HHS Public Access

Author manuscript

Ear Hear. Author manuscript; available in PMC 2016 January 11.

Published in final edited form as:

Ear Hear. 2015 ; 36(6): e326–e335. doi:10.1097/AUD.0000000000000186.

Fast, Continuous Audiogram Estimation using Machine Learning

Xinyu D. Song[1], **Brittany M. Wallace**[2], **Jacob R. Gardner**[3], **Noah M. Ledbetter**[1], **Kilian Q. Weinberger**[3], and **Dennis L. Barbour**[1]

[1]Laboratory of Sensory Neuroscience and Neuroengineering, Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, U.S.A.

[2]Program in Audiology and Communication Sciences, 660 S. Euclid Ave., Campus Box 8042, St. Louis, MO 63110

[3]Department of Computer Science & Engineering, Bryan Hall, CB 1045, 1 Brookings Drive, Saint Louis, MO, USA 63130

Abstract

Objectives—Pure-tone audiometry has been a staple of hearing assessments for decades. Many different procedures have been proposed for measuring thresholds with pure tones by systematically manipulating intensity one frequency at a time until a discrete threshold function is determined. The authors have developed a novel nonparametric approach for estimating a continuous threshold audiogram using Bayesian estimation and machine learning classification. The objective of this study is to assess the accuracy and reliability of this new method relative to a commonly used threshold measurement technique.

Design—The authors performed air conduction pure-tone audiometry on 21 participants between the ages of 18 and 90 years with varying degrees of hearing ability. Two repetitions of automated machine learning audiogram estimation and 1 repetition of conventional modified Hughson-Westlake ascending-descending audiogram estimation were acquired by an audiologist. The estimated hearing thresholds of these two techniques were compared at standard audiogram frequencies (i.e., 0.25, 0.5, 1, 2, 4, 8 kHz).

Results—The two threshold estimate methods delivered very similar estimates at standard audiogram frequencies. Specifically, the mean absolute difference between estimates was 4.16 ± 3.76 dB HL. The mean absolute difference between repeated measurements of the new machine learning procedure was 4.51 ± 4.45 dB HL. These values compare favorably to those of other threshold audiogram estimation procedures. Furthermore, the machine learning method generated threshold estimates from significantly fewer samples than the modified Hughson-Westlake procedure while returning a continuous threshold estimate as a function of frequency.

Please send correspondence to: Dr. Dennis Barbour, Department of Biomedical Engineering, Washington University, One Brookings Dr., Campus Box 1097, Uncas Whitaker Hall Room 200E, St. Louis, MO 63130, U.S.A., Tel. (314) 935-7548, Fax. (314) 935-7448, dbarbour@biomed.wustl.edu.

X.D.S designed the experiments, developed the algorithm, analyzed data, and wrote the main paper; B.M.W. designed and conducted the experiments; J.R.G. refined the machine learning algorithm; N.M.L., K.Q.W. and D.L.B. provided professional expertise and critical revision. All authors discussed the results and implications and commented on the manuscript at all stages.

Conclusions—The new machine learning audiogram estimation technique produces continuous threshold audiogram estimates accurately, reliably, and efficiently, making it a strong candidate for widespread application in clinical and research audiometry.

Introduction

The procedure typically followed for clinical audiogram estimation currently is pure-tone audiometry (PTA) using the modified Hughson-Westlake (HW) procedure (Hughson & Westlake 1944), which was proposed as a standard for audiological testing decades ago (Carhart & Jerger 1959). As detailed by ANSI, the procedure proceeds one frequency at a time with the presentation of a tone at a sequence of intensities determined by the listener's most recent response. In a common variant, the first intensity delivered is at a level audible to the listener, and the level is reduced in fixed-size increments until the listener no longer responds. The intensity is then increased by a smaller fixed-size increment until the listener again responds. This procedure is repeated for several "reversals" (Franks 2001; American National Standards Institute 2004; American Speech-Language-Hearing Association 2005).

In parallel to the development of adaptive conventional approaches like the one described above, automated audiometry methods play a role in clinical audiometry with the earliest form designed by George von Békésy in the late 1940s (Békésy 1947). Békésy's proposed automated audiogram, often referred to as "Békésy audiometry," implemented a method of adjustment, giving listeners control of an attenuator used to identify the intensity at which they could not hear the presented stimulus. Additionally, many computerized audiometric methods designed to ensure consistency and save labor have been developed, with some employing a method of adjustment similar to Békésy's technique but most using a method of limits resembling the HW algorithm (Ho et al. 2009; Margolis et al. 2010; Swanepoel et al. 2010; Mahomed et al. 2013). Even with ready access to powerful digital computing technology today, however, computerized automated audiometry sees relatively little use in clinical diagnostic settings, with most audiograms still obtained manually (Vogel et al. 2007).

A recent exhaustive review and meta-analysis was conducted of techniques developed for automated threshold audiometry (Mahomed et al. 2013). A wide range of automated techniques produced audiograms generally comparable to manual audiograms, with an absolute average difference of 4.2 dB HL and a standard deviation of 5.0 dB HL (n = 360). Test-retest reliability among these automated methods demonstrated an absolute average difference of 2.9 dB HL and a standard deviation of 3.8 dB HL (n = 80). As a comparison, manual threshold audiometry in the reported studies produced an absolute average difference of 3.2 dB HL and a standard deviation of 3.9 dB HL (n = 80). These studies indicate that computerized automation of pure-tone audiometry procedures yields threshold audiograms comparable in value and test-retest reliability to conventional manual procedures.

Adaptive techniques, such as those described in (Mahomed et al. 2013) and the HW procedure itself, share the common feature of systematically manipulating pure-tone intensity one frequency at a time until a threshold value at each of the sampled frequencies

is determined. Disadvantages of such an approach include 1) multiple stimuli with high or low probabilities of detection must be presented for each test frequency; 2) identical or nearly identical stimuli are presented repeatedly near threshold; and 3) stimulus presentation sequences have a large degree of predictability, which facilitates the intentional subversion of test results by noncooperative listeners.

To address primarily the first shortcoming above, several methods have been developed by psychophysicists seeking optimal sampling methods (Leek 2001). For instance, the parameter estimation by sequential testing (PEST) method (Taylor & Creelman 1967; Hall 1981) adjusts step size dynamically to systematically narrow stimulus parameter ranges down to values of interest, while maximum likelihood methods (Pentland 1980; Watson & Pelli 1983) sequentially select points most likely to be informative using current estimates of the psychometric function. Bayesian methods have also been applied to the problem of optimal sampling for psychometric functions, typically by constructing a posterior estimate of the function given the existing data and selecting the next sample point based upon some optimality criterion, such as maximizing information gain (King-Smith et al. 1994; Kontsevich & Tyler 1999; Lesmes et al. 2006; Remus & Collins 2008; Kujala 2011; Shen & Richards 2013). Techniques inspired by these methods have been applied in auditory threshold estimation on a per-frequency basis and have demonstrated threshold estimates consistent with traditional sampling techniques (Green 1992; Formby et al. 1996; Leek et al. 2000). Particularly noteworthy is a dynamic Bayesian technique that guides optimal sampling across a range of frequencies and intensities using interfrequency relationships derived from a database of candidate audiometric patterns (Özdamar et al. 1990).

To address the second shortcoming above, techniques such as Békésy audiometry and Audioscan® systematically sweep tone stimuli through multiple frequencies (Békésy 1947; Meyer-Bisch 1996; Ishak et al. 2011). These continuous-audiogram estimation techniques, particularly Audioscan®, are able to identify various hearing pathologies that cannot always be detected by discrete pure-tone audiometric approaches (Jerger 1960; Zhao et al. 2002; Zhao et al. 2014). Despite this advantage, however, these techniques do not currently see substantial use in the clinic. The main reason for this is the substantially lengthened testing time required compared to conventional PTA, particularly with sweep rates that are comfortable for listeners (Ishak et al. 2011). Furthermore, substantial engagement by the listener is required, which could lead to inefficient acquisition, inaccuracies, and/or intentional misrepresentation.

A new technique that can generate continuous audiogram estimates through efficient deployment of test stimuli could potentially combine the advantages of Bayesian and sweep audiometry. As an added bonus, if this method were less predictable than conventional methods, noncooperative listeners would be revealed.

Machine learning (ML) is a field of computation employing principled methods to subdivide complex parameter spaces into informative categories. It encompasses a powerful set of tools for performing efficient data-driven inference on complex spaces or processes (Bishop 2006; Hastie et al. 2009; Murphy 2012) By merging classification methods from machine learning with techniques for optimal Bayesian estimation and effective sampling procedures

from psychophysics, we propose to simultaneously speed acquisition, increase accuracy and increase test sensitivity for PTA. To this end, we have designed a machine learning audiogram estimation procedure that finds hearing thresholds continuously across frequency while efficiently sampling the psychometric space.

Materials & Methods

Machine learning algorithm

An algorithm employing Gaussian process (GP) regression (Rasmussen & Williams 2006) was used to construct tone detection audiogram estimates from human listeners in real time. GP regression is a nonparametric Bayesian machine-learning technique that performs rapid and accurate estimation of multidimensional functions. GPs can be thought of as an extension of the multivariate Gaussian distribution to infinitely many random variables. Each single variable represents the range of possible values an output function can take when evaluated at a particular input and is Gaussian-distributed with some mean and variance. In the context of the tone-detection audiogram, the input values are the frequency and intensity of presented pure tones, and the output function is an individual's probability of detecting the given tone (i.e. the individual's psychometric profile across frequency-intensity space).

A GP is fully specified by 1) its mean function, which describes the central tendency of the overall output function and 2) its covariance function, which describes the relationship between any pair of output function values. Upon conditioning on a set of observed data, the GP can produce a posterior probability distribution (often called "predictive posterior") for some set of inputs. The mean of this posterior distribution (distinct from the GP mean *function*) represents the GP's best prediction of output function values at the new inputs. The posterior variance, on the other hand, reflects the GP's inherent uncertainty in estimate quality, or the GP's "confidence" about its prediction of the function value at each set of inputs.

Variable space—For audiogram estimation, the input variables are the frequency and intensity of presented pure tones. The GP was trained to predict the probability of a listener's tone detection as a function of these variables, which takes on continuous values between 0 and 1 over all combinations of frequency and intensity.

Covariance function—The GP covariance function describes how variables change with one another, and generally describes the "smoothness" of the GP function in each dimension. For this application, constraints that reflected prior knowledge about psychometric functions were incorporated into the covariance functions. Most crucially, the probability of listener detecting a tone is monotonically increasing as a function of tone intensity, but need not have an explicit dependence upon frequency except that the overall function is continuous. To reflect this scenario, separate covariance functions are used for the frequency and intensity dimensions: a monotonically increasing linear kernel in intensity, and a more flexible squared exponential (SE) kernel in frequency. To ensure that the GP returned a probability estimate, values were transformed with a cumulative Gaussian likelihood function in intensity.

Calculation of hyperparameters—4 covariance function hyperparameters were used for this model: a Gaussian-distributed noise parameter (allowing the GP to be more robust to false positives and negatives), an amplitude for the SE kernel, a characteristic length scale for the SE kernel, and a slope for the linear kernel. Hyperparameters are updated (learned) automatically following each response by minimizing the negative log marginal likelihood of these hyperparameters with respect to the sampled data.

Calculation of predictive posterior—Following each new tone presentation, the data sampled up to that point were used to compute the predictive posterior probability distribution. Both the mean posterior probability of detection at any frequency/intensity combination and an uncertainty (posterior variance) for this estimate were computed. Figure 1A shows an example of the posterior mean during data acquisition for one audiogram estimate and Figure 1B shows the corresponding posterior variance.

Informative sampling—After initializing with a few pseudorandom samples, only points deemed to be highly informative to the estimate were selected for subsequent samples. We adopted a strategy known as *uncertainty sampling* in which at each algorithm iteration, the next chosen sample point was one whose class identity (i.e. heard or unheard) was the most uncertain (Lewis & Catlett 1994; Lewis & Gale 1994; Settles 2009). Based upon the calculated variance function for the previous iteration, the frequency/intensity pair corresponding to the highest value in the variance function was selected for the next point to sample (Figure 1B). If multiple points were tied for maximum variance, a point was selected at random from this set. After determining the listener's response, the posterior distribution was updated for the next iteration (updated posterior mean shown in Figure 1C). The cycle of hyperparameter estimation, posterior calculation, and uncertainty sampling was repeated until convergence criteria were met, as detailed in *Experimental procedure*.

Our technique shares similarities with previously described Bayesian techniques for estimation of psychometric functions. Those techniques also build a posterior probability distribution from existing data to estimate the psychometric function (Kontsevich & Tyler 1999; Lesmes et al. 2006; Remus & Collins 2008; Kujala 2011; Shen & Richards 2013). Additionally, the strategy of uncertainty sampling we have adopted for selecting informative sample points resembles strategies that successively pick points to minimize some cost, such as expected variance (King-Smith et al. 1994) or entropy (Kontsevich & Tyler 1999; Lesmes et al. 2006; Shen & Richards 2013). Our technique is able to generate a multivariate psychometric function in frequency-intensity space in much the same way that, for instance, the *qPvC* technique is able to generate a psychometric function in noise-contrast space (Lesmes et al. 2006).

A key departure relative to the other Bayesian techniques described previously is that ours is a nonparametric technique. Unlike the general shape of the auditory filter, for instance, which can be accurately characterized by a 3-parameter equation (Shen & Richards 2013), the shape of the audiogram varies significantly between individuals and cannot be similarly parameterized. Instead, we adopted a nonparametric approach that only assumes tone-detection thresholds are generally similar for very close frequencies (i.e. the audiogram is continuous). This assumption is reflected in the SE covariance kernel chosen for the GP in

the frequency dimension, which enforces a general smoothness (Rasmussen & Williams 2006). The tone detection probability estimate is produced by a posterior estimate of the function values given the observed data and learned hyperparameters, rather than the more typical case of optimizing over a set of parameters to best fit the observed data (Lesmes et al. 2006; Shen & Richards 2013; Shen et al. 2014).

The current method perhaps most resembles the technique of (Özdamar et al. 1990), which also samples across a range of frequencies and intensities and informs estimates of one frequency using information from nearby frequencies. Rather than selecting between candidate audiogram patterns, however, the current method incorporates prior beliefs about psychometric functions into the covariance function, essentially expanding the number of candidate patterns to all possible patterns under the given covariance function and hyperparameters.

Participants

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

Experimental procedure

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

Manual HW audiometry—A conventional audiogram was conducted by an audiologist according to accepted standards (American National Standards Institute 2004; American Speech-Language-Hearing Association 2005). Each listener was instructed to raise his or her hand upon detection of a presented pure-tone stimulus. Hearing ability was assessed at standard audiogram frequencies (0.25, 0.5, 1, 2, 4, and 8 kHz), with the possible intensity

ranging from −20 to 100 dB HL in a minimum of 5-dB increments. For an individual frequency, a pure tone was first presented at an audible intensity based upon the audiologist's clinical judgment, then reduced in 10-dB increments until the listener failed to respond. Henceforth, the intensity was increased in 5-dB increments following detected tones and decreased in 10-dB increments following undetected tones. The threshold for that frequency was determined by the lowest-intensity tone to elicit a response in at least 2 of 3 ascending trials. The manual audiogram was conducted separately for left and right ears. This manual method is the modified Hughson-Westlake ascending-descending procedure and is referred to here as HW audiometry {Carhart, #159; Katz, 2009 #225}.

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

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

Analysis

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

the intensity dimension, there could be a maximum of only 1 transition point at each frequency. The threshold values at each frequency therefore become a continuous (in frequency) estimate of the listener's threshold audiogram.

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

Results

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

Samples obtained during both the manual HW and automated ML methods are shown in Figure 2 for 1 representative listener, with the final audiogram estimates shown as superimposed lines. Note that the HW method searched each standard audiogram frequency across a number of intensities, with several repeat presentations of specific stimuli. The ML procedure, on the other hand, sampled across a more diverse set of frequencies with no repeats.

The degree of similarity among the different audiogram estimates for each ear is readily apparent, despite the differences in sampling procedure for each. The skilled audiologist was able to rapidly discover reversals and spent the most time probing right around threshold. A

less skilled individual may have spent more time sampling points farther from threshold. These examples concisely demonstrate the utility of the HW procedure in trained hands and help explain why it is still in use many decades after its development.

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

Table 2 shows the results of evaluating the accuracy of the ML audiogram at standard audiogram frequencies relative to the HW audiogram averaged across all listeners and estimation runs. For the 6 standard audiogram frequencies, the mean estimated threshold difference was −0.011 ± 5.61 dB HL, the mean absolute estimated threshold difference was 4.16 ± 3.76 dB HL, the median absolute estimated threshold difference was 3.00 dB HL with an interquartile range of 5.00 dB HL, and the percent 5-dB difference in threshold estimates was 66.25. These values compare favorably with historical differences in audiogram estimation methodologies (Gosztonyi Jr. et al. 1971; Schmuziger et al. 2004; Ishak et al. 2011; Mahomed et al. 2013). Judging from the relatively low percent 5-dB difference yet comparable mean absolute difference, the ML procedure appears to produce somewhat more outlier estimates at individual frequencies than methods that estimate directly at those frequencies. Naturally, it is possible that the outliers in this case arise from the HW procedure.

The ML audiogram's clinically relevant performance was evaluated by classification of audiogram results into conventional categories of hearing loss (normal, mild, moderate, severe, and profound) using the pure-tone average (Stach 2008; Katz et al. 2009). The categorical classifications produced by ML and HW audiogram estimates in our listeners were in agreement 95.0% of the time, and the disagreements in pure-tone average classification resulted in adjacent clinical categories. This result provides further evidence that the ML audiogram generates information that is clinically equivalent to the conventional HW audiogram using current standards.

Table 3 shows the results of evaluating the test-retest reliability of the automated ML audiogram at standard frequencies averaged across all listeners and estimation runs. Across all frequencies, the mean signed difference between automated audiogram runs was 0.75 ± 6.29 dB HL, the mean absolute difference between runs was 4.51 ± 4.45 dB HL, and the median absolute difference between runs was 3.00 dB HL with interquartile range 4.00 dB HL. These values are comparable to previously reported absolute test-retest differences for manual audiometry: 3.2 ± 3.9 dB HL (Fausti et al. 1990; Swanepoel et al. 2010; Mahomed

et al. 2013). This degree of similarity in final estimate between runs where the different initial randomization led to nonoverlapping probe stimuli in the two cases indicates the robustness of the ML procedure.

Figure 4 shows the accuracy of the ML procedure as a function of algorithm iteration, or equivalently, the number of samples collected. This post-hoc analysis was performed by constructing an ML threshold audiogram estimate from the posterior distribution after each iteration of the GP algorithm and then evaluating the absolute difference from the final HW threshold audiogram at the six standard audiogram frequencies. Figures 4A, B show this trend for two representative ears, and Figure 4C shows the accuracy as a function of algorithm iteration averaged across all listeners and GP algorithm runs that terminated following 36 iterations. In both the individual data and the population data, the accuracy of the ML algorithm tended to improve systematically as a function of iteration. The ML estimate tends to achieve close to its final absolute difference value in only 20 samples or so. In some cases the difference function becomes shallow quickly but remains at some positive value (e.g., Figure 4A). It is possible that this outcome originated from a systematic misestimate in the HW procedure instead of the ML procedure. Finally, note that the first 10 iterations show little systematic improvement in estimate quality, which is caused by the random sampling at this early stage before the informative sampling procedure begins.

The normalized GP posterior variance is shown as a function of ML algorithm iteration in Figure 5. At each iteration, the normalized variance was calculated by summing each value in the posterior variance, which spans values [0, 1], and dividing by the total number of values. Figures 5A, B show this trend for the same runs as in Figure 4, and Figure 5C shows this trend averaged across all listeners and runs. In general, the normalized posterior variance tends to decrease as a function of iteration, implying that the ML audiogram produces a less uncertain (more confident) estimate with an increasing number of samples. This function alone or in combination with other factors could therefore be used to evaluate the overall quality of an estimate.

Discussion

We have introduced a novel automated PTA audiogram estimation technique exploiting recent developments in machine learning. This method is able to provide a nonparametric yet continuous estimate of a listener's pure tone detection probability across all combinations of tone frequency and intensity. To our knowledge, this is the first approach capable of doing so. Furthermore, the procedure we have developed is designed to deliver each successive tone at the frequency and intensity of maximum uncertainty in the estimate up to that point. The result is an estimate of the complete psychometric function as a function of all variables. In the case of PTA, this is the probability of detecting a tone at any particular frequency and intensity, which we refer to as the tone detection audiogram. Any particular contour of the detection audiogram reflects a constant probability of tone detection, or threshold. In this study we will specifically refer to the 0.707 contour of the tone detection audiogram as the threshold audiogram.

While obtaining threshold estimates at all frequencies and achieving accuracy comparable to other algorithms, the ML audiogram consistently required significantly fewer samples to do so. Conventional HW approaches query frequencies individually and obey a rigorous rule for selecting tone intensities, depending at any frequency only upon the accumulated number of reversals, the intensity of the last tone presented and the last listener response. In practice, this means that many samples collected by the conventional HW approach are not particularly informative, e.g., several relatively loud intensities in a row that the listener is very likely to hear. In contrast, the GP algorithm successively selects sample points evaluated by uncertainty sampling to be maximally informative at that point in time about the perceptual space. The rapid accumulation of relevant information in the ML audiogram case is demonstrated clearly in Figure 4, where the accuracy of the GP algorithm approaches reasonable values many iterations before the algorithm eventually terminates. Such rapid convergence simply cannot be accomplished with the HW approach because of its rigid sampling criteria, in either manual or automated form.

Another unique property of our ML audiogram procedure is that an estimate of accuracy is automatically included with each newly computed posterior. In general, both the estimate error (in this case, the correspondence with the HW estimate) and normalized posterior variance decrease as a function of algorithm iteration. The trend in accuracy appears more reliable than the trend in variance: additional samples will typically generate a more accurate estimate of the audiogram because there is more information about the function space. The ML procedure could possibly generate a low-variance yet inaccurate audiogram estimate with very few samples by either underfitting or overfitting, which is responsible for the dramatic drop in GP variance shown in the first 5 samples of Figure 5A. Multiple methods exist to deter underfitting or overfitting (Murphy 2012); the simplest is perhaps to enforce a minimum number of iterations while ensuring that the algorithm is still sampling widely, which was deployed in the current experiment. After the first few iterations, the steady decrease in error implies that with more samples, we can achieve even more accurate audiogram estimates. As Figure 4 suggests, however, this is likely only necessary for individuals whose GP estimates do not converge quickly.

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

A related situation is a listener who responds inconsistently, to which the ML procedure is sensitive. Figure 6 shows an example of one listener who provided inconsistent results by falling asleep during the ML audiogram procedure. The ML threshold audiogram deviated

from the HW threshold audiogram obtained for the same ear (Figure 6A), and the inconsistency in responses that produced this result can be seen in Figure 6B. Note that sample points very close in intensity/frequency space elicited different responses, which is physically unrealistic. The ML procedure produced a threshold audiogram estimate that attempted to best match this inconsistent data. It can also be seen from Figure 6C that the ML algorithm hit the ceiling on the number of allowable iterations for that ear, 64, due to high posterior variance. Figure 6C further reveals that the normalized posterior variance did not generally decrease as a function of iteration; in fact, following iteration 15, the normalized variance gradually increased. The normalized variance may sometimes dramatically increase when the GP hyperparameters change substantially due to a particularly informative sample, but a gradual increase in normalized variance indicates that obtained samples may be of poor quality because each additional sample is making the posterior less, rather than more, well-defined. Sufficient native quantification therefore appears to exist within the ML procedure to signal when a poor estimate is being obtained, in which case an alternate audiogram estimation strategy may be pursued.

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

As indicated in Table 2, mean threshold estimates corresponded closely between the predictable, sequential HW procedure and the unconstrained, roving ML procedure. In general, both one-interval and two-interval detection and discrimination tasks have shown elevated thresholds when one or more stimulus parameters are roved (Berliner & Durlach 1973; Mori & Ward 1992; Amitay et al. 2005; Mathias et al. 2010; Bonino et al. 2013). This is widely interpreted to be an attentional rather than a purely perceptual phenomenon because roving under masked conditions leads to observations best described by informational rather than energetic masking. The lack of threshold elevation with the roving ML stimulus presentations in the current study is therefore somewhat surprising. Our unmasked detection condition may have contributed to the similarity in thresholds. Other potential mitigating factors include our delivery of relatively long tones (Ward 1991), relatively long inter-stimulus intervals (Berliner & Durlach 1973) and, perhaps most significantly, repeated tone presentations (Kidd et al. 2003; Burk & Wiley 2004; Leibold & Bonino 2009; Guest et al. 2010).

Recall that the ML estimation procedure presented here uses no information about actual audiograms other than the expectation that only one threshold exists for each frequency and that the threshold is a continuous function of frequency. That potentially leaves room for additional modifications to improve the accuracy, precision, efficiency and robustness of the algorithm. One logical improvement to the accuracy of the algorithm is to further expand the frequency range from which the ML algorithm may sample. From Table 2 it is apparent that the greatest discrepancy between the HW and ML procedures occurred at 250 Hz and 8000 Hz, and the least discrepancy occurred at 1 kHz and 2 kHz. This most likely means that ML estimation at the extremes of the sampled frequencies is adversely affected by edge effects. This limitation apparently exists despite the observation that many samples are taken near the edge frequencies (c.f., Figure 2). Correcting this limitation would undoubtedly increase overall accuracy of the procedure as well as efficiency and could be accomplished in several ways. The most obvious solution would be to sample frequencies during the ML procedure at frequencies lower than 250 Hz and higher than 8000 Hz. This and other improvements are currently being evaluated.

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

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

Conclusion

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

Acknowledgments

We would like to acknowledge the assistance of William Clark and Mitchell Sommers, who supported the conceptualization and design of these studies.

Funding for this project was provided by NIH grants T35 DC008765, T32 NS073547 and R01 DC009215, as well as the Center for Integration of Medicine and Innovative Technology (CIMIT).

The authors have filed a provisional patent application describing the novel audiogram estimation procedure.

References

American National Standards Institute. Methods for manual pure-tone threshold audiometry. ANSI. 2004; 3:21.

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

Amitay S, Hawkey DJ, Moore DR. Auditory frequency discrimination learning is affected by stimulus variability. Percept Psychophys. 2005; 67:691–698. [PubMed: 16134462]

Békésy GV. A new audiometer. Acta Oto-Laryngologica. 1947; 35:411–422.

Berliner JE, Durlach NI. Intensity perception. IV. Resolution in roving-level discrimination. J Acoust Soc Am. 1973; 53:1270–1287. [PubMed: 4712555]

Bishop, CM. Pattern recognition and machine learning. New York: Springer; 2006.

Bonino AY, Leibold LJ, Buss E. Effect of signal-temporal uncertainty in children and adults: tone detection in noise or a random-frequency masker. J Acoust Soc Am. 2013; 134:4446. [PubMed: 25669256]

Burk MH, Wiley TL. Continuous versus pulsed tones in audiometry. Am J Audiol. 2004; 13:54–61. [PubMed: 15248804]

Carhart R, Jerger J. Preferred method for clinical determination of pure-tone thresholds. Journal of Speech & Hearing Disorders. 1959

Fausti SA, Frey R, Henry J, et al. Reliability and validity of high-frequency (8–20 kHz) thresholds obtained on a computer-based audiometer as compared to a documented laboratory system. Journal of the American Academy of Audiology. 1990; 1:162–170. [PubMed: 2132600]

Formby C, Sherlock L, Green DM. Evaluation of a maximum likelihood procedure for measuring pure-tone thresholds under computer control. J Am Acad Audiol. 1996; 7:125–129. [PubMed: 8652865]

Franks, J. Occupational exposure to noise: evaluation, prevention and control. Dortmund: World Health Organization; 2001. Hearing measurement; p. 183-232.

Gosztonyi RE Jr, Vassallo LA, Sataloff J. Audiometric reliability in industry. Archives of Environmental Health: An International Journal. 1971; 22:113–118.

Green DM. A maximum-likelihood method for estimating thresholds in a yes–no task. The Journal of the Acoustical Society of America. 1992; 93:2096–2105. [PubMed: 8473622]

Guest D, Kent C, Adelman JS. Why additional presentations help identify a stimulus. J Exp Psychol Hum Percept Perform. 2010; 36:1609–1630. [PubMed: 20919780]

Hall JL. Hybrid adaptive procedure for estimation of psychometric functions. The Journal of the Acoustical Society of America. 1981; 69:1763–1769. [PubMed: 7240589]

Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning. Springer; 2009.

Ho ATP, Hildreth AJ, Lindsey L. Computer-assisted audiometry versus manual audiometry. Otology & Neurotology. 2009; 30:876–883. [PubMed: 20179426]

Houlsby N, Huszár F, Ghahramani Z, et al. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745. 2011

Hughson W, Westlake H. Manual for program outline for rehabilitation of aural casualties both military and civilian. Trans Am Acad Ophthalmol Otolaryngol. 1944; 48:1–15.

Ishak WS, Zhao F, Stephens D, et al. Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry. Audiological Medicine. 2011; 9:40–46.

Jerger J. Bekesy audiometry in analysis of auditory disorders. Journal of Speech, Language, and Hearing Research. 1960; 3:275–287.

Katz, J.; Medwetsky, L.; Burkhard, R., et al. Handbook of Clinical Audiology. 6th. Lippincott Williams & Wilkins; 2009.

Kidd G Jr, Mason CR, Richards VM. Multiple bursts, multiple looks, and stream coherence in the release from informational masking. J Acoust Soc Am. 2003; 114:2835–2845. [PubMed: 14650018]

King-Smith PE, Grigsby SS, Vingrys AJ, et al. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. Vision research. 1994; 34:885–912. [PubMed: 8160402]

Kontsevich LL, Tyler CW. Bayesian adaptive estimation of psychometric slope and threshold. Vision Res. 1999; 39:2729–2737. [PubMed: 10492833]

Kujala, JV. Descriptive and normative approaches to human behavior. World Scientific; 2011. Bayesian adaptive estimation: a theoretical review; p. 123-159.

Leek MR. Adaptive procedures in psychophysical research. Percept Psychophys. 2001; 63:1279–1292. [PubMed: 11800457]

Leek MR, Dubno JR, He NJ, et al. Experience with a yes–no single-interval maximum-likelihood procedure. The Journal of the Acoustical Society of America. 2000; 107:2674–2684. [PubMed: 10830389]

Leibold LJ, Bonino AY. Release from informational masking in children: effect of multiple signal bursts. J Acoust Soc Am. 2009; 125:2200–2208. [PubMed: 19354396]

Lesmes LL, Jeon ST, Lu ZL, et al. Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. Vision research. 2006; 46:3160–3176. [PubMed: 16782167]

Levitt H. Transformed up-down methods in psychoacoustics. The Journal of the Acoustical Society of America. 1971; 49:467–477. [PubMed: 5541744]

Lewis, DD.; Catlett, J. Heterogeneous uncertainty sampling for supervised learning; Proceedings of the eleventh international conference on machine learning; 1994. p. 148-156.

Lewis, DD.; Gale, WA. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. Springer-Verlag New York, Inc.; 1994. A sequential algorithm for training text classifiers; p. 3-12.

Mahomed F, Eikelboom RH, Soer M. Validity of automated threshold audiometry: A systematic review and meta-analysis. Ear and hearing. 2013; 34:745–752. [PubMed: 24165302]

Margolis RH, Glasberg BR, Creeke S, et al. AMTAS: Automated method for testing auditory sensitivity: Validation studies. International journal of audiology. 2010; 49:185–194. [PubMed: 20109081]

Mathias SR, Micheyl C, Bailey PJ. Stimulus uncertainty and insensitivity to pitch-change direction. J Acoust Soc Am. 2010; 127:3026–3037. [PubMed: 21117752]

Meyer-Bisch C. Audioscan: a high-definition audiometry technique based on constant-level frequency sweeps-A new method with new hearing indicators. International journal of audiology. 1996; 35:63–72.

Mori S, Ward LM. Intensity and frequency resolution: masking of absolute identification and fixed and roving discrimination. J Acoust Soc Am. 1992; 91:246–255. [PubMed: 1737875]

Murphy, KP. Machine learning: a probabilistic perspective. MIT press; 2012.

Özdamar Ö, Eilers RE, Miskiel E, et al. Classification of audiograms by sequential testing using a dynamic Bayesian procedure. The Journal of the Acoustical Society of America. 1990; 88:2171–2179. [PubMed: 2269733]

Pentland A. Maximum likelihood estimation: The best PEST. Perception & Psychophysics. 1980; 28:377–379. [PubMed: 7465322]

Rasmussen, CE.; Williams, CKI. Gaussian Processes for Machine Learning. Cambridge, MA: The MIT Press; 2006.

Remus JJ, Collins LM. Comparison of adaptive psychometric procedures motivated by the Theory of Optimal Experiments: Simulated and experimental results. The Journal of the Acoustical Society of America. 2008; 123:315–326. [PubMed: 18177161]

Roy N, McCallum A. Toward optimal active learning through monte carlo estimation of error reduction. ICML, Williamstown. 2001

Schmuziger N, Probst R, Smurzynski J. Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones. Ear and hearing. 2004; 25:127–132. [PubMed: 15064657]

Settles, B. Computer Sciences Technical Report. Madison: University of Wisconsin; 2009. Active learning literature survey.

Shen Y, Richards VM. Bayesian adaptive estimation of the auditory filter. The Journal of the Acoustical Society of America. 2013; 134:1134–1145. [PubMed: 23927113]

Shen Y, Sivakumar R, Richards VM. Rapid estimation of high-parameter auditory-filter shapes. The Journal of the Acoustical Society of America. 2014; 136:1857–1868. [PubMed: 25324086]

Stach, B. Clinical audiology: an introduction. Delmar: Cengage Learning; 2008.

Swanepoel DW, Mngemane S, Molemong S, et al. Hearing assessment—reliability, accuracy, and efficiency of automated audiometry. TELEMEDICINE and e-HEALTH. 2010; 16:557–563. [PubMed: 20575723]

Taylor MM, Creelman CD. PEST: Efficient estimates on probability functions. The Journal of the Acoustical Society of America. 1967; 41:782–787.

Vogel DA, McCarthy PA, Bratt G, et al. The clinical audiogram: its history and current use. Commun Disord Rev. 2007; 1:81–94.

Ward LM. Informational and neural adaptation curves are asynchronous. Percept Psychophys. 1991; 50:117–128. [PubMed: 1658723]

Watson AB, Pelli DG. QUEST: A Bayesian adaptive psychometric method. Perception & psychophysics. 1983; 33:113–120. [PubMed: 6844102]

Zhao F, Stephens D, Meyer-Bisch C. The audioscan: a high frequency resolution audiometric technique and its clinical applications. Clinical Otolaryngology & Allied Sciences. 2002; 27:4–10. [PubMed: 11903364]

Zhao F, Stephens SDG, Ishak WS, et al. The characteristics of Audioscan and DPOAE measures in tinnitus patients with normal hearing thresholds. International journal of audiology. 2014; 53:309–317. [PubMed: 24495275]

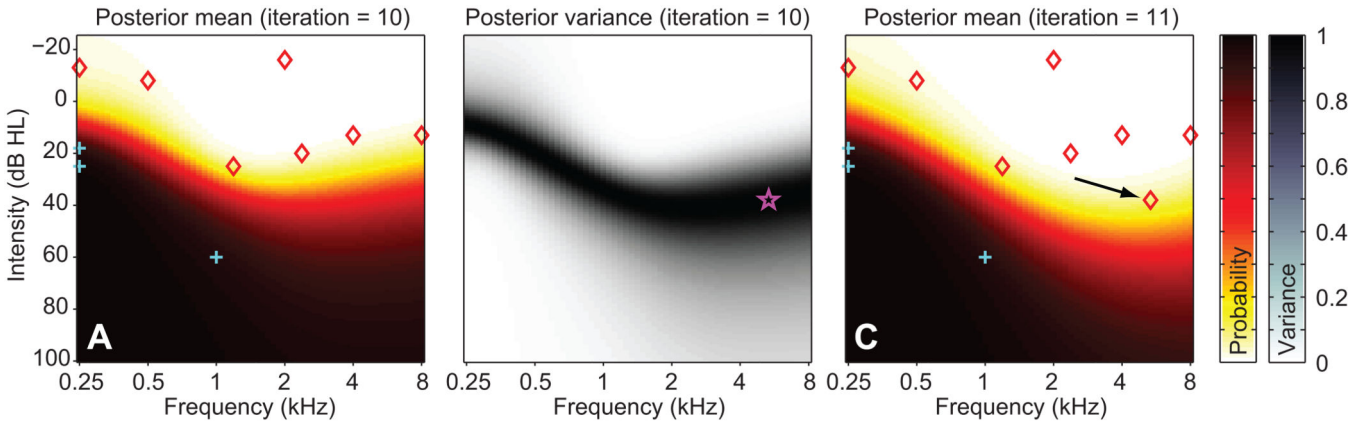

Figure 1.

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

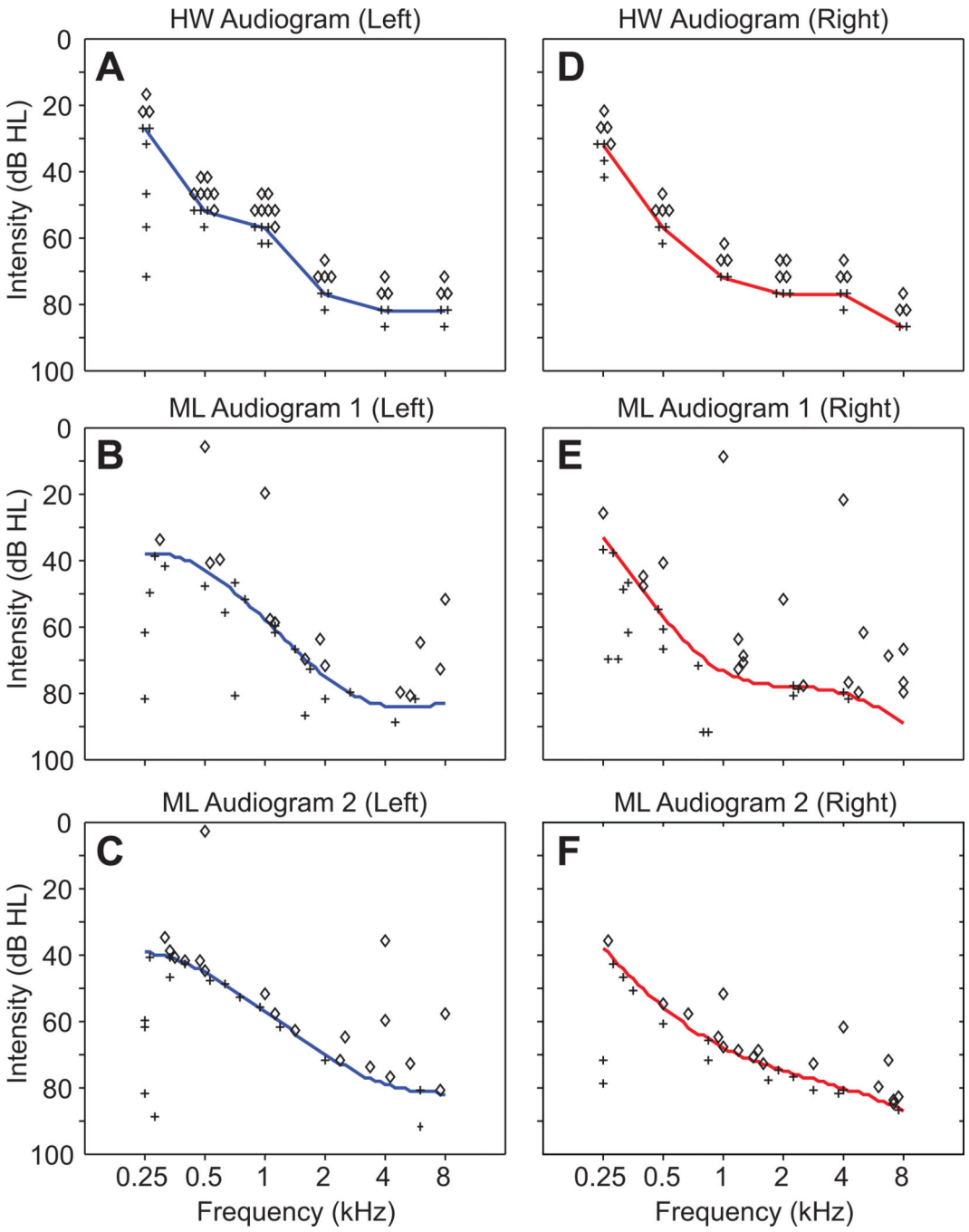

Figure 2.

Sample plots of left- and right-ear audiograms obtained and samples conducted for a representative listener (Listener 4) using the manual HW technique (A, D), the first run of the ML algorithm (B, E), and the second run of the ML algorithm (C, F). Marks represent the frequencies and intensities of the stimuli that were presented, with pluses denoting listener detections and diamonds denoting misses. The superimposed curves are the final audiogram estimates produced by each technique. Note that the small displacements along

the frequency axis *in (A) and (B) only* are to make repeat stimuli more visible and do not reflect actual deviations in the frequency of presented tones.

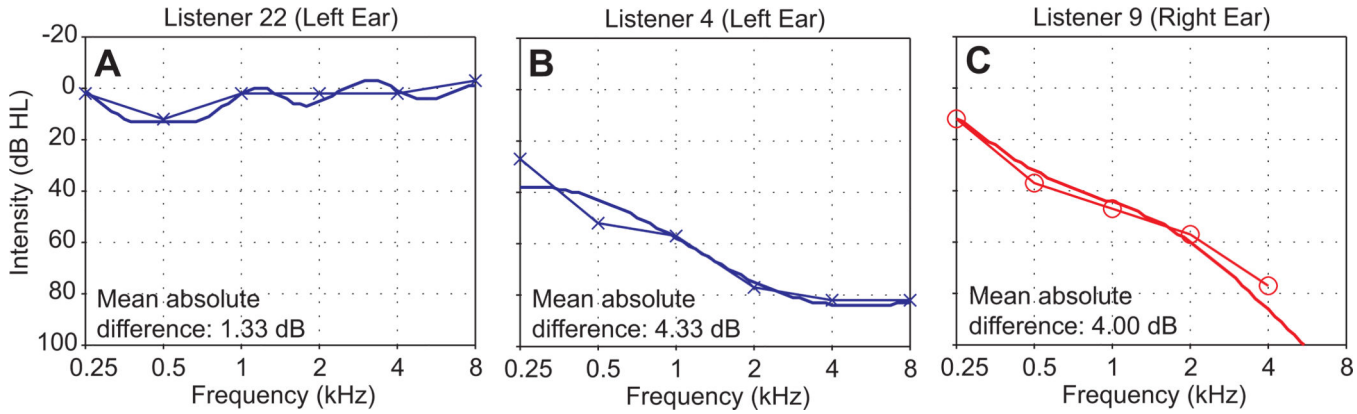

Figure 3.

Sample plots of ML audiogram results for (**A**) an ear with relatively normal hearing; (**B**) an ear with sloping high-frequency hearing loss; and (**C**) an ear with a no-response at 8000 Hz. "X" and "O" marks denote values estimated from the manual HW audiogram (connected by straight lines). The superimposed curves show the results from the automated ML audiogram.

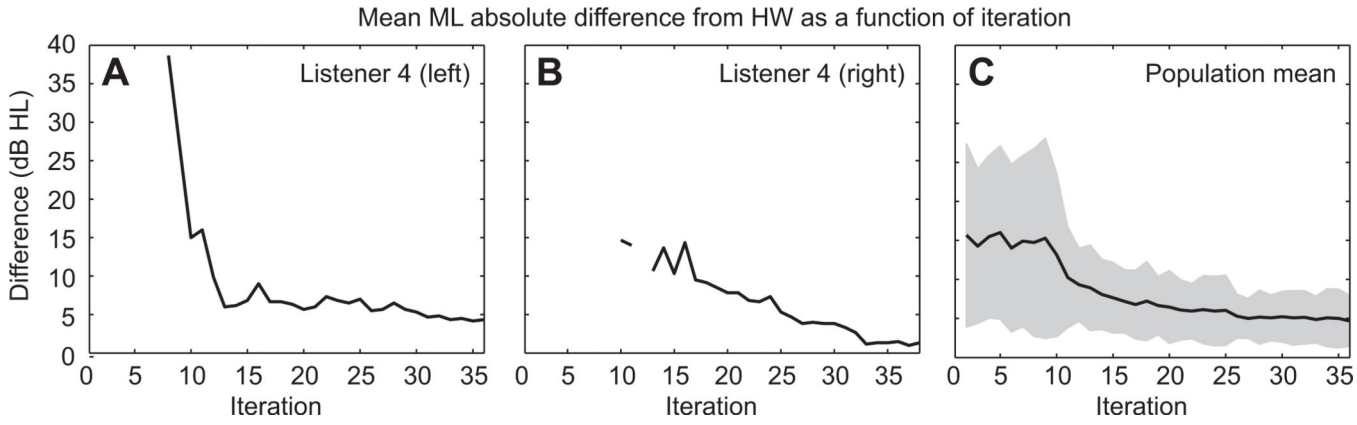

Figure 4.

Cumulative agreement between automated ML and manual HW audiograms as a function of GP algorithm iteration. Mean absolute difference was calculated by obtaining the current ML estimate of the threshold audiogram at each iteration during one run, then calculating the absolute difference between that estimate and the HW threshold audiogram, averaged across all 6 audiogram frequencies. (**A**) and (**B**) show examples for two ears (Listener 4, the same listener as in Figure 2), and (**C**) shows this trend averaged across all runs where the ML audiograms terminated at 36 iterations (53 of 80 runs). Blank areas denote points at which the ML procedure did not produce a posterior mean with a clear boundary, so error could not be assessed (but in practice is very high). Gray shading on (C) indicates ±1 standard deviation from the mean.

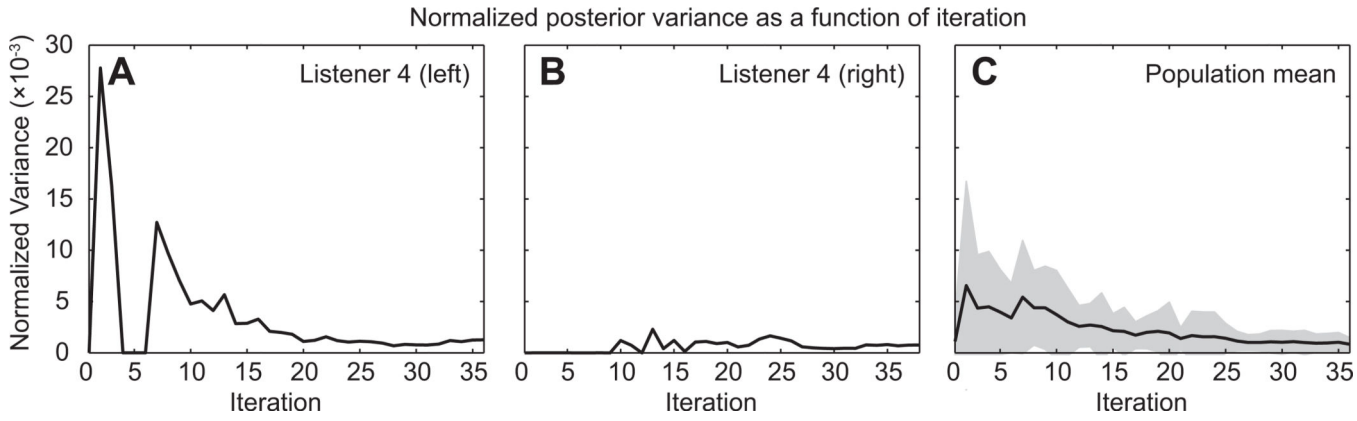

Figure 5.

Normalized posterior variance as a function of algorithm iteration. Normalized posterior variance was calculated by dividing the sum all values in the variance function at each iteration by the total size of the variance function matrix. (**A**) and (**B**) show examples for two ears (Listener 4, the same listener as in Figure 2), and (**C**) shows this trend averaged across all listeners whose ML audiograms terminated at 36 iterations. Gray shading on (C) indicates ±1 standard deviation from the mean.

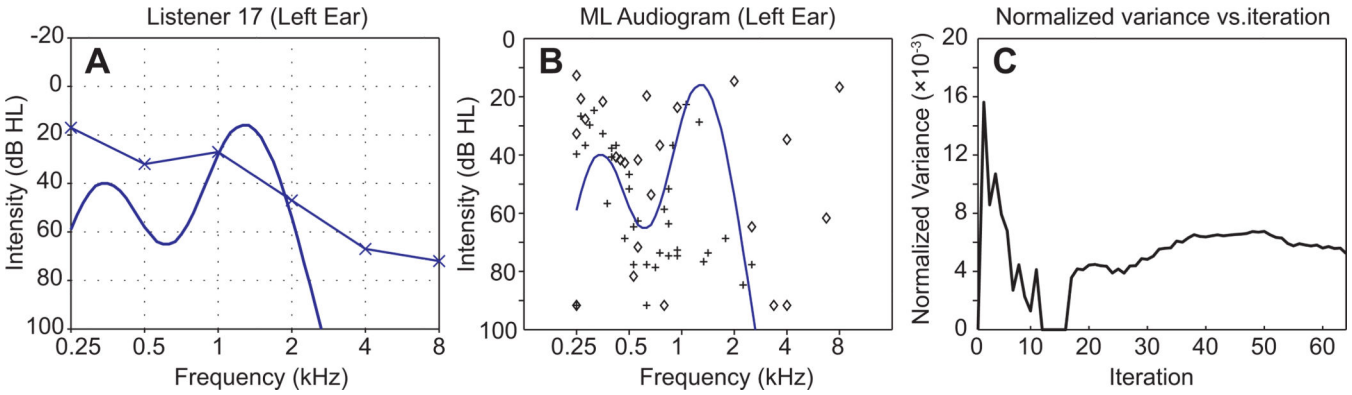

Figure 6.

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

Table 1

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

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

Table 2

Differences between the ML audiogram estimate and the HW estimate.

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

Table 3

Test-retest reliability of ML audiogram.

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