

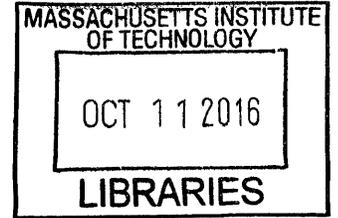
Neural and perceptual correlates of closed-loop sensorimotor training: basic and applied studies

by

Jonathon Whitton

B.A., Northern Kentucky University

Au.D., University of Louisville



ARCHIVES

Submitted to the Department of Health Sciences and Technology in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

© 2016 Massachusetts Institute of Technology. All rights reserved

Signature redacted

Signature of Author.....

Department of Health Sciences and Technology

Aug 29, 2016

Signature redacted

Certified by....

Daniel Polley, Ph.D.

Associate Professor of Otolaryngology, Harvard Medical School

Signature redacted

Accepted by.....

Emery N. Brown, MD, Ph.D.

Director Harvard-MIT Program in Health Sciences and Technology Professor of Computational Neuroscience and Health Sciences and Technology

Neural and Perceptual Correlates of Closed-Loop Sensorimotor Training: Basic and Applied Studies

by

Jonathon Whitton

Submitted to the Department of Health Sciences and Technology
on Aug 29, 2016 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Health Sciences and Technology

ABSTRACT

The global hearing healthcare field is faced with two principal challenges. First, the demand for basic audiometric testing services far exceeds the capacity of trained clinicians even in high income countries, and this supply/demand mismatch is expected to worsen secondary to population aging. Next, once patients are identified as having a hearing loss, the treatments that are provided (hearing aids) do not sufficiently address their primary complaint, namely that they have trouble hearing in noisy environments. To begin to address the first problem, we executed a proof-of-concept study to ask whether mobile consumer electronics could be used to replace manually performed clinic-based testing with self-directed hearing measurements from home. We found that self-administered home hearing measurements were largely equivalent to standard clinical measures. To begin to address the second problem (hearing in noise challenges of patients), we performed three additional experiments. Inspired by promising findings of enhanced visual attention following action videogame training, we developed a closed-loop *audiomotor* training application and asked if playing a game that focused on tone in noise discriminations would provide generalized benefit for speech recognition in noise abilities. In young normally hearing adults, closed-loop training for one month provided a 12 percentage point improvement in speech understanding in noise scores. Next, we recruited older adults who wore hearing aids to play a similar closed-loop training game and observed a 10 percentage point enhancement of speech recognition in noise abilities secondary to gameplay, suggesting that this training could be coupled with standard treatments to improve patient outcomes. Finally, we studied the neurophysiological correlates of audiomotor signal in noise training in a rodent model, where we observed enhanced resistance to noise suppression in auditory cortical neurons following three months of training, perhaps contributing to the perceptual benefits that we observed in human subjects.

Thesis Supervisor: Daniel Polley, Ph.D.

Title: Associate Professor of Otolaryngology, Harvard Medical School

Table Contents

1. Chapter 1: Introduction	
1.1. Limited access to basic services.....	4
1.2. The potential of mobile technology use to increase efficiency of hearing assessment.....	4
1.3. The challenge of hearing in noisy environments for individuals living with hearing loss.....	5
1.4. Limitations of hearing aid to provide perceptual benefit in noisy environments.....	6
1.5. Potential rehabilitative uses of sensory learning protocols to improve speech in noise perception.....	7
1.6. References.....	10
2. Chapter 2: Validation of a self-administered audiometry application: An equivalence study	
2.1. Introduction.....	16
2.2. Materials and Methods.....	18
2.3. Results.....	23
2.4. Discussion and Conclusion.....	26
2.5. Figures.....	29
2.6. References.....	32
2.7. Supplemental Materials.....	36
3. Chapter 3: Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise	
3.1. Introduction.....	55
3.2. Materials and Methods.....	57
3.3. Results.....	63
3.4. Discussion.....	75
3.5. Figures.....	81
3.6. References.....	87
3.7. Supplemental Materials.....	97
4. Chapter 4: Closed-loop audiomotor training enhances the perception of speech in noise: A randomized double-blinded placebo-controlled trial	
4.1. Introduction.....	106
4.2. Materials and Methods.....	108
4.3. Results.....	124
4.4. Discussion.....	134
4.5. Figures.....	141
4.6. References.....	148
4.7. Supplemental Materials.....	165

Chapter 1 Introduction

1.1 Limited access to basic hearing services

Hearing loss is the leading cause of moderate to severe disability worldwide and ranks as the sixth leading cause of burden of disease (1) in high income countries. As an age-related impairment, most of those affected by hearing loss are older adults (2), a demographic that is expected to increase by 60% in the US over the next two decades (3). Thus, the assessment and treatment of hearing loss represent major public health concerns. The shortfall in available services is a global challenge, and even a high-income country with established hearing care infrastructure like the US predicts a two-fold separation between supply of trained clinicians and demand for basic hearing services in the coming decades (4). Logically these public health challenges point towards the need for more efficient ways to assess hearing.

1.2 The potential of mobile technology use to increase efficiency of hearing assessment

Behavioral measurements of hearing are currently performed manually by clinicians in sound treated rooms. For this reason, the number of basic hearing tests that can be performed are limited by *i.*) available clinician hours for basic hearing tests and *ii.*) available sound treated rooms equipped with specialized hardware. To address the hearing assessment bottleneck an approach must decouple the auditory measurements from clinician time and specialized facilities. Behavioral hearing tests follow standardized protocols, making them amenable to automation (5). In fact, Georg von Békésy automated hearing testing nearly 70 years ago (6). Nonetheless, automated

assessments have not been used in clinical assessment of hearing (4). Contemporary smartphones and tablet computers possess sufficient processing power to run sophisticated measurements of hearing. Using consumer electronics to make automated measurements of hearing offers the potential for massive parallel clinical assessments, which would dramatically reduce clinician time and the clinical footprint dedicated to routine procedures. For these reasons, several groups have developed and validated mobile hearing screening applications (7–10). However, while these applications occupy an important space in the hearing health landscape, they do not address the supply demand mismatch since knowing someone failed a hearing test provides limited information for clinical decision making. To address the growing need for audiometric assessments, self-administered mobile hearing tests need to provide test results that are statistically equivalent to the same measurements performed in the clinic environment using established protocols. Mobile hearing assessments have not been previously developed for, or tested with this level of experimental scrutiny; this was the purpose of experiments described in Chapter 2.

1.3 The challenge of hearing in noisy environments for individuals living with hearing loss.

Extracting signals amongst a background of noise is a universal perceptual problem. In industrialized societies, auditory signal and noise challenges abound, where distractors can impede communication with friends at social gatherings, instruction from teachers in classrooms, or transmission of information via a cellular phone (11, 12). However with age, the level of auditory distraction that can be effectively ignored is reduced (13, 14), and a third of middle aged to older adults report great difficulty

following conversations in noisy environments (15). It is not surprising then that most individuals (74%) who are referred to otolaryngologists and audiologists for hearing problems list their primary complaint as difficulty understanding speech in noisy environments (16). So while disorders of hearing can arise from a number of pathological states beginning in middle ear structures and extending to the central nervous systems, the “real world” consequence of this pathophysiology generally manifests as impaired perception in social environments. Hearing aids represent the only available treatment for the most common type of permanent hearing loss, and while amplification devices provide perceptual improvements in quiet environments, they are of limited utility in noisy situations. For this reason, understanding speech in noisy environments remains a primary complaint of individuals living with hearing loss even after treatment with hearing aids (17, 18). In this sense, the current treatments for hearing loss, while valuable as audibility aids, largely miss the point of the most common patient complaint since signals in noisy environments are usually loud already (no need for further amplification, 19) .

1.4 Limitations of hearing aids to provide perceptual benefit in noisy environments

That amplification does not remedy the hearing in noise problems of individuals with sensorineural hearing loss is not surprising given the physiologic changes that are associated with cochlear and primary afferent damage (20). The dysfunction of outer hair cells alters basilar membrane mechanics, resulting in broadened peripheral (21, 22) and central (23) filters. For this reason, ‘signal’ and ‘noise’ information are more likely to interact in the same peripheral channels (24), proving weakened, noisy signals to

central neurons for further processing. Additionally, varying degrees of primary afferent degeneration may further degrade the information available to central neurons for fine temporal coding, which may be important when following signals in modulated noise (25–28). Amplifying an acoustic input that represents the summed mixture of ‘signal’ and ‘noise’ energy will not resolve either of these issues. Furthermore, most individual who are living with hearing loss are older adults (2), and older adults demonstrate an impaired ability to ignore distracting information across sensory modalities (29–31, 14, 32). This age-related deficit in inhibitory control may also contribute to the perceptual difficulty that older adults with hearing loss experience in noisy environments and amplification of the acoustic signal would not be expected to help with this perceptual problem.

1.5 Potential rehabilitative uses of sensory learning protocols to improve speech in noise perception

Perceptual skills can improve with practice, but for decades it was observed that this improvement was highly specific to the demands of the task on which someone trained (33, 34). For example, over the course of several training sessions, the physical difference between two tones carriers that is required for an individual to reliably hear them as different is expected to decrease by 30-40%; however, simply shifting the discrimination test to a higher frequency range will abolish this learned effect (35). Known as the “curse of specificity,” the observed idiosyncrasy of sensory and cognitive learning has historically limited their applied utility in academic, occupational, and health settings (36). Over the past two decades, several elements of training protocols that encourage modest transfer of learning have been identified (37–41), but recent studies

of closed-loop sensorimotor learning have provided evidence that far transfer of learning may be possible with training (36). Specifically, accumulating evidence suggests that playing action video games affords improved visual selective attention abilities across a variety of tasks (42–44). Similar far transfer of learning has also been observed in individuals who have experienced extensive musical training, where professional musicians demonstrate enhanced pitch processing (45, 46) as well as speech recognition in noise abilities (47). While these latter studies of closed-loop audiomotor learning (i.e. musical training) are intriguing when considering the utility of this approach to improve perception for the hearing impaired, as cross-sectional studies, they do not provide causative evidence of the benefits of audiomotor training. Furthermore, both action video games and musical experience are complex tasks (48) and the training protocols are not parametrically controllable, making study of the “active ingredients” of this type of training infeasible. To address issues of causation in audiomotor learning and reduce the number of variables in this type of training task, we developed an audiomotor training game that possessed only two shared features of action video games and professional musicianship: *i.*) the employment of interference resolution in game play, which has been associated with plasticity in attentional control networks (49) and *ii.*) a closed-loop game mechanic; wherein sensory stimuli act as continuous feedback signals to predictive motor commands (50). From a theoretical standpoint, this type of closed-loop mechanic would be expected to efficiently drive neuromodulatory nuclei that are critical for enabling plasticity in adult sensory cortex (51–55). In experiments reported in chapter 3, we tested whether playing the simple audiomotor task would improve the speech recognition in noise abilities of young, normally hearing

adults as well as examining the neurological correlates of training in a rodent model. Based on the positive results of these initial experiments, we then tested whether the closed-loop audiomotor training approach would confer speech recognition in noise benefits to a population with impaired hearing abilities, namely older adults who used hearing aids. This experiment, which is described in chapter 4, was performed in the context of a randomized double-blinded placebo-controlled trial.

1.6 References

1. Organization WH (2008) *The Global Burden of Disease: 2004 Update* (World Health Organization, Geneva).
2. Lin FR, Niparko JK, Ferrucci L (2011) Hearing Loss Prevalence in the United States. *Arch Intern Med* 171(20):1851–1852.
3. US Census Bureau (2014) US Census Bureau 2014 National Population Projections. Available at: <http://www.census.gov/population/projections/data/national/2014/downloadablefiles.html> [Accessed January 1, 2016].
4. Margolis RH, Morgan DE (2008) Automated pure-tone audiometry: an analysis of capacity, need, and benefit. *Am J Audiol* 17(2):109–113.
5. Carhart R, Jerger J (1959) A preferred method for clinical determination of pure-tone thresholds *J Spee Hear Res* 24(4):330–345.
6. von Bekesy G (1947) A new audiometer. *Acta Otolaryngol* 35(5-6):411–422.
7. Yeung J, et al. (2013) The new age of play audiometry: prospective validation testing of an iPad-based play audiometer. *J Otolaryngol Head Neck Surg* 42(1):1-7.
8. Abu-Ghanem S, et al. (2015) Smartphone-based audiometric test for screening hearing loss in the elderly. *Eur Arch Oto-Rhino-Laryngology*. 273(2):333-339.
9. Handzel O, et al. (2013) Smartphone-based hearing test as an aid in the initial evaluation of unilateral sudden sensorineural hearing loss. *Audiol Neurotol*. 18(4): 201-207
10. Swanepoel DW, Myburgh HC, Howe DM, Mahomed F, Eikelboom RH (2014) Smartphone hearing screening with integrated quality control and data management. *Int J Audiol* 53(12):841–849.
11. McDermott JH (2009) The cocktail party problem. *Curr Biol* 19(22):R1024–1027.
12. Bregman AS (1990) *Auditory Science Analysis: The Perceptual Organization of Sound* (The MIT Press, Cambridge).
13. Füllgrabe C, Moore BCJ, Stone MA (2014) Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Front Aging Neurosci* 6:1-25.
14. Tun PA, O’Kane G, Wingfield A (2002) Distraction by competing speech in young and older adult listeners. *Psychol Aging* 17(3):453–467.
15. Davis AC (1989) The prevalence of hearing impairment and reported hearing disability among adults in Great Britain. *Int J Epidemiol* 18(4):911–917.

16. Hind SE, et al. (2011) Prevalence of clinical referrals having hearing thresholds within normal limits. *Int J Aud* 50(10):708-716.
17. Kochkin S (2010) MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *Hear J* 63(1):19–27.
18. Kochkin S (2000) MarkeTrak V: “Why my hearing aids are in the drawer”: The consumers’ perspective. *Hear J* 53(2):34-41.
19. Bentler RA, Duve MR (2000) Comparison of hearing aids over the 20th century. *Ear Hear* 21(6):625–639.
20. Liberman MC, Rosowski JJ, Lewis R (2010) Physiology and Pathophysiology. *Shuknecht’s Pathology of the Ear*, eds Merchant S, Nadol J (Peoples Medical Publishing House USA, Shelton, CT), pp 97-136.
21. Liberman MC, Dodds LW (1984) Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves. *Hear Res* 16(1):55–74.
22. Kiang NYS (1970) Auditory-nerve activity in cats with normal and abnormal cochleas. *Ciba Foundation Symposium on Sensorineural Hearing Loss*, eds Wolstenholme GEW, Knight J (Churchill, London), pp 241–273.
23. de Villers-Sidani E, et al. (2010) Recovery of functional and structural age-related changes in the rat primary auditory cortex with operant training. *Proc Natl Acad Sci U S A* 107(31):13900–13905.
24. Henry KS, Heinz MG (2012) Diminished temporal coding with sensorineural hearing loss emerges in background noise. *Nat Neurosci* 15(10):1362–1364.
25. Makary CA, Shin J, Kujawa SG, Liberman MC, Merchant SN (2011) Age-Related Primary Cochlear Neuronal Degeneration in Human Temporal Bones. *Jaro-Journal Assoc Res Otolaryngol* 12(6):711–717.
26. Furman AC, Kujawa SG, Liberman MC (2013) Noise-induced cochlear neuropathy is selective for fibers with low spontaneous rates. *J Neurophysiol* 110(3):577–586.
27. Kujawa SG, Liberman MC (2009) Adding Insult to Injury: Cochlear Nerve Degeneration after “Temporary” Noise-Induced Hearing Loss. *J Neurosci* 29(45):14077–14085.
28. Moore BCJ (2008) The Role of Temporal Fine Structure Processing in Pitch Perception, Masking, and Speech Perception for Normal-Hearing and Hearing-Impaired People. *Jaro-Journal Assoc Res Otolaryngol* 9(4):399–406.
29. Gazzaley A, Cooney JW, Rissman J, D’Esposito M (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nat Neurosci* 8(10):1298–1300.

30. Hasher L, Zacks RT (1988) Working Memory, Comprehension, and Aging: A Review and a New View. *Psychol Learn Motiv* 22:193–225.
31. Gazzaley A, et al. (2008) Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proc Natl Acad Sci U S A* 105(35):13122–13126.
32. Milham MP, et al. (2002) Attentional control in the aging brain: Insights from an fMRI study of the Stroop task. *Brain Cogn* 49(3):277–296.
33. Ball K, Sekuler R (1982) A specific and enduring improvement in visual motion discrimination. *Science* 218(4573):697–698.
34. Fiorentini A, Berardi N (1980) Perceptual learning specific for orientation and spatial frequency. *Nature* 287(5777):43–44.
35. Wright BA, Fitzgerald MB (2005) Learning and generalization of five auditory discrimination tasks as assessed by threshold changes. *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, eds Pressnitzer D, de Cheveigne A, McAdams S, Collet L (Springer, New York), pp 510–516.
36. Green CS, Bavelier D (2008) Exercising your brain: a review of human brain plasticity and training-induced learning. *Psychol Aging* 23(4):692–701.
37. Ahissar M, Hochstein S (1997) Task difficulty and the specificity of perceptual learning. *Nature* 387(6631):401–406.
38. Hung S-C, Seitz AR (2014) Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *J Neurosci* 34(25):8423–31.
39. Harris H, Gliksberg M, Sagi D (2012) Generalized Perceptual Learning in the Absence of Sensory Adaptation. *Curr Biol* 22(19):1813–1817.
40. Jeter PE, Doshier BA, Liu SH, Lu ZL (2010) Specificity of perceptual learning increases with increased training. *Vision Res* 50(19):1928–1940.
41. Xiao LQ, et al. (2008) Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Curr Biol* 18(24):1922–1926.
42. Green CS, Bavelier D (2006) Enumeration versus multiple object tracking: the case of action video game players. *Cognition* 101(1):217–245.
43. Green CS, Bavelier D (2003) Action video game modifies visual selective attention. *Nature* 423(6939):534–537.
44. Green CS, Bavelier D (2007) Action-video-game experience alters the spatial resolution of vision. *Psychol Sci* 18(1):88–94.
45. Ruggles DR, Freyman RL, Oxenham AJ (2014) Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLoS One* 9(1):1-8.

46. Micheyl C, Delhommeau K, Perrot X, Oxenham AJ (2006) Influence of musical and psychoacoustical training on pitch discrimination. *Hear Res* 219(1-2):36–47.
47. Swaminathan J, et al. (2015) Musical training, individual differences and the cocktail party problem. *Sci Rep* 5:1-10.
48. Green CS, Bavelier D (2012) Learning, Attentional Control, and Action Video Games. *Curr Biol* 22(6):R197–R206.
49. Anguera JA, et al. (2013) Video game training enhances cognitive control in older adults. *Nature* 501(7465):97–101.
50. Whitton JP, Hancock KE, Polley DB (2014) Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise. *Proc Natl Acad Sci U S A* 111(25):E2606–E2615.
51. Parikh V, Kozak R, Martinez V, Sarter M (2007) Prefrontal acetylcholine release controls cue detection on multiple timescales. *Neuron* 56(1):141–154.
52. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593–1599.
53. Steinberg EE, et al. (2013) A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* 16(7):966–973.
54. Letzkus JJ, et al. (2011) A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* 480(7377):331–335.
55. Pi HJ, et al. (2013) Cortical interneurons that specialize in disinhibitory control. *Nature* 503(7477):521–524.

Chapter 2: Validation of a self-administered audiometry application: An equivalence study¹

Abstract

Objective: To compare hearing measurements made at home using self-administered audiometric software against audiological tests performed on the same subjects in a clinical setting

Study Design: Prospective, cross-over equivalence study

Methods: In experiment 1, adults (N = 19) with varying degrees of hearing loss performed air-conduction audiometry, frequency discrimination, and speech recognition in noise testing at home with an automated tablet application and also in sound-treated clinical booths with an audiologist. The accuracy and reliability of computer-guided home hearing tests were compared to audiologist administered tests. In experiment 2, the reliability and accuracy of pure-tone audiometric results was examined in a separate cohort across a wider variety of clinical settings (N = 11).

Results: Remote, automated audiograms were statistically equivalent to manual, clinic-based testing from 500 to 8000 Hz ($P \leq .03$); however, 250 Hz thresholds were elevated when collected remotely. Remote and sound-treated booth testing of frequency discrimination thresholds and speech recognition scores were equivalent ($P \leq 5 \times 10^{-5}$). In a second sample who had received clinic-based testing in a variety of settings (experiment 2), automated, remote testing was equivalent to manual sound booth testing throughout the typical audiometric range ($P \leq .04$).

¹ Data described in this chapter were previously published in Whitton et al (2016) Validation of a self-administered audiometry application: An equivalence study. *Laryngoscope*

Conclusion: These data represent a proof of concept that several self-administered, automated hearing measurements are statistically equivalent to manual measurements made by an audiologist in the clinic. The demonstration of statistical equivalency for these basic behavioral hearing tests points toward the eventual feasibility of monitoring progressive or fluctuant hearing disorders outside of the clinic to increase the efficiency of clinical information collection.

2.1 Introduction

Diagnosing and treating hearing loss represents a formidable public health challenge that is expected to worsen in the near future. An estimated 48 million individuals currently live with hearing loss (HL) in the US, and this number is expected to double over the next 20 years due to changing population demographics (1, 2). Audiological service providers in the US are already in high demand, and the mismatch between service providers and patients who need assessment and treatment is expected to grow to unsustainable levels in the coming decade (2, 3). Widening the lens beyond the US, of the estimated 360 million people who live with disabling HL worldwide, most live in regions where audiological services are scarce (4).

To address the mismatch between supply and demand in the US hearing services market, the efficiency or size of the current cadre of hearing healthcare providers must increase by a factor of 2 over the next decade(2, 3). There is no indication that this gap in patient care will be filled by a doubling of trained personnel(2, 3). Logically, this points toward the central importance of increases in clinical efficiency as the most plausible means to solve the impending supply and demand dilemma. Some savings are likely to be found in developing more efficient standards for the diagnosis and management of HL. Audiologic behavioral tests represent a direct measure of hearing ability and currently require substantial time costs from doctorally trained clinicians because they are performed manually, making this form of information collection a reasonable target for increased efficiency.

Audiologic behavioral tests can be divided into three classes: *i.*) absolute detection thresholds, *ii.*) feature discrimination thresholds and, *iii.*) speech recognition testing. As the “gold standard” test of hearing sensitivity, measurements of absolute detection thresholds (i.e. the pure tone audiogram) are the most commonly performed audiological behavioral tests. The latter two classes of behavioral tests are employed in some cases to obtain supplementary information to pure tone audiograms (e.g. multi-level word recognition testing (5, 6)) but also when characterizing disorders that have little or no correlation with basic audibility, such as auditory neuropathy spectrum disorder and auditory processing disorder(7–14). The administration procedures for all three classes of audiological behavioral tests are standardized to ensure comparable results across testing facilities, making them amenable to automation and improved clinical efficiency (3, 15, 16).

Margolis and Morgan(3) estimated that by automating 80% of pure tone audiograms currently performed by clinicians, each audiologist could see an additional 139 patients annually (1.5 million patients total), which would only partially close the gap between supply and demand in the US for basic hearing services. But the Margolis and Morgan capacity estimates were based on the need for testing to be performed in available sound treated booths with specialized equipment. The processing power of smartphones and tablet computers that are now ubiquitous in both developed and emerging nations (17, 18) make it feasible to distribute applications that could perform sophisticated, automated audiological testing with consumer grade hardware outside of sound treated booths. If remote and clinic-based audiology tests were ever to be considered interchangeable for the eventual purposes of differential diagnosis or

monitoring response to treatment, it must first be established whether remote and clinic-based tests are statistically equivalent. This level of rigorous quantitative scrutiny has not been applied to automated, remote hearing assessment (19–26). This was the purpose of the present study.

We programmed an interactive application for tablet computers (Surface Pro 2, Windows 8.1 operating system). Using consumer-grade headphones, this application provided a means to measure pure-tone audiograms, frequency discrimination thresholds, and speech in noise recognition scores from participants of varying ages and hearing abilities. We chose these three measures because each test represents one of the three classes of audiological behavioral tests (absolute detection, feature discrimination, and speech recognition) and thus affords a proof of concept for automated, remote testing across the behavioral test battery. The accuracy of these remote tests was assessed by comparing home-based results to measurements made in the same subjects by an audiologist in a clinical sound booth. Our aim was to test whether the results of home-based, self-administered audiology tests were equivalent to the same tests administered by an audiologist in a sound-treated booth.

2.2 Materials and Methods

Participants

All procedures were approved by the Human Studies Committee at Massachusetts Eye and Ear Infirmary and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Informed consent was obtained from each participant. For experiment 1, nineteen individuals (7 female)

participated in the study, presenting with various degrees of hearing loss (**Fig 1A**), ranging in age from 25 to 82 years, and reporting a variety of auditory complaints (**Fig S1**). Two subjects only completed the first two time points of the study. For experiment 2, we recruited 21 additional subjects to test the repeatability and generalization of the audiometric findings when clinic-based testing was performed by various ENT and audiology clinics (see **Supplemental Methods** for more detail). Most of the participants in experiment 2 reported chronic tinnitus and presented with hearing thresholds that ranged from normal hearing to profound HL.

Procedures

For experiment 1, participants were asked to complete pure-tone audiometric, frequency discrimination, and speech in noise recognition testing in a sound treated room under the care of a licensed audiologist on two separate visits (clinic visits 1 and 2). Between clinic visits, participants were given a tablet computer and headphones, shown how to login and open the application, and asked to complete testing from home on two occasions (home visits 1 and 2). Explanations for how to use the software were communicated through text and images in the software; subjects were not provided with any detailed instructions by the experimenters. The tablet microphone was used to measure noise levels before each session. Our experimental design allowed us to characterize the accuracy of automated home testing twice for each subject (home test 1 versus clinic test 1 and home test 2 versus clinic test 2). After establishing accuracy using statistical equivalence testing (27), we asked whether the reliability of home testing was significantly different from clinic-based testing by comparing the means and variances of test-retest differences (home 2 versus home 1 and clinic 2 versus clinic 1).

The means and standard deviations for all difference testing described in this study can be found in **Supplemental Tables 1 and 2**. In experiment 2, we measured the accuracy of remote audiometric tests in a separate sample of patients to manual audiograms that they had received under the care of various US audiologists, providing a test of the generalizability of the remote testing approach and a replication of the main findings in experiment 1. Tablet-based software was developed with the Unity game engine and scripted in C#.

Diagnostic Tests

Pure-tone audiometry

For clinic-based testing, a licensed audiologist measured detection thresholds following standard clinical procedures (28) in a sound-treated booth. Contralateral masking was employed according to the optimized masking approach (29) when detection thresholds between ears differed by 35 dB or greater based on the reported interaural attenuation for the TDH-39 headphone (30). For remote testing, participants interacted with an automated tablet audiogram application using the Bose AE2i consumer-grade circumaural headphones (see Supplemental Methods and **Fig S2** for description of equipment calibration). These headphones do not provide active noise cancellation. Although the equipment and reliance on computer guidance were different from the standard approach for threshold measurements, the testing algorithm (28) followed the same rules as a clinician. Specifically, tones were presented for 1 sec with an interstimulus interval of 3-7 sec. Responses (indicated by virtual button press) were considered hits if they occurred within a response window of 2.5 sec. The tone level

was decreased by 10 dB following hits and increased by 5 dB following misses. Threshold was defined as the lowest level that evoked a hit response on 2 of 3 ascending runs or 3 of 6 runs if there was no concordance after 3 runs.

Frequency discrimination thresholds

For both clinic and remote testing, frequency discrimination thresholds were measured diotically through an interactive two-alternative forced choice software interface. The center frequency was roved around 2000 Hz, and the threshold was adaptively measured using standard psychophysical procedures (see Supplemental Methods for details, (31)).

Speech in noise thresholds

Speech in noise thresholds were estimated using two lists of 35 monosyllabic words from the Northwestern University 6 lists (32) presented diotically at 70 dB HL. Our software allowed participants to initiate trials wherein they heard a female speaker produce monosyllabic words while 6-talker babble played in the background. Participants were then cued to respond, and their voices were recorded via the tablet microphone. Recorded responses were transmitted to Massachusetts Eye and Ear Infirmary computer servers and scored offline. For clinic-based testing, an audiologist listened to their responses through the talk back on the audiometer and scored whether or not each word was correct. 50% thresholds were computed using the Spearman-Kärber equation.

Statistical Testing

We quantified the accuracy of automated home-based hearing assessments by performing statistical equivalency tests using the two one sided testing procedure (TOST, (27)). The TOST requires a defined clinical equivalency margin (i.e. difference in measurement that would not be clinically significant). The shorthand equivalency margin for clinic-based pure tone audiometry is considered to be +/- 10 dB (33–36). The clinical margin of equivalence can also be empirically defined from the test-retest difference for any diagnostic measure. Using this approach, we conservatively defined equivalency as an audiometric difference value that fell within the measurement margin of error (80% confidence) as defined by a previous audiometric reliability study (37) and our own clinical test-retest dataset for speech and frequency discrimination. Cases where the 90% confidence interval (CI) for accuracy (clinic vs. home) falls within the clinical equivalency margin provide a visual proxy for the TOST equivalence hypothesis. *P* values for all TOST statistics are listed in **Supplemental Table 3**. We tested for differences in reliability between remote and clinic-based testing by comparing test-retest differences for each condition using mixed-model ANOVA for multiple comparisons and two-sample *t*-tests for paired comparisons. We also compared test-retest variances for remote and clinic-based measurements with two-sample *F*-tests. Data were log transformed for statistical testing when they were not normally distributed (normality tested with one-sample Kolmogorov-Smirnov tests). A *post hoc* power analysis indicated the study was adequately powered ($\beta=0.2$) to test for equivalence ($\alpha=0.025$) for all reported comparisons except for ≤ 250 and ≥ 12500 Hz pure tone thresholds. Importantly, a result was only considered statistically significant if criterion significance level ($P \leq 0.025$) was met for all measurement conditions. For example, a

home-based audiometric threshold was not considered equivalent to a clinic-based threshold unless significance was met for both left and right ears during both the first and second home to clinic comparisons. We viewed this method as the most conservative approach to data analysis.

2.3 Results

Pure-tone audiometry collected remotely with this application is largely equivalent to clinic-based testing.

We measured pure tone thresholds at home and in the clinic from subjects with a wide range of hearing thresholds (**Fig1A**), ages (**Fig S1A**), histories of auditory impairment, and computer tablet experience (**Fig S1B-C**). The mean differences between home and clinic testing were small and fell within the clinical equivalency margin from 500-8000 Hz (**Fig 2A and Supplemental Table 4**, $N = 19$ participants). This finding was repeatable, with the same pattern emerging from the home versus clinic comparisons at tests 1 and 2 in both left and right ear comparisons. However, very low frequency thresholds (≤ 250 Hz) were slightly but consistently elevated when collected at home. Low frequency threshold elevation could be attributed to elevated levels of low frequency background noise in the home environment (**Fig1B** bottom) and was only observed in subjects with thresholds in the normal range (**Supplemental Fig 3**). We next compared the test-retest reliability of home versus clinic audiogram measurements. The difference scores for each testing environment are plotted in **Fig 2B** and demonstrate considerable overlap between measurements with no repeatable

significant differences in means ($P \geq 0.44$, group and group x frequency interaction terms) or variances ($P \geq 0.09$, **Supplemental Table 1**).

Automated, remote measurement of frequency discrimination and word recognition in noise is equivalent to clinic-based measurements.

We did not expect a slight threshold elevation at very low frequencies to have any influence on discrimination and recognition abilities that are measured at sound levels well above ambient room noise. Indeed, we observed that home tests were statistically equivalent to the clinic-based measurements for suprathreshold tests of frequency discrimination (**Fig3A**) and speech recognition in noise (**Fig3B**) across both testing repetitions (home 1 versus clinic 1 and home 2 versus clinic 2, $N = 19$ participants). We next compared the test-retest reliability of home versus clinic frequency discrimination and speech recognition in noise threshold measurements. The difference scores for each testing environment are plotted in **Fig3A-B** (right side) and reveal considerable overlap between measurements, with no statistical difference in means or variances (frequency discrimination, $P=0.6$ means & $P=0.6$ variances; speech recognition in noise, $P=0.09$ means & $P=0.5$ variances).

Age predicts absolute perceptual scores, but not the accuracy of remote audiological testing

Hearing difficulties are most prevalent in middle-aged and older adults (1). Thus, solutions for diagnostic efficiency increases must be amenable to implementation in this population. While older adults are adopting mobile technologies at higher rates, they still lag behind younger adults (38). Based on this discrepancy, one could speculate that

older adults might be less able to accurately self-administer hearing measures using mobile devices and applications. We enrolled a diverse sample in order to assess the predictive power of participant age on test results (**Fig S1A**). Subject age could explain a significant amount of the variability in audiometric thresholds and speech perception in noise thresholds (accounting for 38% and 52% of the variability respectively, $P=0.005$ for both). However, neither age nor degree of HL was a significant factor when comparing measurements made in clinical settings versus unsupervised testing at home ($R^2 \leq 0.17$, $P \geq 0.12$ for all associations). Furthermore, accuracy was not worse for any of the audiologic tests when individuals who did not own tablets were compared to tablet owners ($P \geq 0.67$ for all comparisons). Within our sample, there was no evidence that participant age, hearing status, or tablet ownership conferred a disadvantage in diagnostic accuracy on any of the automated, remote audiological tests that were administered.

Generalization of the home testing approach to a different sample

As a final step, we conducted a second experiment wherein we tested the generalizability of remote testing in a separate sample of patients (N=21) who had previously received clinical testing from other audiology and ENT clinics (see Supplemental Methods). As an example, **Fig 4A** shows that an audiogram collected at a Midwestern ENT clinic from a patient who presented with sudden sensorineural HL in her left ear (solid lines) was captured with a high degree of accuracy using tablet-based software at that patient's home two days later (broken lines). When comparing clinic-based audiograms from all patients' medical records to automated, remote test results,

we observed equivalency from 500-8000 Hz, replicating the results from experiment 1 (Fig 4B).

2.4 Discussion

There are three classes of behavioral audiologic tests that are used for hearing assessment (absolute detection, feature discrimination, and speech recognition). We chose a single test from each class and assessed whether measurements made from home with an automated, tablet application using consumer-grade headphones were statistically equivalent to manual measurements made by an audiologist in the clinic. Absolute detection thresholds collected remotely were equivalent to clinic-based measures across frequencies that convey speech information (500 to 8000 Hz). Frequency discrimination and speech recognition thresholds were equivalent when measured remotely and in the clinic.

Validation studies for automated, remote hearing testing have predominately focused on the sensitivity and specificity of audiograms or speech in noise recognition tests to identify individuals with elevated detection thresholds (19–26). One exception was a study that reported the differences between tone detection thresholds obtained with an iOS application in patients' homes versus manual measurements made in the clinic (39). Though statistical equivalency was not tested in that study, the reported data were qualitatively similar to our findings.

While the behavioral tests examined in this study were as accurate when performed remotely by an application as when an audiologist manually collected them in the clinic, several obstacles prevent the realization of interchangeable remote and clinic-

based test results. *i.*) When hearing sensitivity was normal, accuracy of remote audiogram measurements was reduced for 250 Hz tones, likely as a consequence of environmental noise contamination. *ii.*) The results of air-conducted tests must be interpreted with caution in the absence of initial otoscopic examination. *iii.*) Without coupling bone-conduction to air-conduction measurements, the data from remote audiograms do not afford clinicians the necessary information to characterize “type” in addition to “severity” of hearing loss. However, we do not believe that these obstacles are insurmountable. Active (40) and passive noise (41) reduction techniques have been shown to reduce the elevation of low frequency thresholds that are measured outside of sound treated booths, and are incorporated in consumer grade circumaural and in ear headphones. Additionally, smartphone compatible equipment exists to remotely obtain otoscopic images and transmit these data to clinicians for review (42), and while the widespread release of Google Glass has been suspended, it served as an example of consumer grade hardware that stimulated the cochlea via bone-conduction pathways (43, 44). The experiments reported here suggest that without further hardware solutions, the remote testing approach could serve as a means to monitor patients with known pathology or as an initial screening, wherein normal scores would be interpretable, but measured loss would provide only screening level information. If remote testing were ever to move beyond screening to provide true diagnostic-grade measurements, combined hardware and software solutions will need to reduce ambient low frequency noise levels below approximately 0 dB HL at the tympanic membrane(45, 46) and provide information concerning external and middle ear transmission.

Automated, unsupervised audiometry is a feasible, accurate, and reliable approach for the measurement of tone detection thresholds, frequency discrimination abilities, and word recognition in noise scores. These three behavioral tests represent each major class of audiologic behavioral measures. This study provides a proof of concept that automated, remote testing in relatively quiet environments can provide equivalent accuracy and reliability to clinic-based measures across a battery of audiometric behavioral tests.

2.5 References

1. Lin FR, Niparko JK, Ferrucci L (2011) Hearing Loss Prevalence in the United States. *Arch Intern Med* 171(20):1851–1852.
2. Windmill IM, Freeman B a. (2013) Demand for Audiology Services: 30-Yr Projections and Impact on Academic Programs. *J Am Acad Audiol* 24(5):407–416.
3. Margolis RH, Morgan DE (2008) Automated pure-tone audiometry: An analysis of capacity, need, and benefit. *Am J Audiol* 17(2):109–113.
4. Organization WH (2008) *The Global Burden of Disease: 2004 Update* (World Health Organization, Geneva).
5. Jerger J, Jerger S (1967) Psychoacoustic Comparison of Cochlear and VIIIth Nerve Disorders. *J Speech Lang Hear Res* 10(4):659-688.
6. Jerger J, Jerger S (1971) Diagnostic significance of PB word functions. *Arch Otolaryngol* 93(6):573–580.
7. Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2012) Why Middle-Aged Listeners Have Trouble Hearing in Everyday Settings. *Curr Biol* 22(15):1417–1422.
8. Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2011) Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proc Natl Acad Sci U S A* 108(37):15516–15521.
9. Strelcyk O, Dau T (2009) Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *J Acoust Soc Am* 125(5):3328–3345.
10. Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG (2015) Individual Differences Reveal Correlates of Hidden Hearing Deficits. *J Neurosci* 35(5):2161–2172.
11. Narne VK (2013) Temporal Processing and Speech Perception in Noise by Listeners with Auditory Neuropathy. *PLoS One* 8(2):1-11.
12. American Academy of Audiology (2010) *Diagnosis, treatment and management of children and adults with central auditory processing disorder*. Reston, VA: American Academy of Audiology.
13. American Speech-Language Hearing Association (2005) *(Central) Auditory Processing Disorders [Technical Report]*. Rockville, MD: American Speech-Language-Hearing Association.
14. Whitton JP, Polley DB (2011) Evaluating the Perceptual and Pathophysiological Consequences of Auditory Deprivation in Early Postnatal Life: A Comparison of Basic and Clinical Studies. *JARO-Journal Assoc Res Otolaryngol* 12(5):535–547.
15. Mahomed F, Swanepoel DW, Eikelboom RH, Soer M (2013) Validity of Automated Threshold Audiometry: A Systematic Review and Meta-Analysis. *Ear Hear* 34:745–752.

16. Rudmose W (1963) Automatic audiometry. *Modern Developments in Audiology*, ed Jerger J (Academic Press, New York), pp 30–75.
17. Smith A, McGeeney, K, Duggan, M, Rainie, L, Keeter, S. The smartphone difference. Washington DC: Pew Research Center; 2015. Available at: <http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/>
18. Wike R, Simmons K, Poushter J, et al. Emerging nations embrace internet, mobile technology. Washington DC: Pew Research Center; 2014. Available at: <http://www.pewglobal.org/2014/02/13/emerging-nations-embrace-internet-mobile-technology/>
19. Abu-Ghanem S, et al. (2015) Smartphone-based audiometric test for screening hearing loss in the elderly. *Eur Arch Oto-Rhino-Laryngology* 273(2):333-339.
20. Bexelius C, et al. (2008) Evaluation of an internet-based hearing test—comparison with established methods for detection of hearing loss. *J Med Internet Res* 10(4):1-10.
21. Smits, C., Kapteyn, Theo, Houtgast T (2004) Development and validation of an automatic speech-in-noise screening test by telephone. *Int J Audiol* 43:15-28.
22. Smits C, Merkus P, Houtgast T (2006) How we do it: The Dutch functional hearing-screening tests by telephone and internet. *Clin Otolaryngol* 31(5):436-440.
23. Watson CS, Kidd GR, Miller JD, Smits C, Humes LE (2012) Telephone screening tests for functionally impaired hearing: current use in seven countries and development of a US version. *J Am Acad Audiol* 23(10):757–767.
24. Jansen S, Luts H, Wagener KC, Frachet B, Wouters J (2010) The French digit triplet test: a hearing screening tool for speech intelligibility in noise. *Int J Audiol* 49(5):378–387.
25. Meyer C, et al. Investigation of the actions taken by adults who failed a telephone-based hearing screen. *Ear Hear* 32(6):720–731.
26. Vlaming MSMG, Mackinnon RC, Jansen M, Moore DR (2014) Automated Screening for High-Frequency Hearing Loss. *Ear Hear* 35:667–679.
27. Walker E, Nowacki AS (2011) Understanding equivalence and noninferiority testing. *J Gen Intern Med* 26(2):192-196.
28. Carhart R, Jerger J (1959) A preferred method for clinical determination of pure-tone thresholds *J Speech Hear Disord.* 24(4):330-345.
29. Turner RG (2004) Masking redux. I: An optimized masking method. *J Am Acad Audiol* 15(1):17–28.
30. Killion MC, Wilber LA, Gudmundsen GI (1985) Insert earphones for more interaural attenuation. *Hear Instruments* 36(2):1-2.
31. Levitt H (1971) Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49(2):467–477.
32. Wilson RH, Burks C a (2005) Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol. *J Rehabil Res Dev* 42(6):839–852.

33. OSHA. Occupational noise exposure: Hearing conservation amendment; Final Rule. 1983:9738-9785.
34. NIOSH. Compendium of materials for noise control. 1980:80-116.
35. Eikelboom R, Swanepoel DW, Motakef S, Upson G (2013) AMTAS Clinical Validation. *Int J Audiol* 52:342–349.
36. Lemkens N, et al. (2002) Interpretation of pure-tone thresholds in sensorineural hearing loss (SNHL): a review of measurement variability and age-specific references. *Acta Otorhinolaryngol Belg* 56(4):341–352.
37. Landry J, WB G (1999) Pure-tone audiometric threshold test-retest variability in young and elderly adults. *J Speech-Language Pathol Audiol* 23(2):74–80.
38. Smith A. Older adults and technology use. Washington DC: Pew Research Center; 2014. Available at: <http://www.pewinternet.org/2014/04/03/older-adults-and-technology-use/>
39. Foulad a., Bui P, Djalilian H (2013) Automated Audiometry Using Apple iOS-Based Application Technology. *Otolaryngol - Head Neck Surg* 149(5):700–706.
40. Bromwich MA, Parsa V, Lanthier N, Yoo J, Parnes LS (2008) Active noise reduction audiometry: a prospective analysis of a new approach to noise management in audiometric testing. *Laryngoscope* 118:104–109.
41. Maclennan-Smith F, Swanepoel DW, Hall JW (2013) Validity of diagnostic pure-tone audiometry without a sound-treated environment in older adults. *Int J Audiol* 52(2):66–73.
42. CellScope company website. CellScope Inc. <https://www.cellscope.com/>. Accessed April 9, 2015.
43. Kupersmidt H, Blanka L (2014) Indication of quality for placement of bone conduction transducers. US patent 20140363002 A1. June 9, 2014.
44. Heiman A, Haiut M, Yehuday U (2013) Equalization and power control of bone conduction elements. US patent 20140363033 A1. Dec 30, 2013.
45. Frank T, Williams DL (1993) Ambient noise levels in audiometric test rooms used for clinical audiometry. *Ear Hear* 14(6):414–422.
46. Hawkins JE, Stevens SS (1950) The Masking of Pure Tones and of Speech by White Noise. *J Acoust Soc Am* 22(1):6-13.
47. Gurgel RK, Jackler RK, Dobie R a., Popelka GR (2012) A New Standardized Format for Reporting Hearing Outcome in Clinical Trials. *Otolaryngol - Head Neck Surg* 147(5):803–807.

2.6 Figures

Figure 1

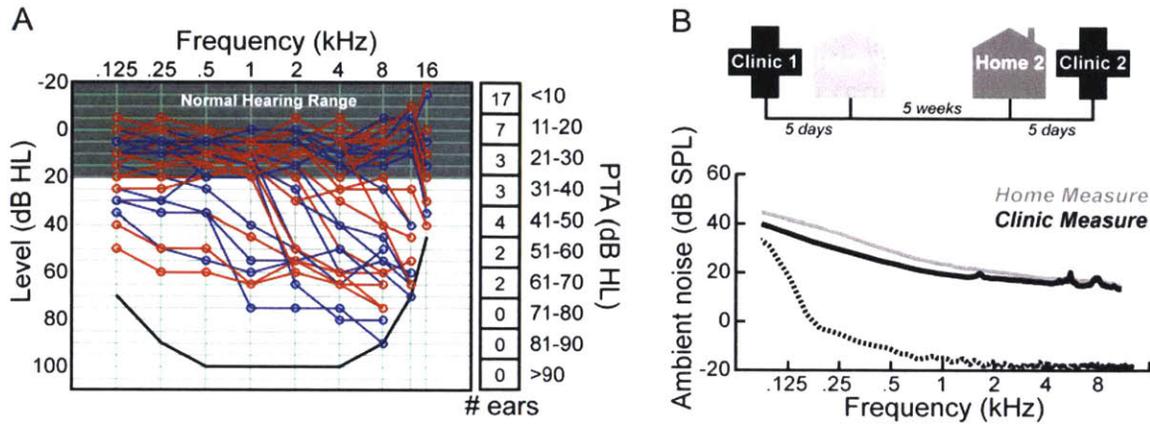


Fig 1. (A, left) Air-conduction thresholds from 19 subjects collected in the clinic (red = right ear, blue = left ear). (A, right) Distribution of pure tone averages (PTA .5 – 2 kHz) in the sample plotted according to AAO-HNS recommendations (47, B, top) Study design. (B, bottom) Average ambient noise measurements made at the subjects' homes (gray line) and in the sound-treated clinical booth (black line) with the tablet computer. Actual sound-treated clinical booth noise (black broken line) measured with a high-quality microphone and signal analyzer revealed that the measurement noise floor of the tablet exceeded all clinic and some home ambient noise measurements (see Supplemental Methods for details).

Figure 2

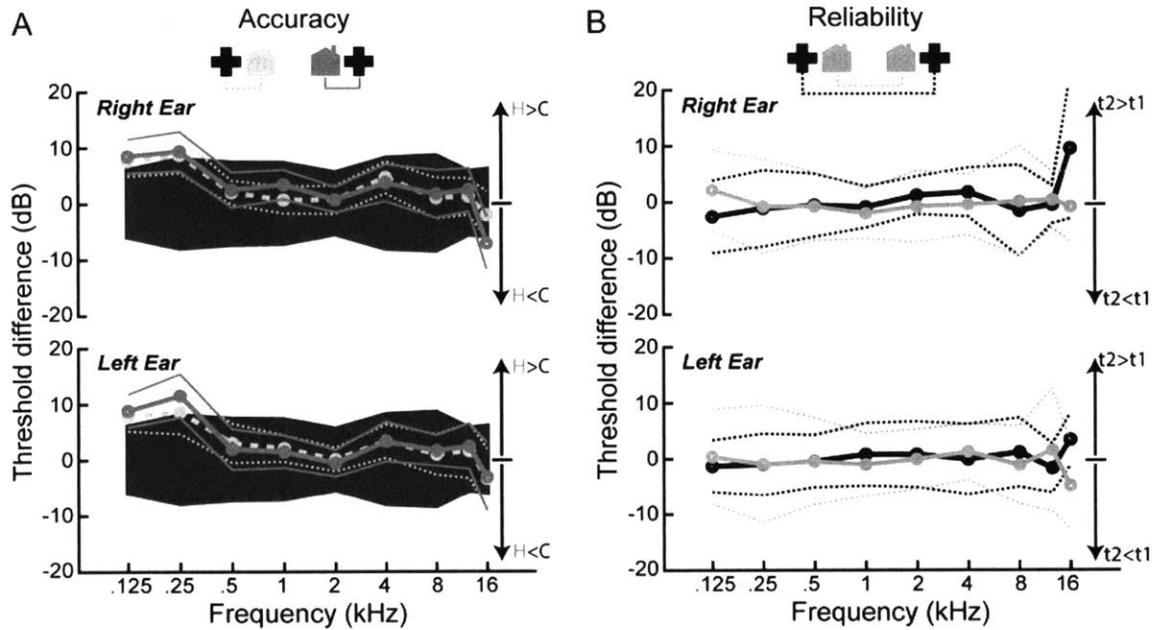


Fig 2. (A) Means and 90% confidence intervals are plotted for differences between home and clinic testing. Difference scores were measured twice (Home 1 – Clinic 1 indicated by broken lines and Home 2 – Clinic 2 indicated by darker solid lines). The clinical equivalence margin defines the upper and lower bounds of the dark gray area. (B) Means and standard deviations are plotted for differences between testing sessions 1 and 2 made at home (“H”, gray) and by an audiologist at the clinic (“C”, black). Measurements made for the right (A&B, top) and left (A&B, bottom) ears were plotted and analyzed separately.

Figure 3

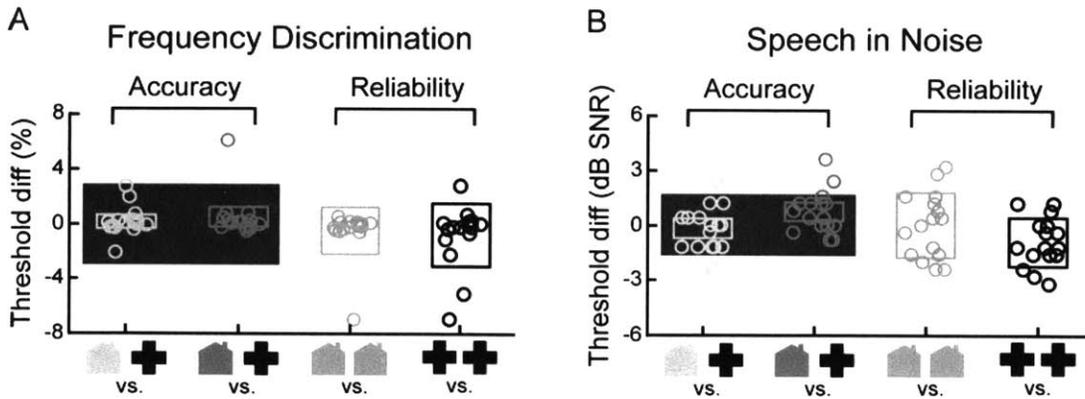


Fig 3. (A, left) Differences between home and clinic frequency discrimination testing are plotted for individual subjects (gray circles). The 90% confidence intervals for differences in home and clinic testing for tests 1 and 2 are indicated by gray boxes. The clinical equivalence margin defines the upper and lower bounds of the dark gray area. (A, right) Measurement differences between tests 1 and 2 made by subjects at home (gray circles) and by an audiologist at the clinic (black circles). Mean differences \pm 1 standard deviation are indicated by boxes for each comparison. (B) The same plotting conventions described for (A) are used to depict the accuracy and reliability of speech in noise testing. Directions of threshold change for accuracy and reliability comparisons match the convention used in Fig 2.

Figure 4

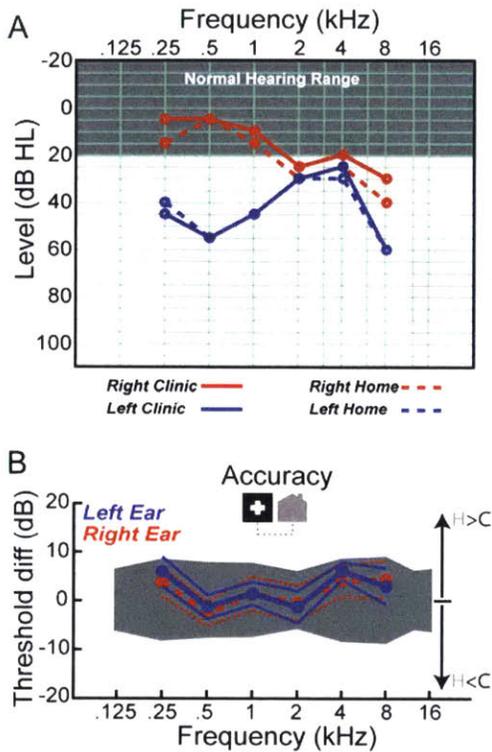
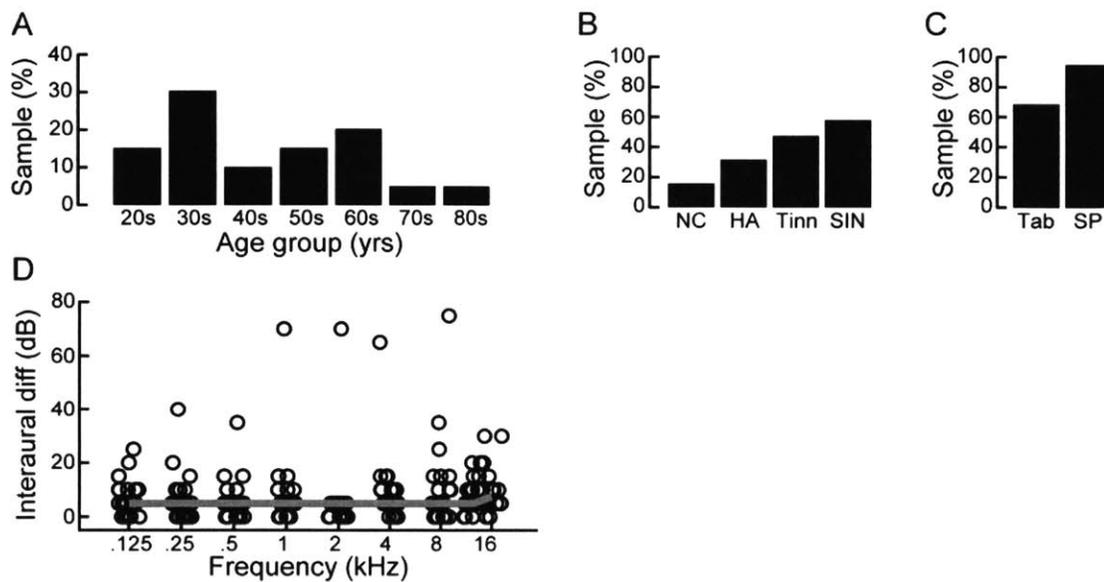


Fig 4. (A) Audiograms for a patient diagnosed with sudden sensorineural HL in the left ear. The solid lines (red = right ear and blue = left ear) represent measurements made by an audiologist at a Midwestern ENT clinic, and broken lines indicate automated measurements made with a tablet from the patient's home two days later. (B) Means and 90% confidence intervals are plotted for differences between automated home testing and clinical audiograms from 21 patients' medical records. The left ear and right ear data were analyzed separately but are plotted on the same figure (left = lighter gray, right = darker gray). The clinical equivalence margin defines the upper and lower bounds of the dark gray area, as per Fig 2A.

2.7 Supplemental Materials

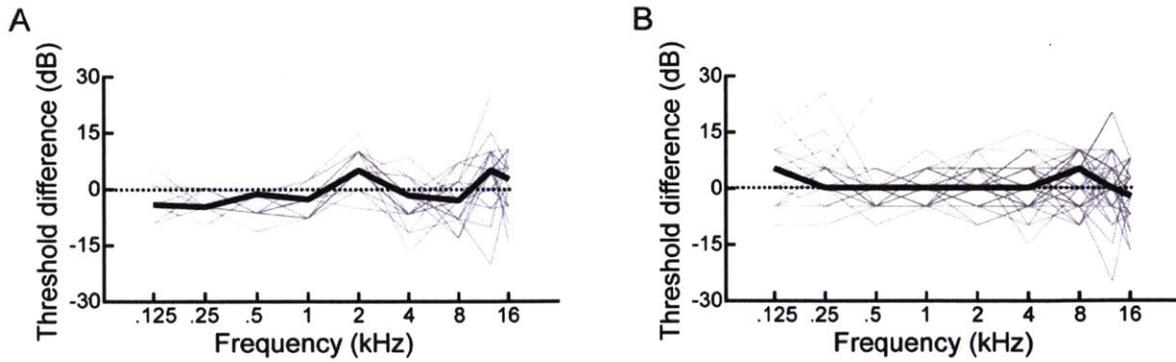
2.7.1 Supplemental Figures

Supplemental Figure 1



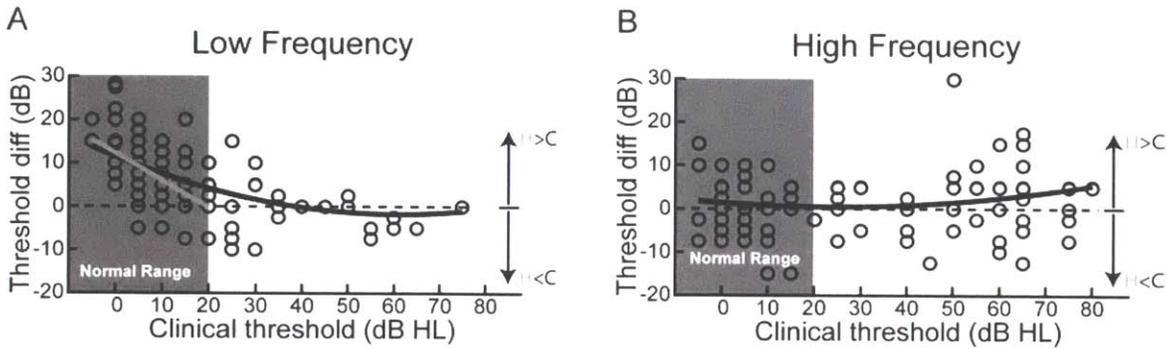
SuppFig1. (A) Distribution of participant age in the sample for experiment 1. (B) Distribution of subject complaints (NC = no complaints, HA = hearing aid use, Tinn = tinnitus, and SIN = speech-in-noise), and (C) mobile technology adoption (Tab = tablet, SP = smartphone). (D) Asymmetry of audiometric thresholds for each subject plotted as a function of frequency. Gray line indicates mean asymmetry across the sample.

Supplemental Figure 2



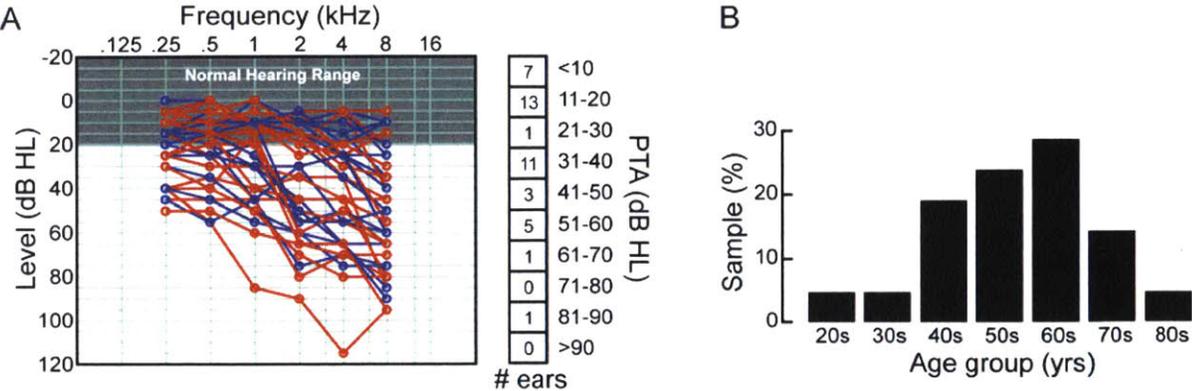
SuppFig2. (A) Difference between automated thresholds collected with consumer grade headphones and manual clinical thresholds collected with clinic grade headphones for individual subjects (thick black line = mean difference). RETSPL values were corrected using the KEMAR transfer function described in the Supplemental Methods (N = 10). (B) Same type of comparison as (A) but including another sample of 12 subjects with RETSPL values defined by the behavioral calibration detailed in the Supplemental Methods (thick black line = mean difference).

Supplemental Figure 3



SuppFig3. (A) Difference (Diff) between home (H) and clinical (C) thresholds plotted as a function of clinical threshold for low frequency carriers (≤ 1000 Hz). (B) Same as (A) but for mid to high frequency carriers (≥ 2000 Hz). Black lines indicate a quadratic fit to the data. The gray line in (A) represents a linear fit of data within the normal hearing range. Directions of threshold change for accuracy and reliability comparisons match the convention used in Fig 2 from the main text.

Supplemental Figure 4



SuppFig4. (A) Clinic-based audiograms that were obtained from patient medical records for 21 individuals who participated in experiment 2 (red = right ear and blue = left ear). (A, right) Distribution of pure tone averages (PTA .5 – 2 kHz) in the sample plotted according to AAO-HNS recommendations. (B) Distribution of participant age in the sample for experiment 2.

2.7.2 Supplemental Tables

Supplemental Table 1: Comparisons of home and clinic-based audiograms: signed and unsigned mean and standard deviations of difference scores.

Signed Differences in Audiometric Thresholds		Carrier Frequency (Hz)								
		125	250	500	1000	2000	4000	8000	12500	16000
H1vC1 L Ear	<i>Mean</i>	7.63	8.42	2.89	1.94	0.00	3.23	0.85	1.54	-3.44
	<i>Std Dev</i>	6.53	9.73	8.71	5.88	5.27	7.38	8.47	10.24	7.91
H1vC1 R Ear	<i>Mean</i>	7.37	8.42	1.84	0.53	0.53	4.47	0.79	1.07	-2.19
	<i>Std Dev</i>	6.74	8.00	5.33	5.98	6.21	7.05	9.02	6.84	7.07
H2vC2 L Ear	<i>Mean</i>	8.71	11.37	1.87	1.44	-0.59	3.24	1.67	2.44	-3.03
	<i>Std Dev</i>	6.88	9.11	8.14	6.75	5.56	8.28	5.88	7.76	7.36
H2vC2 R Ear	<i>Mean</i>	8.24	9.12	2.35	3.24	0.59	3.53	1.56	2.26	-7.19
	<i>Std Dev</i>	7.06	8.15	7.31	6.11	5.27	7.66	9.61	7.36	7.50
H2vH1 L Ear	<i>Mean</i>	0.16	-1.15	-0.63	-1.25	-0.31	1.02	-1.37	1.25	-5.00
	<i>Std Dev</i>	8.39	10.38	7.72	5.60	5.31	4.99	6.96	10.90	7.75
H2vH1 R Ear	<i>Mean</i>	2.19	-0.63	-0.63	-1.88	-0.63	-0.31	0.33	0.60	-0.63
	<i>Std Dev</i>	7.06	8.34	6.02	4.43	6.29	5.31	9.72	5.00	6.23
C2vC1 L Ear	<i>Mean</i>	-1.47	-1.18	-0.59	0.59	0.59	-0.29	0.94	-1.92	3.12
	<i>Std Dev</i>	4.60	5.46	4.64	5.56	5.83	6.24	6.12	4.35	4.58
C2vC1 R Ear	<i>Mean</i>	-2.65	-1.18	-0.59	-0.88	1.18	1.76	-1.47	-0.42	9.44
	<i>Std Dev</i>	6.40	6.74	5.56	3.64	3.32	4.31	8.06	3.34	12.36
Generalization* L Ear	<i>Mean</i>	NaN	3.81	-2.62	1.43	-0.48	4.05	4.29	NaN	NaN
	<i>Std Dev</i>	NaN	6.31	5.39	7.44	7.73	6.82	8.26	NaN	NaN
Generalization R Ear	<i>Mean</i>	NaN	5.71	-1.43	1.30	-1.43	5.82	2.78	NaN	NaN
	<i>Std Dev</i>	NaN	6.57	5.28	5.11	7.10	3.82	7.58	NaN	NaN

Absolute Differences in Audiometric Thresholds											
					Carrier Frequency (Hz)						
		125	250	500	1000	2000	4000	8000	12500	16000	
H1vC1 L Ear	Mean	8.68	10.53	6.58	4.57	3.68	5.34	6.15	7.35	7.35	
	Std Dev	4.96	7.24	6.25	4.06	3.67	5.96	5.69	7.02	3.79	
H1vC1 R Ear	Mean	8.55	10.00	5.26	4.65	4.21	5.69	6.65	5.99	6.84	
	Std Dev	5.05	6.88	5.32	3.75	3.77	5.75	5.46	6.13	3.32	
H2vC2 L Ear	Mean	8.71	11.96	6.88	4.97	3.53	6.76	3.67	5.78	6.56	
	Std Dev	6.88	8.28	4.43	4.65	4.24	5.57	4.81	5.51	3.67	
H2vC2 R Ear	Mean	8.82	9.71	5.88	5.59	4.12	7.06	7.81	6.07	8.44	
	Std Dev	6.26	7.39	4.76	3.91	3.18	4.35	5.47	4.43	5.86	
H2vH1 L Ear	Mean	6.09	8.02	5.63	3.75	3.44	4.15	5.37	7.08	5.00	
	Std Dev	5.55	6.36	5.12	4.25	3.97	2.77	4.42	8.11	7.75	
H2vH1 R Ear	Mean	5.31	6.25	3.75	3.13	5.00	3.44	7.00	3.24	4.37	
	Std Dev	4.99	5.32	4.65	3.59	3.65	3.97	6.49	3.74	4.17	
C2vC1 L Ear	Mean	3.24	3.53	3.53	4.12	4.12	4.41	4.69	3.46	4.37	
	Std Dev	3.51	4.24	2.94	3.64	4.04	4.29	3.86	3.15	3.20	
C2vC1 R Ear	Mean	4.41	4.71	4.12	2.65	1.76	3.53	6.18	2.08	9.44	
	Std Dev	5.27	4.83	3.64	2.57	3.03	2.94	5.16	2.57	12.36	
Generalization* L Ear	Mean	NaN	5.71	3.57	5.24	5.24	5.95	7.14	NaN	NaN	
	Std Dev	NaN	4.55	4.78	5.36	5.58	5.15	5.82	NaN	NaN	
Generalization R Ear	Mean	NaN	7.14	3.81	3.46	4.76	6.32	5.64	NaN	NaN	
	Std Dev	NaN	4.89	3.84	3.91	5.36	2.86	5.65	NaN	NaN	
H1vC1 = Home1 - Clinic 1											
H2vC2= Home2 - Clinic 2											
H2vH1 = Home 2 - Home 1											
C2vC1 = Clinic 2 - Clinic 1											
*Generalization experiment refers to home threshold - threshold data from medical record in 11 separate subjects.											
125, 12500, and 16000 Hz were not tested for the measurements in their medical records, so no comparison could be made between the home and clinical tests at those frequencies.											

Supplemental Table 2: Comparisons of home and clinic-based frequency discrimination thresholds and speech in noise scores: signed and unsigned mean and standard deviations of difference scores.

Differences in Suprathreshold Measures					
		FD signed	FD abs	SIN signed	SIN abs
H1vC1	<i>Mean</i>	0.22	0.61	-0.18	0.93
	<i>Std Dev</i>	1.03	0.85	1.2	0.76
H2vC2	<i>Mean</i>	0.51	0.66	0.75	1
	<i>Std Dev</i>	1.54	1.47	1.17	0.95
H2vH1	<i>Mean</i>	-0.52	0.66	0.03	1.48
	<i>Std Dev</i>	1.73	1.68	1.79	0.93
C2vC1	<i>Mean</i>	-0.81	1.32	-0.9	1.3
	<i>Std Dev</i>	2.29	2.01	1.33	0.92
H1vC1 = Home1 - Clinic 1					
H2vC2= Home2 - Clinic 2					
H2vH1 = Home 2 - Home 1					
C2 v C1 = Clinic 2 - Clinic 1					
FD = Frequency Discrimination					
SIN = Speech in Noise					

Supplemental Table 3: P values for all TOST analyses of signed difference scores.

P values associated with Two One Sided Testing in manuscript					
	Measure	Comparison			
		Left Ear		Right Ear	
		H1vC1*	H2vC2*	H1vC1	H2vC2
	125	0.87	0.94	0.82	0.91
	250	0.61	0.94	0.63	0.74
	500	<i>0.02</i>	<i>0.01</i>	<i>1.89E-04</i>	<i>7.66E-03</i>
Audiogram	1000	<i>6.82E-04</i>	<i>1.76E-03</i>	<i>8.10E-05</i>	<i>0.01</i>
Experiment 1	2000	<i>1.56E-04</i>	<i>1.33E-03</i>	<i>1.59E-03</i>	<i>8.80E-04</i>
	4000	<i>6.19E-03</i>	<i>0.02</i>	<i>0.02</i>	<i>0.02</i>
	8000	<i>1.11E-03</i>	<i>3.11E-04</i>	<i>9.35E-04</i>	<i>6.57E-03</i>
	12500	0.09	0.1	0.02	0.08
	16000	0.21	0.19	0.002813	0.69
		Left Ear		Right Ear	
		H1vC1		H1vC1	
	250	<i>4.37E-03</i>		0.08	
Audiogram	500	<i>4.79E-04</i>		<i>3.57E-05</i>	
Experiment 2	1000	<i>1.25E-03</i>		<i>2.44E-05</i>	
Generalization	2000	<i>4.38E-03</i>		<i>9.52E-03</i>	
	4000	<i>8.35E-03</i>		<i>0.01</i>	
	8000	<i>0.02</i>		<i>2.55E-03</i>	
		Diotic	Diotic		
		H1vC1	H2vC2		
	FD	<i>1.31E-06</i>	<i>5.27E-05</i>		
	SIN	<i>9.14E-07</i>	<i>4.77E-05</i>		
*The TOST produces two p values (for the comparison with the higher and lower end of the clinical equivalency range). We followed the convention of reporting the higher of the two p values here.					
Results that met statistical significance are italicized.					
FD = Frequency Discrimination					
SIN = Speech in Noise					

Supplemental Table 4. Percentage of ears with home thresholds values within the clinical equivalence margin.

% Thresholds within clinical test-retest limit									
Audiogram									
Carrier Frequency (Hz)									
	125	250	500	1000	2000	4000	8000	12500	16000
H1vC1	76%	71%	95%	97%	100%	92%	86%	86%	94%
H2vC2	74%	56%	91%	97%	97%	85%	87%	92%	87%
H2vH1	91%	84%	94%	97%	97%	97%	90%	96%	93%
C2vC1	97%	91%	100%	100%	97%	100%	94%	100%	94%
Generalization*	NaN	88%	93%	93%	95%	95%	85%	NaN	NaN
Suprathreshold				H1vC1 = Home1 - Clinic 1					
	FD	SIN		H2vC2= Home2 - Clinic 2					
H1vC1	94%	88%		H2vH1 = Home 2 - Home 1					
				C2 v C1 = Clinic 2 - Clinic 1					
H2vC2	94%	88%		FD = Frequency Discrimination					
				SIN = Speech In Noise					
H2vH1	94%	75%		*Generalization experiment refers to home threshold - data					
				from medical record in 11 separate subjects.					
C2vC1	81%	81%		Clinical Limit Audiogram = within 10 dB					
				Clinical Limit FD = within 2.2%					
				Clinical Limit SIN = within 1.7 dB					

2.7.3 Supplemental Methods

Noise measurements

Microphone sensitivity on a single Microsoft Surface Pro 2 was compared with a Brüel and Kjær type 2670 1/4 in microphone in response to a broadband noise presented in the free field. Ambient sound pressure levels were measured through the microphone on the tablet computer for a 2s period prior to all measurements made at home or in the clinical sound booth. These recordings were analyzed offline to detect and remove contamination by acoustic transients (e.g. coughing, moving the tablet, and speaking). The noise measurements in the clinic sound booth were higher than expected, suggesting to us that these measurements might represent the noise floor of the tablet (self-generated noise and microphone sensitivity). We verified this hypothesis by measuring noise levels in the sound booth using a Hewlett-Packard 35660A signal analyzer coupled to a Brüel and Kjær type 4134 1/2 in microphone. This equipment configuration had a very low measurement noise floor and afforded true dB SPL measurements of ambient sound levels across the frequency range of interest. The acoustic noise floor of the sound booth measured using this approach is plotted in **Fig 1B** and confirms that tablet measurements made in the sound booth reflect the tablet's noise floor, as the acoustic noise floor in the sound-treated booth was substantially lower across the measured frequency range.

Equipment Calibration

Four sets of Bose AE2i headphones were coupled to a Knowles Electronics Mannequin for Auditory Research (KEMAR). A pair of Etymotic ER-11 ½ in

microphones recorded sound levels of a click stimulus at the approximate positions of the tympanic membranes. The transfer function of the Bose AE2i headphones was defined as the mean function across 8 measurements (4 sets of headphones x 2 ears).

Clinical audiometers express sound level in dB Hearing Level (dB HL). This is a behaviorally normalized scale, wherein the frequency-specific, average thresholds measured in dB Sound Pressure Level (dB SPL) across a group of normal hearing individuals is defined as 0 dB HL. These behavioral data are then used to create a conversion table between dB SPL to dB HL. This table is called the Reference Equivalent Threshold Sound Pressure Level (RETSPL) and is specific to each headphone. In the case of circumaural headphones, the RETSPLs defined for two different headphones in the American National Standards Institute (ANSI) 2010 audiometer specifications (1) reflect the threshold equivalent dB SPL measured on an artificial ear (type 1 coupler) with an applied force of 10 N. Generalizing the RETSPLs to the consumer grade headphones used in this study would be expected to pose at least two problems. *i.)* While calibrating with 10 N of force reflects the actual test conditions of clinical headphones, it overestimates the force of consumer grade headphones, which are built with comfort rather than acoustic variability in mind. This is expected to cause overestimates of low-frequency sound level due to leakage from the consumer grade products. *ii.)* The differences in headphone speaker locations and orientations will cause variability in sound pressure measurements. To address these problems, we characterized the transfer function between measurements made with a Larson Davis AEC 101 artificial ear using ANSI specifications and the sound level measured at the eardrum when worn on the head as approximated by KEMAR using the Bose and TDH

39 headphones. We were then able to use the difference in these transfer functions to correct the RETSPL values defined for the TDH 39 headphones in the ANSI 2010 audiometer specifications and apply them to our Bose headphones. Using this as a means to define dB HL, we recruited 10 subjects, who did not participate in any other studies, to complete supervised manual testing using clinical equipment and to complete automated testing with the consumer-grade equipment used in this study. Both clinical and automated tests were performed in a sound treated booth. We observed good agreement between clinical and automated thresholds with the Bose headphones using the KEMAR corrected RETSPLs (**Supplemental Fig 2A**). However, low frequency thresholds were somewhat lower and mid frequencies were elevated with the Bose headphones. We believed that this difference was likely a product of our inability to fully capture the acoustic variability that accompanied differing head/ear anatomy and headphone placement in our limited measurements.

To further examine the difference between thresholds obtained with the clinical headphones and consumer grade headphones using the KEMAR corrected RETPLs we computed a behavioral calibration from these 10 subjects that would equalize thresholds across measurements made with clinical and consumer-grade headphones. These RETSPL values were then used when collecting automated thresholds from an additional group of 12 subjects in a sound treated booth with the Bose headphones (**Supplemental Fig 2B**). When comparing these thresholds to the clinical thresholds, we found good agreement across the audiometric range. These RETSPL values were then used to convert dB SPL to dB HL for the Bose headphones in the home testing equivalence study reported in the main text. It is important to note that these 10

individuals whose thresholds were used to compute RETSPLs for the Bose headphones did not participate in the diagnostic testing described in the main text.

Diagnostic Tests

Clinic Testing

Two clinic testing sessions were performed under the supervision of a licensed Audiologist an average of 7+/- 2 weeks apart. The same clinician performed the first and second testing sessions except in 3 cases, where a different audiologist performed the second testing. The audiologist provided instruction before testing and ensured that the test environment was free of distractions (e.g. active cell phone). All testing was performed in a sound-treated research booth.

Absolute detection thresholds

A licensed Audiologist measured detection thresholds following the modified Hughson-Westlake procedure (2) in a sound-treated research booth. Signals were generated using an Interacoustics AC40 or Madsen Astera clinical audiometer and delivered through supra-aural headphones (TDH-39 or DD 45) for tones in the clinical audiometric range (125-8000 Hz) and circumaural headphones (Sennheiser HD200) for high frequency tones (12000 & 16000 Hz). Contralateral masking was employed according to the optimized masking approach³ when detection thresholds between ears differed by 35 dB or greater based on the reported interaural attenuation for the TDH-39 headphone (4) (a 15 dB air-bone gap was conservatively assumed in the computation as bone conduction threshold were not measured).

Frequency discrimination thresholds

Frequency discrimination thresholds were measured through an interactive user interface on the tablet while wearing the Bose AE2i headphones. Subjects saw two speaker icons on the left and right of the screen. On each trial they were given a prompt to listen and then heard two amplitude modulated tones (duration = 1.2s) that were separated by a 0.5s interstimulus interval. The subject was asked to indicate which tone was higher in pitch by pressing the left or right speaker. The center frequency and intensity of the tones were roved around 2000 Hz (± 200 Hz) and 36 dB Sensation Level (± 6 dB) to remove off-task listening cues. The two-down-one-up procedure (5) was used to converge on the 70.7% correct point. The frequency difference between the two intervals was initially 400 Hz (20% of the center frequency) and changed by a factor of 1.5 for the first 5 reversals, decreasing to a factor of 1.2 for the last 7 reversals. The geometric mean of the last 6 reversals was used to compute the run value. Four runs were repeated and the median value across runs was used to define the participant's threshold.

Speech in noise thresholds

Speech in noise threshold estimates were made for two 35 word lists from the Words In Noise test materials (6) diotically presented at 70 dBHL over the Bose AE2i headphones. Our software allowed participants to initiate trials wherein they heard a female speaker produce monosyllabic words while 6-talker babble played in the background at signal-to-noise ratios that varied from 24 to 0 dB SNR in 4 dB steps (level of target speech was reduced and babble remained at a constant level).

Participants were then cued to respond, and their voices were recorded via the microphone on the tablet. An audiologist listened to their responses through the talk back on the audiometer and scored whether or not each word was correct in real time. Answers were later verified offline by comparing the scoring of the recorded response to the Audiologist's score sheet measured. Thresholds (50% correct) were computed using the Spearman-Kärber equation.

Home testing

Two unsupervised home testing sessions were performed by each participant an average of 5 +/- 3 weeks apart. Before the first testing session began, the software application instructed each subject on how to make reliable measurements from their home environment (e.g. choose a quiet location, do not chew gum or use any candy during testing).

The frequency discrimination and speech in noise tasks were performed at the participant's home using the same procedures as were described for the clinical tests, except only offline scoring of voice recordings could be performed with the speech in noise test. Absolute detection thresholds were measured with an automated audiogram application on the tablet, while wearing the Bose AE2i consumer grade headphones. While the equipment was different and the test was automated rather than manually performed by a professional, the algorithm² followed the same rules as a clinician. Contralateral masking was employed according to the optimized masking approach (3) when detection thresholds between ears differed by 20 dB or greater based on our measurement of the interaural attenuation for the Bose AE2i headphones (a 15 dB air-

bone gap was conservatively assumed in the computation as bone conduction threshold were not measured).

Managing “no response” outcomes in audiometric thresholds

At carrier frequencies above 8 kHz, many individuals did not respond to sound at the maximum SPL. Specifically, “no response” outcomes were recorded at presentation level limits 26% of the time for 12.5 kHz carriers and 50% of the time at 16 kHz carriers. We found a similar pattern for the home-based testing (“no response” at limits 21% of the time for 12.5 kHz carriers and 52% of the time for 16 kHz carriers). For home testing, there were also two occasions wherein an 8 kHz threshold could not be measured at the equipment limits. While shared “no response” outcomes (84% of examples) between the clinic and home tests offer qualitative agreement between methods, this categorical outcome could not be included in a numerical statistical equivalence analysis. This was considered to be the most conservative analysis approach.

Interaction between Hearing Level and Threshold Accuracy

A key limitation of the accuracy of home-based audiometry was a modest, but persistent elevation of low-frequency thresholds compared to clinical measures. Since noise levels at subjects’ homes were higher for low frequency sounds, we reasoned that low frequency noise was the limiting factor associated with threshold accuracy. This hypothesis generated 2 predictions. 1.) Low frequency thresholds should only be elevated at home if the clinical thresholds were low to begin with. In other words, if subjects presented to the clinic with low frequency hearing loss, then the environmental

noise in the home environment would not be expected to affect threshold accuracy. 2.) This interaction between clinical thresholds and accuracy should be limited to low frequency carriers since environmental noise in the mid to high frequency range were similar in the home and sound treated booth when measured with the tablet microphone. We tested these predictions by plotting home threshold elevation as a function of clinical hearing loss for low frequency carriers (≤ 1000 Hz, **Supplemental Figure 3A**) and high frequency carriers (≥ 2000 Hz, **Supplemental Figure 3B**). We found that while there was a strong dependence of threshold elevation on amount of hearing loss for low frequency carriers, there was no relationship between amount of hearing loss and threshold elevation for high frequency carriers. We also noted that this relationship was restricted to clinical hearing levels below 20 dB HL (x intercept of linear fit = 19.6 dB HL), the limit of normal hearing sensitivity. Thus, if subjects presented with thresholds within normal limits, a 0.63 dB elevation in home thresholds was expected for every 1 dB improvement in clinical thresholds relative to the normative limits. In this sense, low frequency noise contamination is only a concern for subjects with low frequency sensitivity in the normal range.

Generalization of Audiogram Accuracy to a Separate Patient Population

In experiment 1, individual subjects were tested by an audiologist in a clinical setting and at home, using an automated testing software and consumer grade headphones. A more demanding test of the equivalence for home-based, automated audiometry is a comparison with manual audiograms performed at patients' home clinics, outside the context of our research study. We had an opportunity to make such a comparison for experiment 2. As part of a separate study in our lab, we had enrolled a

cohort of patients who were living with tinnitus and/or hearing loss and did not use hearing aids. These subjects signed release forms that allowed us to requisition their audiological records from their home audiology or ENT clinics. They also completed home audiometric testing with the software described in experiment 1. All subjects had undergone audiological testing at their home clinic within a year and a half of entering our study (**Supplemental Fig 4**). We computed the difference between their home-based and clinical audiograms and tested their equivalency (**Fig 4B**).

References

1. American National Standards Institute. *American National Standard Specification for Audiometers*. Vol S3.6-2010. 2010.
2. Carhart R, Jerger J. A Preferred Method for Clinical Determination of Pure-Tone Thresholds. *J Speech Hear Disord*. 1959;24(4): 330-345.
3. Turner RG. Masking redux. I: An optimized masking method. *J Am Acad Audiol*. 2004;15(1):17-28.
4. Killion MC, Wilber LA, Gudmundsen GI. Insert earphones for more interaural attenuation. *Hear Instruments*. 1985;36(2):1-2.
5. Levitt H. Transformed up-down methods in psychoacoustics. *J Acoust Soc Am*. 1971;49(2):467-477.
6. Wilson RH, Burks C a. Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol. *J Rehabil Res Dev*. 2005;42(6):839-852.
7. Gurgel RK, Jackler RK, Dobie R a., Popelka GR. A new standardized format for reporting hearing outcome in clinical trials. *Otolaryngol Head Neck Surg*. 2012;147(5):803-807.

Chapter 3: Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise²

Abstract

All sensory systems face the fundamental challenge of encoding weak signals in noisy backgrounds. Though discrimination abilities can improve with practice, these benefits rarely generalize to untrained stimulus dimensions. Inspired by recent findings that action video game training can impart a broader spectrum of benefits than traditional perceptual learning paradigms, we trained adult humans and mice in an immersive audio game that challenged them to forage for hidden auditory targets in a two-dimensional soundscape. Both species learned to modulate their angular search vectors and target approach velocities based on real time changes in the level of a weak tone embedded in broadband noise. In humans, mastery of this tone in noise task generalized to an improved ability to comprehend spoken sentences in speech babble noise. Neural plasticity in the auditory cortex of trained mice supported improved decoding of low-intensity sounds at the training frequency and an enhanced resistance to interference from background masking noise. These findings highlight the potential to improve the neural and perceptual salience of degraded sensory stimuli through immersive computerized games.

² Data presented in this chapter were previously published in Whitton et al (2016) Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise. *Proceedings of the National Academy of Sciences* 111(25):E2606-E2615

3.1 Introduction

Efficient search for resources is critical to the survival of most species. As such, foraging represents a conserved, adaptive behavior that drives decision making under the types of naturalistic contexts for which brains have evolved. Efficient foraging involves the dynamic integration of sensory cues, memory, and the costs and values associated with foraging decisions (1-3). The sensory cues used to guide foraging can be either discrete or gradient-based. For instance, moths, dogs and humans navigate odor gradients using characteristic casting and zig-zagging behaviors in response to dynamic somatosensory and olfactory cues (4-6). While the successful execution of these behaviors would be expected to strongly rely on the integration of rapidly changing, weak and noisy sensory information, previous work has primarily focused on computations involved in cost/value decisions related to the exploration/exploitation tradeoff (1, 2, 7-9), rather than whether and how foraging behavior is refined through learned associations between these dynamic sensory cues and reinforcement signals (but see, 10-12).

Accumulating evidence suggests that sensory learning in mature animals reflects the coordinated activation of sensory brain areas and neuromodulatory control nuclei (13). Of these neuromodulatory systems, cholinergic and dopaminergic neurons in the nucleus basalis and ventral tegmental area, respectively, have been observed to code cognitive operations of cue detection (14, 15) and reward prediction (16) associated with behaviorally relevant sensory stimuli and to subserve learning in complex sensory-

guided tasks (17). Theoretically, these learning systems are maximally engaged by tasks that require the continuous interplay of sensory cues, dynamically updated motor action programs, and neuromodulatory feedback as occurs during the naturalistic process of sensory guided foraging. This “closed-loop” approach to perceptual training has very little in common with traditional perceptual learning studies; wherein, isolated and unpredictable stimuli are presented at low rates with sparse, temporally distant feedback signals and training improvements typically do not generalize beyond the specific practice materials. By contrast, sensory-guided foraging shares many characteristics with exploration-based, immersive sensorimotor learning tasks such as musical training (18, 19) and action video game play (20-22), which appear to promote highly generalizable improvements in sensory perception (18, 20, 21). Training protocols that engage learning circuits at high rates and result in generalized improvements in sensory perception offer appealing therapeutic options for perceptual disorders that have traditionally been considered untreatable (23, 24), making the study of sensory guided foraging behavior both theoretically interesting and clinically relevant.

Much like our foraging ancestors, the modern urbanite faces the challenge of guiding his/her behaviors using noisy, dynamic sensory cues. Examples of these conditions abound in the auditory domain, where distractors can impede communication with friends at social gatherings, instruction from teachers in classrooms, or transmission of information via a cellular phone. As the extraction of weak signals from background distractors represents a universal perceptual problem, presenting in the hearing impaired and typically hearing alike (25-30), it offers a good test case for the malleability of perceptual skill following practice on an auditory foraging task.

Using a combination of psychophysical measurements and *in vivo* neurophysiological recordings in humans and mice respectively, we examined *i.*) whether subjects could improve their efficiency on a closed-loop auditory foraging task requiring them to continuously discriminate changes in the level of a faint sound embedded in masking noise, *ii.*) the behavioral strategies used to solve the foraging task, *iii.*) if, in humans, learning in the context of foraging transferred to untrained tests of speech recognition in the presence of distractors, and *iv.*) if, in mice, foraging experience altered the neural representation of target signals and distractors in primary sensory cortex. We found that while both humans and mice learned to improve their foraging efficiency with practice, disparate behavioral strategies were employed both within and across species. Furthermore, behavioral improvements on the foraging game were associated with improved speech perception in noise abilities in humans and enhanced neural representation of weak, noisy signals in primary auditory cortex of mice.

3.2 Materials and Methods

Auditory Foraging Task Procedures

All procedures performed with humans were approved by the Human Studies Committee at Massachusetts Eye and Ear Infirmary and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. 20 young adult subjects with audiologically-confirmed normal hearing sensitivity participated in this study. Subjects were randomly assigned to train on the auditory foraging task ($n = 10$, 8 female, mean age = 28 yrs \pm 3) for one month (30 min per day

for 5 days per week) or to be passively exposed to the training stimuli (n= 10, 6 female, mean age = 26 yrs \pm 3) over the same time period. Humans controlled the movements of an avatar in a 2-dimensional virtual arena using a gamepad (Logitech F310) in the context of a custom audio game. The game was downloaded on the participants' laptop PCs and circumaural headphones (Razer Carcharias) were provided. Humans used audio feedback (level of a continuous 500 Hz tone) to guide their avatar to a location associated with the lowest sound level. A broadband masker of \sim 65 dB SPL (calibrated at their initial visit) was played continuously as a distractor. Like the mice, human subjects received no verbal instructions about the goals of the game; rather they learned to forage for rewards (points) through trial and error. All data from training sessions and passive listening were automatically uploaded to our servers at the Massachusetts Eye and Ear Infirmary using a secure file transfer protocol.

All procedures performed with mice were approved by the Animal Care and Use Committee at Massachusetts Eye and Ear Infirmary and followed the guidelines established by the National Institutes of Health for the care and use of laboratory animals. 8 male CBA-CaJ mice, aged 6 weeks, were water restricted and their light dark cycles were reversed. 4 mice were chosen to train on the auditory foraging task, and the other 4 served as passively exposed controls. Passive exposure was implemented through yoking; while one mouse was training, their yoked counterpart was placed in an elevated listening chamber situated inside the training arena. The auditory foraging task for mice was similar to the game played by humans except that it occurred in a physical (rather than virtual) space (40 x 65 cm, sound treated, training arena with overhead tweeter) and utilized a 16 kHz (rather than 500 Hz) carrier frequency for the tone. The

position of the mouse relative to the target was monitored with a webcam (Creative Labs) and custom software. If the mouse was able to navigate to the target location (a 14 cm diameter circle) and remain within this space for 2 seconds, the auditory stimulus stopped, indicating that the mouse could return to the water spout for a reward of variable magnitude. Mice generally completed 40 to 50 trials per day. Mice were trained on the foraging task for approximately three months.

Auditory Foraging Task Data Analysis

For the moment-by-moment behavioral analysis, we divided behavioral traces from training trials into movement vectors that were subsampled every 0.3 s. At any given time point, the optimal bearing towards the target could be calculated between the forager's current position and the target location. By subtracting the forager's actual movement vector at each 0.3 s behavioral 'moment' from the ideal bearing, we were able to represent, with search trajectories, how the forager's movements deviated from the optimal search vector. Toward target bias of search trajectories was quantified as the mean *cosine* of difference vectors across each trial (max possible value = 1). SNR bias was quantified in a similar fashion using the absolute value of the *sine* of the vectors subtracted from random performance, defined as the *sine* of an average vector angle of $\pi/4$ (max possible value = .707). Speed target bias and SNR bias were quantified in the same manner as described for the angular measures. The difference between these two methods was that the length of each behavioral response vector was defined as the speed at which the animal or avatar was moving (rather than 1). All vectors were binned into 1 of 16 categories from 0 to $15\pi/8$ rad in increments of $\pi/8$. The mean speed was calculated for all response vector categories and then normalized

before the cosine or sine of the vectors was determined. Due to binning effects, maximum possible speed SNR bias was 0.63.

Tests of Learning Transfer

All testing was performed in a sound-treated research booth. We tested whether learning on the foraging task transferred outside of the task demands by making pre- and post-intervention measurements of signal in noise perception using both tonal and speech stimuli. Tone in noise detection thresholds were measured using a two-interval, two-alternative forced choice procedure. Stimuli were generated and the testing protocol was implemented using the *SoundGen* system (82) to adaptively identify the threshold for 79% response accuracy (83). Thresholds were measured for tones with carrier frequencies of 250, 375, 500 and 750. Speech perception in noise was measured using a standard clinical assessment tool called the Quick Sentence in Noise Test (84, QuickSIN) that is meant to assess real-world hearing in noise abilities. This test requires that a listener extract and repeat a sentence (with low predictability) spoken by a target female speaker in the presence of four-talker babble at increasingly difficult SNRs.

Neurophysiological Recording Procedures

Trained (n = 4) and passively exposed (n =4) mice were anesthetized (120 mg/kg ketamine and 12 mg/kg xylazine) and a scalpel was used to make a 4 x 3 mm (rostrocaudal x mediolateral) craniotomy over the right auditory cortex. A 16-channel silicone probe (177 μm^2 contact area, 50 μm contacts on each of four shanks, 125 μm between shanks, NeuroNexus) was inserted orthogonal to the cortical surface to record multiunit responses from the middle cortical layers (0.3 – 0.5 mm). To ensure cortical

recordings were performed equivalently in passively exposed and trained mice, we first identified the caudal, low-frequency boundary and then proceeded to make three progressive recordings along the tonotopic axis of auditory cortex.

All acoustic stimuli were generated using a 24 bit Digital to Analog Converter (National Instruments Model PXI-4461) and delivered to the left ear of the mouse via custom miniature acoustic assemblies. Frequency Response Areas (FRA) were measured at each recording site by pseudorandomly presenting tone pips (50 ms duration and 5 ms cosine squared onset and offset ramps) with carrier frequencies of 4 to 48.5 kHz in 0.12 octave steps at intensity levels from 0 to 80 dB SPL in 5 dB steps. Additionally, we also measured FRAs, using these same procedures, under 4 continuous background noise conditions (noise = 40-70 dB SPL).

Neurophysiological Data Analysis

The raw response traces were digitized at 32-bit, 24.4 kHz (RZ5 BioAmp Processor, Tucker Davis Technologies) and stored in binary format. All subsequent analyses were performed in Matlab (Mathworks) using custom scripts. The signals were first notch filtered at 60 Hz and bandpass filtered at 300–5000 Hz with a fifth-order Butterworth filter. Multiunit spikes were identified adaptively as voltage deflections that exceeded 4.5 standard deviations from the mean recorded activity (10s running average). The boundaries of the FRAs (FRA mask) were defined objectively in most cases (85, see SI Methods).

CF was defined as the tone frequency that evoked the highest spike rate at the threshold level of the recording site (inflection point of the Spike Level function). Best

level (BL) was defined as the stimulus level presented at the site's CF that evoked the highest spike rate. Monotonicity was objectively defined as the slope of a linear regression between either the threshold or BL inflection point in a rate level function and its response to the highest stimulus level presented (further elaborated in SI Methods). The index of neural signal-to-noise ratio was defined as $\frac{S-N}{S+N}$, where S is the average response (sp/s) of each site to tones presented at 16 kHz (\pm .12 oct) and 35-60 dB SPL and N represents the average spike rate at that site recorded over a 0.1 s window beginning 150 ms before stimulus onset. The first derivatives of smoothed rate level functions (5 point median filter) were approximated using a 5 point centered numerical algorithm

$$F'(x) = \frac{-F(x+2h)+8F(x+h)-8F(x-h)+F(x-2h)}{12h} + O(h^4)$$

where, x is the signal level and h is the change in signal level. For each signal level (10 -70 dB SPL), the mean absolute value of the approximate first derivative was calculated. Fisher Information functions were computed for each recording site using the methods described by Dean et al (39). The quantification of Fisher Information is detailed in the SI Methods. The PSTH classifier model (40) compared the Euclidean distances between the population single trial spike train elicited by a given stimulus level (5 dB range) of a 16 kHz (\pm .24 octave) tonal stimulus to response templates created for that level and neighboring stimulus levels (levels 10 dB higher and lower than the actual stimulus level). The spike train is classified as being generated in response to the stimulus level from which its distance is minimal. The classifier model is elaborated in the SI Methods.

3.3 Results

Humans and mice learn to forage in a soundscape for hidden rewards

Human participants played an auditory foraging game for approximately one half hour per day (33 ± 1 min) over the course of one month. The objective of the game was to use a remote-controlled avatar to search a 2-dimensional, virtual soundscape for the location of a hidden target before time expired (**Fig. 1A top**). The target location varied randomly from trial to trial. Visual search cues were not provided. The only available cue to locate the hidden target came from the level of a 500 Hz tone presented in a constant level of broadband masking noise. To make the task perceptually demanding, the level of the tone relative to distractor (signal-to noise-ratio, SNR) was decreased in real time according to the Euclidean distance between the subject and the hidden target location (**Fig. 1A**). Mice were engaged in a live-action version of the game played by humans; wherein, they also foraged within a 2-dimensional soundscape to find the location of a low SNR target and receive a water reward (**Fig. 1A bottom**). Thus, mice and humans learned to develop adaptive movement strategies that would reveal subtle changes in SNR, allowing them to find the virtual target location and receive reward.

Consistent with observations in insects (4) and mammals (5, 31) (including humans) moving along odor gradients, we rarely observed direct paths to the target location. Rather, we typically noted sweeping initial searches that were ultimately refined as the subjects closed on the target location. On some trials, these paths resembled the casting and zigzagging search strategies employed by insects operating

on sporadic cues and partial information (4, **Fig. 1B**). Over the course of training (humans = 1 month, mice = 3 months), both species learned to find the auditory target location more successfully ($n = 10$ humans, $P = 5 \times 10^{-7}$; $n = 4$ mice, $P = 2.9 \times 10^{-3}$, Friedman test; **Fig. 1C**) and to identify the target location more quickly (humans, $P = 3 \times 10^{-5}$; mice, $P = 3 \times 10^{-3}$, Friedman test; **Fig. 1D**). For humans, but not mice, the reduced time to target was accompanied by a decrease in average travel distance per trial (humans, $P = 3 \times 10^{-4}$; mice, $P = 0.13$, Friedman test; **Fig. 1E**). By contrast, in mice, but not humans, search speed progressively increased over training (humans, $P = 0.34$; mice, $P = 6 \times 10^{-4}$, Friedman test; **Fig. 1F**). This double dissociation between adaptive changes in pathlength and speed led us to hypothesize that the humans and mice solved the foraging task differently.

Humans and mice employ different strategies to solve the auditory foraging task

To delineate the strategies used by humans and mice in this task, we analyzed their moment-by-moment behavioral decisions by dividing behavioral traces from training trials (**Fig. 2A-C far left in black**) into movement vectors that were sampled every 0.3 s (**Fig. 2A-C colored arrows**). At any given time point, the optimal bearing towards the target could be calculated between the forager's current position and the target location (**Fig. 2D**). By subtracting the forager's movement vector at each 0.3 s behavioral 'moment' from the ideal bearing, we were able to represent, with search trajectories, how the forager's movements deviated from the optimal trajectory. For both species, we found that search trajectories were fairly randomly distributed early in training (**Fig. 2F-H top**). By later stages in training, we found that search trajectories in humans were generally biased toward the target (**Fig. 2F**) and, in many cases, along

the most informative SNR vector within the sound gradient (i.e., greatest increase or decrease in tone amplitude per unit distance, **Fig. 2G**). We quantified target bias as the degree to which subjects moved along any angle that took them closer to the target (**Fig. 2D**, magenta) and SNR bias as the degree to which subjects selected movement trajectories that provided the greatest SNR change per unit distance (either directly toward or away from the target, **Fig. 2D**, cyan). We found that all human foragers were more likely to increase their angular target bias over the course of training ($P = 2 \times 10^{-6}$, Friedman test; **Fig. 2I,J, top & Fig. S1**). This class of search strategy typically began with high-speed, wide excursions and multiple turns to likely reveal the general flow of the gradient, followed by a winding, slower local search that was most often directed toward the target (**Fig. 2A**). In addition, 6/10 human subjects also developed an SNR bias over the course of training ($P = 5 \times 10^{-3}$, Friedman test; **Fig. 2J & Fig. S1**). These subjects essentially performed coordinate descent optimization by creating orthogonal excursions along the axes within the soundscape associated with the steepest slopes in the SNR gradient (**Fig. 2B & G top**). Improved use of either strategy allowed human subjects to identify the hidden target with a reduced path length (**Fig. 1E**). As befitting their relatively constant pathlength over training, mice did not exhibit an improvement in target or SNR bias. If anything, their search trajectories became more random over the course of training. (Target bias, $P = 0.31$; reduction in SNR bias, $P = 0.02$, Friedman test; **Fig. 2H,K top and Fig. S1**).

Their improved success in the foraging task (**Fig. 1C**) and overall increase in running speed (**Fig. 1F**) suggested that mice used an alternate gradient-based strategy to solve the foraging task. When navigating a light gradient, *Chlamydomonas nivalis*, a

species of green algae, has been observed to be directly photokinetic, modulating its speed in a graded fashion relative to the “ideal” angle toward a light source (32). We tested whether mice might use a similar gradient-based strategy by calculating the running speed of each mouse with respect to the angular deviation from the optimal bearing (**Fig. 2E**). At early stages of training, running speed was not modulated by the mouse’s chosen angle. However, over the course of training, mice learned to increase their running speed when moving towards the target and along bearings associated with the most pronounced SNR changes (Target bias, $P = 3 \times 10^{-4}$; SNR bias, $P = 5 \times 10^{-4}$, Friedman test; **Fig. 2C,H bottom & K bottom as well as Fig. S1**). Humans were not observed to modulate their running speed by either strategy (Target bias, $P = 0.79$; SNR bias, $P = 0.42$, Friedman test; **Fig. 2I,J**). These findings suggested that humans and mice used different types of adaptive foraging strategies; humans learned to bias their search trajectories towards the target and in most cases also along the steepest slopes in the SNR gradient, whereas mice continued to move along a variety of angles but selectively increased their running speed according to real time changes in SNR.

Foraging strategies depend on local sensory environment

As a final step, we asked how foraging strategies learned over a period of weeks were dynamically coordinated over the course of a single trial. We first characterized whether target and SNR angular biases observed in human foragers depended on sensory context by measuring each type of bias according to position within the overall SNR gradient (**Fig 3A-B and Fig. S2**). Well-trained human subjects exhibited target bias at all SNRs (**Fig. 3A top**). Human subjects who performed coordinate descent optimization in the soundscape (Strategy B) demonstrated SNR bias in their search

trajectories at low and high SNRs. There was a dip in SNR bias at moderate SNRs that roughly coincided with the peak of the angular target bias function, suggesting that foragers who used gradient descent strategies may have flexibly switched between gradient orientation at the lowest and highest SNRs and target bias at intermediate SNRs (SNR effect, $P = 9.1 \times 10^{-11}$, ANOVA; **Fig. 3A-B top**). Contrasting trials where subjects successfully located the target within the allotted time (solid lines) versus those where they did not (dashed lines), revealed that failures in successful foraging were distinguished by strategic search differences within a region close to the low SNR target (enclosed by vertical, red lines, Target Bias, Correct vs. Failed \times SNR interaction, $P = 4.3 \times 10^{-6}$; SNR bias, Correct vs. Failed \times SNR interaction, $P = 0.69$, ANOVA; **Fig. 3A,B top**).

Although our trial-level analysis suggested that mice did not use angular target bias as a search strategy in the foraging task, our more detailed SNR-based analysis revealed that mice did, in fact, employ this strategy, albeit only at low SNRs that were local to the target ($n = 4$, SNR effect, $P < 3 \times 10^{-16}$, ANOVA; **Fig. 3C top**). At higher SNRs, we found that mouse running speed was modulated with a combination of angular target and SNR bias (Speed Target Bias; SNR effect, $P < 3 \times 10^{-16}$, Speed SNR bias; SNR effect, $P < 3 \times 10^{-16}$, ANOVA; **Fig. 3C,D bottom**). Similar to the human subjects, we noted that failures in foraging success for mice were associated with strategic differences restricted to a low SNR region local to the target (Target Bias, Correct vs. Failed \times SNR interaction, $P < 3 \times 10^{-16}$; Speed Target Bias, Correct vs. Failed \times SNR interaction, $P = 1.2 \times 10^{-10}$; Speed SNR bias, Correct vs. Failed \times SNR interaction, $P = 0.03$, ANOVA).

Finally, to further examine the dependence of foraging strategy on sensory cues, we analyzed the overall search speed as a function of SNR (**Fig. 3E,F**). Across both species, we found that search speed decreased at the lowest SNRs on successful trials (humans, SNR effect, $P < 3 \times 10^{-16}$; mice, SNR effect, $P < 3 \times 10^{-16}$, ANOVA; **Fig. 3E,F**). Importantly, unsuccessful trials were characterized by a failure to modulate search speed with sensory cues for these same SNRs (humans, Correct vs. Failed \times SNR interaction, $P = 0.04$; mice, Correct vs. Failed \times SNR interaction, $P < 3 \times 10^{-16}$, ANOVA; **Fig. 3E,F**).

To summarize, across both species, we observed that at favorable SNRs, a high speed search was conducted, driven by either choosing search trajectories that were biased toward the target and steepest changes in the SNR gradient (humans) or modulating running speed with angular target and SNR bias (mice). As the foragers moved to lower SNRs (local to the target), slower, systematic exploration dominated the search strategy, representing a strategy switch from speed modulation to choosing more accurate search trajectories in mice. Across all foragers, the behavioral strategies used to successfully solve the foraging task were only disrupted at the lowest SNRs on failed trials, suggesting that failures in slow, systematic, local exploration at locations providing the most degraded sensory feedback accounted for limitations in trial level success.

Learning on the auditory foraging task transfers to an untrained speech perception task

We next asked whether increased proficiency in the auditory foraging game generalized to other measures of auditory perception. Psychophysical tests were performed on human subjects who had been randomly assigned either to train on the foraging game for one month or had been passively exposed to game stimuli for the same time period (**Fig. 4A**). First, we assessed near transfer (**Fig. 4A middle**) by measuring detection thresholds for pure tones (250, 375, 500, 750 Hz) presented in the presence of a simple broadband masker before and after foraging or passive listening. We observed increases in response thresholds for trained subjects compared to passively exposed controls that were specific to the frequency channels (based on peripheral excitation patterns) used in the foraging task (Group x Test frequency interaction, $P = 0.05$; 250 Hz, $P = 1.38$; 375 Hz, $P = 0.24$; 500 Hz, $P = 0.17$; and 750 Hz, $P = 0.04$, ANOVA followed by *post hoc* two sample *t* tests with Holm-Bonferroni correction for multiple comparisons; **Fig. 4B**). This finding of stimulus-specific elevated detection thresholds following a task that relied primarily on stimulus discriminations is consistent with previous observations (33, 34).

We next tested whether observed improvements in using weak tones in noise to guide behavior in the auditory foraging task also transferred to more ethologically relevant situations such as understanding speech in noisy environments (**Fig. 4A,C & Fig. S3A-C**). Toward this goal, we administered the Quick Sentence in Noise Test, a common clinical tool used to assess real-world speech in noise perception. We found that subjects who trained on the foraging task improved their word recognition scores at the most difficult SNR tested (0 dBSNR) by an average of 12% (**Fig. 4C**). This represented a significant improvement compared to the passive exposure group (Group

x SNR interaction, $P = 8 \times 10^{-4}$; 0 dB SNR, $P = 1 \times 10^{-3}$, ANOVA followed by *post hoc* two sample *t* tests with Holm-Bonferroni correction for multiple comparisons; **Fig. 4C & Fig. S3A,B**) with large effect sizes measured at both +5 and 0 dBSNR (effect sizes = 0.8 and 1.9 respectively; Hedges' *g*). The small speech in noise improvements demonstrated by the passively exposed group are expected based on learning that occurs during the pretest evaluation (35). Combined with the stimulus-specific modulation of tone detection thresholds reported above, these far transfer findings demonstrate that training on an auditory foraging task is associated with both stimulus-specific and generalized learning effects.

We then examined whether improved comprehension of the most degraded speech in noise samples could be predicted from individual differences in game play strategy. While more traditional measures such success rate or time to target were not significantly predictive of learning transfer (**Fig. S4**), we found that dynamic search behaviors, specifically in low SNR conditions close to the target, were significantly predictive of improved processing of highly degraded speech (0 dBSNR, **Fig. 4D,E**). Specifically, subjects who learned to slow their search speeds and to move along the steepest slope of the SNR gradient when within 5dB of the target demonstrated the greatest improvement in speech comprehension in high levels of background speech babble (Low SNR Speed, $R = -0.78$, $R^2 = .60$ $P = 0.03$; SNR Bias at Low SNRs, $R = 0.77$, $R^2 = 0.60$, $P = 0.03$, Pearson's correlation with Holm-Bonferroni correction for multiple comparisons; **Fig. 4D,E**). Thus, the same search strategies that differentiated successful versus failed foraging trials (**Fig. 3B,E**) were also associated with the highest generalized improvement in speech comprehension. While foraging at low SNRs (local

to the target) could be accomplished using rapid motor excursions that were guided less by fine sensory cues, it was the subjects who slowed their search speeds (perhaps integrating noisy information over longer time periods) and guided their searches by the weak, noisy sensory cues who showed the most benefit on the transfer task.

Interestingly, by contrast to several previous auditory training studies that employed the same speech transfer measure, improvement on this untrained task secondary to training was not correlated with pre-training performance (35-37, $R = .14$, $R^2 = .02$, $P = 0.69$, Pearson's correlation).

Improved foraging ability is associated with a reorganized cortical representation of weak tones.

We examined the neural correlates of learning on the auditory foraging task by making unit recordings from the auditory cortex (A1) of mice that were trained on the task or passively exposed to the same auditory stimuli but did not participate in the task. We hypothesized that training on the task would be associated with an altered representation of trained stimulus features, such that the representational salience of weak, noisy inputs would be enhanced. We first collected frequency response areas (FRA) in both groups of mice by presenting pure-tones with pseudorandomly varied frequencies between 4 and 48 kHz and intensity levels from 0 to 80 dB sound pressure level (SPL, **Fig. 5A**). Training was associated with a marked over-representation of characteristic frequencies (CF, the preferred frequency at threshold) near the 16 kHz training frequency compared to passively exposed controls ($n = 151$ neural recording sites from 4 trained mice, $n = 180$ neural recording sites from 4 passively exposed mice, $P = 6 \times 10^{-4}$, two-sample Kolmogorov-Smirnov test, **Fig. 5B**). This proportional increase

in frequency tuning was not limited to 16 kHz, but rather extended a half octave above and below the training frequency. Because the foraging task emphasized recognition of subtle variations in tone level, we examined the encoding of sound frequency across the full range of levels encountered in the task.

In passively exposed control mice, increasing the tone level above threshold was associated with a monotonic increase in firing rate and little change in best frequency (BF), as has previously been described in rodent A1 (38, **Fig. 5A,C,D**). In trained mice, we observed that many rate-level functions were non-monotonic, decreasing their firing rate in response to high level stimuli ($P = 9 \times 10^{-13}$, two-sample t test; **Fig. 5C,D**) and, accordingly, were often best driven by relatively faint tone levels, near the target intensity range in the foraging task ($P = 3 \times 10^{-10}$, two-sample Kolmogorov-Smirnov test, **Fig. 5E**). By contrast to units recorded in passively exposed controls, many FRAs recorded in trained mice “leaned”, such that best frequency shifted downward by nearly one octave across the range of sound levels tested here (Group effect, $P < 3 \times 10^{-16}$; Level effect, $P < 3 \times 10^{-16}$; Group \times Level interaction, $P < 3 \times 10^{-16}$, ANOVA; **Fig. 5A,F**). Finally, we plotted the mean normalized neural response across all recording sites in order to characterize how the combination of the described distortion in frequency tuning, increase in non-monotonicity in level tuning, and the interaction of frequency tuning with presentation level might alter the representation of sounds in the absolute frequency/intensity coordinates. We found that in the trained animals, population neural activity maximized responsiveness across the frequency-intensity range of the target (**Fig. 5G,H**).

Neural responses in trained animals are resistant to suppression by continuous background noise

The SNR foraging task places a premium on suppressing the distraction imposed by the masking noise as well as enhancing the representational salience of low-level tones at the target frequency. To better understand how A1 responses were modified according to both of these perceptual demands, we derived tonal receptive fields under a background of continuous broadband masking noise ranging from 40-70 dB SPL. In passively exposed mice, increasing masker amplitude suppressed tone-evoked spiking, elevated thresholds, and restricted the range of frequency tuning (**Fig. 6A & Fig. S5**). However, FRAs measured in trained mice were more resistant to noise degradation at levels matching the background distractor intensities encountered in the foraging task (40 - 50 dB SPL, **Fig. 6A & Fig. S5**). We next asked whether this reduced suppression of neural responses to tones in the presence of a continuous distractor might also result in an improved neural SNR. Thus, at each recording site, we calculated the ratio between the response to the target signal (low level, 16 kHz tones) and the response to the continuous distractor. We found that the neural SNR index was significantly higher in the trained than passively exposed animals in the quiet condition and in the presence of low (40-50 dB SPL) but not high (60-70 dB SPL) noise levels (Group effect, $P = 0.18$; Noise effect, $P < 3 \times 10^{-16}$; Group x Noise level interaction, $P = 1 \times 10^{-4}$; Quiet, $P = 3 \times 10^{-3}$; Low Noise, $P = 9 \times 10^{-4}$; High Noise, $P = 0.02$, ANOVA followed by *post hoc* two-sample *t* tests with Holm-Bonferroni correction for multiple comparisons; **Fig. 6B**). Further analysis revealed that this improvement in neural SNR was largely due to a decrease of the neural response to the ongoing white noise stimulus across noise

levels, while the response to target signals were equivalent between the groups (Noise Response, Group effect, $P = 2 \times 10^{-5}$; Noise level effect, $P = 0.12$; Group x Noise level interaction, $P = 0.30$; Signal Response, Group effect, $P = 0.1$; Noise level effect, $P < 3 \times 10^{-16}$; Group x Noise level interaction, $P = 0.32$; **Fig. 6C**).

SNR foraging enhances the neural coding of weak signals

As a final step to characterize changes in the cortical representation of task-relevant acoustic parameters, we analyzed rate-level functions at the training frequency under varying levels of background noise. In passively exposed mice, the steeply sloping region of the rate-level functions shifted according to the masking noise level. Under levels of masking noise encountered in the training task (50 dB), this shift reduced the availability of dynamic firing rate cues for tone levels associated with the target (**Fig. 6D & Fig. S6**). By contrast, in trained mice, we found that the steepest slopes of the firing rate functions remained inside the range of weak signal levels that served as targets in the foraging task regardless of the masker level (**Fig. 6D & Fig. S6**). This relationship is captured by the first derivative of the rate-level function, which confirmed significantly steeper growth of response across weak signal levels in trained mice compared to passively exposed controls (Group effect, $P = 1 \times 10^{-3}$; SNR effect, $P = 3 \times 10^{-16}$; Group x SNR interaction, $P = 0.3 \times 10^{-16}$; -15 to 0 dB SNR, $P < 3 \times 10^{-6}$, ANOVA followed by *post hoc* two-sample *t* tests with Holm-Bonferroni correction for multiple comparisons; **Fig. 6E top & Fig. S7**). Often, the steeply sloping region of a growth function contains the most information for coding differences between stimuli because the contrast between neural responses to similar physical stimuli is high and the variability in trial-by-trial responses is low. This can be expressed quantitatively

using Fisher Information (39) for neural responses obtained from trained and passively exposed mice. Under low noise conditions, Fisher Information was low for weak signals in passively exposed mice, reaching a maximum at levels just above the masker. By contrast, the Fisher Information function peaked at weak signal levels in the trained mice, perhaps supporting the perceptual demands of the auditory foraging task (**Fig. 6E bottom and Fig. S7**).

To test whether task-related plasticity conferred any adaptive benefit to sound coding, we used an *in silico* PSTH based classifier (40) to decode tone stimulus intensity across the populations of neurons recorded in trained or passively exposed mice. In this template matching model, the neural response is classified as belonging to the stimulus class to which its Euclidean distance is shortest. We found that the classification of sound level using the neural data from the trained animals was superior to that of the passive controls under low noise conditions, indicating that the representational plasticity in trained animals supported an improved neural code for stimulus properties encountered in the foraging task (Quiet, $P = 0.12$; Low Noise, $P = 6 \times 10^{-3}$; High Noise, $P = 0.14$, Bootstrapped Permutation Test for Difference in Means with Holm-Bonferroni correction for multiple comparisons; **Fig 6F**).

3.4 Discussion

By tapping into evolutionarily conserved behavior, we were able to compare learning on a closed-loop audiomotor task in two commonly used species for neuroscience research. We expected that movement trajectories would coalesce around the general direction of the target as subjects learned to use sensory cues to

guide foraging behavior. While this “target bias” in foraging strategy was employed to some degree by all human subjects over the course of training, we also found that some subjects also learned to restrict their search trajectories to the steepest, most informative slope in the SNR gradient (both toward and away from the target). A similar strategy has been described in the echolocating Egyptian fruit bat when “locking” to a target and has been computationally shown to provide optimal discriminatory feedback for localization at the expense of detection (41). Employment of an SNR bias strategy suggests that as sensory information accumulated, most humans built a detailed model of the sensory search space. Evidence of similar modeling of the search space was not generally observed in mice. We found that mice increased angular target bias only at low SNRs, during slow search on the foraging task. By contrast to humans, most of their foraging efficiency improvements were attributable to an increased running speed when moving toward the target, a phenomenon which has previously been observed in green algae during phototaxis (32).

Following one month of training on an auditory foraging task with simple acoustic stimuli, we observed significant transfer of learning to an untrained task of speech recognition in the presence of 4-talker babble that was well predicted by game performance. Learning transfer to a more complex signal in noise task was surprising given that stimulus specificity has been repeatedly associated with sensory learning since the seminal report of Fiorentini and Berardi (42) over 3 decades ago. However, recent studies have cast doubt on the inviolate specificity of perceptual learning, suggesting that the particular training methodology may influence the degree of learning transfer (43-48). For example, experience with action video game video games has

been associated with accelerated learning of non-native phonetic contrasts (49) and enhanced visual abilities on tasks ranging from useful field of view to contrast sensitivity (20, 21, 23, 24). The key elements of action video game play that lead to appreciable transfer of visual learning are not yet clearly understood. However, the varied perceptual demands in these tasks are congruent with many of the conditions that promote learning transfer on traditional perceptual learning paradigms (45-47, 50). Musicianship represents yet another form of sensorimotor learning that shares a number of qualities with the auditory foraging game (e.g. audiomotor feedback that is both immediate and directional) and has recently been associated with generalized enhancement of auditory skills (18, 19, 51). Interestingly, musicians have been shown to outperform non-musicians on the same speech in noise perception test administered in the present study, with years of experience positively correlating with better performance (18, but see 52). Thus, it is plausible that due to the dynamic nature of the discriminanda, which, like roving reference paradigms, offers no “standard” reference stimulus, or because of the immersive game-based sensorimotor approach, learning in this auditory foraging task transferred to challenging listening contexts that were dissimilar, both acoustically and cognitively, from the conditions of the training task.

Another possible explanation for the observed learning transfer is that training to extract signals from noise might represent a more generalizable skill than the fine feature discrimination that is typically trained in perceptual learning studies. Evidence for this notion comes from a recent study in the visual system, which found that human participants who were trained to discriminate the orientation, motion or displacement of random dot stereograms in the presence of visual distractor noise, demonstrated

learning transfer to both untrained stimulus dimensions (53). While distractor stimuli in that study were similar across training and transfer test conditions, our results indicate that transfer effects can also be observed when both the stimulus and distractor in the transfer tests are more complex than the training stimuli (speech and 4-talker speech babble versus a pure tone and broadband continuous noise).

Perceptual improvements conferred by both traditional learning paradigms and action video gameplay are thought to arise from reductions of internal noise and filtering of external noise (54, 55), increased efficiency (24, 56), and improved probabilistic inference (57-59). Pertinent to the experiments reported here, probabilistic inference was measured using a tone in noise lateralization task in the experiments reported by Green and colleagues (57), demonstrating some cross-modal transfer of learning to auditory signal in noise perception following video game play. In many of these studies, neural plasticity associated with training, expressed either as induced bias or increased connection strength, was localized to connections between higher cortical areas that update movement representations based on dynamic sensory information (54, 55, 57).

We explored the neural correlates of training in A1, a comparatively early stage of cortical processing where unit responses are known to be strongly modulated by auditory associative learning (10, 11, 60-65). We noted that neural responses of trained mice were globally suppressed relative to passive controls. However, response suppression was far more robust for tone frequencies far from the target or for broadband continuous maskers (**Fig. 6C**), resulting in a relative enhancement of target signal representation. Differential suppression of neural activity in primary auditory cortex has also been observed in ferrets during engagement in a signal in noise

detection task (66), with suppression scaling indirectly with SNR and directly with performance. Similar findings have been reported in early auditory (62, 67, 68), visual (69, 70), and somatosensory cortices (71) of primates trained to discriminate targets from distractors. While the training studies mentioned here, as well as the currently reported experiment, suggest that learning to extract signals from noise alters the relative neural representation of task-specific targets and distractors in primary auditory cortex (perhaps explaining behavioral improvements on trained tasks and stimulus specific changes in tone-detection thresholds), the transfer effects observed in our study as well as another (53) suggest an additional stimulus-general effect of training, perhaps via response plasticity in sensory-motor brain areas (72, 73) or fronto-parietal networks involved in sensory distractor suppression (22, 74-76) . The latter possibility could be tested across species by making pre-and post-foraging training unit recordings in the primary auditory cortex of awake, behaving mice and recording steady-state auditory evoked potentials in behaving humans to examine attentional modulation of target and distractor responses.

The ability to improve generalized, auditory signal in noise perception through a learning paradigm makes it an appealing therapeutic for certain clinical populations. There are an estimated 48 million individuals living with hearing impairment in the US alone (77). Even after treatment with hearing aids or cochlear implants, these individuals present with substantial deficits when attempting to extract target speech signals from background talkers. There are several potential factors that contribute to this difficulty, some of which are associated with peripheral pathology (e.g. reduced spectral resolution of auditory filters, 78), and others with impaired central processing

(79-81). As the need to quickly and reliably extract signals from background noise is ubiquitous in work, educational and social contexts, improved signal in noise extraction in these listening environments may improve quality of life for these individuals.

3.5 Figures

Figure 1

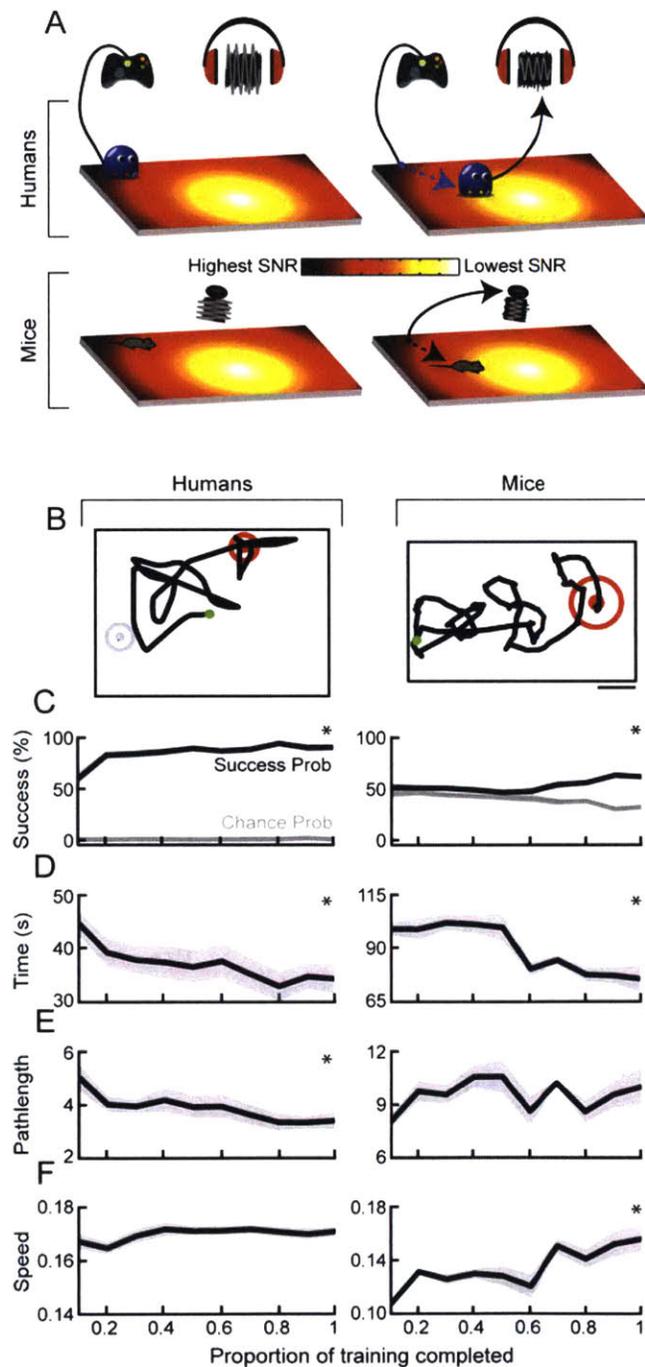


Fig 1. Humans and mice learn to use dynamic auditory cues to locate hidden targets. (A top) Humans played a video game wherein the movements of an avatar were controlled with a gamepad, while (A bottom) mice trained in a physical behavioral arena. The heat map corresponds to SNR. (B) A representative trial for a human and a mouse illustrates casting and zig-zagging behaviors along the sound gradient. Filled and open circles indicate the center and perimeter of the target (red) and "dummy" (gray) zones. Green dots indicate position at the start of the trial. Scale bar = 10 cm and ~ 3.5 cm for the mouse and human arena, respectively. Time spent in "dummy" targets provides the basis for calculating target identification by chance alone (C) Percentage of trials in which humans (left) and mice (right) located the target within the time constraints across the training period. (D) Time and (E) length of path required to complete trial as a function of training time. Pathlength taken to reach the target was normalized by the diagonal distance of the training arena. (F) Likewise, running speed is reported as normalized distance per second and plotted as a function of training time. Line plots reflect mean \pm SEM. Significant learning effects are indicated with an asterisk in the upper right hand corner of the plot.

Figure 2

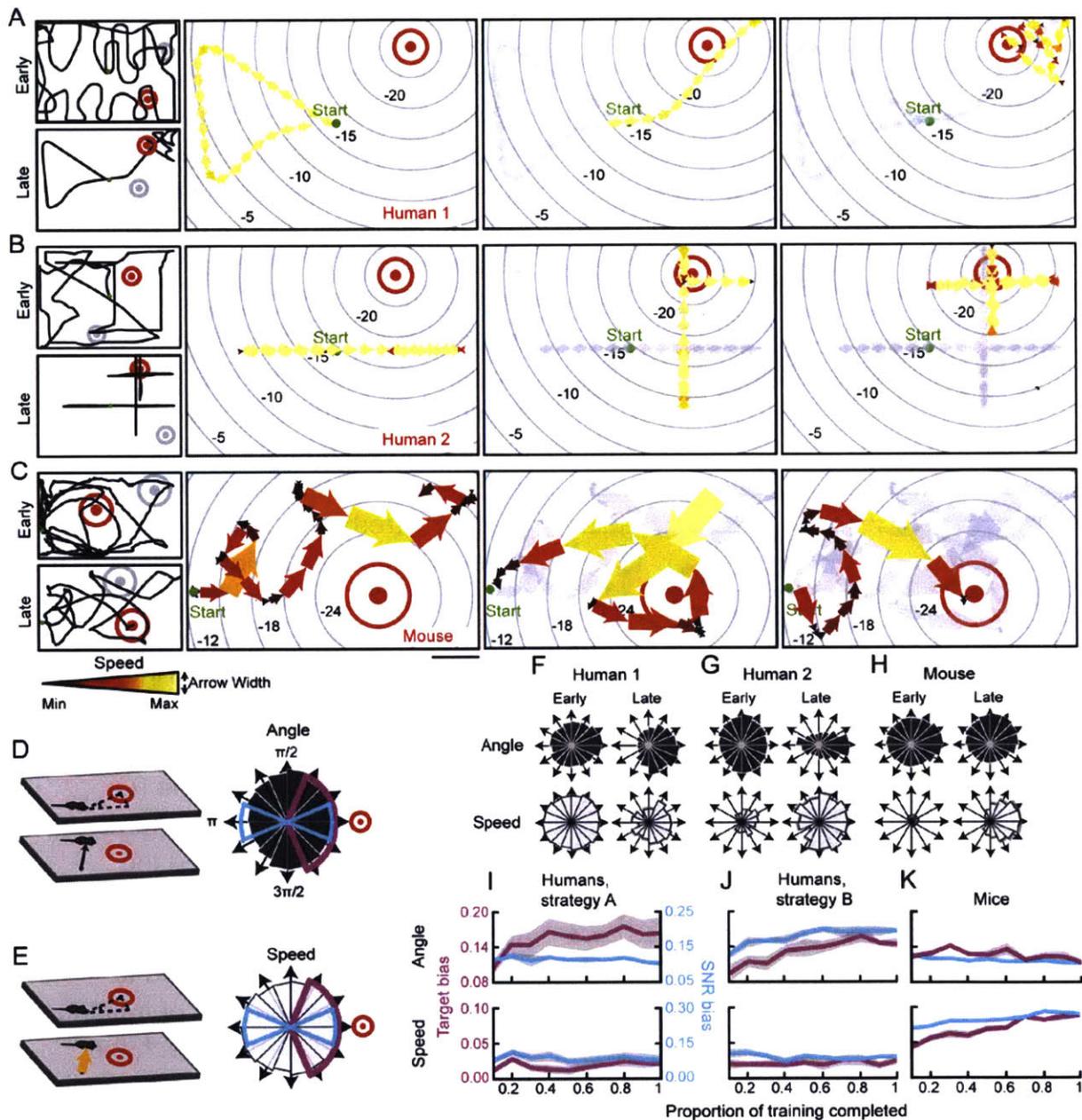


Fig 2. Adaptive sensory-guided foraging strategies emerge with practice. (A-C) (Far left) Individual early and late training trials for two humans and a mouse. (Right), Movement speed and trajectory from sequential epochs of the corresponding “late” exemplar trial (time progresses from left to right). Concentric circles demarcate the mapping of auditory SNR onto the two-dimensional training arena. Direction of arrowheads reflect trajectory, color of arrows represents search speed normalized at the trial level, and arrow size reflects search speed normalized across all 3 examples. Gray arrows are superimposed from the previous time epoch(s). Scale bar = 10 cm and ~ 3.5 cm for the

mouse and human arena, respectively. Filled and open circles indicate the center and perimeter of the target (red) and “dummy” (gray) zones. Green dots indicate position at the start of the trial. (D) The difference between actual trajectory and the ideal bearing is calculated every 0.3 s. Adaptive search strategies could emphasize movements toward the target (target bias, magenta) or along the steepest slope of the SNR gradient, (SNR bias, cyan). (E) Like angular target and SNR bias, normalized search speed can also be expressed across movement trajectories. (F-G) Normalized distributions of (*top*) search trajectories and (*bottom*) speed modulation early versus late in training for the two example human subjects and one mouse. Speed axis bar = 0.13 to 0.19 d/s in humans and 0.09 to 0.21 d/s in the mouse, mean speed (white foreground), SEM (gray background). (I-K) Target and SNR bias in movement trajectories (*top*) or speed (*bottom*) for human subjects who employed foraging strategy A, (I, n = 4) or B (J, n = 6), and all trained mice (K, n = 4) plotted as a function of training time. Line plots reflect mean \pm SEM.

Figure 3

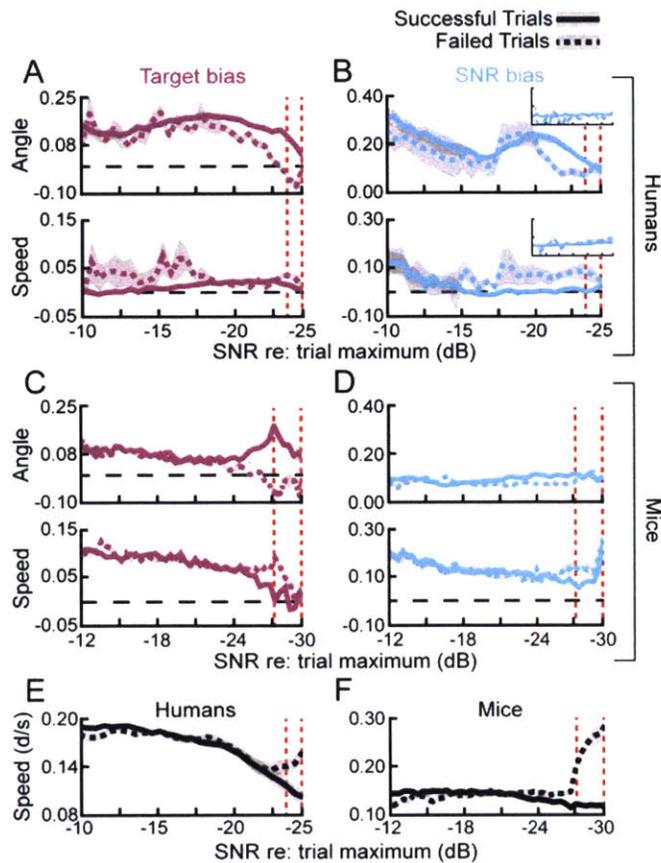


Fig 3. Foraging strategy is modulated by local sensory context. This visualization breaks down the overall foraging biases plotted in **Fig 2** according to spatial position within the SNR training arena. For all plots, behavioral data are shown from well-trained subjects (second half of training) according to spatial proximity to the target, expressed as SNR. Broken vertical red lines indicate target SNRs. Solid and broken lines reflect data from successful (i.e., rewarded) and failed trials, respectively. Broken horizontal black bars indicate unbiased foraging behavior. (A,C) Target bias in angular search trajectory (top) and speed (bottom) for all humans (A) and mice (C). (B,D) SNR bias in angular search trajectory (top) and speed (bottom) for humans (B) and mice (D). SNR bias is plotted separately for subjects utilizing Strategy B versus those that did not (Strategy A, inset). (E-F) Overall search speed is plotted as a function of distance from target in humans (E) and mice (F). The unit of measurement for speed (d/s) is distance traveled, normalized to the diagonal length of the training arena, per second. Data are plotted as mean \pm SEM.

Figure 4

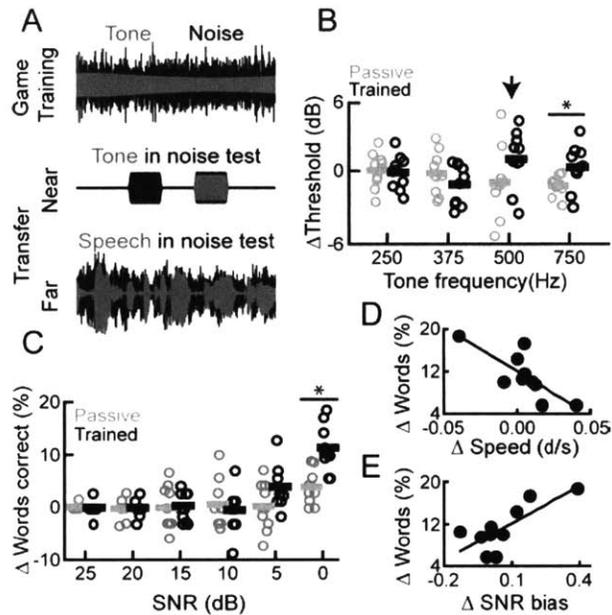


Fig 4. Learned improvements in auditory-guided foraging generalize to distinct listening tasks. (A) The physical stimulus waveform used in the foraging task (top) is similar to the stimulus used for the tone in noise task (middle) but dissimilar to the test of speech perception in multi-talker babble (bottom). (B) Change in tone detection thresholds (Post – Pre) assessed at four tone frequencies (training frequency indicated by arrow). (C) Change in words correctly recognized for the speech in noise task (Post – Pre) plotted according to target speaker SNR. (D-E) Correlation between change in search speed (D) and SNR bias (E) at locations between 0 and 5 dB SNR and improved performance on the speech in noise test at 0 dB SNR. Horizontal lines in scatter plots reflect group means. Asterisks indicate statistically significant differences between groups.

Figure 5

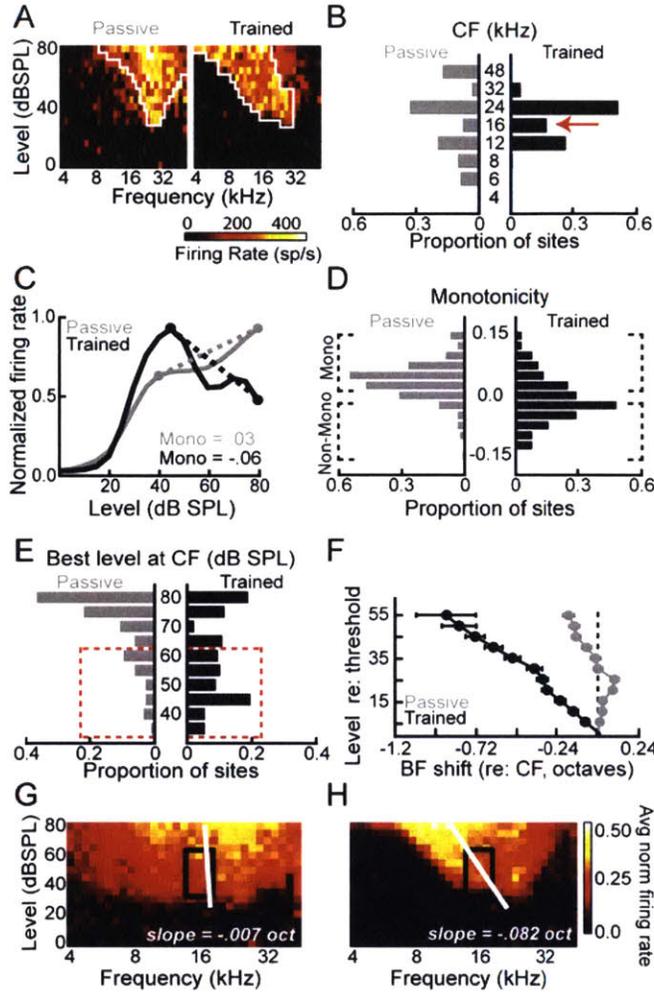


Fig 5. Low SNR foraging is associated with a reorganized cortical representation of the target stimulus. (A) Frequency response areas (outlined in white) were delineated at each recording site by measuring multiunit spikes (sp) to tone pips presented at 527 frequency x level combinations. (B) Distribution of characteristic frequencies (CF) measured from trained (black) and passively exposed control (gray) mice. Arrow indicates training frequency. (C) Representative multiunit rate-level functions from a trained and passively exposed mouse. Monotonicity (mono) is calculated from the slope of the linear fit line applied to the high-intensity region of the rate level function (dashed lines). (D) Histogram of monotonicity values recorded in trained and passively exposed mice. Brackets delineate monotonic versus low-intensity tuned, non-monotonic.slope values.

(E) Histogram of preferred sound level measured in trained and passively exposed recording sites. Red lines demarcate levels used in the foraging task. (F) Best frequency plotted separately for each intensity cross-section within the FRA (adjusted to the CF and threshold for each recording site). (G-H) The normalized FRA averaged across all recording sites for passively exposed (G) and trained (H) mice. The frequency and level range of the target tone from the foraging task is designated by the black rectangles. The white line depicts the linear regression fit to the best frequency at each sound level. Error bars represent SEM.

Figure 6

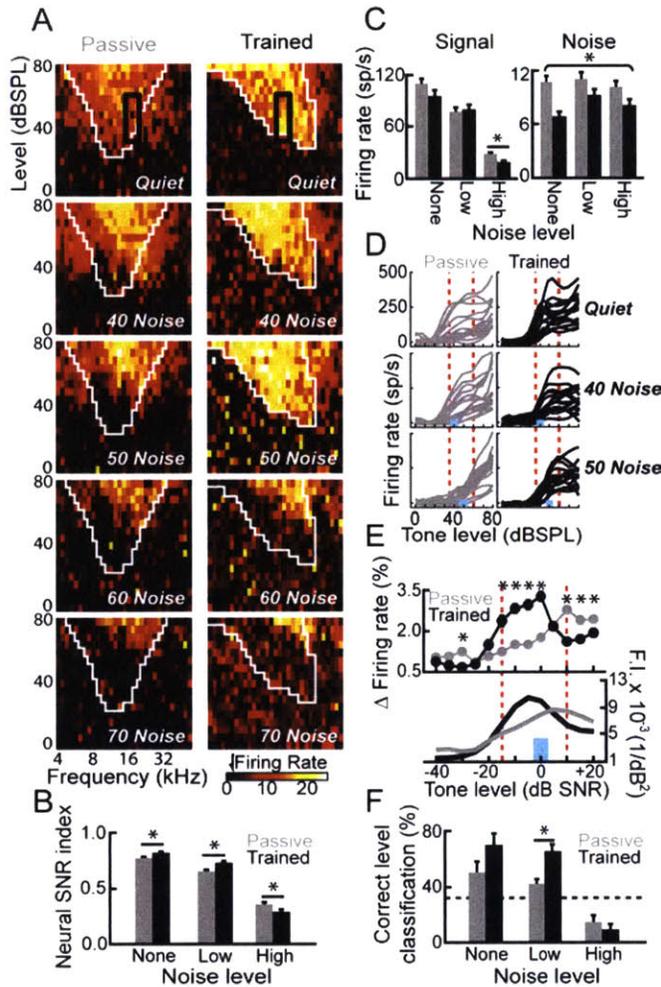


Fig 6. Auditory foraging enhances cortical encoding of weak tones embedded in noise. (A) Example FRAs recorded from a single A1 site in a passively exposed (left) and trained (right) mouse in the presence of varying levels of continuous masking noise. White outline designates receptive field boundaries without masking noise. Black rectangle and bracket indicate the frequency/intensity of target and the masking noise level in the training task, respectively. (B) Neural SNR in quiet, low (40-50 dB), and high levels (60-70 dB) of masking noise. (C) Firing rate measured during the response to the target tone (signal) versus a pre-stimulus period (noise). (D) Example 16 kHz rate-level functions. (E) Mean absolute change in normalized firing rate between neighboring sound levels (Top) and mean Fisher Information (bottom) for the target frequency at various signal to noise ratios (50 dB masking noise). For D-E, vertical red lines and cyan rectangles indicate the target intensity and masking noise level, respectively. (F) *In silico* classification of tone level from individual trials in passive and trained neural recordings. Broken black line indicates chance classification. All data reflect mean \pm SEM. Significant pairwise differences and group effects (C) are indicated with asterisks.

various signal to noise ratios (50 dB masking noise). For D-E, vertical red lines and cyan rectangles indicate the target intensity and masking noise level, respectively. (F) *In silico* classification of tone level from individual trials in passive and trained neural recordings. Broken black line indicates chance classification. All data reflect mean \pm SEM. Significant pairwise differences and group effects (C) are indicated with asterisks.

3.6 References

1. Stephens D & Krebs J (1987) *Foraging theory* (Princeton University Press, Princeton) p 239.
2. Charnov EL (1976) Optimal foraging, the marginal value theorem. *Theor Popul Biol.* 9(2):129-136.
3. Greggers U & Menzel R (1993) Memory dynamics and foraging strategies of honeybees. *Behav. Ecol. Sociobiol.* 32(1):17-29.
4. Kennedy JS (1983) Zigzagging and casting as a programmed response to wind-borne odor - a review. *Physiol. Entomol.* 8(2):109-120.
5. Porter J, *et al.* (2007) Mechanisms of scent-tracking in humans. *Nat. Neurosci.* 10(1):27-29.
6. Thesen A, Steen JB, & Doving KB (1993) Behavior of dogs during olfactory tracking. *J. Exp. Biol.* 180:247-251.
7. Kolling N (2012) Neural mechanisms of foraging. *Science* 336(6077):95-98.
8. Hayden BY, Pearson JM, & Platt ML (2011) Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.* 14(7):933-939.
9. Kvitsiani D, *et al.* (2013) Distinct behavioural and network correlates of two interneuron types in prefrontal cortex. *Nature* 498(7454):363-366.
10. Bao SW, Chang EF, Woods J, & Merzenich MM (2004) Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. *Nat. Neurosci.* 7(9):974-981.
11. Polley DB, Heiser MA, Blake DT, Schreiner CE, & Merzenich MM (2004) Associative learning shapes the neural code for stimulus magnitude in primary auditory cortex. *Proc. Natl. Acad. Sci. USA* 101(46):16351-16356.

12. Bergan JF, Ro P, Ro D, & Knudsen EI (2005) Hunting increases adaptive auditory map plasticity in adult barn owls. *J. Neurosci.* 25(42):9816-9820.
13. Seitz A & Watanabe T (2005) A unified model for perceptual learning. *Trends in Cognitive Sciences* 9(7):329-334.
14. Parikh V, Kozak R, Martinez V, & Sarter M (2007) Prefrontal acetylcholine release controls cue detection on multiple timescales. *Neuron* 56(1):141-154.
15. Froemke RC, *et al.* (2013) Long-term modification of cortical synapses improves sensory perception. *Nat. Neurosci.* 16(1):79-88.
16. Schultz W, Dayan P, & Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593-1599.
17. Leach ND, Nodal FR, Cordery PM, King AJ, & Bajo VM (2013) Cortical cholinergic input is required for normal auditory perception and experience-dependent plasticity in adult ferrets. *J. Neurosci.* 33(15):6659-6671.
18. Parbery-Clark N (2009) Musician enhancement for speech-in-noise. *Ear and Hearing* 30(6):653-661.
19. Kraus N & Chandrasekaran B (2010) Music training for the development of auditory skills. *Nature Reviews Neuroscience* 11(8):599-605.
20. Green CS & Bavelier D (2003) Action video game modifies visual selective attention. *Nature* 423(6939):534-537.
21. Li RJ, Polat U, Makous W, & Bavelier D (2009) Enhancing the contrast sensitivity function through action video game training. *Nat. Neurosci.* 12(5):549-551.
22. Anguera JA, *et al.* (2013) Video game training enhances cognitive control in older adults. *Nature* 501(7465):97-101.

23. Li JR, *et al.* (2013) Dichoptic training enables the adult amblyopic brain to learn. *Curr. Biol.* 23(8):R308-R309.
24. Li RW, Ngo C, Nguyen J, & Levi DM (2011) Video-game play induces plasticity in the visual system of adults with amblyopia. *PLoS Biol.* 9(8).
25. Sperling AJ, Lu ZL, Manis FR, & Seidenberg MS (2005) Deficits in perceptual noise exclusion in developmental dyslexia. *Nat. Neurosci.* 8(7):862-863.
26. Ziegler JC, Pech-Georgel C, George F, Alario FX, & Lorenzi C (2005) Deficits in speech perception predict language learning impairment. *Proc. Natl. Acad. Sci.* 102(39):14110-14115.
27. Kim SH, Frisina RD, Mapes FM, Hickman ED, & Frisina DR (2006) Effect of age on binaural speech intelligibility in normal hearing adults. *Speech Communication* 48(6):591-597.
28. Ruggles D, Bharadwaj H, & Shinn-Cunningham BG (2012) Why middle-aged listeners have trouble hearing in everyday settings. *Curr. Biol.* 22(15):1417-1422.
29. Kidd G, Arbogast TL, Mason CR, & Walsh M (2002) Informational masking in listeners with sensorineural hearing loss. *Jaro* 3(2):107-119.
30. Marrone N, Mason CR, & Kidd G (2008) The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *J. Acoust. Soc. Am.* 124(5):3064-3075.
31. Khan AG, Sarangi M, & Bhalla US (2012) Rats track odour trails accurately using a multi-layered strategy with near-optimal sampling. *Nature Communications* 3.
32. Hill NA & Hader DP (1997) A biased random walk model for the trajectories of swimming micro-organisms. *J. Theor. Biol.* 186(4):503-526.

33. Sabin AT, Eddins DA, & Wright BA (2012) Perceptual learning evidence for tuning to spectrotemporal modulation in the human auditory system. *J. Neurosci.* 32(19):6542-6549.
34. Fitzgerald MB & Wright BA (2005) A perceptual learning investigation of the pitch elicited by amplitude-modulated noise. *J. Acoust. Soc. Am.* 118(6):3794-3803.
35. Song JH, Skoe E, Banai K, & Kraus N (2012) Training to improve hearing speech in noise: Biological mechanisms. *Cereb. Cortex* 22(5):1180-1190.
36. Henderson Sabes J & Sweetow RW (2007) Variables predicting outcomes on listening and communication enhancement (lace) training. *Int J Audiol* 46(7):374-383.
37. Olson AD, Preminger JE, & Shinn JB (2013) The effect of lace dvd training in new and experienced hearing aid users. *J Am Acad Audiol* 24(3):214-230.
38. Polley DB, Read HL, Storace DA, & Merzenich MM (2007) Multiparametric auditory receptive field organization across five cortical fields in the albino rat. *J. Neurophysiol.* 97(5):3621-3638.
39. Dean I, Harper NS, & McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. *Nat. Neurosci.* 8(12):1684-1689.
40. Foffani G & Moxon KA (2004) PSTH-based classification of sensory stimuli using ensembles of single neurons. *J. Neurosci. Methods* 135(1-2):107-120.
41. Yovel Y, Falk B, Moss CF, & Ulanovsky N (2010) Optimal localization by pointing off axis. *Science* 327(5966):701-704.
42. Fiorentini A & Berardi N (1980) Perceptual-learning specific for orientation and spatial-frequency. *Nature* 287(5777):43-44.

43. Jeter PE, Doshier BA, Liu SH, & Lu ZL (2010) Specificity of perceptual learning increases with increased training. *Vision Res.* 50(19):1928-1940.
44. Ahissar M & Hochstein S (1997) Task difficulty and the specificity of perceptual learning. *Nature* 387(6631):401-406.
45. Xiao LQ, *et al.* (2008) Complete transfer of perceptual learning across retinal locations enabled by double training. *Curr. Biol.* 18(24):1922-1926.
46. Harris H, Gliksberg M, & Sagi D (2012) Generalized perceptual learning in the absence of sensory adaptation. *Curr. Biol.* 22(19):1813-1817.
47. Zhang JY, *et al.* (2010) Rule-based learning explains visual perceptual learning and its specificity and transfer. *J. Neurosci.* 30(37):12323-12328.
48. Resnik J, Sobel N, & Paz R (2011) Auditory aversive learning increases discrimination thresholds. *Nat. Neurosci.* 14(6):791-796.
49. Lim SJ & Holt LL (2011) Learning foreign sounds in an alien world: Video game training improves non-native speech categorization. *Cognitive Science* 35(7):1390-1405.
50. Green CS & Bavelier D (2012) Learning, attentional control, and action video games. *Curr. Biol.* 22(6):R197-R206.
51. Wong PCM, Skoe E, Russo NM, Dees T, & Kraus N (2007) Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nat. Neurosci.* 10(4):420-422.
52. Ruggles DR, Freyman RL, & Oxenham AJ (2014) Influence of musical training on understanding voiced and whispered speech in noise. *PLoS ONE* 9(1):e86980.

53. Chang DHF, Kourtzi Z, & Welchman AE (2013) Mechanisms for extracting a signal from noise as revealed through the specificity and generality of task training. *J. Neurosci.* 33(27):10962-10971.
54. Doshier BA & Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc. Natl. Acad. Sci.* 95(23):13988-13993.
55. Doshier BA & Lu ZL (1999) Mechanisms of perceptual learning. *Vision Res.* 39(19):3197-3221.
56. Levi DM (2005) Perceptual learning in adults with amblyopia: A reevaluation of critical periods in human vision. *Dev. Psychobiol.* 46(3):222-232.
57. Green CS, Pouget A, & Bavelier D (2010) Improved probabilistic inference as a general learning mechanism with action video games. *Curr. Biol.* 20(17):1573-1579.
58. Bavelier D, Green CS, Pouget A, & Schrater P (2012) Brain plasticity through the life span: Learning to learn and action video games. *Annual Review of Neuroscience, Vol 35* 35:391-416.
59. Bejjanki VR, Beck JM, Lu ZL, & Pouget A (2011) Perceptual learning as improved probabilistic inference in early sensory areas. *Nat. Neurosci.* 14(5):642-U139.
60. Recanzone GH, Schreiner CE, & Merzenich MM (1993) Plasticity in the frequency representation of primary auditory-cortex following discrimination-training in adult owl monkeys. *J. Neurosci.* 13(1):87-103.

61. Polley DB, Steinberg EE, & Merzenich MM (2006) Perceptual learning directs auditory cortical map reorganization through top-down influences. *J. Neurosci.* 26(18):4970-4982.
62. Blake DT, Heiser MA, Caywood M, & Merzenich MM (2006) Experience-dependent adult cortical plasticity requires cognitive association between sensation and reward. *Neuron* 52(2):371-381.
63. Engineer CT, *et al.* (2014) Speech training alters tone frequency tuning in rat primary auditory cortex. *Behav. Brain Res.* 258:166-178.
64. Edeline JM, Pham P, & Weinberger NM (1993) Rapid development of learning-induced receptive-field plasticity in the auditory-cortex. *Behav. Neurosci.* 107(4):539-551.
65. David SV, Fritz JB, & Shamma SA (2012) Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl. Acad. Sci.* 109(6):2144-2149.
66. Atiani S, Elhilali M, David SV, Fritz JB, & Shamma SA (2009) Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. *Neuron* 61(3):467-480.
67. Blake DT, Strata F, Churchland AK, & Merzenich MM (2002) Neural correlates of instrumental learning in primary auditory cortex. *Proc. Natl. Acad. Sci.* 99(15):10114-10119.
68. Beitel RE, Schreiner CE, Cheung SW, Wang XQ, & Merzenich MM (2003) Reward-dependent plasticity in the primary auditory cortex of adult monkeys

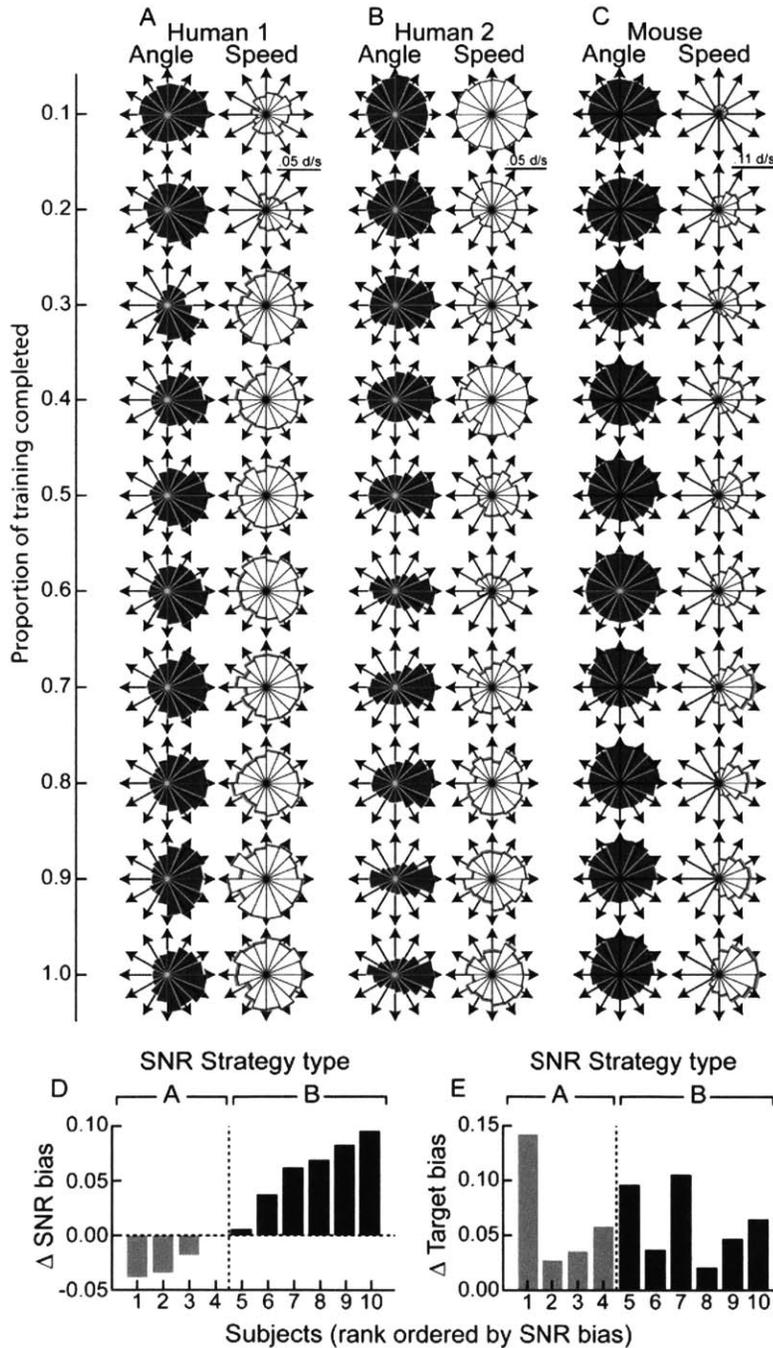
- trained to discriminate temporally modulated signals. *Proc. Natl. Acad. Sci.* 100(19):11070-11075.
69. Chen Y, *et al.* (2008) Task difficulty modulates the activity of specific neuronal populations in primary visual cortex. *Nat. Neurosci.* 11(8):974-982.
 70. Boudreau CE, Williford TH, & Maunsell JHR (2006) Effects of task difficulty and target likelihood in area v4 of macaque monkeys. *J. Neurophysiol.* 96(5):2377-2387.
 71. Spingath EY, Kang HS, Plummer T, & Blake DT (2011) Different neuroplasticity for task targets and distractors. *Plos One* 6(1).
 72. Znamenskiy P & Zador AM (2013) Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. *Nature* 497(7450):482-485.
 73. Law CT & Gold JI (2008) Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat. Neurosci.* 11(4):505-513.
 74. Zanto TP, Rubens MT, Thangavel A, & Gazzaley A (2011) Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat. Neurosci.* 14(5):656-661.
 75. Bavelier D, Achtman RL, Mani M, & Focker J (2012) Neural bases of selective attention in action video game players. *Vision Res.* 61:132-143.
 76. Mishra J, Zinni M, Bavelier D, & Hillyard SA (2011) Neural basis of superior performance of action video game players in an attention-demanding task. *J. Neurosci.* 31(3):992-998.
 77. Lin FR, Niparko JK, & Ferrucci L (2011) Hearing loss prevalence in the united states. *Archives of Internal Medicine* 171(20):1851-1852.

78. Florentine M, Buus S, Scharf B, & Zwicker E (1980) Frequency-selectivity in normally-hearing and hearing-impaired observers. *Journal of Speech and Hearing Research* 23(3):646-669.
79. Doherty KA & Lutfi RA (1999) Level discrimination of single tones in a multitone complex by normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 105(3):1831-1840.
80. Buran BN, *et al.* (2014) A sensitive period for the impact of hearing loss on auditory perception. *The Journal of Neuroscience* 34(6):2276-2284.
81. de Villers-Sidani E, *et al.* (2010) Recovery of functional and structural age-related changes in the rat primary auditory cortex with operant training. *Proc. Natl. Acad. Sci.* 107(31):13900-13905.
82. Naber M (2008) Soundgen a web services based sound generation system for the psychoacoustics laboratory. Master of Electrical Engineering and Computer Science (Massachusetts Institute of Technology, Cambridge).
83. Levitt H (1971) Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.* 49(2):467-477.
84. Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, & 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. *J. Acoust. Soc. Am.* 116(4):2395-2405.
85. Guo W, *et al.* (2012) Robustness of cortical topography across fields, laminae, anesthetic states, and neurophysiological signal types. *J. Neurosci.* 32(27):9159-9172.

3.7 Supplemental Materials

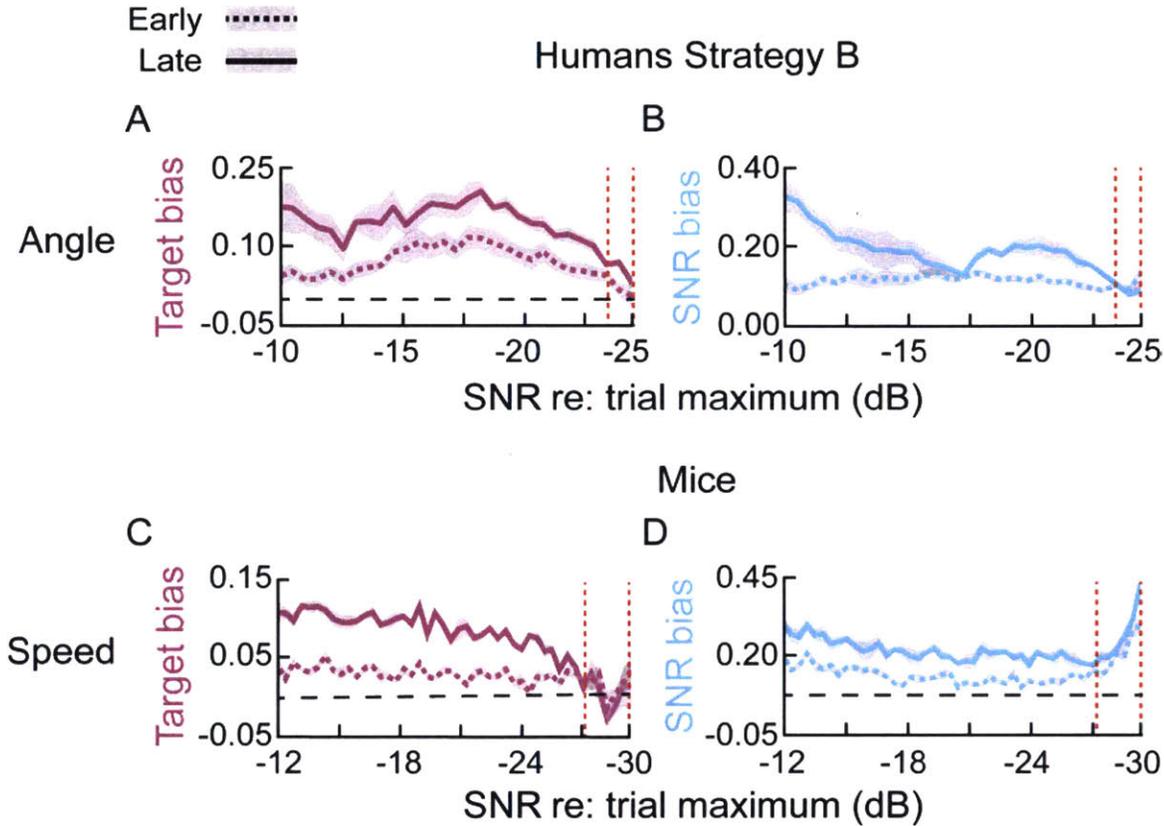
3.7.1 Supplemental Figures

Supplemental Figure 1



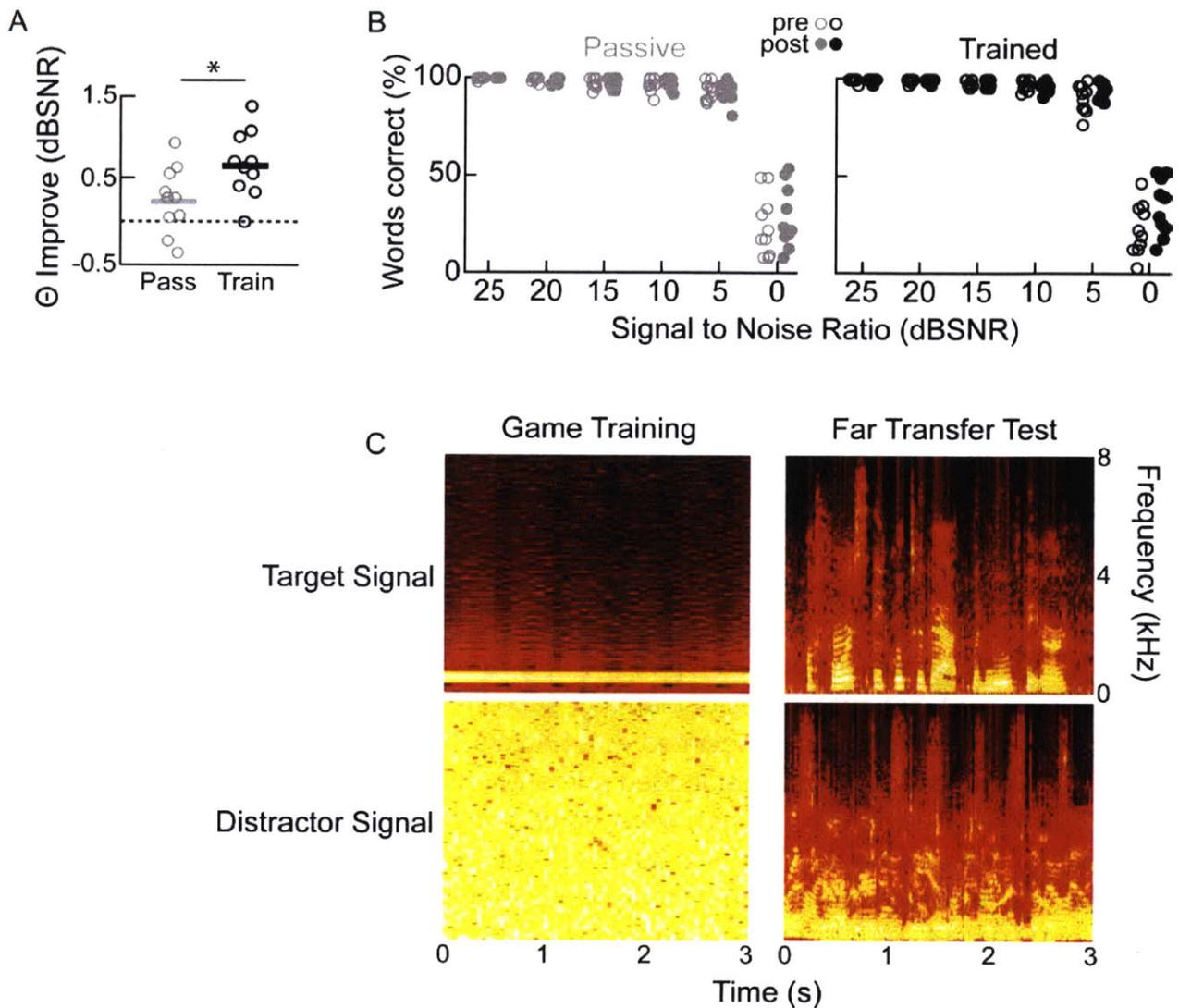
Supp Fig 1. Tracking the development of distinct foraging strategies over the course of training. (A-C) Distributions of trajectories (filled bars) and speed modulation (unfilled bars) at various stages of training in three example subjects. Gray outline reflects SEM. Human subjects 1 and 2 are representative of foraging strategy A (target bias) and B (SNR bias), respectively. (D) Foraging strategy for each human subject rank-ordered by the change (Late–Early) in SNR bias over the course of training. (E) Change in target bias based on the same subject ordering in (D) illustrates that change in target bias was unrelated to change in SNR bias, though subject 1 does show the largest negative change in SNR bias and positive change in target bias.

Supplemental Figure 2



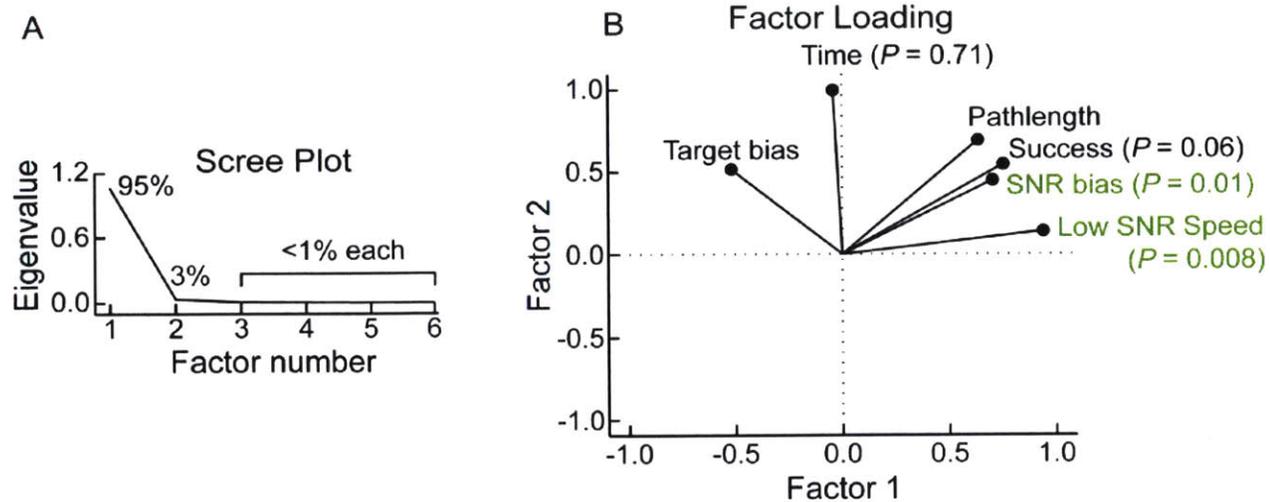
Supp Fig 2. Humans and mice increased their use of adaptive foraging strategies across SNR conditions in the task. (A-D) Comparing early (first 10%) versus late (last 10%) behavioral runs, reveals that improvements in adaptive, sensory guided foraging behaviors were observed at both low and high SNRs for (A-B) humans who used strategy B and (C-D) mice. Note that these figures are plotted in the same fashion as those found in Fig. 3 (search bias vs. SNR). However, in this case, early and late training trials are compared rather than successful and failed trials in well trained subjects.

Supplemental Figure 3



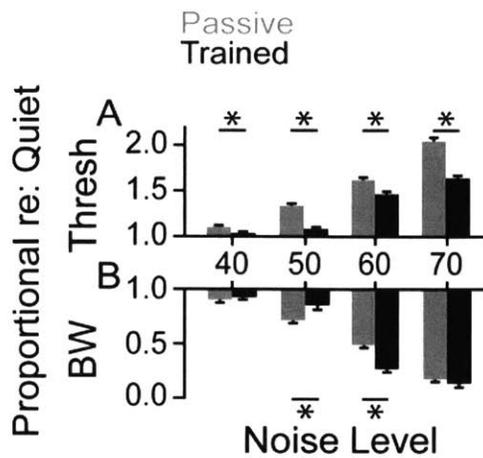
Supp Fig 3. Foraging training also improves speech in noise threshold. (A) The threshold improvements reported for each subject using the clinical scoring methods described in the QuickSIN manual also demonstrate a significant improvement following training ($P = 0.013$, Bootstrapped Permutation Test for Difference in Means) with a large effect size (1.04 Hedges' g). The black broken line reflects no change in performance. The horizontal gray and black lines represent group means. (B) Absolute QuickSIN scores for Passive and Trained subjects provide a numerical basis for the change in speech processing described in Fig. 4c. (C) Spectrograms illustrate the difference in spectro-temporal complexity of target and distractor stimuli used in the foraging game versus the speech in noise perception task.

Supplemental Figure 4



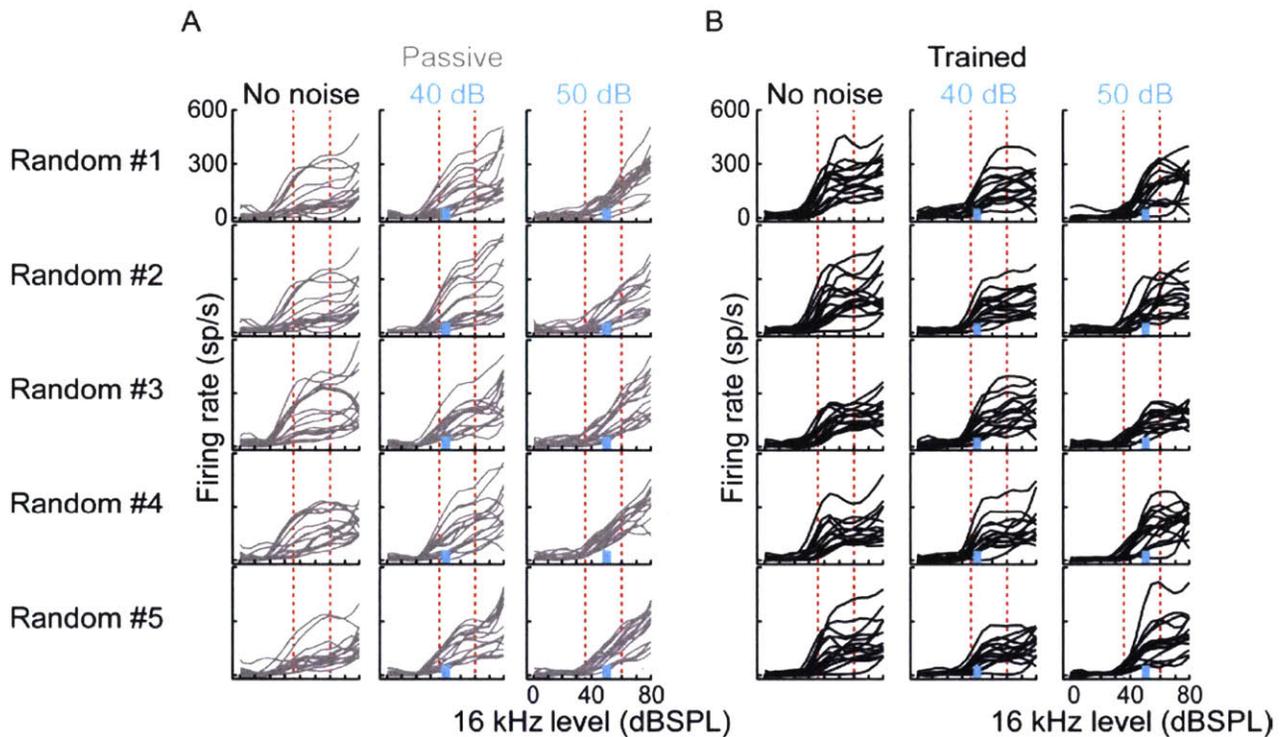
Supp Fig 4. Relationships between changes in foraging behaviors and the far transfer task. (A) The six behaviors that changed over training described in Figs. 1 and 3 were subjected to dimensionality reduction using principle components analysis. We found that the first two factors accounted for 98% of the variance. (B) Subsequent factor analysis revealed that two indices of learning, speed at low SNRs and time to solve the task, were nearly exclusively associated with Factors 1 and 2, respectively. We then found that behavioral markers that loaded primarily onto Factor 1 (e.g., speed at Low SNRs and SNR bias near the target) were significantly correlated with speech perception in noise ($R = -0.77$, $R^2 = 0.60$, $P = 0.008$; $R = 0.77$, $R^2 = 0.60$, $P = 0.01$, respectively), whereas markers that loaded onto Factor 2 or a mixture of both factors were not ($R = 0.14$, $R^2 = 0.02$, $P = 0.71$ and $R = 0.62$, $R^2 = 0.38$, $P = 0.06$ for Time and Success, respectively). In summary, this analysis revealed that after Holm-Bonferroni correction for multiple (4) comparisons, the SNR bias and search speed at low SNRs were predictive of transfer effects ($P = .03$ in both cases, **Fig 4D-E**).

Supplemental Figure 5



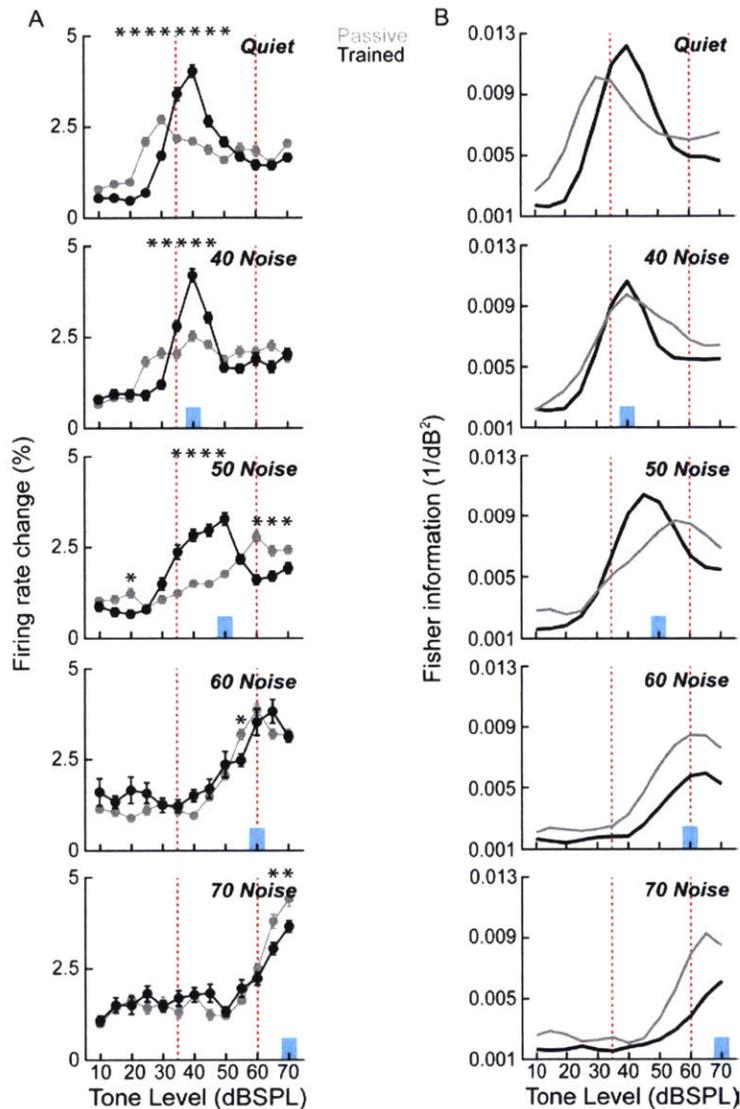
Supp Fig 5. (A-B) Threshold (thresh) elevation (A) and narrowing of (BW) frequency tuning (B) at 60 dB SPL with the addition of masking noise. Data points closer to the x-axis indicate greater similarity to the response in quiet (i.e., enhanced resistance to masking noise interference). Data points reflect mean values and error bars represent SEM. Group x Noise level interactions were statistically significant for the threshold and bandwidth data ($P < 2 \times 10^{-3}$). Asterisks indicate statistical significance assessed with *post hoc* two-sample *t* tests with Holm-Bonferroni correction for multiple comparisons ($P < 0.05$).

Supplemental Figure 6



Supp Fig 6. Additional examples of 16 kHz rate-level functions from Trained and Passively exposed A1 recordings. Recording sites that were highly responsive to 16 kHz tones (max firing rate ≥ 100 sp/s) were randomly selected with replacement from each group under each noise conditions 5 consecutive times. The top row of rate-level functions is replotted from **Fig. 6d**, to illustrate that the difference depicted in the main text was representative. Red broken lines enclose the stimulus levels associated with reward in the training task. Cyan rectangles on the x axes depict the level of the continuous noise masker.

Supplemental Figure 7



Supp Fig 7. Training increases coding information for weak signals presented in noise. (A) Change in absolute firing rate per unit change in signal magnitude was calculated (Methods) for data collected in quiet and 4 levels of background noise. Red broken lines enclose the stimulus levels that predicted reward in the training task. Cyan rectangles on the x axes depict the level of the continuous noise masker. (B) Using the same color scheme, the average population Fisher information is plotted for quiet and noise conditions. Increases in masking level shifted peaks in rate-level function slopes and Fisher information to the right for the data collected from passively exposed mice. In general, the peaks of these functions were at or slightly above the masking levels (as functions

flattened across weak signal levels). By contrast, at low masking levels, the peaks in firing rate function slopes as well as Fisher information remained within the weak levels that predicted reward in the foraging task (i.e. below the masker level) for the data collected from trained mice. At high noise levels (60–70 dB SPL), data from the trained and passively exposed mice were nearly identical. Data points reflect mean values and error bars represent SEM. Significance testing performed with *post hoc* two-sample *t* tests with Holm-Bonferroni correction for multiple comparisons, asterisks indicate $P \leq .05$.

Chapter 4. Closed-loop audiomotor training enhances the perception of speech in noise: A randomized double-blinded placebo-controlled trial

Abstract

Sensorimotor training with action videogames provides a transfer of learning to a wide range of selective attention and feature detection tasks. Boosting generalized perceptual processing abilities through action videogame training could have far-ranging therapeutic potential for individuals with sensory processing disorders. However, the essential features of effective action games are unknown and the possible influence of placebo effects in these and other cognitive intervention studies have been called into question. Action videogames ‘close the loop’ between motor actions and sensory feedback, but also present graphic, emotionally arousing stimuli and intense multisensory stimulation. We programmed a tablet-based audiomotor action game that captured the closed-loop aspect of action gaming and tested whether it was sufficient to drive generalized perceptual improvements beyond those found with a ‘placebo’ auditory working memory game. We randomly assigned hearing-impaired older adults to eight weeks of closed-loop training with synthetic sounds or memory training with speech stimuli and assessed low-level auditory feature processing, speech processing, and attentional control before training, just after training, and at a two-month follow-up. We found that working memory training provided an effective placebo but imparted no benefits on psychophysical measures of memory or hearing ability. Closed-loop training provided significant, but short-term, gains on various measures of sentence comprehension in background noise, the most common hearing complaint of older age for which assistive devices offer little benefit. Thus, training in a simplified action audio

game can provide generalized benefits for real-world perceptual challenges in sensory-impaired elderly subjects that cannot be attributed to placebo effects.

4.1 Introduction

Perceptual abilities can be enhanced with practice, but these benefits often do not extend beyond the narrow constraints of the training task (1–3). While this specificity has afforded some clues concerning the processing stages at which neural plasticity may occur following certain types of perceptual training (4–7), it has also limited the practical utility of perceptual learning protocols to provide benefits outside of the laboratory. However, more recent experiments involving sensorimotor training approaches like action video game play and musicianship have suggested that broad transfer of learning to selective attention (8–13) and feature detection tasks (14–16) is possible following behavioral intervention. Training protocols that drive generalized perceptual improvements hold intriguing therapeutic potential for the rehabilitation of individuals with sensory impairments (17–20), but the degree to which far transfer of learning in psychological intervention studies are attributable to placebo effects is unknown, since subject expectations have not been controlled in these studies (21–26).

The world's population is growing, aging and becoming increasingly urban (27–29). Levels of 'background' environmental noise have risen steadily in industrialized areas, which can pose a challenge for aural speech communication. Following a conversation in noisy, public spaces is particularly difficult for older adults, the fastest growing age group in the United States (30), sixty percent of whom live with sensorineural hearing loss (SNHL) and struggle to accurately process target sounds with even moderate levels of distraction encountered in the average restaurant (31–34). Hearing aids are the only widely available treatment available for chronic SNHL. Modern hearing aids effectively amplify quiet sound and restore audibility, but offer little benefit

for the thornier problem of selectively enhancing the audibility of the target speaker and not the background noise (35). Thus, the only widely available treatment has limited efficacy for the average older adult whose primary complaint is an inability to follow a conversation in background noise (36–39). This prompted us to ask whether action game training, which provides generalized gains in visual processing to individuals with low vision (16, 17, 19, 20) might also be useful for improving speech comprehension in high levels of background noise to older subjects with sensorineural hearing loss who use hearing aids.

We programmed a tablet-based audiomotor ‘action’ training game and tested whether game learning would transfer to improved speech in noisy environments. We identified two components of action video games that we hypothesized would be important for driving improved attentional control and central sound processing: *i.*) interference resolution, which has been associated with an adaptive plasticity in attentional control networks (18) and *ii.*) a closed-loop game mechanic; wherein sensory stimuli act as continuous feedback signals to guide fluid motor commands (40–42). From a theoretical standpoint, closing the loop between continuous sensory input and motor output greatly increases the number of reinforced sensory-guided decisions per unit time over conventional perceptual learning tasks and would be expected to more efficiently drive neuromodulatory nuclei that are critical for enabling plasticity in adult sensory cortex (43–47).

In order to address rising concerns about non-specific effects in this class of behavioral intervention study, we also developed a placebo control game to act as a psychological “sugar pill” for comparison with the closed-loop audiomotor training task.

The placebo control used in this study was a speech-based auditory memory game that shared the design properties of the closed-loop training game. While some types of memory games have been linked to far-transfer effects ((48–50) but see (25, 51–53)), there is no reason to believe that memory training would afford improved understanding of speech in noisy environments for adults with SNHL (54, 55). Nonetheless, we reasoned that the adaptive speech-based challenge in the memory game coupled with older adults’ positive expectations concerning the benefits of “cognitive training” (56) would elicit high expectations for training effects, matching those evoked by the experimental, closed-loop game.

Older adults were randomly assigned to play either the auditory memory (*Mem*) or closed-loop (*C-L*) audio tracking game for two months. All aspects of the study (randomization, testing, and training) were managed by the subjects’ interactions with tablet devices that they used from home, allowing the researchers and subjects to remain blinded to group assignments. Speech perception, feature detection, and attentional control abilities were assessed before and after intervention using automated, mobile psychophysical tests that we have previously validated for use on tablet computers at home (57). While both groups of older adults demonstrated robust learning on their assigned tablet games and responded with matched expectations and beliefs of training benefit (confirming placebo control), we observed a transfer of learning to speech recognition in noise tasks following *C-L* but not *Mem* training.

4.2 Materials and Methods

Subject Recruitment and Retention

All procedures were approved by the Human Studies Committee at the Massachusetts Eye and Ear Infirmary and the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Informed consent was obtained from each participant. For this study, we recruited 50 to 89 yr old adults living with mild to severe sensorineural hearing loss who used hearing aids full-time in both ears. Most subjects were referred to this study by their clinician. Forty eight individuals were consented into this study and completed a screening visit to assess intelligence (Wechsler's Abbreviated Scale of Intelligence II(58)), cognition (Montreal Cognitive Assessment(59)), depression (Geriatric Depression Scale(60)), medical history (103 item questionnaire), pure-tone detection thresholds (audiologist administered in sound treated booth), and hearing aid performance. Thirty-six adults met inclusion criteria based on their scores in the initial screening and were invited to return for baseline assessment and stratified randomization into the study (**Fig. S1**). Hearing aid fit verification was performed by placing a small microphone in the ear canal, measuring the output of the hearing aid to various signal levels, and comparing the acoustic output of the hearing aid to gain prescriptions (NAL-NL2, (61)) based on the individual's hearing loss (Audioscan, Axiom system). Sixty-nine percent of the subjects' hearing aids were fit within 5 dB of the prescriptions and the remainder was between 5 and 10 dB of prescribed gain. Verification equipment was unavailable to make measurements for four subjects. Thirty-two individuals returned to the clinics to undergo baseline assessment. Twenty-three participants completed baseline assessments, training, and post-assessment. Twenty-one of these subjects also completed the final two month follow-up testing session. Thus, eight subjects began the

study but did not complete training or post-assessment. Of these, two dropped out before completing the initial baseline assessment. Four individuals dropped out after completing the baseline assessment and a few days of training. Two subjects were dismissed by the study leader due to lack of compliance over the first month of the study.

Testing and Training Equipment

Subjects performed all behavioral testing and game training on individually assigned Microsoft Surface Pro 2 tablets. The interactive software that was used for both games and the behavioral testing was developed as a Windows Store App using the Unity game engine and side-loaded onto the tablets. Audio stimuli were presented through a Dell AX 210 speaker that was connected to the tablet and placed at an approximate distance of 1 m and azimuthal position of 0° relative to the subject. Subjects performed a calibration at each home testing/training session to ensure reliable positioning of the speaker throughout this four month study. Briefly, the calibration program launched the native front-facing camera on the tablet and provided the subjects with visual guides to adjust the position of the speaker relative to the tablet device. Once the speaker was properly placed, the subject touched the screen to transmit an image of the test/training setup to the secure servers at MEEI for offline review. The native microphone on the tablet was also used by the custom application to make ambient noise level measurements in the home environment. If noise levels exceeded 60 dB A, the participant was locked out of the software, provided with a warning about the excessive noise levels in the test environment, and prompted to find a quieter location for testing. Average ambient noise levels measured by the tablet were

43 ± 4 dB A. Subjects used their experimental tablets to establish a wireless internet connection from their home environments at the beginning of the study. Data collected during testing and gameplay were automatically encrypted and uploaded using a secure file transfer protocol. Subjects did not receive compensation for their participation in this study, though parking fees were reimbursed and we did enter participants in a lottery to win one of five tablets at the conclusion of the study.

Auditory Training Games

Subjects were randomly assigned to play the auditory memory game (n= 11, 5 female, mean age = 70 yrs ± 11) or the closed-loop audio tracking game (n= 13, 10 female, mean age = 70 yrs ± 7) for two months. Both the auditory memory and closed-loop audiomotor tasks were embedded in a puzzle game. Subjects earned puzzle pieces by successfully executing their assigned tasks and were able to use them to reconstruct paintings by well-known artists or their own photos. The skins and graphics for each game environment were identical. Group randomization was stratified based on the subjects' baseline performance on the Quick Sentence-in-Noise test. This stratification was performed because baseline performance on speech recognition tests has prognostic value in speech training studies (62, 63), though we found no evidence of this in our generalization study (40). Group randomization was automated by an algorithm that cumulatively tracked baseline sentence in noise scores for randomized subjects. Therefore, members of the research team were not involved in randomization and did not know which game the subjects would play. Subjects were only exposed to the game that they were randomly assigned to play, and all instructions for gameplay and testing were provided through video tutorials as well as assisted play in the

application. In this way, all aspects of the daily testing and training activities in which the subjects engaged were addressed by the user interface of the software application. However, four subjects (two from each game group), had trouble using the games from home. These subjects returned to the clinic and were provided with a 30 min coaching session by a study staff member. Importantly, these staff members were not involved in data scoring or analysis, allowing us to maintain double-blinding during the study. Presentation levels for game sounds were tailored to the aided hearing sensitivity of each subject (i.e. baseline audiogram and loudness discomfort levels measured with hearing aids on).

Auditory memory game: We developed an auditory memory game as a control intervention for this study. The choice to develop a speech-based auditory memory game was based on our expectation that it would have good face-validity as an intervention for speech recognition abilities, it would promote auditory memory learning, but it would provide no generalized speech recognition benefit for this study population (54). During the task, subjects heard one or more strings which took the form of “Ready *name* go to *color number* now.” The *name*, *color*, and *number* for each string were randomly selected from eight, four, and eight possibilities respectively. After the subjects were presented with the auditory string, some number of labeled virtual elements slowly emerged on the screen, each with a non-overlapping 0.5 to 1.5 s delay. The subjects’ task was to identify virtual elements on the screen that corresponded to the *name*, *color*, and *number* that they had previously heard and then to connect these elements to create a composite object. Game difficulty adaptively changed to maintain perceptual challenge by incrementing the number of distractor

elements, the speed of elements, the number of phrases that were spoken, or the number of phrases that required responses. We also administered “yardstick” conditions periodically throughout the study to track learning under the same perceptual and cognitive demands (two name-color-number strings presented). We used publically available recordings of the Coordinate Response Measurements corpus generated by eight different speakers to create all of the task-related auditory stimuli for this game (64).

Each time that a player successfully matched a *name-color-number* string, they were then required to position it in a target location in order to generate a new puzzle piece. After they had generated a sufficient number of pieces, they were taken to a new sub-game screen where they were asked to use the pieces that they had earned to construct a puzzle (**Fig. S3**). As with any jigsaw puzzle, the subjects were provided with patterned and geometrical cues for the correct positioning of each puzzle piece. To provide feedback for the subjects’ positioning of the virtual pieces, the tiles would “snap into place” when they were placed in the correct location. While subjects performed this spatial reasoning task, they were simultaneously presented with stimuli from the closed-loop audiomotor game (see below). Specifically, they heard the audio generated by a random search in an auditory gradient. Because the sound was not tied to their motor activities, the participants generally perceived it as a distractor stimulus, and it served as a control condition for passive exposure to the same stimuli used in the audiomotor learning game.

Closed-loop audiomotor game: The audio tracking task used in this study was initially inspired by sensory-guided foraging behaviors in rodents (65–70) and refined

based on our own experiments involving closed-loop audiomotor learning (40). The basic closed-loop audiomotor method involves the establishment of an acoustic gradient that is mapped to some physical or virtual space. Subjects explore the space and their searches reveal the manner in which their motor behaviors parametrically alter the stimulus attributes. This information can then be iteratively used by the forager to identify hidden spatial targets more efficiently (40). In the auditory tracking task used in this study, subjects were aware that the outline of a polygon was hidden somewhere on the screen. They were required to use either a stylus or their finger to identify the location of the polygon and trace the outline of its shape. An auditory gradient was established relative to the individual lines comprising the shape, and as a subject moved his/her stylus through the gradient, either the level, frequency, or modulation rate of the sound was changed logarithmically with the subject's distance from the shape outline. Subjects adaptively learned to use these real time cues to reveal the shape with less error over the course of training (**Fig. 1E**) Game difficulty increased over the course of the study in two ways. The signal to noise ratio adaptively changed based on the subjects' performance and its maximum value decreased after every puzzle completion. Additionally, the complexity of shapes (defined by number of vertices) was increased following each puzzle completion. We also administered "yardstick" conditions periodically throughout the study to track learning under the same perceptual demands (-18 dB SNR). The dependent measurement that we used to define learning on the yardstick conditions was audio tracking error, which is defined as the perpendicular Euclidean distance between a player's current position and the nearest line segment, $\text{Tracking error} = \sqrt{(\text{trace}X - \text{line}X)^2 + (\text{trace}Y - \text{line}Y)^2}$.

Based on the accuracy of the subject's tracing, they were awarded time to complete two sub-games. The first sub-game was a gradient-based search task, similar to that used in our previous training studies in rats, mice and young adult human subjects with normal hearing ((40, 69)) (**Fig. S4A**). Subjects were required to move the puzzle piece to a target location on the display. The target location was invisible, but a circular audio gradient was established that logarithmically varied audio stimulus attributes with Euclidean distance from the target. Once the subject thought that they had found the correct location, they would release the virtual puzzle piece. If the subject was correct the piece would remain in place, but if they were incorrect (outside the rewarded area), the piece would fall to the bottom of the screen and a new target area and gradient would be randomly generated. If sufficient time remained on the countdown timer, the subject began a second sub-game that challenged them to rotate the virtual puzzle piece around a central axis to achieve the correct orientation (like a combination lock, **Fig. S4G**). There were no visual cues concerning the correct orientation of the puzzle piece, but the audio stimulus would rapidly change its value from a reference sound to a target sound when the piece was rotated into the correct orientation. If the subject rotated beyond the correct point or released the piece prior to the correct point, a new target orientation was selected and the user was permitted to try again until the time expired.

The three game mechanics were packaged into three “worlds” defined by the sound feature that subjects were asked to discriminate: pitch, level and amplitude modulation. The stimuli used for each world consisted of amplitude modulated pure tones, spectrotemporally modulated ripple noises, and tone clouds. The levels of all

stimuli varied from 20-40 dB sensation level (dB SL) and were limited by the measured loudness discomfort levels of each subject. Minimum sensation level and maximum tolerable level were defined for each subject across a range of pure tone frequencies prior to the start of training. The carrier frequencies of tones varied from 125-8000 Hz. The modulation rates of tones varied from 2-32 Hz. Spectrotemporally modulated ripple noise was synthesized from sinusoidal components with frequencies spaced from 354-5656 Hz in 0.05-octave steps. Ripple density varied from 0.5 to 3 cycles per octave, and modulation velocity varied from 4-12 Hz. For the tone clouds, 50 ms tone pips were randomly selected from a uniform distribution that varied in bandwidth from 0.25 to 1.5 octaves. The level of each tone pip in the cloud was roved by ± 6 dB. All game signals were presented while 1-6 talker babble played in the background. Background speech materials were generated by concatenating a subset of IEEE sentences (sentences 361-720) presented by 20 different talkers from the Pacific Northwest/Northern Cities corpus (71). The subset of IEEE sentences that we used as distractors in this study were selected such that they did not overlap with the IEEE sentences used in the Quick Sentence-in-Noise Test (sentences 1-360).

Psychophysical tasks to test transfer of learning

All behavioral testing was self-directed. Subjects interacted with a custom software interface to perform alternative forced choice, reaction time, and open response tasks. Subjects wore their hearing aids and were asked to use their typical settings for all testing and training performed in the study. Each behavioral task began with instructions and practice trials. In the practice trials, the perceptual demands were kept at an “easy” level and subjects were given feedback concerning the accuracy of

their responses (an exception was the speech recognition tasks). Subjects were required to achieve a minimum performance level to assure that basic procedural aspects of the task were learned before the testing blocks began. In previous experiments, we found that home testing using this software interface provided results that were statistically equivalent to manual testing in sound treated rooms (57).

Speech recognition in noise: For all speech recognition in noise tasks, subjects were asked to repeat a target talker who produced either a word or a sentence. After the speaker finished, the subjects touched a virtual button on the tablet screen to activate the tablet's native microphone and record his/her verbal responses. These responses were saved as encrypted raw binary files, transmitted wirelessly to secure servers, converted to .wav files, and scored offline by a blinded experimenter. Word recognition in noise testing was conducted using the Words In Noise test ((72) (WIN)). The WIN is a clinical test that consists of monosyllabic words from the Northwestern University 6 corpus spoken by a female talker while 6-talker babble was played continuously in the background. 35 monosyllabic words were presented at SNRs that varied from 24 to 0 dB SNR in 4 dB steps. We administered a unique randomization of WIN lists 1 and 2 at each time point in the study.

Sentence recognition in noise was assessed using two clinical tests, the Quick Sentence-in-Noise test (73) (QuickSiN) and the BKB Sentences-in-Noise Test (74) (BKBSiN). For both the QuickSiN and the BKBSiN, target sentences were presented while 4-talker babble played continuously in the background. The main difference between the two tests is that the QuickSiN employed sentences with low linguistic context (IEEE sentences) while the BKBSiN test used sentences that contained high

linguistic context (Bamford-Kowal-Bench sentences). The signal to noise ratio of the QuickSiN varied from 25 to 0 dB SNR in 5 dB steps, while the signal to noise ratio of the BKBSiN varied from 21 to -6 dB SNR in 3 dB steps. Four unique sentence lists (QuickSiN) or two unique list pairs (BKBSiN) were administered at each testing time point in the study. Additionally, two list pairs of the BKBSiN were measured while the subject was not wearing hearing aids during the pretest visit to establish the amount of benefit provided by the subjects' hearing aids (**Fig. 5A-B**). The lists that were used for aided and unaided testing were randomly selected, as was their presentation order.

Memory Assessment: Subjects were tested with the Letter Number Sequencing test (Wechsler's Adult Intelligence Scale III(75)). All audio stimuli were pre-recorded and shared with us by Dr. Adam Gazzaley's laboratory at University of California, San Francisco. Audio stimuli consisted of increasingly long strings of letters and numbers spoken by a male talker. An experimenter at our research facility administered the test in a clinical sound booth. The experimenter initiated each trial with a virtual button press on the tablet. After the stimulus presentation, the native microphone of the tablet was used to record the responses of the subjects. The subjects were asked to verbally respond by repeating all elements of the string with the numbers first in ascending order, followed by the letters in alphabetical order. The experimenter scored the participants' responses online in order to determine when testing was to terminate (following 3 incorrect responses at the same memory load level). However, the actual scoring of the data that are presented in this manuscript was performed by a blind experimenter after the .wav files of the participants' responses were uploaded to our servers. It should be noted that due to the subjects' deficient speech processing and the

auditory-only presentation mode of the stimuli, subjects often made phonemic confusions, even under conditions of low memory load. For this reason, we scored their responses in two ways, strict and loose. For strict scoring, any phonemic mistake was counted as incorrect. For loose scoring, confusions that involved up to two of the distinctive features of phonemic categories were tolerated as hearing errors (e.g. place and voice onset time errors). Training effects in the study were identical regardless of our scoring method. In this manuscript we reported data that reflected the loose scoring method because we believed that it was a better approximation of working memory abilities.

Competing Digits: Subjects were initially familiarized with a male speaker (fundamental frequency = 115 Hz) as he spoke 120 digits in relative quiet. On each trial, the male speaker produced a string of four randomly selected digits (digits 1-9, excluding the bisyllabic '7') with 0.68 s between the onset of each digit stimulus. The subjects used a virtual keypad on the tablet to enter with the digits spoken by the target speaker. After familiarization, two additional talkers were introduced (male, fundamental frequency = 90 Hz; female, fundamental frequency = 175 Hz) as distractors. These speakers produced randomly selected distracting digits with target-matched onset times. The only contingency was that two speakers could not produce the same digit at once, otherwise the digit produced by each speaker was selected at random. The target speaker was presented at 65 dB SPL. Four hundred and twenty eight digits were presented at 0 dB SNR (target and distractors were presented at the same level), and ninety two digits were presented at 3 dB SNR (the target was 3 dB higher in level than the distractors). We observed that 32% of the subjects performed at chance level in the

more challenging 0 dB SNR condition during baseline testing. By contrast, only 8% of the sample performed at chance levels in the 3 dB SNR condition. To avoid floor effects in our analysis, we focused on the 3 dB SNR condition. We analyzed performance on the digits task in two ways. First, we asked how often the subjects correctly identified the first digit in each stream. We viewed performance on this condition as analogous to monosyllabic word recognition in noise task. Next, we asked how often the subjects correctly identified all four digits in a stream. We viewed performance under this condition as analogous to a sentence recognition in noise task.

Audio/Visual Stroop: The Stroop effect (76) has been studied extensively in psychological research over the past 8 decades as a measure of inhibitory control (77). For all versions of the Stroop tasks, subjects are asked to attend to a stimulus and then report the identity of a specific attribute of that stimulus while ignoring other stimulus attributes. In some cases the “distractor” stimulus attributes are congruent with the target stimulus attribute and in other cases they are incongruent. The congruency of target and distractor stimulus attributes has a marked effect on reaction times (RT) with responses to congruent conditions occurring ~250 ms sooner than responses to incongruent trials. A neutral condition is also presented wherein there is no relationship (congruent or not) between the target and distractor stimulus features. The neutral condition provides a control measurement for processing speed and can be used to compute normalized Stroop interference $\frac{\text{Incongruent RT (s)} - \text{Congruent RT (s)}}{\text{Neutral RT (s)}}$.

We measured performance on a visual and audio Stroop task in this study. In the visual Stroop task, subjects were visually presented with the text, “Red”, “Blue”, and

“Legal” in a random vertical location on the screen (letter height = 3.5 cm, white background). The color of the word was either red or blue. This created three conditions, color-letter congruency, color- letter incongruency, and a neutral condition (the word “Legal”). Likewise, the audio Stroop employed three words (“High”, “Low,” and “Day”) that were either spoken with a low fundamental frequency (180 Hz) or a high fundamental frequency (280 Hz). The three words were spoken by the same female talker, and the TANDEM-STRAIGHT vocoder was used to synthesize these three vocalizations and shift the fundamental frequency up and down to create high and low pitch versions of each word (78). Subjects began each trial by placing their thumbs in two circles that were positioned on each side of the tablet screen, midway along the vertical axis. After a 0.5-2 s delay, an audio or visual stimulus was presented and two virtual response buttons appeared just above and below each thumb fixation circle. The participants were required to select one of the two responses as quickly and accurately as possible. Their reaction times were recorded as the latency of the first of the two thumb responses.

Before each trial began, either a visual or an audio masker was presented to cue the trial and wash out stimulus recency effects. The visual masker was a grid of 39 individually colored squares (grid dimensions 16.9 x 4.2 cm) that was positioned 3.6 cm from the top of the tablet screen. The color of each element in the grid was randomly selected to be red, blue, green, or yellow with a refresh rate of 4 Hz. The audio masker consisted of 15 tones that were each 50 ms in duration and presented with an interstimulus interval of 0.18 s at a level of 60 dB SPL. The carrier frequency of each

tone was randomly selected from an interval of values that ranged from 500 to 8000 Hz. The duration of the video and audio maskers were 1 s each.

To compute average reaction times for the congruent, incongruent, and neutral conditions, each word-color and word-pitch combination was repeated 30 times over the course of 3 training blocks. Testing was complete once each individual had accrued at least 40 correct responses for each of the three congruency conditions (i.e. incongruent, congruent, and neutral). The reaction time of correct responses was analyzed. If a subject's accuracy score was \leq chance probability, his/her reaction time data were not used in the analysis. This criterion resulted in two subjects' (one from each group) reaction time data being removed from the analysis.

Frequency Modulation Detection: The subjects were initially exposed to the perceptual experience of frequency modulation (FM) through an interactive slider that they manipulated to increase and decrease the excursion depth of a frequency modulated tone. High excursions were labeled as 'squiggly' to allow the subjects to associate the sound with a label that could be used when completing the 2-interval 2-alternative forced choice FM detection task. After initial familiarization, two tones (carrier frequency = 1000 Hz, duration = 1 s, level = 55 dB SL) were presented to subjects with an interstimulus interval of 0.5 s. Frequency modulation was applied at a rate of 2 Hz to one of the two tones (random order). A quasi-sinusoidal amplitude modulation (6 dB depth) was applied to both tones to reduce cochlear excitation pattern cues (79). The subject was asked to indicate whether the first or second tone was frequency modulated ('squiggly'). The two-down-one-up procedure was used to modulate the frequency excursion magnitude in order to converge on the 70.7% correct

point (80). The frequency excursion of the FM tone was initially set to 75 Hz and changed by a factor of 1.5 for the first 5 reversals, decreasing to a factor of 1.2 for the last 7 reversals. The geometric mean of the last 6 reversals was used to compute the run value. A minimum of 3 runs were collected. The coefficient of variation across runs was computed online. If the coefficient of variation was > 0.2, additional runs were collected until this criterion was met or six runs had been collected, whichever came first. The median threshold across all runs collected was used to define the participant's FM detection threshold.

Analysis of proportional data: We primarily analyzed the speech recognition data by computing correct scores over challenging SNRs for each test. Across both groups, performance changed as a function of noise level over a similar range of SNRs. We used SNRs from the steeply sloping portion of the psychometric function to assess performance (WIN, 12-16 dB SNR; QuickSiN, 5-10 dB SNR; BKBSiN, 0-6 dB SNR). Scores were expressed as rationalized arcsine units (81) (RAU) since proportional scores generally violate several assumptions of parametric statistical tests for values near zero and one. We also performed clinical scoring for each speech test by computing a non-adaptive threshold using the Spearman-Kärber equation (**Fig. 5D**). To summarize our findings for the words and sentence tests, we computed a change index, $\frac{Post\ score(\%) - Pre\ score(\%)}{Post\ score(\%) + Pre\ score(\%)}$. We used this index to compare the patterns of results between the open-set speech tasks and the digit streaming task. We used *Hedges' g* to compute treatment effect sizes for the speech recognition and digits streaming tasks at the post and two-month follow-up assessment periods. We computed 95% confidence intervals around the effect size using a bootstrapping approach (82).

Statistical testing

Normality of data distributions was assessed using the Shapiro-Wilk test and q-q plots. In the cases of normal distributions, parametric tests were employed to test for statistical significance. Statistical significance of group intervention effects was assessed by performing repeated measures ANOVA. Analysis of covariance using the subjects' pretest scores as the covariate has been recommended for analysis in clinical studies (83). Therefore, we also performed analysis of outcome measures using ANCOVA. These data are reported in Table S1 and the pattern of significant results are identical to those reported in the manuscript. Other between group comparisons were tested using two-sample *t* tests, and within group differences were assessed using paired-sample *t* tests. Group comparisons between non-normally distributed data were made using the Wilcoxon rank sum test. Correlations were quantified using Pearson's linear correlation coefficient and corrected for multiple comparisons when appropriate (Holm-Bonferroni).

4.3 Results

Older adults become more proficient at playing their tablet-based auditory working memory and closed-loop audio tracking games.

We enrolled older adults ($\bar{x} = 70$ yrs) living with mild to severe sensorineural hearing loss in a randomized, double-blinded placebo-controlled study (**Fig 1A, Fig S1-2**). Participants were randomized to play either an auditory memory (*Mem*) or closed-loop auditory tracking (*C-L*) game for approximately 3.5 hours per week for 8 weeks while wearing their hearing aids (*Mem* group, $n = 11$, $\bar{x} = 31$ total hrs; *C-L* group, $n = 13$,

\bar{x} = 35 total hrs; z = -1.3, P = 0.2, Wilcoxon Rank-Sum). All subjects were long-term, bilateral hearing aid users (mean period of hearing aid use = 7 yrs).

The overarching objective of both games was to reconstruct broken jigsaw puzzles using a touchscreen interface on a tablet computer. In order to accomplish this goal in the *Mem* game, subjects were required to attend to sentences spoken by different talkers and identify the items that they had heard following a 3 -16 s delay (**Fig. 1B top**). Subjects indicated their responses by connecting virtual elements on the screen to create combination objects that represented the keywords in each sentence (*name, color, number, Fig. 1B midde & bottom*). The sentence objects were then dropped on targets corresponding to the vertices of hidden puzzle pieces that were subsequently revealed and then used to reconstruct virtual jigsaw puzzles in a sub-game (**Fig. S3**). During the puzzle construction sub-game, stimuli from the other training game (closed-loop audio tracking) were played as background sound that was not connected to the task on which they were operating (i.e. open-loop passive stimulus exposure). The difficulty level of the game adaptively changed such that advancing to later puzzle boards presented the challenge of higher memory loads, longer delay periods, and additional distractor elements (**Fig. 1B middle and bottom**).

Subjects trained on the closed-loop auditory tracking game (*C-L*) were aware that a shape was hidden on the screen (**Fig. 1B top**). Using a virtual pencil, they were required to trace the outline of the shape without visual cues. Like an odor trail, the game generated an auditory gradient projected onto the contour of the hidden shape. Subjects used their finger movements to explore the soundscape. As the subject passed his/her virtual pencil through the dynamic regions of the gradient, either the

level, frequency, or amplitude modulation rate of a tonal sound was instantaneously updated with the subject's distance from the outline of the shape. Shapes were randomly generated and placed on each trial. Early in training, subjects tracked audio gradients at favorable signal to noise ratios (SNR), and while they were able to use the continuous audio feedback to detect some general proximity to the shape, they were unable to employ the subtle information that was provided by the gradient to accurately track its outline (**Fig. 1B middle**). However, following eight weeks of training, subjects were able to use the real time audio feedback to adjust their tracing trajectory with less error despite the presence of continuous multi-talker background speech babble that was louder than the target signal (**Fig. 1B bottom**). Following successful tracing of a shape, subjects engaged in two sub-games wherein they used both gradient cues and rapid stimulus changes to place the puzzle piece that they had traced into the correct location on a puzzle board (**Fig. S4**).

Based on subject performance, the difficulty levels of both games increased over the course of training, defined by increased memory load in the *Mem* game and reduced SNR in the *C-L* game (**Fig. 1D**, $F = 2.2$, $P = 0.03$, RMANOVA; **Fig. 1D**, $F = 9.2$, $P = 3 \times 10^{-10}$, RMANOVA). While adjusting the difficulty level of the task so as to maintain a fixed level of trial failures (i.e., adaptive tracking) is a hallmark of behavioral paradigms that drive plasticity in early sensory cortex (84), we also wished to characterize task-related learning on a constant stimulus set. To this end, we occasionally administered 'yardstick' trials to characterize learning on each task, using stimulus parameters that provided a fixed level of perceptual difficulty. The *Mem* group demonstrated a ~ 20% improvement over the two-month training period on the yardstick

condition (**Fig. 1E**, $F = 14.2$, $P = 2 \times 10^{-13}$, RMANOVA). Likewise, subjects who trained on the *C-L* game learned to use subtle gradient changes to reduce their tracking error by ~ 20% (**Fig. 1C & 1G**, $F = 7.4$, $P = 2 \times 10^{-8}$, RMANOVA), a result that is similar in magnitude to observations in a rodent scent tracking task (85). In addition to the behavioral improvements described on the primary games, learning was also observed on the sub-games for the both the *Mem* and *C-L* training groups (**Fig. S3-4**).

While the subjects were engaged in home-based game training, they responded to several questionnaires that gauged their *i.*) expectations for hearing improvements as a consequence of gaming, *ii.*) their experience playing their assigned game, and *iii.*) their perception of hearing benefit over the course of training (**Fig S5**). There were no differences between the ratings given by either group for any of these measurements ($P > 0.56$ for all comparisons). As a whole, both groups were tentatively optimistic that their games would improve their hearing early in the study and about half believed that their hearing had improved during the study. Therefore, both the *C-L* and *Mem* tasks became more difficult as subjects advanced in their game play, subjects showed comparable levels of learning in both tasks, and subjects' game play experience and expectations for benefits through training were matched. As such, the *Mem* game meets the high standard for a placebo control (21).

Learning on the closed-loop audio tracking game, but not the memory game, transferred to untrained sentence-in-noise tasks.

Given the robust learning demonstrated by older adults who had played the *C-L* and *Mem* games, we next asked whether learning with the stimuli encountered in the

training games (**Fig. 2A**) transferred to improved performance on untrained stimuli encountered in clinical tests of speech recognition in noise (**Fig. 2B**). Subjects completed three speech recognition-in-noise tests that are commonly used in audiology clinics. One of the tests required subjects to listen to monosyllabic words at various SNRs and to provide a verbal response that was recorded via the microphone on their tablet (**Fig. 2C**) (72). The other two speech tests were performed with identical procedures, but employed sentence materials that possessed either low-levels of linguistic context (e.g., “Dimes showered down from all sides”, **Fig. 2D**) or high-levels of linguistic contextual cues (e.g., “The janitor swept the floor”, **Fig. 2E**) (73, 74).

Speech intelligibility declined precipitously across a narrow range of SNRs in all tests (**Fig. 2C-E left, red vertical lines**). As an example, speech processing accuracy changed by ~50% in the high-context sentence test (**Fig. 2E**) as the ratio between the target speaker level varied from 0-6 dB above the din of background speaker babble, a perceptual condition that would be expected to occur regularly in an average restaurant (33). We focused our analysis on these challenging noise levels where speech intelligibility was greatly diminished but not impossible (**Fig. 2C-E left, red vertical lines**). We observed that word recognition at difficult SNRs (**Fig. 2C right**) did not improve secondary to either training approach ($F = 2.68$, $P = 0.12$, Time Effect, $F = 0.14$, $P = 0.71$, Group x Time Interaction, Repeated Measures ANOVA). By contrast, *C-L* training significantly improved sentence recognition-in-noise abilities at challenging SNRs when compared to *Mem* training (**Fig. 2D-E right, low-context sentence**, $F = 5.5$, $P = 0.03$; **high-context sentence**, $F = 6.3$, $P = 0.02$, Group x Time Interaction, Repeated Measures ANOVA). Thus, there was a selective improvement in sentence, but not word

recognition tasks that was restricted to older adults who had trained on the *C-L* game (**Fig. 2F**, words, $t = -0.03$, $P = 0.98$; sentences, $t = 2.2$, $P = 0.04$, two-sample t test).

Why were training benefits observed for tests of sentence recognition in noise, but not individual words? The procedural and perceptual demands of each task were matched, but there were at least two differences between the materials that were used to make these measurements: *i.*) Sentence tests (low and high context) presented listeners with some degree of semantic and syntactic cues, while the word tests were devoid of linguistic context. *ii.*) As compared to words, which last only a few hundred milliseconds (**Fig. 2B left**), sentence tests required that listeners distribute their attention to a noisy, but predictable spectro-temporal pattern (speaker) distributed over timescales of several seconds (**Fig. 2B right**). Either or both of these differences could have contributed to the transfer of learning to sentence testing following *C-L* training.

To differentiate between these possibilities, we administered a competing digits streaming test. This involved the presentation of four randomly selected digits by a target talker while two distractors simultaneously produced competing digit streams (**Fig. 2G**). The subject was asked to use a virtual keypad to reproduce the target digit stream. Like the sentence tests, this task involved distributed attention to a predictable target over longer timescales, but in contrast to sentence tests, it was devoid of linguistic context. Furthermore, the test could be scored as a word or sentence task by analyzing whether the subjects' responses to the first digit were correct (word scoring) or if they correctly identified the entire sequence of digits (sentence scoring). We found that performance improvements on the competing digit streams task mirrored our observations of improved sentence recognition following *C-L* training, suggesting that

these improvements were not dependent on linguistic context, but were, perhaps, specific to tasks that required distributed attention to predictable signals over longer timescales—like conversational speech (**Fig. 2H**, first digit, $t = 0.01$, $P = 0.99$; digit stream, $t = 2.2$, $P = 0.04$, two-sample t test). As further support of this possibility, we found that digits streaming improvements were significantly correlated with sentence in noise (SiN) recognition benefits ($R = 0.72$, $P = 0.01$, Pearson's correlation coefficient).

Generalized benefits in sentence in noise recognition were not dependent on improved low-level sensory processing or working memory capacity.

Behavioral measurements that assess low-level spectro-temporal processing abilities and high-level cognitive functions, such as working memory, have been shown to predict sentence recognition in noise performance for individuals with and without audiometrically confirmed hearing loss (86–94). We measured baseline spectro-temporal processing ability by assessing sensitivity to periodic fluctuations in the frequency of a pure tone (frequency modulation, FM), a task that is believed to depend on the accurate encoding of rapid timing information ((95), but see (96)). We also measured working memory (WM) capacity by administering the Letter Number Sequencing test (75), which involves the repetition and ordering of alphanumeric strings that progressively increase in length. Consistent with previous reports, we found that both basic spectro-temporal processing abilities and auditory working memory capacity were significantly correlated with baseline SiN recognition (FM, $R = 0.44$, $P = 0.04$; WM, $R = 0.45$, $P = 0.04$, Pearson's correlation coefficient corrected for multiple comparisons with the Holm-Bonferroni method, **Fig. 3A-B left**). Because FM detection thresholds and WM capacity were both predictive of pre-training speech processing scores and C-

L training improved speech processing, transitive logic suggests that one or both training tasks might also improve FM detection and WM. However, we found that the SiN recognition improvements observed in this study were not accompanied by generalized improvements in frequency modulation detection thresholds or working memory capacity, in spite of the fact that *Mem* training improving auditory working memory abilities within the confines of the game (FM, $F = 0.31$, $P = 0.58$; WM, $F = 0.18$, $P = 0.68$, Group x Time Interaction, Repeated Measures ANOVA).

Measures of inhibitory control do not vary with training but do predict degree of learning transfer.

Younger adults have a superior ability to ignore certain types of irrelevant information compared to older adults (97–102). At the physiologic level, a failure to suppress the neural representation of distractor signals in early levels of cortical processing has been associated with downstream, age-related visual memory limitations under conditions of distraction (103, 104). We assessed the relationship between inhibitory control abilities and sentence recognition under conditions of distraction by administering visual and audio versions of the classic Stroop task (76) (**Fig. 3C**). While the visual color-word version of the Stroop task has been studied extensively (for review see (77)) and been specifically used to measure age-related inhibitory control deficits (97–101), relatively few investigators have examined the Stroop effect in the auditory modality (105–109). We observed that the pattern of congruency-related reaction times that are consistently reported for visual color-name congruency conditions were also observed for audio pitch-name congruency and the magnitudes of the Stroop effects in each modality scaled together (**Fig. 3D & Fig. S6A**).

Regarding associations between inhibitory control and SiN abilities, we found that performance on neither the audio nor visual Stroop tasks was predictive of baseline SiN recognition (Audio: $R = 0.10$, $P = 0.61$; Visual: $R = 0.17$, $P = 0.78$, Pearson's correlation coefficient corrected for multiple comparisons with Holm- Bonferroni method, **Fig. 3E-F left**). Likewise, we did not find that normalized performance on either version of the Stroop task improved secondary to game intervention (Audio: $F = 1.04$, $P = 0.32$; Visual, $F = 0.17$, $P = 0.68$, Group x Time Interaction, Repeated Measures ANOVA, **Fig. 3E-F right**), though reaction times decreased significantly (~50-100 ms) for both *C-L* and *Mem* groups (**Fig. S5C-D**).

While Stroop interference effects did not change with improvements in SiN recognition, we did find that baseline inhibitory control abilities as assessed with either the audio or visual version of the Stroop task predicted the amount of transfer that we observed on the SiN recognition tasks following 2 months of *C-L* training (Combined Stroop: $R = 0.62$, $P = 0.01$, Pearson's correlation coefficient, **Fig. 3K & S6C**). Specifically, subjects with the best inhibitory control (smallest effect of distractor congruence on reaction times) demonstrated the greatest benefit on sentence in noise comprehension secondary to *C-L* training.

Improvements in closed-loop audio tracking skills predicts the degree of enhanced speech processing

For any complex sensory game, there are many possible aspects of a subject's experience that could predict degree of transfer. Simple metrics like time on task or overall task completion rate were not expected to predict transfer of learning secondary

to *C-L* gaming (40). Rather, we expected that the features of improved game play that most accurately predicted far transfer to improved speech processing would share certain key features with the core challenges of processing sentences in noise. Although we used synthetic, parametric stimuli such as modulated tones or ripple noise in the *C-L* tracing task, the game required subjects to continuously discriminate subtle changes in target stimuli and predict upcoming signal changes during ‘bouts’ of tracing that lasted several seconds while suppressing distracting background speech. In this respect, tracing the outline of a hidden shape using dynamic audio cues seems share many common features with sentence processing in noise. In support of this argument, we observed that improved audio tracing accuracy over the course of *C-L* training predicted the improved ability to correctly process a target speaker in the presence of high level background speech babble (**Fig. 4C**, $R = 0.60$, $P = 0.03$, Pearson’s correlation coefficient).

Audiomotor training temporarily enhances speech perception at background noise levels where hearing aids provide little benefit.

As this study was performed in an elderly patient population who used hearing aids to compensate for their sensorineural hearing loss, we asked how the benefit imparted by *C-L* training compared to that received from their hearing aids. As expected, we found that hearing aids improved sentence recognition ability in relatively quiet conditions with high SNRs (**Fig. 5A-B**). However, under noisier conditions typical of social environments, their hearing aids provided little benefit ($\bar{x} = 4 \pm 8\%$. **Fig. 5A-B**). By contrast, eight weeks of *C-L* training provided approximately three times the benefit of their hearing aids in the impoverished SiN environments commonly encountered in

noisy, social settings (**Fig. 5A-B**, $t = -5.19$, $P = 3 \times 10^{-4}$, paired sample t test). Thus, by coupling the use of a sensory prosthesis with *C-L* training, speech perception benefits were extended to the challenging listening environments that represent the chief complaint of patients (34, 39). The SiN recognition tests employed in this study have known psychometric properties, which are used to define a “clinically significant” difference in test scores for an individual patient ((73, 74) critical difference at 80% confidence). Using this clinical definition, we found that eight weeks of *C-L* training imparted a speech processing benefit that would be judged as “clinically significant” to more than half of the subjects, whereas, only 1/11 (9%) of those trained on the *Mem* game met this stringent criterion (**Fig. 5D**).

As a final step, we computed the effect sizes for *C-L* relative to *Mem* treatment at the post-training and two month follow-up tests. While we observed large effect sizes for both SiN recognition tests as well as the digits streaming task at the conclusion of training (≥ 0.85 , *Hedges' g*), the difference did not persist when tested in the absence of further intervention two months later (**Fig. 5 D-F**, (*Hedges' g* ≤ 0.13). These results suggest that, at least with this sensory-impaired elderly population, generalized benefits of *C-L* gaming may require some degree of continued training for optimal effect maintenance, as is expected for most types of cognitive (110, 111) or skill learning (112, 113).

4.4 Discussion

We programmed a suite of tablet-based self-administered auditory training and psychoacoustic tasks that could be used by older adults ($\bar{x} = 70$ yrs) with sensorineural

hearing loss from home. Subjects became more proficient at playing their auditory working memory or a closed-loop audio tracking games and expressed equivalent expectations that each would improve their hearing abilities (**Fig. 1, Fig. S5**). Following two months of training, subjects who played the *C-L*, but not the *Mem*, game exhibited significant transfer of learning to tests that required them to process speech in high levels of background noise. This transfer of learning was specific to perceptual tasks that involved tracking a signal over time because learning from *C-L* training *i.*) did not generalize to an isolated word recognition in noise test that was procedurally identical to the sentence tests, but *ii.*) did transfer to a digit streaming task that was devoid of linguistic context (**Fig. 2**). While SiN recognition deficits of older adults with SNHL are thought to result from broadened cochlear tuning (114), deficient spectro-temporal processing (88, 90, 115), and impaired cognitive control abilities (94, 116), we found no evidence that improvements in SiN recognition that followed *C-L* training were accompanied by perceptual enhancements in spectro-temporal processing, memory capacity, or inhibitory control measures (**Fig. 3**).

The particular pattern of improvements on sentence-level speech processing without any transfer to either low-level sound processing or higher cognitive control or memory processes, suggested that the *C-L* training may have improved subjects' ability to distribute attention over longer timescales—a central requirement for maintaining a conversation in social environments. Interestingly, baseline inhibitory control measured in the visual or auditory modality predicted the degree to which learning transferred to SiN recognition secondary to *C-L* gaming (**Fig. 3K**). In the *C-L* game, the subjects who learned to use subtle and noisy acoustic cues to more accurately trace objects also

received the greatest benefit on SiN recognition, directly linking game strategy and transfer effects (**Fig. 4**). Collectively, these findings suggested that the best candidates for C-L training benefits might be identifiable before training begins and that greater improvements in speech processing might be possible by developing games that reward participants to actively discriminate variations in local stimulus features over extended time scales.

C-L training imparted a benefit for sentence comprehension in high levels of background noise that exceeded the benefit of hearing aids alone by a factor of three (**Fig. 5A-B**). Although the gains may seem modest in their overall magnitude (~14% more words comprehended in a given sentence), it is worth keeping in mind that gains occurred within the “intelligibility cliff” of SNRs, where even a small change in the number of correctly understood words can have a disproportionate effect on overall comprehension and communication experience. However, these effects did not persist for two months in the absence of further intervention (**Fig. 5D**). Whether game “booster sessions” as have been previously used in training studies of older adults (110, 117) would allow long-term maintenance of “real world” hearing in noise benefits, and what the optimal “booster” schedule would be remain open questions.

Controlling for expectations and experimenter interactions in training studies

Randomized, double-blind placebo study designs provide the highest standard of control because they account for subject expectations and Hawthorne effects that arise from extended subject-experimenter interactions, yet are rarely employed by sensory training studies (21–24). As a “sugar pill” training interface to create a matched

expectation of improved speech processing abilities did not exist, we developed a control game that we believed would possess good face-validity and capitalized on the general public's contemporary, positive beliefs concerning memory training (56). We hypothesized that if a control game required subjects to act on increasingly difficult speech sequences that challenged their memory abilities, they might believe that their hearing would be improved by playing the game. This supposition was confirmed through their answers on our questionnaires administered partway through training (**Fig. S5**). Several groups have observed that certain types of working memory training, such as the dual *n*-back task, generalized to improvements in a broader set of cognitive skills, including inhibitory control (48, 49), and fluid intelligence (50). However, others have failed to replicate some or all of these findings (25, 51–53).

The purpose of the memory training group was to control for expectation effects in the primary and secondary outcome measures. Interestingly, we found that placebo effects in this study were limited to subjective reports of improved hearing (**Fig. S5C-D**). Following *Mem* training, there was no observed benefit for overall sentence recognition in noise using psychophysical measurements ($\bar{x} = -.03\%$, **Fig. 2F**). This contrasts with a report of improvement equivalents of 5 to 10 points on standardized intelligence tests that were completely attributable to placebo effects and selection bias (23). Several key methodological differences in the present study could explain the differential magnitude of the placebo effect across these studies, including recruitment procedures, testing intervals, and familiarity of primary outcome measures (i.e. listening and responding to speech vs. solving unfamiliar puzzles). This substantial difference in placebo magnitude

across studies highlights the nuances underlying expectation effects in behavioral intervention studies (22).

Making a good game better

We isolated two features of action video games that we hypothesized were important for driving cortical processing towards plastic states through the engagement of neuromodulatory nuclei (43, 44, 46) and attentional control systems (118): *i.*) interference resolution and *ii.*) a closed-loop mechanic. Other attributes of action video games including immersion (which ranked low in our games, **Fig S5B**), multi-tasking, complex sensory stimuli, multifarious reward schedules, high speeds, substantial motor loads, and emotionally salient content (119) were explicitly not implemented here because we set out to test the hypothesis that a closed-loop mechanic that included interference resolution would be sufficient to drive generalized gains in speech processing. Game design features that were left out of our *C-L* training application may be important for driving wider apparent task generalization through plasticity in neural networks that underlie improved probabilistic inference and allow players to more rapidly learn new tasks (120–122).

On-task perceptual learning persists for months following the cessation of training in younger (2) and older adults (18, 123). However, the persistence of transfer effects following sensory or cognitive training is less certain. Studies which have examined generalized cognitive benefits in younger adults following action video game play have not generally reported on the persistence of transfer effects (10–12, 124–128), though one study reported improved contrast sensitivity several months after

game training (16). Pertinent to this study, reported transfer effects to cognitive control tasks following action video game play are weaker or non-existent in older adults, even at the immediate post training test (129–133), which may reflect a poor match between the game preferences of older adults and the experience provided by action video games (129, 134). Nonetheless, when transfer effects to a working memory task were observed in older adults who had trained on an action video game, they did not persist without continued intervention (111). In this study, we found that the transfer effects to auditory streaming did not persist for 2 months without continued *C-L* training, suggesting that the neural underpinnings of the behavioral changes described here are not maintained without continued practice provided by the game (consistent with (111)). This does not come as a surprise, as practice is also required to maintain proficiency in other audiomotor skills, such as musicianship in humans or courtship song in birds (135, 136).

There is no reason to believe that the design and implementation of the *C-L* game studied here represented the optimal solution to enhance auditory signal in noise perception. Given that fairly subtle properties of sensory training tasks can dramatically alter the degree of generalized learning (137–139), there are likely many attributes of the currently tested *C-L* game that could be optimized to enhance desired transfer effects. Nonetheless, in the context of a double-blinded placebo controlled trial, we observed clear transfer of game learning to real world listening situations that pose significant communication and social problems for more than 60% of older adults. Hearing loss represent a major public health concern (32), affecting 360 million individuals worldwide and has been linked to a 24% increased risk for incident cognitive

impairment (140), perhaps due to social isolation in response to impaired hearing. As such, the maintenance and rehabilitation of social hearing abilities represents an important challenge for neuroscientists, computer-programmers, engineers, and clinicians. Here we have shown that a neuroscience-based game intervention can be paired with a sensory prosthesis to improve patient outcomes by a factor of three. Whether this type of brain-based intervention can be optimized to provide greater benefits in isolation or in conjunction with other neuromodulatory approaches (141, 142) will be important questions to guide future studies.

4.5 Figures

Figure 1

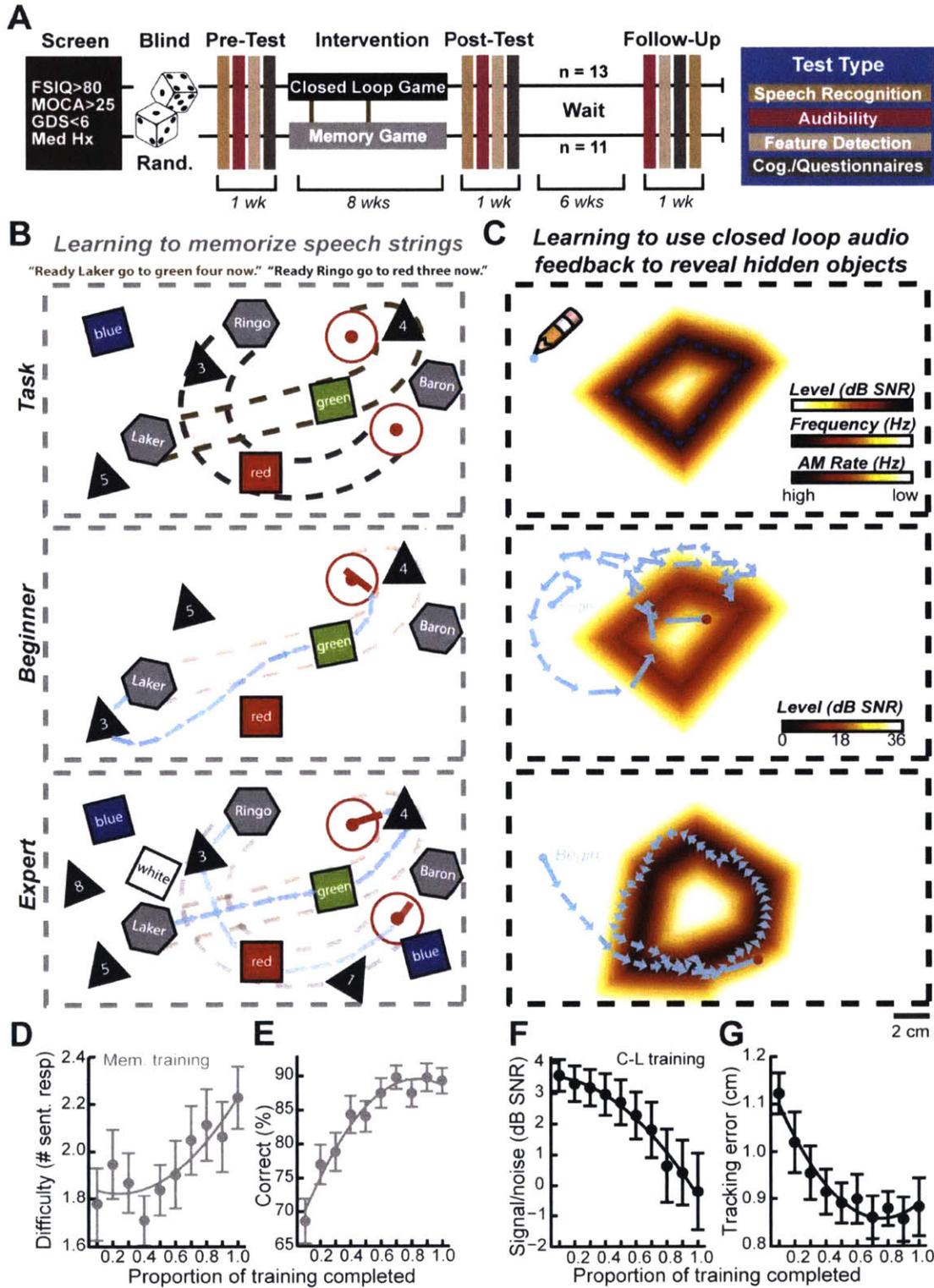


Fig1. Design and implementation of experiments to test generalized performance changes following training with custom-made auditory working memory and audiomotor closed-loop tracing tasks. (A) Older adults that passed an initial screening were randomly assigned to play either an auditory memory or closed-loop auditory tracking game in the context of a randomized control trial. All subjects underwent testing with psychophysical measures or questionnaires before, during, just after, and 2 months after training (color-coded by subject area for ease of visualization). (B) Subjects who trained on the memory game heard one or more sentences before being presented with virtual objects to match on the tablet screen. The correct responses on the task are coded by the colors of the spoken sentence (*top*). Schematized finger movements on the touchscreen from early (*middle*) and late (*bottom*) in training depict more accurate performance and increasing game complexity over time. (C) Subjects who trained on the closed-loop audio tracking game used a virtual pencil to trace the outline of an invisible shape. The only available information to solve the task was an auditory gradient that parametrically varied one of three possible sound features with the subject's distance from the lines on the invisible shape (see colorbars, *top*). Data traces from a subject on day 8 (*middle*) and day 28 (*bottom*) of training, show early failure to use audio gradient information and later mastery of the closed-loop audio tracking task despite more challenging perceptual demands (SNR scale bar in middle panel applies to both). Cyan arrows depict the subject's tracing path, with each arrow representing the distance and trajectory over a single 0.5 s sample period. This subject learned to use real time audio feedback to trace the outline more slowly and more accurately. (D) The working memory task became more difficult over the course of training. Difficulty was defined as memory load, i.e., the number of sentences that required a response. (E) Working memory performance improved over the course of training on a fixed "yardstick" condition repeated throughout training, that required subjects to respond to one of two sentences presented with three additional distractors on the screen. (F) The closed-loop tracing task became more difficult over training as the available SNR cues to identify the outline of the hidden shape dropped closer to zero. (G) Tracking error decreased on the yardstick condition (SNR ~ -18 dB) over the course of training. Circular symbols and errorbars reflect mean \pm s.e.m. Overlying lines are quadratic fits to the group averaged data.

Figure 2

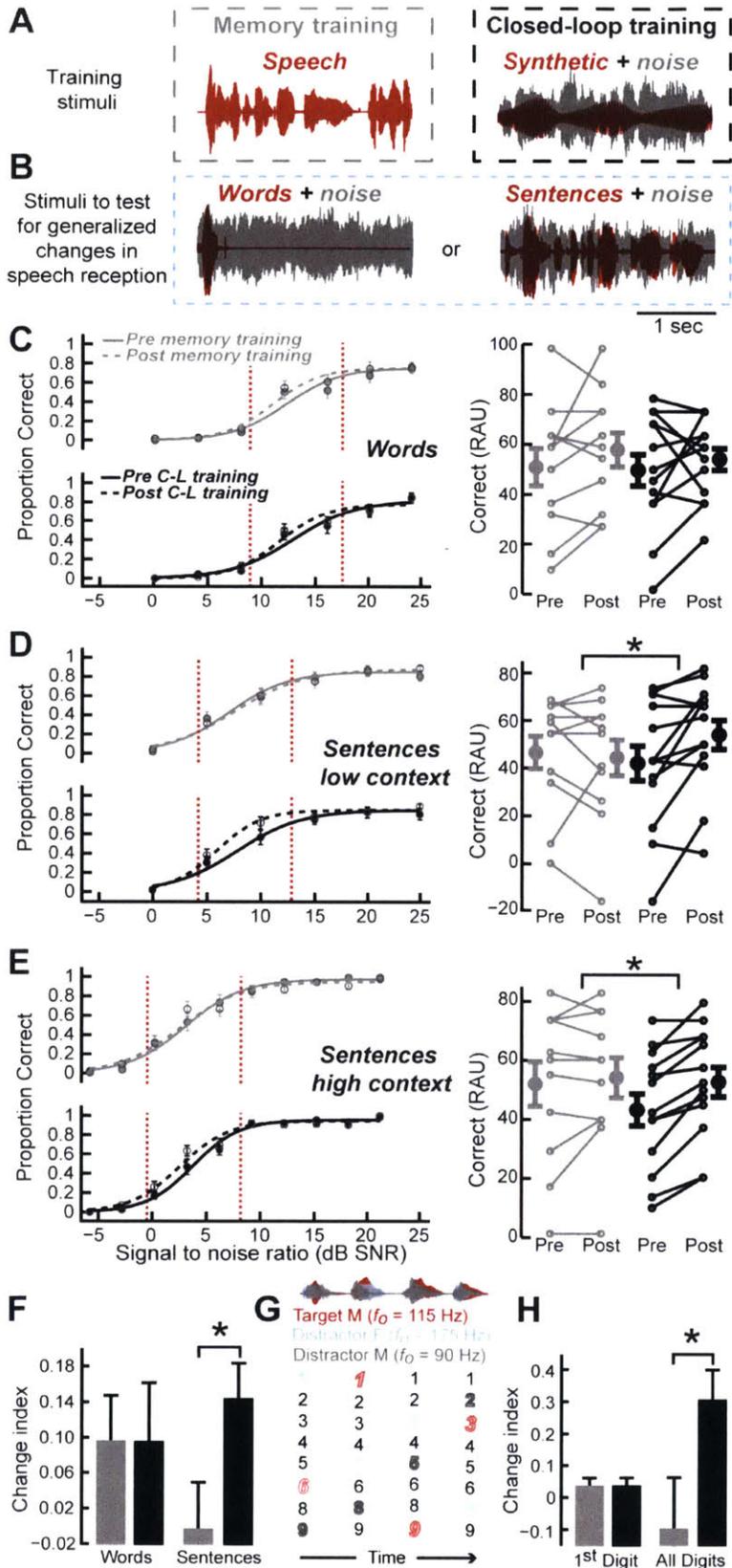


Fig 2. Closed-loop training, but not memory training, was associated with generalized improvements in sentence comprehension in high levels of background noise. (A) Memory gaming involved listening and responding to speech materials (*left*). Closed-loop audio gaming required subjects to track changes in tone features in the presence of background distractors (*right*). (B) Transfer of learning was measured for tests that assessed recognition of target (red waveforms) words (*left*) or sentences (*right*) in the presence of distractor speakers (gray waveforms). (C-E) Recognition of monosyllabic words (C), low-context sentences (D) or high-context sentences (E) was assessed before and after training with the memory (gray) or closed-loop (black) training under increasingly difficult distractor conditions. *Left*, Proportional scores were measured at each SNR and average performance data were fit with a logistic function using constrained maximum likelihood estimation. Recognition performance declined steeply within a restricted set of SNRs (vertical red broken lines). *Right*, Smaller lines reflect individual subject pre and post test scores. The adjacent larger circles and error bars represent mean \pm s.e.m. across each training group. (F) Summary plot of primary outcome measures expressed as change index $\frac{Post\ score(\%) - Pre\ score(\%)}{Post\ score(\%) + Pre\ score(\%)}$ (G) Schematic of digit streaming task. Male target speaker waveform and target digits (red) are depicted alongside distractor male (gray) and female (cyan) speech waveforms and spoken digits. f_o = fundamental frequency (voice pitch). (H) Summary of digit task improvements when the first digit is scored in isolation (similar to a word test, C) or the whole stream of digits is scored (similar to test of sentence comprehension, D-E). Bar plots reflect mean \pm s.e.m. Asterisks indicate statistically significant differences $P < 0.05$.

Figure 3

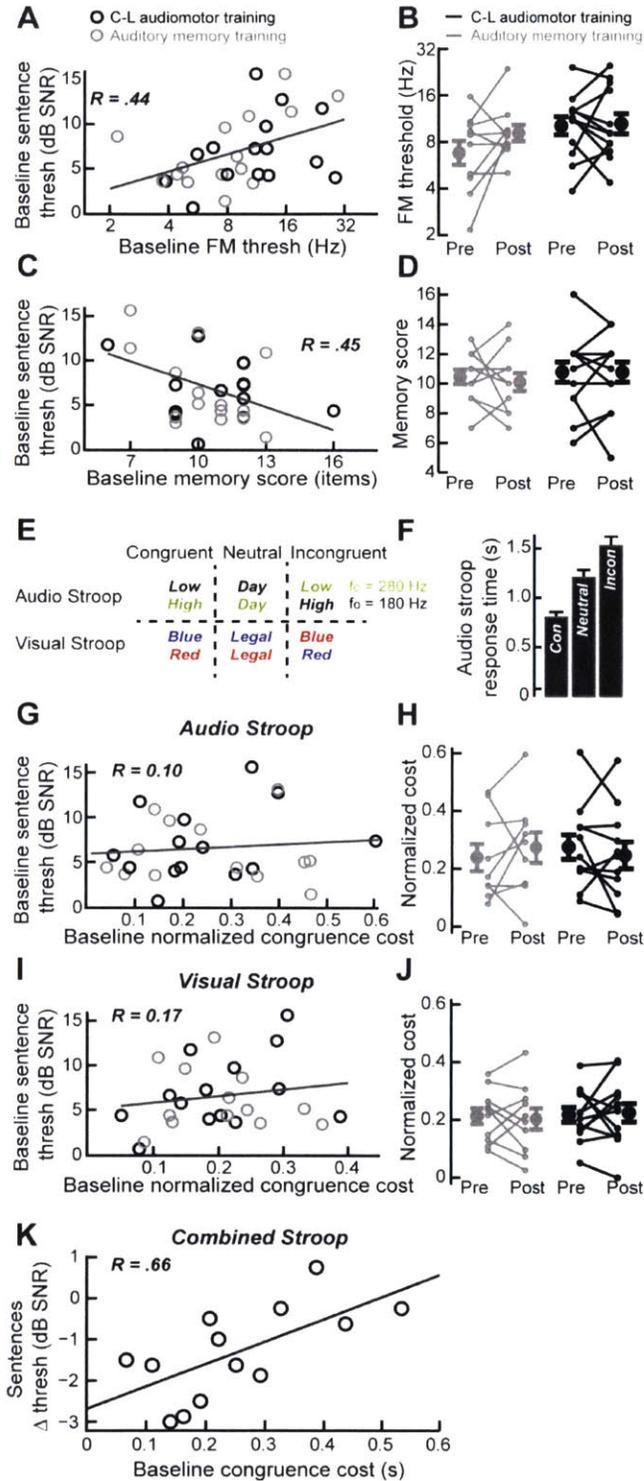


Fig 3. Auditory feature detection and cognitive control do not improve with training, but can predict speech in noise processing abilities. (A-D)

Baseline frequency modulation detection thresholds (A) and working memory scores (C) are correlated with baseline sentence recognition in noise abilities but do not change significantly as a result of training (B and D, gray = memory training, black = closed-loop training). Plotting conventions follow Figure 2. (E)

Schematic of the audio and visual Stroop test conditions. (F) Average reaction times for each congruency condition in the audio Stroop task from a representative subject. (G-J) Baseline audio (G) and visual (I) normalized congruence cost $\frac{\text{Incongruent RT (s)} - \text{Congruent RT (s)}}{\text{Neutral RT (s)}}$, do not

predict baseline sentence recognition in noise and do not change over the course of training. (K) The mean baseline congruence cost averaged between visual and audio Stroop conditions predicts the degree of improvement in sentence processing before and after closed-loop training. Improved speech processing is defined as the change in the SNR at which subjects correctly perceive 50% of the words. Correlation coefficient (R) computed using Pearson's method.

Figure 4

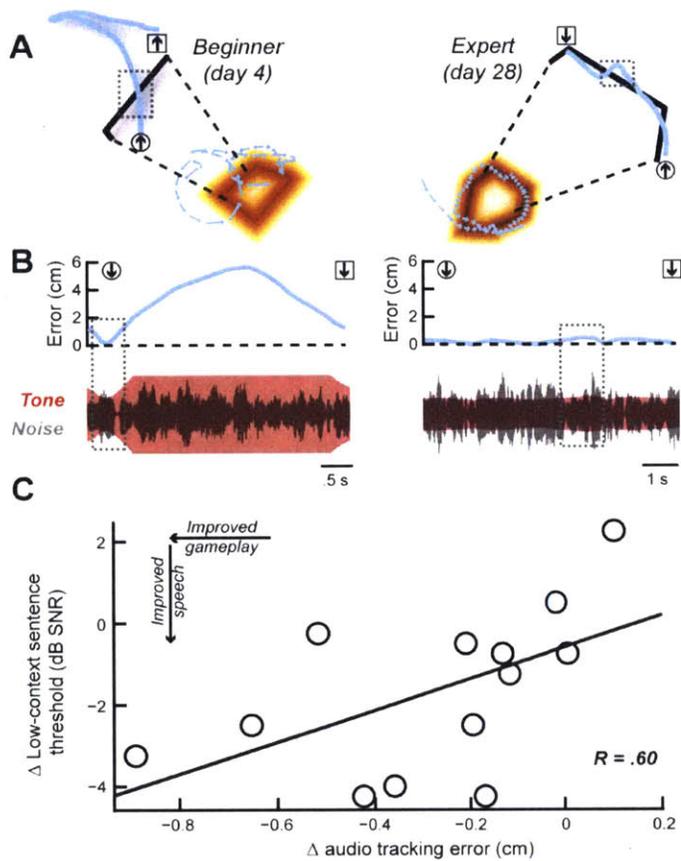


Fig 4. Improvements in speech processing after closed-loop training can be predicted from game play performance. (A) Tracing segments from early and late closed-loop game traces (taken from Fig. 1B and magnified by 3x) illustrate improvements in audio tracking over the course of training. Tracing error, defined as the instantaneous Euclidean distance measured between each point in the virtual pencil trace (cyan) and the shape border (black), is plotted over ~3-7 s tracing period. The beginning and end of the tracing segment are represented by an arrow encased by a circle or square, respectively. (B) Tracing error magnitude is instantly translated into the SNR of the tone in noise. Note that the tone level gradient saturates outside of a 2 cm area surrounding the shape outline (B, left bottom). The gray rectangles in A and B focus on a specific portion of the trace, highlighting the relatively subtle changes in the signal envelope used to refine finger movement trajectory in this well-trained subject (compare B left and right). Signal magnitude (bottom) is plotted logarithmically for visualization. (C) Reductions in audio tracking error over the course of closed-loop training are associated with sentence threshold improvements (lower = better performance). Correlation coefficient (R) computed using Pearson's method.

Figure 5

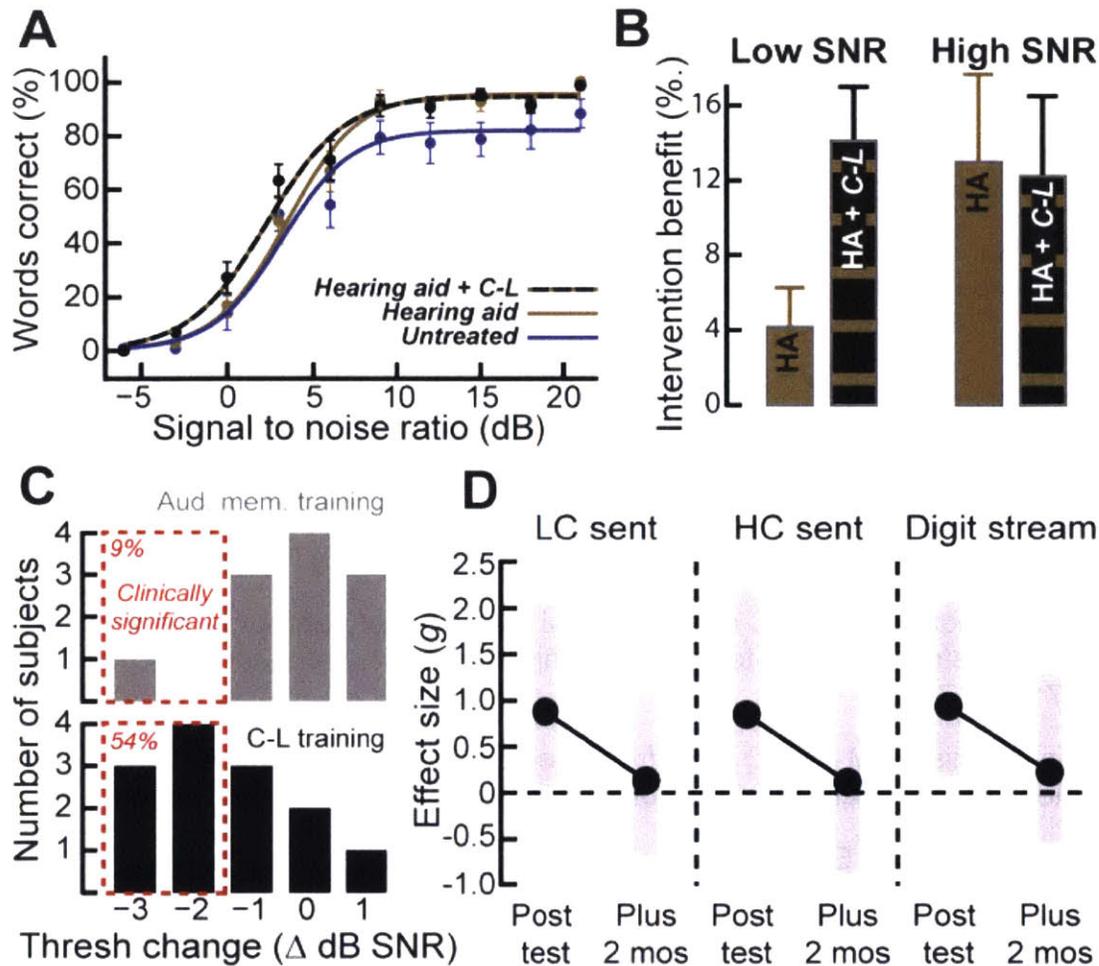


Fig 5. Audiomotor training temporarily enhances speech perception at background noise levels where hearing aids provide little benefit. (A) Recognition of high context sentences was measured at SNRs that ranged from poor to favorable without hearing aids (blue), while using their hearing aids before closed-loop training (gold), and when using a hearing aid after closed-loop audiomotor training (black and gold). **(B)** The average sentence recognition benefit provided by a hearing aid and a hearing aid plus closed-loop training under low (0-6 dB) and high SNR (9 – 20 dB) conditions. HA = hearing aid, C-L = Closed-loop audio training. Error bars represent s.e.m. **(C)** Histograms of sentence recognition threshold changes (negative values = better performance) following memory training (gray) and closed-loop training (black). Threshold changes 1.3 dB SNR or greater are considered “clinically significant” (red dashed lines). **(D)** Closed-loop gaming effect sizes and 95% confidence intervals (gray shaded areas) for sentence tests (LC = low context, HC = high context) and digits streaming at the post-test and at the two month follow-up visit.

4.6 References

1. Fiorentini A, Berardi N (1980) Perceptual learning specific for orientation and spatial frequency. *Nature* 287:43–44.
2. Ball K, Sekuler R (1982) A specific and enduring improvement in visual motion discrimination. *Science* 218(4573):697–698.
3. Wright BA, Buonomano D V, Mahncke HW, Merzenich MM (1997) Learning and generalization of auditory temporal-interval discrimination in humans. *J Neurosci* 17(10):3956–3963.
4. Recanzone GH, Schreiner CE, Merzenich MM (1993) Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J Neurosci* 13(1):87–103.
5. Schoups A, Vogels R, Qian N, Orban G (2001) Practising orientation identification improves orientation coding in V1 neurons. *Nature* 412(6846):549–553.
6. Schwartz S, Maquet P, Frith C (2002) Neural correlates of perceptual learning: a functional MRI study of visual texture discrimination. *Proc Natl Acad Sci U S A* 99(26):17137–17142.
7. Hochstein S, Ahissar M (2002) View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron* 36(5):791–804.
8. Parbery-Clark A, Skoe E, Lam C, Kraus N (2009) Musician Enhancement for Speech-In-Noise. *Ear Hear* 30(6):653–661.
9. Swaminathan J, et al. (2015) Musical training, individual differences and the cocktail party problem. *Sci Rep* 5:1–10.
10. Green CS, Bavelier D (2006) Enumeration versus multiple object tracking: the case of action video game players. *Cognition* 101(1):217–245.
11. Green CS, Bavelier D (2003) Action video game modifies visual selective attention. *Nature* 423(6939):534–537.
12. Green CS, Bavelier D (2007) Action-video-game experience alters the spatial resolution of vision. *Psychol Sci* 18(1):88–94.
13. Başkent D, Gaudrain E (2016) Musician advantage for speech-on-speech perception. *J Acoust Soc Am* 139(3):EL51–EL56.
14. Micheyl C, Delhommeau K, Perrot X, Oxenham AJ (2006) Influence of musical and psychoacoustical training on pitch discrimination. *Hear Res* 219(1-2):36–47.
15. Ruggles DR, Freyman RL, Oxenham AJ (2014) Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLoS One* 9(1):1–8.

16. Li RJ, Polat U, Makous W, Bavelier D (2009) Enhancing the contrast sensitivity function through action video game training. *Nat Neurosci* 12(5):549–551.
17. Li RW, Ngo C, Nguyen J, Levi DM (2011) Video-Game Play Induces Plasticity in the Visual System of Adults with Amblyopia. *PLoS Biol* 9(8):1–11.
18. Anguera JA, et al. (2013) Video game training enhances cognitive control in older adults. *Nature* 501(7465):97–101.
19. Li JR, et al. (2013) Dichoptic training enables the adult amblyopic brain to learn. *Curr Biol* 23(8):R308–R309.
20. Vedamurthy I, et al. (2015) Mechanisms of recovery of visual function in adult amblyopia through a tailored action video game. *Sci Rep* 5:1–7.
21. Boot WR, Simons DJ, Stothart C, Stutts C (2013) The Pervasive Problem With Placebos in Psychology: Why Active Control Groups Are Not Sufficient to Rule Out Placebo Effects. *Perspect Psychol Sci* 8(4):445–454.
22. Green CS, Strobach T, Schubert T (2014) On methodological standards in training and transfer experiments. *Psychol Res* 78(6):756–772.
23. Foughi CK, Monfort SS, Paczynski M, McKnight PE, Greenwood PM (2016) Placebo effects in cognitive training. *Proc Natl Acad Sci* 113(27):7470–7474.
24. Shipstead Z, Redick TS, Engle RW (2012) Is working memory training effective? *Psychol Bull* 138(4):628–654.
25. Melby-Lervåg M, Hulme C (2013) Is working memory training effective? A meta-analytic review. *Dev Psychol* 49(2):270–291.
26. Jacoby N, Ahissar M (2013) What does it take to show that a cognitive training procedure is useful? A critical evaluation. *Prog Brain Res* 207:121–140.
27. Department of Economic and Social Affairs PD (2015) *World population ageing* Available at:
http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf.
28. Department of Economic and Social Affairs PD (2015) *World urbanization prospects: The 2014 revision* Available at:
<https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Report.pdf>.
29. Department of Economic and Social Affairs PD *World population prospects: The 2015 revision, key findings and advance tables*.
30. Colby SL, Ortman JM (2015) *Projections of the size and composition of the U.S. population: 2014 to 2060* Available at:
<http://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143.pdf> [Accessed August 20, 2016].

31. Wilson RH, McArdle R a, Smith SL (2007) An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *J Speech Lang Hear Res* 50(4):844–856.
32. Lin FR, Niparko JK, Ferrucci L (2011) Hearing loss prevalence in the United States. *Arch Intern Med* 171(20):1851–1852.
33. Hodgson M, Steininger G, Razavi Z (2007) Measurement and prediction of speech and noise levels and the Lombard effect in eating establishments. *J Acoust Soc Am* 121(4):2023–2033.
34. Hind SE, et al. (2011) Prevalence of clinical referrals having hearing thresholds within normal limits. *Int J Audiol* 50(10):708–716.
35. Bentler RA, Duve MR (2000) Comparison of hearing aids over the 20th century. *Ear Hear* 21(6):625–639.
36. Shinn-Cunningham BG, Best V (2008) Selective attention in normal and impaired hearing. *Trends Amplif* 12(4):283–299.
37. Kochkin S (2000) MarkeTrak V: “Why my hearing aids are in the drawer”: The consumers’ perspective. *Hear J* 53(2):34–41.
38. Kochkin S (2010) MarkeTrak VIII: Consumer satisfaction with hearing aids is slowly increasing. *Hear J* 63(1):19–27.
39. Davis AC (1989) The prevalence of hearing impairment and reported hearing disability among adults in Great Britain. *Int J Epidemiol* 18(4):911–917.
40. Whitton JP, Hancock KE, Polley DB (2014) Immersive audiomotor game play enhances neural and perceptual salience of weak signals in noise. *Proc Natl Acad Sci U S A* 111(25):E2606–E2615.
41. Mishra J, Gazzaley A (2014) Closed-Loop Rehabilitation of Age-Related Cognitive Disorders. *Semin Neurol* 34:584–590.
42. Kraus N, White-Schwoch T (2015) Unraveling the Biology of Auditory Learning: A Cognitive-Sensorimotor-Reward Framework. *Trends Cogn Sci* 19(11):642–654.
43. Parikh V, Kozak R, Martinez V, Sarter M (2007) Prefrontal acetylcholine release controls cue detection on multiple timescales. *Neuron* 56(1):141–154.
44. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593–1599.
45. Steinberg EE, et al. (2013) A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* 16(7):966–973.
46. Letzkus JJ, et al. (2011) A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* 480(7377):331–335.

47. Pi HJ, et al. (2013) Cortical interneurons that specialize in disinhibitory control. *Nature* 503(7477):521–524.
48. Chein JM, Morrison AB (2010) Expanding the mind's workspace: training and transfer effects with a complex working memory span task. *Psychon Bull Rev* 17(2):193–9.
49. Klingberg T, et al. (2005) Computerized training of working memory in children with ADHD - A randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry* 44(2):177–186.
50. Jaeggi SM, Buschkuhl M, Jonides J, Perrig WJ (2008) Improving fluid intelligence with training on working memory. *Proc Natl Acad Sci U S A* 105(19):6829–6833.
51. Thompson TW, et al. (2013) Failure of working memory training to enhance cognition or intelligence. *PLoS One* 8(5):1–15.
52. Chooi W-T, Thompson LA (2012) Working memory training does not improve intelligence in healthy young adults. *Intelligence* 40(6):531–542.
53. Redick TS, et al. (2013) No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *J Exp Psychol Gen* 142(2):359–379.
54. Ferguson MA, Henshaw H (2015) Auditory training can improve working memory, attention, and communication in adverse conditions for adults with hearing loss. *Front Psychol* 6:1–7.
55. Wayne R V., Hamilton C, Jones Huyck J, Johnsrude IS (2016) Working memory training and speech in noise comprehension in older adults. *Front Aging Neurosci* 8:1–15.
56. Rabipour S, Davidson PSR (2015) Do you believe in brain training? A questionnaire about expectations of computerised cognitive training. *Behav Brain Res* 295:64–70.
57. Whitton JP, Hancock KE, Shannon JM, Polley DB (2016) Validation of a self-administered audiometry application: An equivalence study. *Laryngoscope*:1–7.
58. Wechsler D (2011) WASI II Wechsler abbreviated scale of intelligence, Second edition.
59. Nasreddine ZS, et al. (2005) The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* 53(4):695–699.
60. Yesavage JA, et al. (1983) Development and validation of a geriatric depression screening scale: A preliminary report. *J Psychiatr Res* 17(1):37–49.
61. Keidser G, Dillon H, Carter L, O'Brien A (2012) NAL-NL2 empirical adjustments.

Trends Amplif 16(4):211–223.

62. Song JH, Skoe E, Banai K, Kraus N (2012) Training to improve hearing speech in noise: Biological mechanisms. *Cereb Cortex* 22(5):1180–1190.
63. Henderson Sabes J, Sweetow RW (2007) Variables predicting outcomes on listening and communication enhancement (LACE) training. *Int J Audiol* 46(7):374–383.
64. Bolia RS, Nelson WT, Ericson MA, Simpson BD (2000) A speech corpus for multitalker communications research. *J Acoust Soc Am* 107(2):1065–1066.
65. Stephens D, Krebs J (1987) *Foraging Theory* (Princeton University Press, Princeton).
66. Kennedy JS (1983) Zigzagging and casting as a programmed response to wind-borne odor: A review. *Physiol Entomol* 8(2):109–120.
67. Porter J, et al. (2007) Mechanisms of scent-tracking in humans. *Nat Neurosci* 10(1):27–29.
68. Bao SW, Chang EF, Woods J, Merzenich MM (2004) Temporal plasticity in the primary auditory cortex induced by operant perceptual learning. *Nat Neurosci* 7(9):974–981.
69. Polley DB, Heiser MA, Blake DT, Schreiner CE, Merzenich MM (2004) Associative learning shapes the neural code for stimulus magnitude in primary auditory cortex. *Proc Natl Acad Sci USA* 101(46):16351–16356.
70. Gire DH, Kapoor V, Arrighi-Allisan A, Seminara A, Murthy VN (2016) Mice Develop Efficient Strategies for Foraging and Navigation Using Complex Natural Stimuli. *Curr Biol* 26(10):1261–1273.
71. McCloy DR, et al. (2013) The PN/NC corpus. Available at: <http://depts.washington.edu/phonlab/resources/pnnc/>.
72. Wilson RH, Burks C a (2005) Use of 35 words for evaluation of hearing loss in signal-to-babble ratio: A clinic protocol. *J Rehabil Res Dev* 42(6):839–852.
73. Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, 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. *J Acoust Soc Am* 116(4):2395–2405.
74. Bamford-Kowal-Bench Speech-in-Noise Test (2005).
75. Wechsler D (1997) WAIS III Wechsler Adult Intelligence Scale, Third Edition.
76. Stroop JR (1935) Studies of interference in serial verbal reactions. *J Exp Psychol* 18:643–662.

77. MacLeod CM (1991) Half a century of research on the Stroop effect: an integrative review. *Psychol Bull* 109(2):163–203.
78. Kawahara H, Morise M (2011) Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana* 36(5):713–727.
79. Papakonstantinou A, Strelcyk O, Dau T (2011) Relations between perceptual measures of temporal processing, auditory-evoked brainstem responses and speech intelligibility in noise. *Hear Res* 280(1-2):30–37.
80. Levitt H (1971) Transformed up-down methods in psychoacoustics. *J Acoust Soc Am* 49(2):467–477.
81. Studebaker GA (1985) A rationalized arcsine transform. *J Speech Hear Res* 28(3):455–462.
82. Hentschke H, Stüttgen MC (2011) Computation of measures of effect size for neuroscience data sets. *Eur J Neurosci* 34(12):1887–1894.
83. Egbewale BE, et al. (2014) Bias, precision and statistical power of analysis of covariance in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Med Res Methodol* 14(1):1–12.
84. Engineer ND, et al. (2012) Inverted-U function relating cortical plasticity and task difficulty. *Neuroscience* 205:81–90.
85. Khan AG, Sarangi M, Bhalla US (2012) Rats track odour trails accurately using a multi-layered strategy with near-optimal sampling. *Nat Commun* 3:1–10.
86. Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2011) Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proc Natl Acad Sci U S A* 108(37):15516–15521.
87. Bharadwaj HM, Masud S, Mehraei G, Verhulst S, Shinn-Cunningham BG (2015) Individual Differences Reveal Correlates of Hidden Hearing Deficits. *J Neurosci* 35(5):2161–2172.
88. Mehraei G, Gallun FJ, Leek MR, Bernstein JGW (2014) Spectrotemporal modulation sensitivity for hearing-impaired listeners: Dependence on carrier center frequency and the relationship to speech intelligibility. *J Acoust Soc Am* 136(1):301–316.
89. Bernstein JGW, Summers V, Grassi E, Grant KW (2013) Auditory Models of Suprathreshold Distortion and Speech Intelligibility in Persons with Impaired Hearing. *J Am Acad Audiol* 24(4):307–328.
90. Strelcyk O, Dau T (2009) Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *J Acoust Soc Am* 125(5):3328–3345.

91. Füllgrabe C, Moore BCJ, Stone MA (2014) Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Front Aging Neurosci* 6:1–25.
92. Humes LE, Lee JH, Coughlin MP (2006) Auditory measures of selective and divided attention in young and older adults using single-talker competition. *J Acoust Soc Am* 120(5):2926–2937.
93. Woods WS, Kalluri S, Pentony S, Nooraei N (2013) Predicting the effect of hearing loss and audibility on amplified speech reception in a multi-talker listening scenario. *J Acoust Soc Am* 133(6):4268–4278.
94. Akeroyd MA (2008) Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int J Audiol* 47 Suppl 2:S53–71.
95. Moore BCJ, Sek A (1996) Detection of frequency modulation at low modulation rates: Evidence for a mechanism based on phase locking. *J Acoust Soc Am* 100(4):2320–2331.
96. Whiteford KL, Oxenham AJ (2015) Using individual differences to test the role of temporal and place cues in coding frequency modulation. *J Acoust Soc Am* 138(5):3093–3104.
97. Spieler DH, Balota DA, Faust ME (1996) Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *J Exp Psychol Hum Percept Perform* 22(2):461–479.
98. Comalli PE, Wapner S, Werner H (1962) Interference effects of Stroop color-word test in childhood, adulthood, and aging. *J Genet Psychol* 100:47–53.
99. Panek PE, Rush MC, Slade LA (2012) Locus of the age-Stroop interference relationship. *J Genet Psychol* 145:209–216.
100. Cohn NB, Dustman RE, Bradford DC (1984) Age-related decrements in Stroop Color Test performance. *J Clin Psychol* 40(5):1244–1250.
101. Hartley AA (1993) Evidence for the selective preservation of spatial selective attention in old-age. *Psychol Aging* 8(3):371–379.
102. Hasher L, Zacks RT (1988) Working memory, comprehension, and aging: A review and a new view. *Psychol Learn Motiv* 22:193–225.
103. Gazzaley A, et al. (2008) Age-related top-down suppression deficit in the early stages of cortical visual memory processing. *Proc Natl Acad Sci U S A* 105(35):13122–13126.
104. Gazzaley A, Cooney JW, Rissman J, D'Esposito M (2005) Top-down suppression deficit underlies working memory impairment in normal aging. *Nat Neurosci* 8(10):1298–1300.

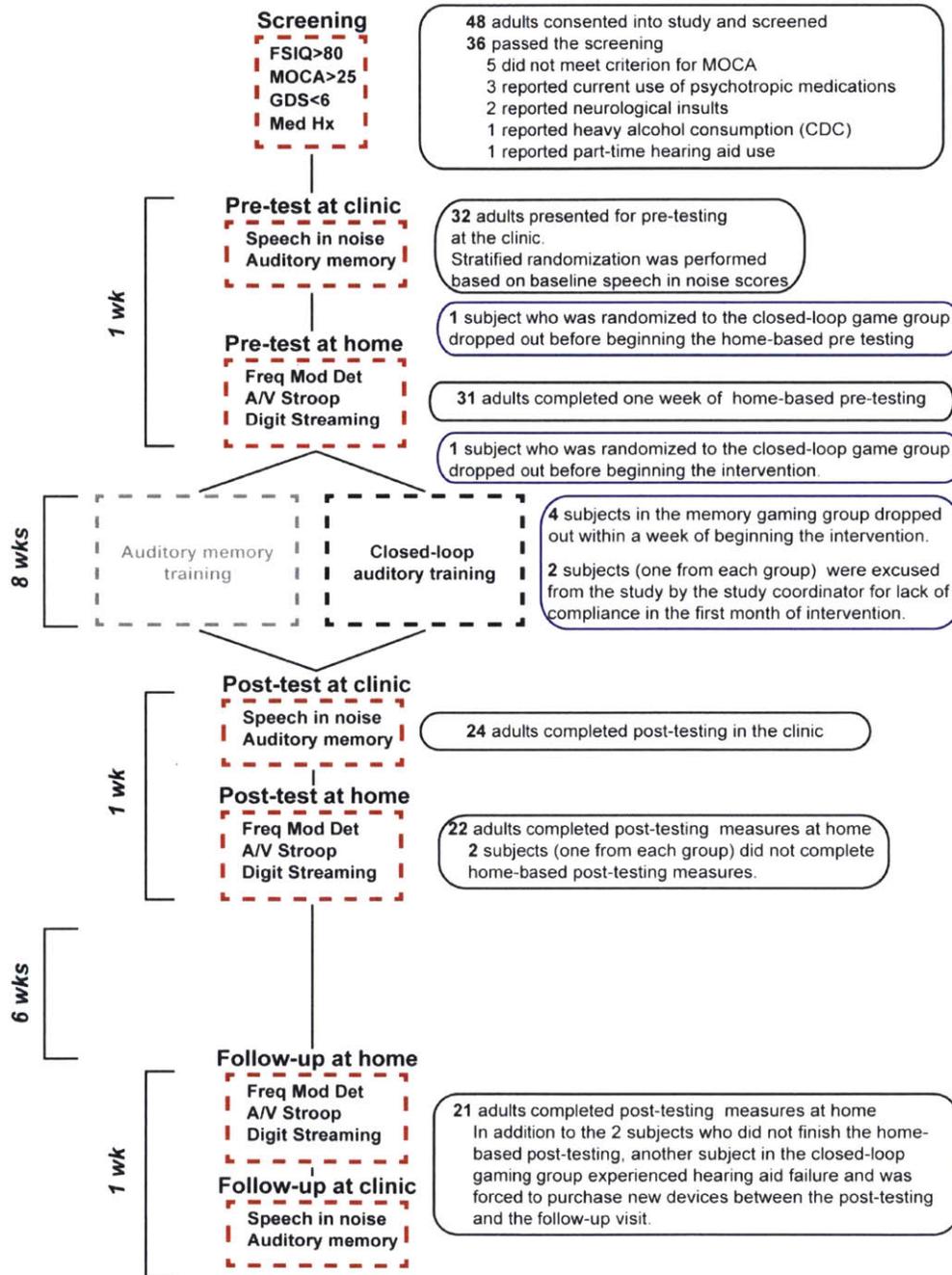
105. Siu KC, Lugade V, Chou LS, van Donkelaar P, Woollacott MH (2008) Dual-task interference during obstacle clearance in healthy and balance-impaired older adults. *Aging Clin Exp Res* 20(4):349–354.
106. Roberts KL, Summerfield AQ, Hall DA (2006) Presentation modality influences behavioral measures of alerting, orienting, and executive control. *J Int Neuropsychol Soc* 12(4):485–492.
107. Amer T, Kalender B, Hasher L, Trehub SE, Wong Y (2013) Do older professional musicians have cognitive advantages? *PLoS One* 8(8):1–8.
108. McClain L (1983) Stimulus-response compatibility affects auditory Stroop interference. *Percept Psychophys* 33(3):266–270.
109. Green EJ, Barber PJ (1983) Interference effects in an auditory Stroop task: congruence and correspondence. *Acta Psychol (Amst)* 53(3):183–194.
110. Ball K, et al. (2002) Effects of cognitive training interventions with older adults: a randomized controlled trial. *JAMA* 288(18):2271–2281.
111. Stern Y, et al. (2011) Space Fortress game training and executive control in older adults: a pilot intervention. *Neuropsychol Dev Cogn B Aging Neuropsychol Cogn* 18(6):653–677.
112. Tunney N, et al. (2009) Aging and Motor Learning of a Functional Motor Task. *Phys Occup Ther Geriatr* 21(3):1–16.
113. Rodrigue KM, Kennedy KM, Raz N (2005) Aging and longitudinal change in perceptual-motor skill acquisition in healthy adults. *J Gerontol B Psychol Sci Soc Sci* 60(4):P174–181.
114. Florentine M, Buus S, Scharf B, Zwicker E (1980) Frequency-selectivity in normally-hearing and hearing-impaired observers. *J Speech Hear Res* 23(3):646–669.
115. Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BCJ (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc Natl Acad Sci U S A* 103(49):18866–18869.
116. Humes LE, Dubno JR (2010) Factors affecting speech understanding in older adults. *The Aging Auditory System*, eds Gordon-Salant S, Frisina RD, Popper AN, Fay RR (Springer New York), pp 211–257.
117. Willis SL, et al. (2006) Long-term effects of cognitive training on everyday functional outcomes in older adults. *JAMA* 296(23):2805–14.
118. Seitz A, Watanabe T (2005) A unified model for perceptual learning. *Trends Cogn Sci* 9(7):329–334.
119. Green CS, Bavelier D (2012) Learning, attentional control, and action video

- games. *Curr Biol* 22(6):R197–R206.
120. Green CS, Pouget A, Bavelier D (2010) Improved probabilistic inference as a general learning mechanism with action video games. *Curr Biol* 20(17):1573–1579.
 121. Bejjanki VR, et al. (2014) Action video game play facilitates the development of better perceptual templates. *Proc Natl Acad Sci U S A* 111(47):16961–16966.
 122. Doshier BA, Lu ZL (1998) Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc Natl Acad Sci U S A* 95(23):13988–13993.
 123. Andersen GJ, Ni R, Bower JD, Watanabe T (2010) Perceptual learning, aging, and improved visual performance in early stages of visual processing. *J Vis* 10(13):1–13.
 124. Li R, Polat U, Scalzo F, Bavelier D (2010) Reducing backward masking through action game training. *J Vis* 10(14):1–13.
 125. Sanchez CA (2012) Enhancing visuospatial performance through video game training to increase learning in visuospatial science domains. *Psychon Bull Rev* 19(1):58–65.
 126. Schubert T, et al. (2015) Video game experience and its influence on visual attention parameters: An investigation using the framework of the Theory of Visual Attention (TVA). *Acta Psychol (Amst)* 157:200–214.
 127. Blacker KJ, Curby KM, Klobusicky E, Chein JM (2014) Effects of action video game training on visual working memory. *J Exp Psychol Hum Percept Perform* 40(5):1992–2004.
 128. Wu S, Spence I (2013) Playing shooter and driving videogames improves top-down guidance in visual search. *Atten Percept Psychophys* 75(4):673–686.
 129. Boot WR, et al. (2013) Video Games as a means to reduce age-related cognitive decline: attitudes, compliance, and effectiveness. *Front Psychol* 4:1–9.
 130. Belchior P, et al. (2013) Video game training to improve selective visual attention in older adults. *Comput Human Behav* 29(4):1318–1324.
 131. Seçer I, Satyen L (2014) Video game training and reaction time skills among older adults. *Act Adapt Aging* 38(3):220–236.
 132. Wang P, et al. (2016) Action video game training for healthy adults: A meta-analytic study. *Front Psychol* 7:1–13.
 133. Basak C, Boot WR, Voss MW, Kramer AF (2008) Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychol Aging* 23(4):765–777.

134. Pearce C (2008) The truth about baby boomer gamers: A study of over-forty computer game players. *Games Cult* 3(2):142–174.
135. Krampe RT, Ericsson KA (1996) Maintaining excellence: Deliberate practice and elite performance in young and older pianists. *J Exp Psychol Gen* 125(4):331–359.
136. Tschida K, Mooney R (2012) The role of auditory feedback in vocal learning and maintenance. *Curr Opin Neurobiol* 22(2):320–327.
137. Hung S-C, Seitz AR (2014) Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *J Neurosci* 34(25):8423–8431.
138. Harris H, Glicksberg M, Sagi D (2012) Generalized Perceptual Learning in the Absence of Sensory Adaptation. *Curr Biol* 22(19):1813–1817.
139. Xiao LQ, et al. (2008) Complete Transfer of Perceptual Learning across Retinal Locations Enabled by Double Training. *Curr Biol* 18(24):1922–1926.
140. Lin FR, et al. (2013) Hearing loss and cognitive decline in older adults. *JAMA Intern Med* 173(4):293–299.
141. Rokem A, Silver MA (2010) Cholinergic enhancement augments magnitude and specificity of visual perceptual learning in healthy humans. *Curr Biol* 20(19):1723–1728.
142. Gervain J, et al. (2013) Valproate reopens critical-period learning of absolute pitch. *Front Syst Neurosci* 7:1–11.
143. Gurgel RK, Jackler RK, Dobie R a., Popelka GR (2012) A new standardized format for reporting hearing outcome in clinical trials. *Otolaryngol -- Head Neck Surg* 147(5):803–807.
144. Norman KL GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers. *Interact Comput* 25(4):278–283.
145. IJsselsteijn WW, et al. (2008) Measuring the experience of digital game enjoyment. *Measuring Behavior*, eds Spink A, et al. (Maastricht), pp 88–89.

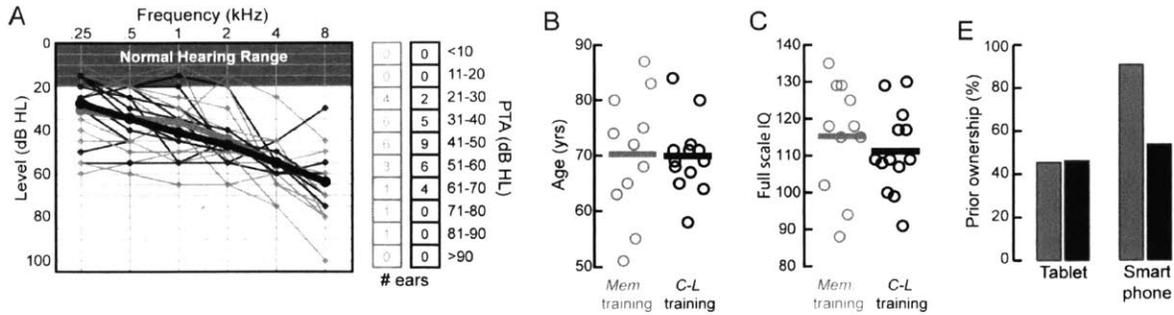
4.7 Supplemental Figures and Tables

Supplemental Figure 1



Supp Fig 1. Study flowchart .(A) Of the thirty-two participants who began the study, twenty-four completed the pretest, intervention, and posttest. Twenty-one participants completed the two month follow-up assessment.

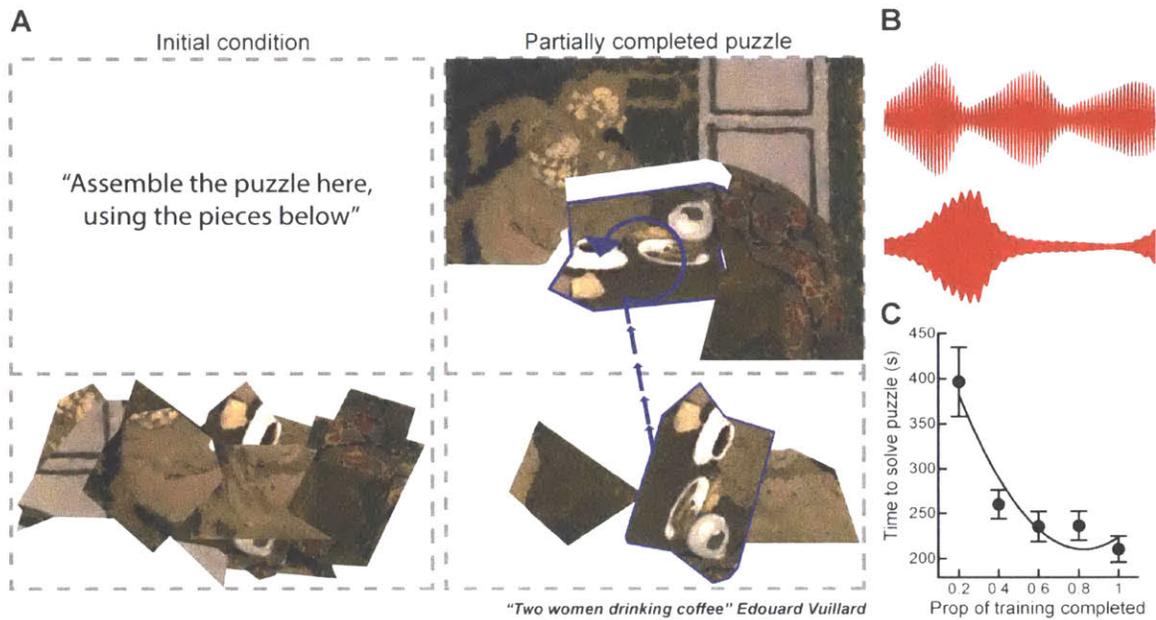
Supplemental Figure 2



Supp Fig 2. Study demographics. (A) Air conduction pure tone detection thresholds were collected by an audiologist in a sound treated booth. Subjects generally presented with mild sloping to moderately-severe sensorineural hearing loss (memory game = gray, closed-loop game = black). Distribution of pure tone averages (PTA .5 – 2 kHz) in the sample plotted according to AAO-HNS recommendations ((143), right). (B) Participant age, (C) full scale IQ and (D), prior technology ownership were well balanced across training groups.

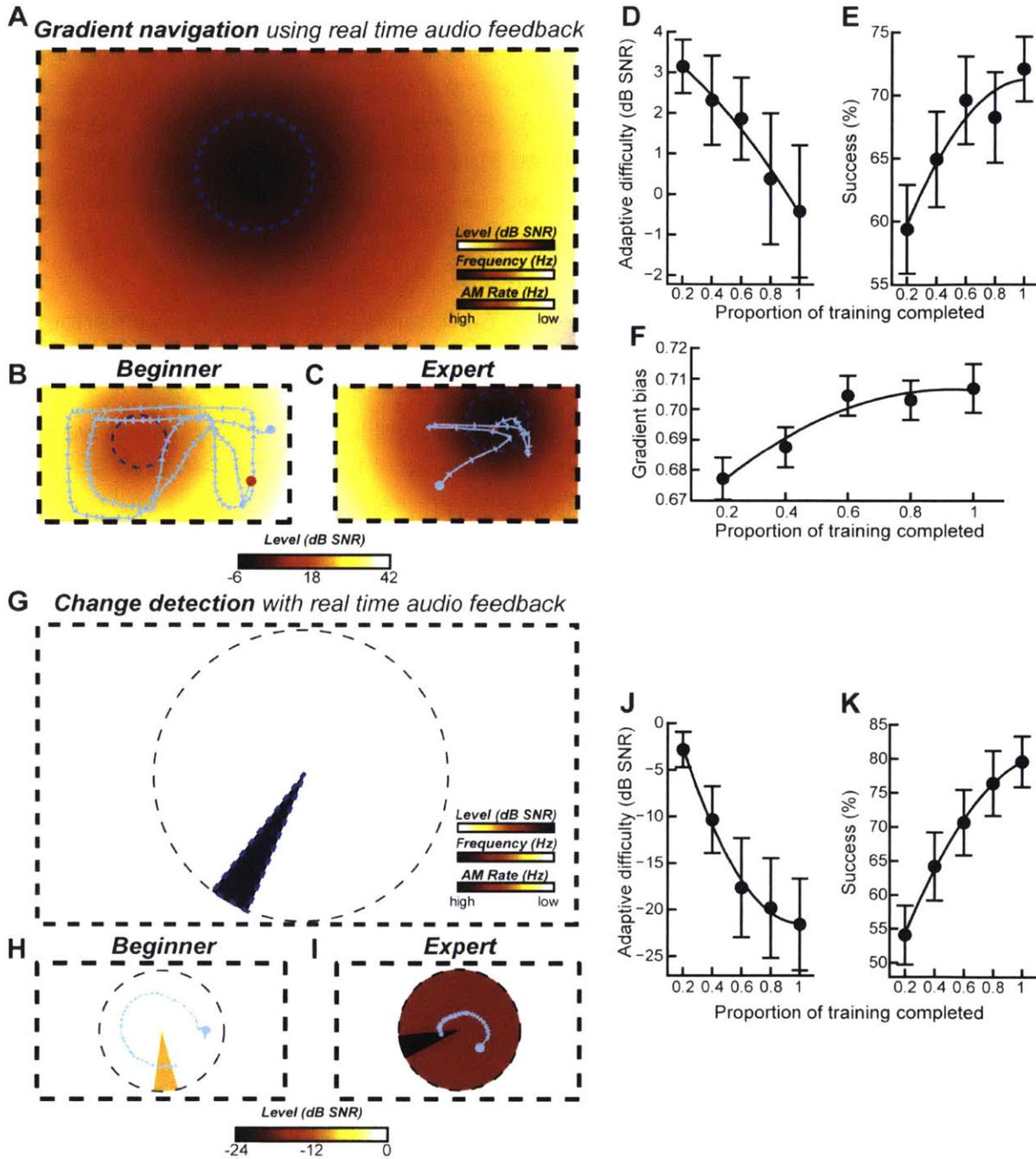
Supplemental Figure 3

Audio memory subgame - complete puzzle while listening to C-L game sounds



Supp Fig 3. Subgame design for the auditory working memory training task. (A) The auditory memory game involved a visuospatial puzzler sub-game. Initially, the puzzle pieces were jumbled at the bottom of the screen (*left*). The participant touched a puzzle piece with their finger, guided it to the correct position on the puzzle board, and then rotated it into the correct orientation (*right*). (B) While the subject performed this task, dynamic tone stimuli from the closed-loop training game were played in the background (red waveforms). It is important to note that while these stimuli were similar to those that were used in the closed-loop game, they were presented in an open-loop context during the memory sub-game. So though the motor behaviors involved in moving the puzzle pieces to the correct location and rotating them into the correct orientation while dynamic sounds continuously played in the background were identical to the conditions of the sub-game tasks in the closed-loop game (**Fig S4**), the subjects' motor behaviors were not correlated with the changes in the sound that they were hearing in the memory sub-game. Rather, subjects were guided by visuospatial cues. (C) Subject performance improved substantially on this puzzle sub-game; the time required to complete a puzzle reduced by nearly 50% over the course of training ($F = 15.2$, $P = 1.2 \times 10^{-7}$, RMANOVA).

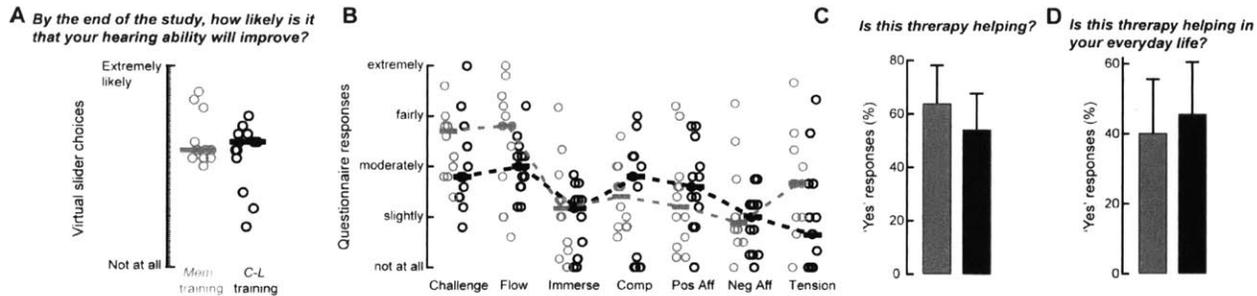
Supplemental Figure 4



Supp Fig 4. Subgame design for the closed-loop audiomotor task. There were two closed-loop audio sub-games. (A) The first game involved navigating audio gradients that varied logarithmically with Euclidian distance from a circular target in order to place a puzzle piece in its correct location. (B-C) While beginners attempted to solve the task through exhaustive searches (B), expert players used the most information rich vectors,

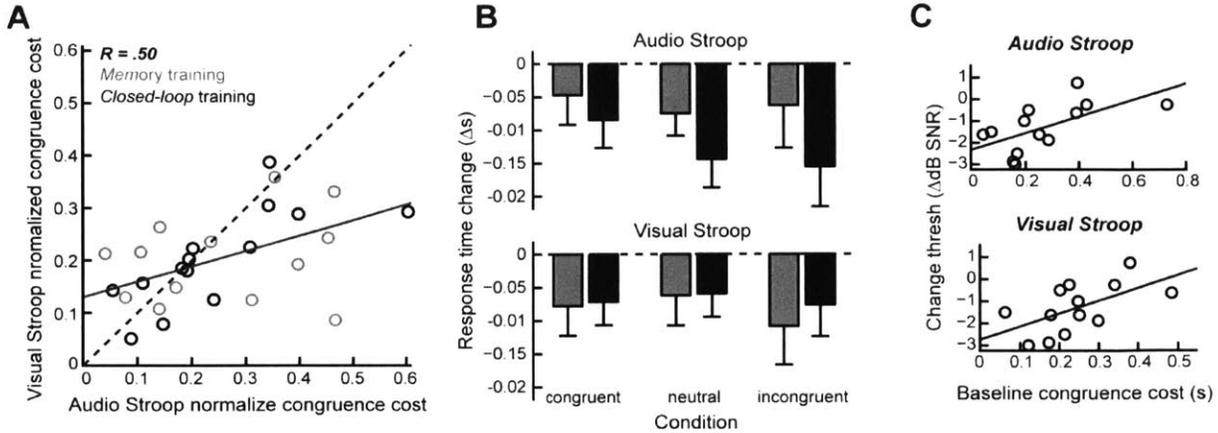
biasing their searches along the highest sloping portions of the gradient (C). Cyan arrows depict the subject's paths, with each arrow representing movement over a 0.5 s time bin. (D) Difficulty in the game adaptively increased via reduction of the SNR (D, $F = 3.72$, $P = 0.01$, RMANOVA). (E) By examining performance on "yardstick" trials (where SNR was fixed at -18 dB), we observed significant success rate improvements with training (E, $F = 6.8$, $P = 2.0 \times 10^{-2}$, RMANOVA). (F) Using the "yardstick" trials, we also examined the degree to which subjects aligned their search vectors with the most informative region of the gradient ($Gradient\ bias = \frac{v \cdot w}{(|v||w|)}$, where v is the player's actual traveled angle and w is the angle between the player and the target). We observed significant increases in gradient bias with training (F, $F = 8.4$, $P = 3.0 \times 10^{-5}$, RMANOVA). Gradient bias values range from 0.64 (chance) to 1 (movements perfectly aligned with the highest sloping portion of the gradient). (G) The second sub-game involved rotating the puzzle piece around a central axis to identify the correct orientation. As the subject rotated the puzzle piece, sound level, pitch, or rate stayed constant until the target angle was reached, at which point the target feature was modulated with a step function. The subjects were required to stop rotation and release the puzzle piece immediately upon detecting this change. (H-J) Rotating beyond this point, as visualized in a trial generated by a beginner (H) resulted in failure and re-randomization of the target orientation. Likewise, releasing the piece before arriving at the target location also resulted in failure. Expert users (I) learned to perform this task accurately at progressively worse SNRs (J, $F = 18.2$, $P = 3.0 \times 10^{-9}$, RMANOVA). (K) By examining performance on "yardstick" conditions (where SNR was fixed at -18 dB), we observed significant success rate improvements with training (K, $F = 16.57$, $P = 1.0 \times 10^{-8}$, RMANOVA).

Supplemental Figure 5



Supp Fig 5. Expectations and game play experience were matched between the auditory memory and closed-loop audiomotor training tasks. (A) After subjects played their training game for a week they were asked to use a virtual slider to rank their expectancy that their hearing would improve as a function of playing their assigned game. Expectations were well matched across training groups (gray = memory game, black = closed-loop game, $z = 0.18$, $P = 0.86$, Wilcoxon rank-sum). (B) At the same time point, we also measured the participants' impressions of their game experience using the Game Experience Questionnaire (144, 145). Responses to questions on the Game Experience Questionnaire are divided into seven experience categories. Both games were rated as moderately to fairly challenging and only slightly immersive. Flow, which involves questions concerning "losing track of time" and "being fully occupied with the game," was also ranked as moderate to fair. (Comp = competence, Pos Aff = positive affect, Neg Aff = negative affect). There were no significant differences between the ratings that the memory and closed-loop games received across categories. The largest non-significant difference between the two groups was found for challenge, with the memory game being rated as more challenging than the closed-loop game ($z = 1.75$, $P = 0.08$ uncorrected, $P = 0.56$ Holm-Bonferroni corrected for multiple comparisons, Wilcoxon rank-sum). (C) After the subjects had trained for 1 month, they were asked whether they felt that the therapy was helping. About half of each group responded affirmatively to that question. (D) After two months of training we asked participants if the therapy was helping in their everyday lives. Around 40% responded affirmatively to this question. There were no significant response differences between groups for either of these questions ($P \geq 0.7$, Fisher's exact test).

Supplemental Figure 6



Supp Fig 6. Estimating cognitive interference with audio and visual Stroop tests. (A) The normalized congruence cost for the visual and auditory versions of the Stroop task were significantly correlated ($R = 0.50$, $P = 8 \times 10^{-3}$, Pearson's correlation coefficient, gray = *Mem.* training, black = *C-L* training). (B) Stroop congruency costs did not decrease as a function of either training game; however, reaction times decreased for all congruency conditions ($P \leq 0.05$ for all conditions, time effect, RMANOVA). (C) Baseline performance on both the audio and visual versions of the Stroop task were significant predictors of learning transfer to the sentence recognition in noise tasks following closed-loop training (C, Audio: $R = 0.62$, $P = 0.048$; Visual: $R = 0.58$, $P = 0.04$, Pearson's correlation coefficient, Holm-Bonferroni corrected for multiple comparisons).

Supplemental Table 1

Analysis of Covariance: Pre vs Post scores

Outcome measure	F value	P value
Words in noise	0.3	0.60
Low context sentences in noise	5.2	0.03
High context sentences in noise	5.1	0.04
Frequency modulation detection	0.0	0.96
Letter number sequencing	0.4	0.65
Audio stroop	0.7	0.40
Visual stroop	0.2	0.67

Supp Table 1. Analysis of Covariance . An alternate analysis of changes in outcome measures following game training was executed by performing ANCOVA using the baseline score as a covariate. The pattern of statistical significance is identical to that obtained using interaction terms of the repeated measures ANOVA reported in the main text.