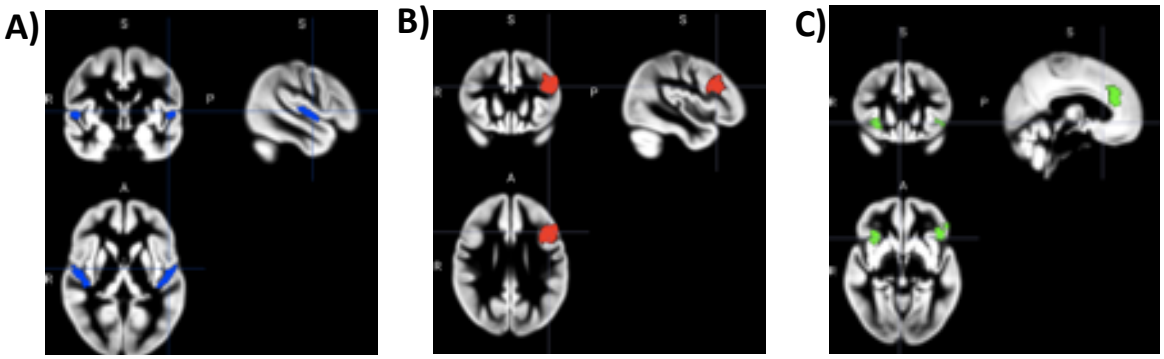

Figure 6. 2: Regions of interest used for neural activity analysis in A) auditory speech network, B) left dorsolateral prefrontal cortex, and C) cingulo-opercular networks.

A binary mask for each region extracted estimates for contrasts of interest from each participant's first-level analyses by averaging over all voxels in each region. These network specific contrast

estimates were used for the individual difference analysis. Given the left dorsal lateral prefrontal cortex region is wholly defined on the left hemisphere of the brain, left brain analysis is depicted in all figures. Additionally, the primary auditory regions were analyzed for whole-brain and left hemisphere activity since left-lateralized differences the inferior frontal gyrus have been identified during speech comprehension tasks (Humphries, Willard, Buchsbaum, & Hickok, 2001; Obleser, Wise, Alex Dresner, & Scott, 2007; Peelle, Troiani, et al., 2010). Whole brain activity was analyzed for the cingulo-opercular. Because the joint speech and memory test only delivered stimuli from dense phonological neighborhoods, only brain activity during the dense neighborhood conditions was included in analysis.

6.2.5 Data Analysis

A Pearson's correlation was calculated to determine if individual differences in the joint speech and memory test were correlated to individual differences in brain activity.

6.3 Results

The mean accuracy above the 0.75 threshold was chosen as a summary measure of the overall performance on AMLPACT due to its high correlation to a widely used speech-in-noise measure, QuickSIN (see Chapter 5). The relationship between the mean activity during the dense in-scanner test and the AMLPACT summary measure, as modeled by the GP posterior distribution, was evaluated with a Pearson's correlation for each region of interest (**Figure 6.3**). None of the correlation coefficients were statistically significant ($p = 0.45$, $p = 0.76$, $p = 0.84$, $p = 0.58$, left hemisphere primary auditory, whole brain primary auditory, left dorsal lateral prefrontal, and whole brain cingulo-opercular parcels respectively). With a small sample size ($N = 11$) insignificant p-values are expected. Disregarding the p-values, a Pearson's correlation coefficient

between 0.3 and 0.6 for psychophysical assessments is often considered a correlation worth consideration. Even with this concession, none of the correlation coefficients would be considered compelling.

Visual inspection of the **Figure 6.3** identifies one potential outlier in the primary auditory region. Removing this participant from the analysis for these regions increased the Pearson's correlation coefficient for the left hemisphere from $r = 0.25$ to $r = 0.51$ and for the whole brain from $r = 0.10$ to $r = 0.41$. While still not statistically significant ($p = 0.13$ and $p = 0.41$ for left and whole brain analysis, respectively), with additional participants included in analysis, the relatively high r-value might result in a significant trend.

The two participants with higher pure-tone averages are indicated in **Figure 6.3** (with arrows) to determine if worse hearing predicted neural activity. Because there were only two participants with mild hearing loss, a formal analysis was not conducted. However, a Pearson's correlation was calculated across all 11 participants to establish the relationship between pure-tone average and brain activity. The correlation coefficients were $r = 0.059$ ($p = 0.86$) for left hemisphere auditory regions, $r = 0.26$ ($p = 0.44$) for whole brain auditory regions, $r = -0.22$ ($p = 0.52$) for the left dorsolateral prefrontal parcel, and $r = 0.39$ ($p = 0.24$) for whole brain cingulo-opercular parcels (**Figure 6.4**).

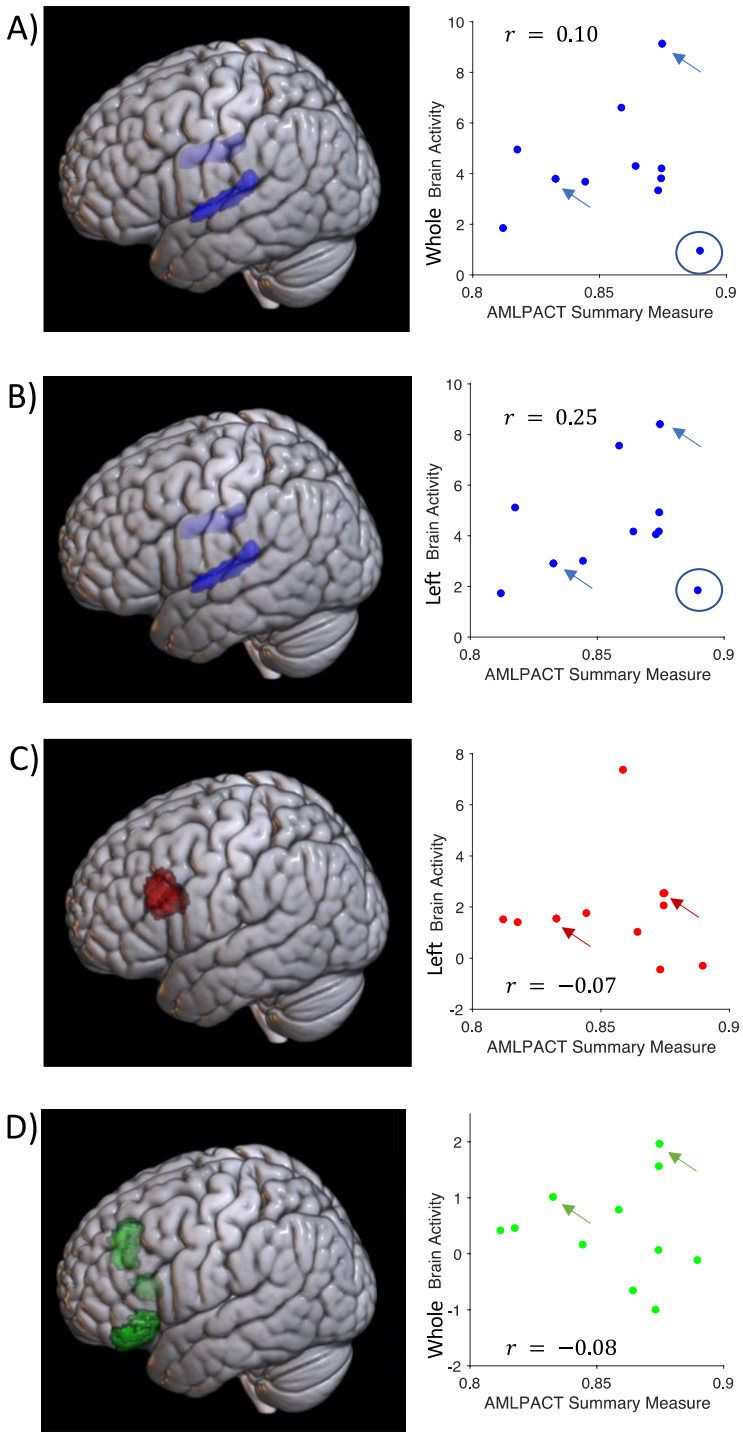

Figure 6. 3: Correlation analysis between individual brain activity and performance on the AMLPACT speech and memory test in regions shown in Figure 6.1. The Pearson's correlation coefficient is recorded for each region of interest. Participants with mild hearing loss (according to their pure-tone averages) are identified with arrows. An outlier was identified in the auditory network and is circled in (A).

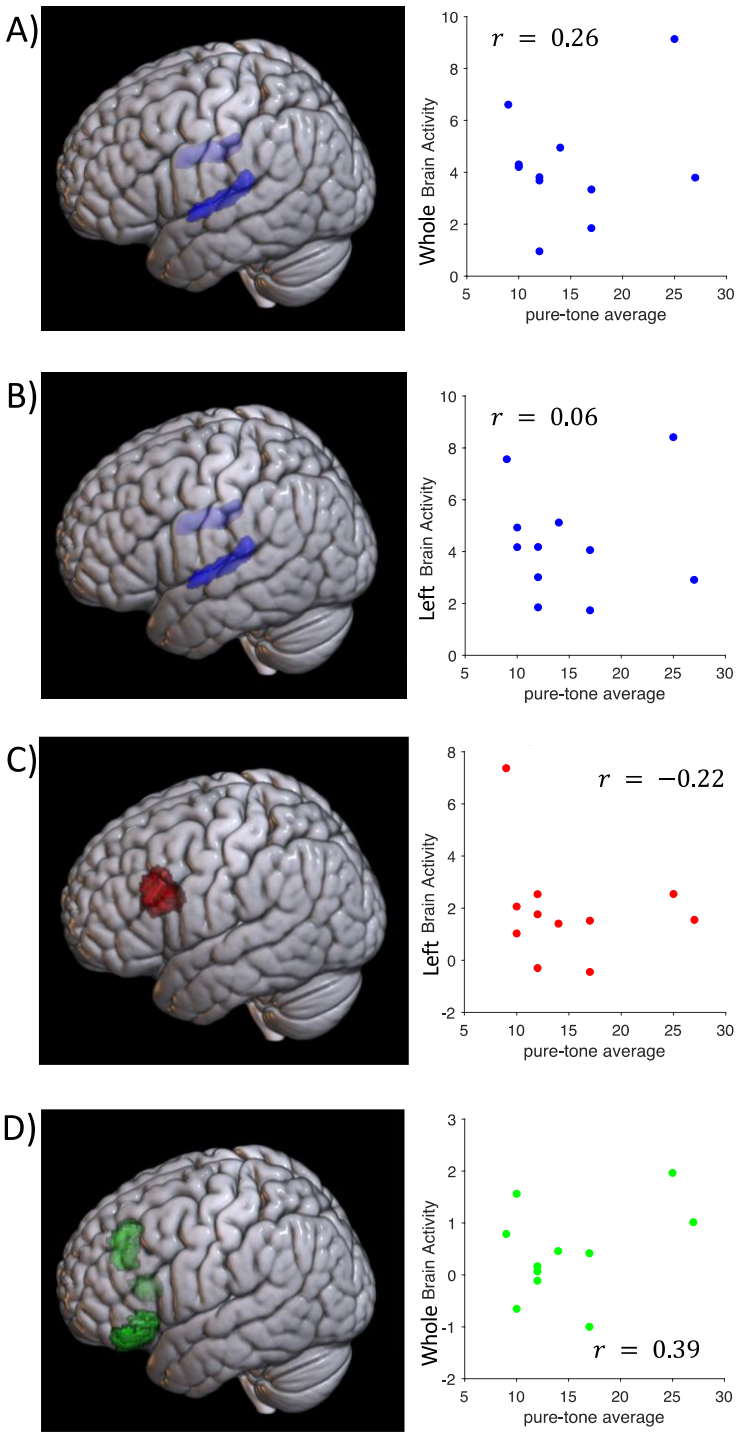

Figure 6. 4: Correlation analysis between individual brain activity and pure-tone averages in regions shown in Figure 6.1. The Pearson's correlation coefficient is recorded for each region of interest.

6.4 Discussion

Performance on AMLPACT did not predict neural activation in the non-auditory regions of the left dorsolateral prefrontal cortex nor the cingulo-opercular network, but it did correlate to activity in the primary auditory cortex when outliers were removed from the comparison.

By presenting words at a large range of SNRs, AMLPACT fully captures the test parameters of the in-scanner task, and in previous analysis (Chapter 5) the summary measure of AMLPACT performance was found to negatively correlate to measures of speech-in-noise ability. Meaning participants with lower speech-in-noise thresholds have higher accuracy on AMLPACT. Participants with lower speech-in-noise thresholds would find the in-scanner assessment less challenging and may rely more fully on core speech networks as opposed to recruiting non-auditory neural support. Successful speech perception results in activation in the primary auditory cortex (Davis & Johnsrude, 2003; Kuchinsky et al., 2016; Narain et al., 2003; Scott Blank, Catrin, Rosen, Stuart, and Wise, Richard J.S., 2000). In this study, only correct trials of the in-scanner test were included in analysis; therefore, it was AMLPACT performance was expected to show a positive correlation with the primary auditory cortex. Removing the outlier, this correlation is robust for whole brain and left hemisphere analysis.

Activity in the left dorsolateral prefrontal cortex did not have high variance, and so it is not unexpected that a significant correlation between that region of interest and AMLPACT performance was not found. One reason for the lack of variability in this region during the in-scanner test might be that most adults in this study had normal hearing or, at most, very slight hearing loss. Previous research has shown a correlation between the degree of hearing loss and increased activation in non-auditory regions that might compensate for decreased activity in

primary auditory regions (Peelle et al., 2011). For the population included in this study, recruitment of additional cognitive resources might not be able to improve their perceptual processing in this task. Another possibility could be that the specific parcel selected in left dorsolateral prefrontal cortex is not active during the in-scanner speech-in-noise assessment. This region was selected because of its robust activity in N-back assessments, but it is possible other regions in the frontoparietal network are active during the in-scanner test.

The cingulo-opercular regions of interest showed individual variability in brain activity. This result could indicate a set of domain-general resources that are uniquely utilized to support performance. Individual differences in attention and error-monitoring, both necessary for verbal N-back tasks, could be reflected in variations of brain activity and motivated the use of AMLPACT as a behavioral measure linking neural activity to variations in the joint speech and memory test. However, there were no real correlations found. This result might be explained, at least in part, by the inclusion of only correct trials in the analysis presented here. The cingulo-opercular network shows higher activity during incorrect trials compared to correct trials on speech perception tasks (Vaden et al., 2013). The insignificant correlation between AMLPACT performance and cingulo-opercular network activity during the in-scanner test was negative. This result is expected given the role of the cingulo-opercular network in error monitoring and salience. Better performance on the joint speech and memory test should negatively correlate to neural activity during correct trials of the speech-in-noise in-scanner test.

The SNR for the in-scanner test was selected to be challenging to participants without sacrificing task performance. Subjects with reduced auditory perception were expected to rely more heavily on domain-general resources in order to maintain high task performance. Previous research has

shown older adults with hearing loss have reduced activation in auditory regions and increased activity in prefrontal regions (Davis & Johnsrude, 2003, 2007; Peelle et al., 2011). Most participants in this study had normal hearing. The two who did have some mild hearing loss did not show drastic differences in neural activation nor AMLPACT performance. This might be because the hearing loss is so mild that the +3 dB SNR did not present much additional challenge compared to normal hearing participants or because of the small number of participants included in this study. Increased activation in working memory regions and decreased activation in auditory regions in normal hearing older adults has only been shown in low accuracy trials (Wong et al., 2009). Despite insignificant correlations, the cingulo-opercular network did show a positive correlation with pure-tone averages. However, primary auditory cortex showed a positive correlation despite previous research consistently stating a negative correlation between hearing ability and neural activation in primary auditory cortex in older adults (Peelle et al., 2011). If an 'outlier' is hand-picked to be removed from the analysis, the correlation does become negative ($r = -0.256$, $p = 0.476$, Pearson's correlation), but there is little justification in selecting that specific data point to remove, other than pretest hypotheses of trends to be expected in the data.

This study assessed only older adults. Including data from young adults might have revealed age-related differences in neural activation patterns and its relationship to AMLPACT performance. It is anticipated that domain-general brain regions associated with verbal working memory tasks (including those assessed here) would contribute more to challenging speech-in-noise scenarios for older adults, even during highly accurate trials, compared to young adults. Activation in primary auditory cortex would be expected to be reduced in older adults compared to younger

adults (Rogers et al., 2020). However, it should be noted that AMLPACT revealed only small age-related differences in performance so the comparison between cohorts might yield similar results.

Generally, the lack of significant findings might be a result of the limitations of the joint speech and memory test outlined in Chapter 5; namely, AMLPACT appears to reflect verbal working memory ability to a greater extent than speech-in-noise ability. Adaptations to AMLPACT that improve its estimation of speech-in-noise measures may better link performance to in-scanner speech-in-noise assessments. As it is currently defined, AMLPACT performance might be more correlated to neural activity during an in-scanner N-back assessment. Additionally, the small number of mostly normal hearing participants included in this study might contribute to the lack of significant findings. As more data are processed, AMLPACT performance-related differences could be evident.

One merit of the joint speech and memory test, AMLPACT, is its ability to assess competing demands on shared neural resources. Future research should focus on validating an AMLPACT test design that equally measures a participant's cognitive and perceptual ability in the domains of interest. Once validated, delivering AMLPACT to participants in the scanner would directly measure individual neural strategies during tasks that better reflect real-world demands.

6.5 Concluding Remarks

Individual differences in the allocation of neural resources during tasks with competing cognitive demands may help explain the interplay between age, cognitive function, and hearing ability. This study examined the connection between individual differences in AMLPACT performance

outside of an MRI scanner and neural activity during a speech-in-noise test in the scanner. The small sample size and possible limitations of the current AMLPACT test design contributed to insignificant correlations found. However, as AMLPACT is adjusted and further de-risked, it could eventually facilitate the building of individual models of complex human behavior to be used in the scanner.

Chapter 7: Summary and Future Directions

7.1 Summary of Findings

This thesis demonstrated the feasibility and utility of a flexible, multidimensional machine learning framework for individual assessments. Chapter 3 evaluated this framework in the four-dimensional input domain of bilateral audiogram estimation (left and right ears, in intensity and frequency). The GP framework was able to estimate hearing ability in left and right ears with the same amount of tone deliveries as it takes to estimate hearing ability in one ear with traditional methods. The increase in efficiency is a result of the framework's ability to exploit shared information across similar domain spaces and to implement active learning techniques to optimize data collection.

Building towards more complex models capable of individual inference in one sitting, the framework was evaluated for dynamically masked audiogram acquisition. Masking represents a complex perceptual test and requires individual-specific customization of a time-consuming protocol in order to achieve accurate threshold estimates for every individual. The GP framework provides a solution to this dilemma for every individual, regardless of hearing ability, and accurately and efficiently models even the most complex hearing abilities with one test. In Chapter 4, the GP framework dynamically masked all audiograms to ascertain true threshold estimates as quickly as unmasked threshold estimation in symmetric hearing individuals, but with substantial efficiency gains in individuals for whom masking represents a significant increase in test time.

Having validated the GP framework for multidimensional, individualized assessment of complex perceptual tests, the framework was extended to assess cognitive and perceptual domains in Chapter 5. The first assessment implemented was a joint speech and memory test. The goal was to successfully model the two dimensional, cognitive and perceptual domain (defined by SNR and memory load). The joint estimator successfully predicted independent measures of speech-in-noise and working memory ability for young and older adults. One advantage of the GP framework is its ability to capture non-linearities and variable interactions as the test is being administered. In the applications tested here for speech and memory, no substantial interactions were revealed. However, the GP framework successfully modeled all trends as data were collected, linear or otherwise. Traditional methods, on the other hand, are limited to constructing models after data are collected and must systematically add predictor variables to develop more complex models. In the behaviors estimated in this thesis the GP framework was able to provide as much inference as a traditional linear regression model given the amount of data collected, with the added advantage of leveraging an active learning technique to optimize queries.

In Chapter 6, performance on the joint speech and memory test was compared to individual neural activity during an in-scanner speech-in-noise test. Activation of non-auditory regions during noisy speech comprehension is thought to support performance in older adults. No significant correlations were found. This might be due to small sample sizes, limitations or discrepancies in the joint test compared to the in-scanner test, or the homogeneous hearing ability of the individuals included in the study. Regardless, future research could mitigate some of these obstacles by administering the joint test in the scanner for a more direct measurement of competing cognitive demands.

The GP framework demonstrates that multidimensional models capable of individual inference does not have to equate with increased test times or a loss in accuracy. By integrating advancements in technology, machine learning, and neuroscience, it is possible to model individual behavior in one assessment. This degree of flexible, efficient, and detailed measurement is not practical with current methods.

7.2 Future Directions

This thesis provides a foundation for a variety of future research questions. In audiometry applications, further testing of bilateral, masked AMLAG in hearing loss populations is needed. AMLAG could also be extended to incorporate conjoint ipsilateral masking. Ipsilateral masking would allow hearing capability to potentially be assessed dynamically in suboptimal acoustic environments.

As was discussed in Chapter 5, the test design of AMLPACT should be further explored. The current implementation provided a measure of speech comprehension ability but would benefit from a model that more fully estimates the speech-in-noise domain. Once confident test designs are configured, AMLPACT can be used to explore the effects of disease, hearing loss, or even speech structure. As previously noted, directly assessing hearing and working memory ability provides a unique opportunity to investigate two of the most predictive measures of cognitive decline. In the one participant tested in Chapter 5 with independently verified cognitive decline, AMLPACT shows promise in evaluating early signs of the disorder even before declines in individual performance on standalone assessments are evident.

AMLPACT is an intriguing option for in-scanner assessment. Since it does not require input or scoring from a test administrator, it presents a feasibility not found in current speech-in-noise tests. A complex test combining multiple cognitive and perceptual dimensions in one scan maximizes the limited and expensive resource of fMRI scanning. Imaging procedures actively joined with AMLPACT assessments will offer a more thorough understanding of how the brain handles challenging cognitive and perceptual tasks that contend for exhaustible resources. Determining how neural resources are allocated in healthy and diseased aging can help create a more accurate framework of lifespan adaptations and lead to earlier diagnosis or interventions.

Individual differences from one test to another might have significant insight to a patient's changing health. Using an original test result or an average healthy result as a prior could allow an active machine learning algorithm to quickly assess if a patient has deviated from past behavior or is outside of an acceptable healthy range. Preliminary research from the Barbour lab has shown that determining hearing categorization can be achieved in only a handful of informative probe tones. This has the potential to dramatically reduce routine screening time in most individuals. Similarly, a participant's N-back performance could be incorporated into the AMLPACT model as a prior distribution. The current implementation of AMLPACT uses a linear regression approach. Because GP linear regression models the entire domain, observations were chosen by the framework to evenly sample the domain. An informative prior would allow greater optimization of test observations, likely improving the predictive power of the GP and better modeling true performance.

All applications of the GP framework delivered a predetermined number of queries to each participant. Developing appropriate stopping criteria will further enable individualization of

assessments. The GP framework has built-in capabilities to offer measures of model variance. These could be exploited to halt assessments once sufficient data have been collected to enable individual inference. Effective stopping criteria must be concurrently developed with capabilities to reconcile participant lapses. To date, lapses have been overcome by additional data collection. As more data are observed, individual discrepancies carry less weight and the GP is able to determine the true estimate. When fewer data are observed, each observation can significantly alter the model, and overly confident estimates are common in early iterations of data collection.

7.3 Concluding Remarks

This thesis applied the GP framework to a broad set of applications in audiometry, speech comprehension, working memory, and neural activity. The generalized nature of the framework and its extensive use of kernel methods enables this large degree of flexibility. As demonstrated here, this framework can be extended to a variety of cognitive and perceptual domains by simply adjusting the GP definitions. In summary, not only are multidimensional, individual assessments practical, but they provide the more informative inference, often in less time than standard approaches.

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twent experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*.

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics*, *8*(1), 1–12.

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

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). Reversal of age-related neural timing delays with training. *Proceedings of the National Academy of Sciences*, *110*(11), 4357–4362.

Anderson, Samira, White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hearing Research*, *300*(June), 18–32.

Aneshensel, C. S. ., Ko, M. J. ., Chodosh, J., & Wight, R. G. (2012). The Urban Neighborhood and Cognitive functions in late middle age. *Health (San Francisco)*, *52*(2), 163–179.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851.

Baddeley, A. (1986). Working memory. In *Working memory.* New York, NY, US: Clarendon Press/Oxford University Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(October), 829–839.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Barbey, A. K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *49*(5), 1195–1205.

Barbour, D. L. (2019). Precision medicine and the cursed dimensions. *Npj Digital Medicine*, *2*(1), 4.

Barbour, D. L., DiLorenzo, J. C., Sukesan, K. A., Song, X. D., Chen, J. Y., Degen, E. A., … Garnett, R. (2019). Conjoint psychometric field estimation for bilateral audiometry. *Behavior Research Methods*, *51*(3), 1271–1285.

Barbour, D. L., Howard, R. T., Song, X. D., Metzger, N., Sukesan, K. A., DiLorenzo, J. C., … Heisey, K. L. (2019). Online Machine Learning Audiometry. *Ear and Hearing*, *40*(4), 918–926.

Barch, D. M., Braver, T. S., Akbudak, E., Conturo, T., Ollinger, J., & Snyder, A. (2001). Anterior Cingulate Cortex and Response Conflict: Effects of Response Modality and Processing Domain. *Cerebral Cortex, 11*(9), 837–848.

Bayes, T., & Price, R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Billings, C. J., Grush, L. D., & Maamor, N. (2017). Acoustic change complex in background noise: phoneme level and timing effects. *Physiological Reports*, *5*(20), e13464.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Reseach*, *8*, 135–160.

Brännström, K. J., & Lantz, J. (2010). Interaural attenuation for Sennheiser HDA 200 circumaural earphones. *International Journal of Audiology*, *49*(6), 467–471.

Braver, T., Cohen, J., Nystrom, L., Jonides, J., Smith, E., & Noll, D. (1997). A Parametric Study of Prefrontal Cortex Involvement in Human Working Memory. *NeuroImage*, *5*, 49–62.

Braver, T., & West, R. (2008). Working Memory, Executive Control and Aging. In *The Handbook of Aging and Cognition* (pp. 311–372).

Brungart, D. S., Sheffield, B. M., & Kubli, L. R. (2014). Development of a test battery for evaluating speech perception in complex listening environments. *The Journal of the Acoustical Society of America*, *136*(2), 777–790.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Cacace, A. T., & McFarland, D. J. (2013). Factors Influencing Tests of Auditory Processing: A Perspective on Current Issues and Relevant Concerns. *Journal of the American Academy of Audiology*, *24*, 572–589.

Carhart, R., & Jerger, J. F. (1959). Preferred Method For Clinical Determination Of Pure-Tone Thresholds. *Journal of Speech and Hearing Disorders*, *24*(4), 330–345.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823.

Chein, J. M., & Fiez, J. A. (2010). Evaluating Models of Working Memory Through the Effects of Concurrent Irrelevant information. *J Exp Psychol Gen*, *139*(1), 117–137.

Chen, H. C., Vaid, J., Boas, D. A., & Bortfeld, H. (2011). Examining the phonological neighborhood density effect using near infrared spectroscopy. *Human Brain Mapping*, *32*(9), 1363–1370.

Cole, M. W., Yarkoni, T., Repovš, G., Anticevic, A., & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *Journal of Neuroscience*, *32*(26), 8988–8999.

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology*, *8*(SEP), 1–16.

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (pp. 62–101). Cambridge: Cambridge University Press.

Cusack, R., Vicente-Grabovetsky, A., Mitchell, D. J., Wild, C. J., Auer, T., Linke, A. C., & Peelle, J. E. (2015). Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. *Frontiers in Neuroinformatics*, *8*, 90.

D'Esposito, M., Aguirre, G. K., Zarahn, E., Ballard, D., Shin, R. K., & Lease, J. (1998). Functional MRI studies of spatial and nonspatial working memory. *Brain Research. Cognitive Brain Research*, *7*(1), 1–13.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, *3*(4), 422–433.

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical Processing in Spoken Language Comprehension. *Journal of Neuroscience*, *23*(8), 3423–3431.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*, 132–147.

DeCaro, R., Peelle, J. E., Grossman, M., & Wingfield, A. (2016). The Two Sides of Sensory – Cognitive Interactions : Effects of Age, Hearing Acuity, and Working Memory Span on Sentence Comprehension. *Frontiers in Psychology*, *7*(236).

Denes, P., & Naunton, R. F. (1951). Masking in Pure-tone Audiometry. *Proceedings of the Royal Society of Medicine*, 790–794.

Dubno, J. R., Eckert, M. A., Lee, F. S., Matthews, L. J., & Schmiedt, R. A. (2013). Classifying human audiometric phenotypes of age-related hearing loss from animal models. *Journal of the Association for Research in Otolaryngology*, *14*, 687–701.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences*.

Edgerton, B. J., & Klood, D. A. (1977). Occlusion effect in bone conduction pure tone and speech audiometry. *Journal of the American Auditory Society*, *2*(4), 151–158.

Edmister, W. B., Talavage, T. M., Ledden, P. J., & Weisskoff, R. M. (1999). Improved auditory cortex imaging using clustered volume acquisitions. *Human Brain Mapping*, *7*(2), 89–97.

Egan, J. P. (1948). Articulation Testing Methods. *Laryngoscope*, *58*, 955–991.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–1335.

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2013). The Brain Dynamics of Rapid Perceptual Adaptation to Adverse Listening Conditions. *The Journal of Neuroscience*, *33*(26), 10688–10697.

Etymotic Research. (2001). *QuickSIN Speech in Noise Manual*. Elk Grove Village, IL: Research Etymotic.

Fechner, G. (1860). *Elements of psychophysics* (and W. Holt, Rinehart, Ed.). New York.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Broca's area. *Current Biology*, *22*(21), 2059–2062.

Fletcher, H. (1929). *Speech and Hearing*. New York: Van Nostrand.

Foo, C., Rudner, M., Rönnberg, J., & Lunner, T. (2007). Recognition of Speech in Noise with New Hearing Instrument Compression Release Settings Requires Explicit Cognitive Storage and Processing Capacity. *Journal of the American Academy of Audiology*, *18*, 618–631.

Füllgrabe, C., & Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology*, *7*(1268).

Gajewski, P. D., Hanisch, E., Falkenstein, M., Thönes, S., & Wascher, E. (2018). What Does the n-Back Task Measure as We Get Older? Relations Between Working-Memory Measures and Other Cognitive Functions Across the Lifespan. *Frontiers in Psychology*, *9*, 1-17e.

Garnett, R., Osborne, M. A., & Hennig, P. (2013). *Active Learning of Linear Embeddings for Gaussian Processes*.

Gates, G. A., Anderson, M. L., McCurry, S. M., Feeney, M. P., & Larson, E. B. (2011). Central auditory dysfunction as a harbringer of Alzheimer's dementia. *Archives of Otolaryngology - Head and Neck Surgery*, *137*(4), 390–395.

Gates, G. a, Anderson, M. L., Feeney, M. P., Susan, M., & Larson, E. B. (2008). Central Auditory Dysfunction in Older People with Memory Impairment or Alzheimer's Dementia. *Arch Otolaryngo Head Neck Surg.*, *134*(7), 771–777.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489.

Gordon, E. M., Laumann, T. O., Adeyemo, B., & Petersen, S. E. (2017). Individual Variability of the System-Level Organization of the Human Brain. *Cerebral Cortex (New York, N.Y. : 1991)*, *27*(1), 386–399.

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., … Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron*, *95*(4), 791-807.e7.

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., … Margulies, D. S. (2015). NeuroVault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*, 8.

Grady, C. (2013). Trends in Neurocognitive Aging. *Nat Rev Neurosci*, *13*(7), 491–505.

Gumus, N. M., Gumus, M., Unsal, S., Yuksel, M., & Gunduz, M. (2016). Examination of Insert Ear Interaural Attenuation (IA)Values in Audiological Evaluations. *Clinical and Investigative Medicine. Medecine Clinique et Experimentale*, *39*(6), 27507.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., … Bowtell, R. W. (1999). "sparse" temporal sampling in auditory fMRI. *Human Brain Mapping*, *7*(3), 213–223.

Hamil, T. . (2016). *Making Masking Manageable*. North Charleston, SC: CreateSpace Independent Publishing Platform.

Hancock, A. B., LaPointe, L. L., Stierwalt, J. A. G., Bourgeois, M. S., & Zwaan, R. A. (2007). Computerized Measures of Verbal Working Memory Performance in Healthy Elderly Participants. *Contemporary Issues in Communication Science and Disorders*, *34*(Fall), 73–85.

Harris, K. C., Dubno, J. R., Keren, N. I., Ahlstrom, J. B., & Eckert, M. A. (2009). Speech recognition in younger and older adults: a dependency on low-level auditory cortex. *J Neurosci.*, *29*(19), 6078–6087.

Heinrich, A., Schneider, B. A., & Craik, F. I. M. (2008). Investigating the influence of continuous babble on auditory short-term memory performance. *The Quarterly Journal of Experimental Psychology*, *61*(5), 735–751.

Heisey, K. L., Buchbinder, J. M., & Barbour, D. L. (2018). Concurrent Bilateral Audiometric Inference. *Acta Acustica United with Acustica*, *104*(5), 762–765.

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

Honey, G. D., Fu, C. H. Y., Kim, J., Brammer, M. J., Croudace, T. J., Suckling, J., … Bullmore, E. T. (2002). Effects of Verbal Working Memory Load on Corticocortical Connectivity Modeled by Path Analysis of Functional Magnetic Resonance Imaging Data. *NeuroImage*, *17*(2), 573–582.

Hood, J. D. (1960). *The Principles and practice of bone conduction audiometry*. *70*(9).

Houlsby, N., Huszar, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian Active Learning for Classification and Preference Learning. *ArXiv Preprint ArXiv:1112.5745*.

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

Humes, L. E., Busey, T. A., Craig, J., & Kewley-port, D. (2013). Are age-related changes in cognitive function driven by age-related changes in sensory processing? *Attention, Perception, and Psychophysics*, *75*(75), 508–524.

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, *7*, 1–16.

Humphries, C., Willard, K., Buchsbaum, B., & Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fMRI study. *NeuroReport: For Rapid Communication of Neuroscience Research*, *12*(8), 1749–1752.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 2–8.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412.

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

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790.

Johnstone, I. M., & Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4237–4253.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., & Koeppe, R. A. (1997). Verbal Working Memory Load Affects Regional Brain Activation as Measured by PET. *Journal of Cognitive Neuroscience*, *9*(4), 462–475.

Kagan, J. (2018). Kinds of individuals defined by patterns of variables. *Development and Psychopathology*, *30*, 1197–1209.

Kane, B. M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin and Review*, *9*, 637–671.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working Memory, Attention Control, and the N-Back Task: A Question of Construct Validity. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(3), 615–622.

Killion, M C, & Niquette, P. A. (2000). What can the pure-tone audiogram tell us about a patient's SNR loss? *The Hearing Journal*, *53*(3), 46–53.

Killion, M C, Wilber, L. A., & Gudmundsen, G. I. (1985). Insert earphones for more interaural attenuation. *Hearing Instruments*, Vol. 36, p. 1:2.

Killion, Mead C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4), 2395–2405.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, Vol. 55, pp. 352–358. US: American Psychological Association.

Kirova, A.-M., Bays, R. B., & Lagalwar, S. (2015). Working Memory and Executive Function Decline across Normal Aging, Mild Cognitive Impairment, and Alzheimer's Disease. *BioMed Research International*, 1–9.

Krecic-Shepard, M. E., Park, K., Barnas, C., Slimko, J., Kerwin, D. R., & Schwartz, J. B. (2000). Race and sex influence clearance of nifedipine: Results of a population study. *Clinical Pharmacology and Therapeutics*, *68*(2), 130–142.

Kuchinsky, S. E., Ahlstrom, J. B., Jr, K. I. V., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*(1), 23–34.

Kuchinsky, S. E., Vaden, K. I. J., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2016). Task-related vigilance during speech recognition in noise for older adults with hearing loss. *Exp Aging Res*, *42*(1), 64–85.

Kwak, S., & Kwak, E. (2007). Auditory Notches in 134 Band Audiograms. *American Academy of Otolaryngology-Head and Neck Surgery*, *137*, P180.

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by "the dress" photograph. *Current Biology*, *25*, R545–R546.

Lamichhane, B., Westbrook, A., Cole, M. W., & Braver, T. S. (2020). Exploring brain-behavior relationships in the N-back task. *NeuroImage*, *212*, 116683.

Landry, J., & Green, W. (1999). Pure-tone audiometric threshold test-retest variability in young and elderly adults. *Journal of Speech-Language Pathology & Audiology*, *23*(2), 74–80.

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., … Petersen, S. E. (2015). Functional system and areal organization of a highly sampled individual brain. *Neuron*, *87*(3), 657–670.

Le Prell, C. G., & Clavier, O. H. (2017). Effects of noise on speech recognition: Challenges for communication by service members. *Hearing Research*, *349*, 76–89.

Lin, F. R., Ferrucci, L., Metter, E. J., An, Y., Zonderman, A. B., & Resnick, S. M. (2011). Hearing Loss and Cognition in the Baltimore Longitudinal Study of Aging. *Neuropsychology*, *25*(6), 763–770.

Liu, C. M., & Lee, C. T. C. (2019). Association of Hearing Loss With Dementia. *JAMA Network Open*, *2*(7), e198112.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology*, *42*, S49–S58.

Magee, M. H., Blum, R. A., Lates, C. D., & Jusko, W. J. (2001). Prednisolone pharmacokinetics and pharmacodynamics in relation to sex and race. *Journal of Clinical Pharmacology*, *41*(11), 1180–1194.

Mahomed, F., Swanepoel, D. W., Eikelboom, R. H., & Soer, M. (2013). Validity of Automated Threshold Audiometry: A Systematic Review and Meta-Analysis. *Ear and Hearing*, *34*(6), 745–752.

Marek, S., & Dosenbach, N. U. F. (2018). The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. *Dialogues in Clinical Neuroscience*, *20*(2), 133–140.

Marek, S., Siegel, J. D., Gordon, E. M., Raut, R. V, Gratton, C., Newbold, D. J., … Dosenbach, N. U. F. (2018). Spatial and Temporal Organization fo the Individual Human Cerebellum. *Neuron*, *100*(4), 977–993.

Margolis, R. H., & Saly, G. L. (2008). Asymmetric hearing loss: Definition, validation, and prevalence. *Otology and Neurotology*, *29*, 422–431.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

Martin, F. N., Armstrong, T. W., & Champlin, C. A. (1994). A Survey of Audiological Practices in the United States. *American Journal of Audiology*, *3*(2), 20–26.

Martin, F. N., & Blosser, D. (1970). Cross hearing - air conduction or bone conduction. *Psychon. Sci.*, *20*(4), 231.

Martin, F. N., Champlin, C. A., & Chambers, J. A. (1998). Seventh Survey of Audiometric Practices in the United States. *Journal of the American Academy of Audiology*, *9*(2), 95–104.

Masterson, E. A., Tak, S., Themann, C. L., Wall, D. K., Groenewold, M. R., Deddens, J. A., & Calvert, G. M. (2013). Prevalence of hearing loss in the United States by industry. *American Journal of Industrial Medicine*, *56*(6), 670–681.

Mattay, V. S., Fera, F., Tessitore, A., Hariri, A. R., Berman, K. F., Das, S., … Weinberger, D. R. (2006). Neurophysiological correlates of age-related changes in working memory capacity. *Neuroscience Letters*, *392*, 32–37.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58*(1), 22–33.

McFarland, D. J. (2017). How neuroscience can inform the study of individual differences in cognitive abilities. *Reviews in the Neurosciences*, *28*(4), 343–362.

Mello, L. A. de, Silva, R. A. M. da, Gil, D., & Ram, S. (2015). Test-retest variability in the pure tone audiometry: comparison between two transducers Variabilidade teste-reteste na audiometria tonal: comparação entre dois transdutores. *Audiol Commun Res. Brazil. Audiol Commun Res*, *2020*(33), 239–45239.

Mencarelli, L., Francesco, N., Davide, M., Arianna, M., Simone, R., Alessandro, R., & Emiliano, S. (2019). Stimuli, presentation modality, and load-specific brain activity patterns during n-back task. *Human Brain Mapping*, *40*(13), hbm.24633.

Meyer, J., Dentel, L., & Meunier, F. (2013). Speech Recognition in Natural Background Noise. *PLoS ONE*, *8*(11), e79279.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information. *Psychological Review*, *63*, 81–97.

Millman, R. E., & Mattys, S. L. (2017). Auditory Verbal Working Memory as a Predictor of Speech Perception in Modulated Maskers in Listeners With Normal Hearing. *Journal of Speech, Language, and Hearing Research*, *60*, 1236–1245.

Monk, A. F., Jackson, D., Nielsen, D., Jefferies, E., & Olivier, P. (2011). N-backer: An auditory n-back task with automatic scoring of spoken responses. *Behavior Research Methods*, *43*(3), 888–896.

Moore, D. R., Edmondson-Jones, M., Dawes, P., Fortnum, H., McCormack, A., Pierzycki, R. H., & Munro, K. J. (2014). Relation between speech-in-noise threshold, hearing loss and cognition from 40-69 years of age. *PLoS ONE*, *9*(9), e107720.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human Primary Auditory Cortex: Cytoarchitectonic Subdivisions and Mapping into a Spatial Reference System. *NeuroImage*, *13*(4), 684–701.

Mueller, S., Wang, D., Fox, M. D., Thomas Yeo, B. T., Sepulcre, J., Sabuncu, M. R., … Liu, H. (2013). Individual Variability in Functional Connectivitivy Architecture of the Human Brain. *Neuron*, *77*(3), 586–595.

Munro, K. J., & Contractor, A. (2010). Inter-aural attenuation with insert earphones. *International Journal of Audiology*, *49*(10), 799–801.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269.

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a Left-lateralized Response Specific to Intelligible Speech Using fMRI. *Cerebral Cortex*, *13*, 1362–1368.

National Institute of Health. (2016). Consideration of Sex as a Biological Variable in NIH-funded Research.

Nyberg, L., Dahlin, E., Stigsdotter Neely, A., & BÄckman, L. (2009). Neural correlates of variable working memory load across adult age and skill: Dissociative patterns within the fronto-parietal network. *Scandinavian Journal of Psychology*, *50*, 41–46.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Obleser, J., Wise, R. J. S., Alex Dresner, M., & Scott, S. K. (2007). Functional Integration across Brain Regions Improves Speech Perception under Adverse Listening Conditions. *Journal of Neuroscience*, *27*(9), 2283–2289.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., … Hickok, G. (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. *Cerebral Cortex*, *20*(10), 2486–2495.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716.

Owen, A. M. (1997). The Functional Organization of Working Memory Processes Within Human Lateral Frontal Cortex : The Contribution of Functional Neuroimaging. *European Journal of Neuroscience*, *9*, 1329–1339.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*, 46–59.

Paivio, A. (2014). Intelligence, dual coding theory, and the brain. *Intelligence*, *47*, 141–158.

Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience*, *12*, 61.

Peelle, J. E. (2018). Listening Effort : How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204–214.

Peelle, J. E., Johnsrude, I. S., & Davis, M. H. (2010). Hierarchical processing for speech in human auditory cortex and beyond. *Frontiers in Human Neuroscience*, *4*, 1–3.

Peelle, J. E., Troiani, V., Grossman, M., & Wingfield, A. (2011). Hearing loss in older adults affects neural systems supporting speech comprehension. *J Neurosci.*, *31*(35), 12638–12643.

Peelle, J. E., Troiani, V., Wingfield, A., & Grossman, M. (2010). Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity. *Cerebral Cortex*, *20*, 773–782.

Peelle, J. E., & Wingfield, A. (2016). The neural consequences of age-related hearing loss. *Trends in Neurosciences*, *39*(7), 486–497.

Pichora-fuller, M. K., Schneider, B. A., & Daneman, M. (2006). *How young and old adults listen to and remember speech in noise*. *593*(1995).

Pichora-Fuller, Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *The Journal of the Acoustical Society of America*, *66*(1333).

Postle, B. R. (2006). Working Memory as an Emergent Property of the Mind and Brian. *Neuroscience*, *139*(1), 23–38.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: the MIT Press.

Read, J. C. A. (2015). The place of human psychophysics in modern neuroscience. *Neuroscience*, *296*, 116–129.

Reuter-Lorenz, P. A., Jonides, J., Smith, E. E., Hartley, A., Miller, A., Marshuetz, C., & Koeppe, R. A. (2000). Age differences in the frontal lateralization of verbal and spatial working memory revealed by PET. *Journal of Cognitive Neuroscience*, *12*(1), 174–187.

Robinson, G. S., & Casali, J. (2003). Speech Communications and Signal Detection in Noise. In E. H. Berger, L. H. Royster, J. D. Royster, D. P. Driscoll, & M. Layne (Eds.), *The Noise Manual* (fifth ed., pp. 567–700). Fairfax: American Industrial Hygiene Association.

Rodriguez-Jimenez, R., Avila, C., Garcia-Navarro, C., Bagney, A., Aragon, A. M. de, Ventura-Campos, N., … Palomo, T. (2009). Differential dorsolateral prefrontal cortex activation during a verbal n-back task according to sensory modality. *Behavioural Brain Research*, *205*(1), 299–302.

Rogers, C. S., Jones, M. S., Mcconkey, S., Spehar, B., Engen, K. J. Van, Sommers, M. S., & Peelle, J. E. (2020). Age-related differences in auditory cortex activity during spoken word recognition. *BioRxiv Preprint*, 1–27.

Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology*, *42*(sup1), 68–76.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., … Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, *7*, 1–17.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology*, *47*(sup2), S99–S105.

Rose, L. T., Rouhani, P., & Fischer, K. W. (2013). The science of the individual. *Mind, Brain, and Education*, *7*(3), 152–158.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., … Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage*, *60*(1), 830–846.

Rudner, M., Foo, C., Rönnberg, J., & Lunner, T. (2007). Phonological mismatch makes aided speech recognition in noise cognitively taxing. *Ear and Hearing*, *28*(6), 879–892.

Rudner, M., Rönnberg, J., & Lunner, T. (2011). Working Memory Supports Listening in Noise for Persons with Hearing Impairment. *J Am Acad Audiol*, *22*, 156–167.

Sadaghiani, S., & D'Esposito, M. (2015). Functional Characterization of the Cingulo-Opercular Network in the Maintenance of Tonic Alertness. *Cerebral Cortex*, *25*(9), 2763–2773.

Sanders, J. W., & Rintelmann, W. F. (1964). Masking in Audiometry. *Archives of Otolaryngology*, *80*, 541–556.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., … Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114.

Scott Blank, Catrin, Rosen, Stuart, and Wise, Richard J.S., S. K. (2000). Identification of a pathway for intelligible speech in the left temporal lob. *Brain*, *123*, 2400–2406.

Shojaeemend, H., & Ayatollahi, H. (2018). Automated audiometry: A review of the implementation and evaluation methods. *Healthcare Informatics Research*, *24*(4), 263–275.

Siegel, J. S., Power, J. D., Dubis, J. W., Vogel, A. C., Church, J. A., Schlaggar, B. L., & Petersen, S. E. (2014). Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Human Brain Mapping*, *35*(5), 1981–1996.

Sklare, D., & Denenberg, L. (1987). Interaural attenuation for Tubephoen. *Ear and Hearing*, Vol. 8, pp. 298–300.

Smith, C. R. (1968). Clinical Masking During Pure Tone Audiometry. *Archives of Otolaryngology*, *88*, 169–170.

Smith, S. L., & Pichora-Fuller, M. K. (2015). Associations between speech understanding and auditory and visual tests of verbal working memory: effects of linguistic complexity, task, age, and hearing loss. *Frontiers in Psychology*, *6*(September), 1–15.

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

Song, X. D., Wallace, B. M., Gardner, J. R., Ledbetter, N. M., Weinberger, K. Q., & Barbour, D. L. (2015). Fast, Continuous Audiogram Estimation using Machine Learning. *Ear and Hearing*, *36*(6), e326-e335.

Spyridakou, C., & Bamiou, D.-E. (2015). Need of speech-in-noise testing to assess listening difficulties in older adults. *Hearing, Balance and Communication*, *13*(2), 65–76.

Stuart, A., Stenstromb, R., Tompkins, C., & Vandenhoff, S. (1991). Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults. *International Journal of Audiology*, *30*(2), 82–90.

Studebaker, G. A. (1964). Clinical Masking of Air- and Bone-Conducted Stimuli. *Journal of Speech and Hearing Disorders*, *29*(1), 23–35.

Swanepoel, D. W., Mngemane, S., Molemong, S., Mkwanazi, H., & Tutshini, S. (2010). Hearing Assessment—Reliability, Accuracy, and Efficiency of Automated Audiometry. *Telemedicine and E-Health*.

Taylor, B. (2003). Speech in noise tests: How and why to include them in your basic test battery. *The Hearing Journal*, *56*(1), 40–44.

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165.

Tun, P. A., Mccoy, S., & Wingfield, A. (2009). Aging, Hearing Acuity, and the Attentional Costs of Effortful Listening. *Psychology and Aging*, *24*(3), 761–766.

Turner, R. G. (2004a). Masking redux. I: An optimized masking method. *Journal of the American Academy of Audiology*, *15*(1), 17–28.

Turner, R. G. (2004b). Masking Redux II: A Recommended Masking Protocol. *J Am Acad Audiol*, *15*(1), 29–46.

Vaden, K. I. J., Kuchinsky, S. E., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2015). Cortical Activity Predicts Which Older Adults Recognize Speech in Noise and When. *Journal of Neuroscience*, *35*(9), 3929–3937.

Vaden, K. I. J., Kuchinsky, S. E., Ahlstrom, J. B., Teubner-Rhodes, S., Dubno, J. R., & Eckert, M. A. (2016). Cingulo-Opercular Function during Word Recognition in Noise for Older Adults with Hearing Loss. *Exp Aging Res*, *42*(1), 67–82.

Vaden, K. I. J., Kuchinsky, S. E., Cute, S. L., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2013). The Cingulo-Opercular Network Provides Word-Recognition Benefit. *The Journal of Neuroscience*, *33*(48), 18979–18986.

Vaden, K. I. J., Teubner-Rhodes, S., Ahlstrom, J. B., Dubno, J. R., & Eckert, M. A. (2017). Cingulo-opercular activity affects incidental memory encoding for speech in noise. *NeuroImage*, *157*, 381–387.

Valente, M. (2009). *Pure-tone audiometry and masking*. Plural Publishing.

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6.

Wallis, S. A., Baker, D. H., Meese, T. S., & Georgeson, M. A. (2013). The slope of the psychometric function and non-stationarity of thresholds in spatiotemporal contrast vision. *Vision Research*, *76*, 1–10.

Wallisch, P. (2017). Illumination assumptions account for individual differences in the perceptual interpretation of a profoundly ambiguous stimulus in the color domain: "The dress." *Journal of Vision*, *17*(4), 1–14.

Wang, M., Yang, P., Wan, C., Jin, Z., Zhang, J., & Li, L. (2018). Evaluating the Role of the Dorsolateral Prefrontal Cortex and Posterior Parietal Cortex in Memory-Guided Attention With Repetitive Transcranial Magnetic Stimulation. *Frontiers in Human Neuroscience*, *12*, 236.

Ward, C. M., Rogers, C. S., Engen, K. J. Van, & Peelle, J. E. (2016). Effects of age, acoustic challenge, and verbal working memory on recall of narrative speech. *Exp Aging Res*, *42*(1), 97–111.

Wayne, R. V., Hamilton, C., Huyck, J. J., & Johnsrude, I. S. (2016). Working memory training and speech in noise comprehension in older adults. *Frontiers in Aging Neuroscience*, *8*(49), 1–15.

Wayne, R. V., & Johnsrude, I. S. (2015). A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. *Ageing Research Reviews*, *23*, 154–166.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening : The Processing of Degraded Speech Depends Critically on Attention. *The Journal of Neuroscience*, *32*(40), 14010–14021.

Wingfield, A., & Grossman, M. (2006). Language and the Aging Brain : Patterns of Neural Compensation Revealed by Functional Brain Imaging. *J Neurophysiol*, *96*, 2830–2839.

Wingfield, A., Mccoy, S. L., Peelle, J. E., Tun, P. A., & Cox, L. C. (2006). Effects of Adult Aging and Hearing Loss on Comprehension of Rapid Speech Varying in Syntactic Complexity. *J Am Acad Audiol*, *17*, 487–497.

Wingfield, A., Tun, P. A., & Mccoy, S. L. (2005). Hearing Loss in Older Adulthood. What It Is and How It Interacts With Cognitive Performance. *Current Directions in Psychological Science*, *14*(3), 144–148.

Wong, P. C. M., Jin, J. X., Gunasekera, G. M., Abel, R., Lee, E. R., & Dhar, S. (2009). Aging and cortical mechanisms of speech perception in noise. *Neuropsychologia*, *47*(3), 693–703.

Yacullo, W. (2015). Clinical masking. In J. Katz (Ed.), *Handbook of clinical audiology.* (pp. 77–111). Philadelphia, PA: Wolters Kluwer Health.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155.