Published in final edited form as:

Ear Hear. 2019 ; 40(4): 918–926. doi:10.1097/AUD.0000000000000669.

Online Machine Learning Audiometry

Dennis L. Barbour¹, Rebecca T. Howard^{1,2}, Xinyu D. Song¹, Nikki Metzger¹, Kiron A. Sukesan^{1,3}, James C. DiLorenzo^{1,3}, Braham R. D. Snyder¹, Jeff Y. Chen¹, Eleanor A. Degen¹, Jenna M. Buchbinder^{1,2}, and Katherine L. Heisey¹

¹Laboratory of Sensory Neuroscience and Neuroengineering, Department of Biomedical Engineering, Washington University in St. Louis, 1 Brookings Drive, Campus Box 1097, St. Louis, Missouri 63130, U.S.A.

²Program in Audiology and Communication Sciences, Washington University School of Medicine, 660 S. Euclid Avenue, Campus Box 8042, St. Louis, MO 63110

³Department of Computer Science and Engineering, Washington University in St. Louis, 1 Brookings Drive, Campus Box 1045, St. Louis, Missouri 63130, U.S.A.

Abstract

Objectives.—A confluence of recent developments in cloud computing, real-time web audio and machine learning psychometric function estimation has made wide dissemination of sophisticated turn-key audiometric assessments possible. The authors have combined these capabilities into an online (i.e., web-based) pure-tone audiogram estimator intended to empower researchers and clinicians with advanced hearing tests without the need for custom programming. The objective of this study is to assess the accuracy and reliability of this new online machine learning audiogram method relative to a commonly used hearing threshold estimation technique also implemented online for the first time in the same platform.

Design.—The authors performed air-conduction pure-tone audiometry on 21 participants between the ages of 19 and 79 years (mean 41, standard deviation 21) exhibiting a wide range of hearing abilities. For each ear, two repetitions of online machine learning audiogram estimation and two repetitions of online modified Hughson-Westlake ascending-descending audiogram estimation were acquired by an audiologist using the online software tools. 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 estimation methods delivered very similar threshold estimates at standard audiogram frequencies. Specifically, the mean absolute difference between threshold estimates was 3.24 ± 5.15 dB. The mean absolute differences between repeated measurements of the online machine learning procedure and between repeated measurements of the Hughson-Westlake procedure were 2.85 ± 6.57 dB and 1.88 ± 3.56 respectively. The machine learning method generated estimates of both threshold and spread (i.e., the inverse of psychometric slope)

continuously across the entire frequency range tested from fewer samples on average than the modified Hughson-Westlake procedure required to estimate 6 discrete thresholds.

Conclusions.—Online machine learning audiogram estimation in its current form provides all the information of conventional threshold audiometry with similar accuracy and reliability in less time. More importantly, however, this method provides additional audiogram details not provided by other methods. This standardized platform can be readily extended to bone conduction, masking, spectrotemporal modulation, speech perception, etc., unifying audiometric testing into a single comprehensive procedure efficient enough to become part of the standard audiologic workup.

Introduction

The most common methodology for measuring hearing thresholds in audiologic patients and research participants is a variant of the method of limits originally introduced by Fechner (Fechner, 1860). This modified Hughson-Westlake audiogram (HWAG) procedure (Carhart & Jerger, 1959; Hughson & Westlake, 1944) is incapable of inferring across frequencies; therefore, it proceeds one frequency at a time. Pure tones are manually delivered at each new frequency and at a sequence of ascending and descending sound levels determined by previous participant responses at that frequency. The test terminates and a threshold is determined at each frequency once a sufficient number of reversals has been achieved (American National Standards Institute, 2004; American Speech-Language-Hearing Association; Franks).

Computerized methods of obtaining auditory thresholds typically follow similar procedures and deliver comparable estimates to HWAG, with absolute differences averaging $4.2 \text{ dB} \pm 5.0 \text{ dB}$ (Mahomed, Eikelboom, & Soer, 2013). These same computerized methods exhibited absolute test-retest differences averaging $2.9 \pm 3.8 \text{ dB}$ compared to $3.2 \pm 3.9 \text{ dB}$ for manual estimation methods in the same participant populations. Computerized pure-tone audiometry procedures can therefore yield hearing thresholds comparable in value and test-retest reliability to conventional manual threshold-estimation procedures.

A new machine learning audiogram (MLAG) technique based on probabilistic classification and active sampling can deliver a fully predictive hearing function estimate in significantly less time than the HWAG method can deliver a coarse estimate of threshold alone (Song, Garnett, & Barbour, 2017; Song, Sukesan, & Barbour, 2018; Song et al., 2015). While this method uses sophisticated computational techniques, the training input into the algorithm is simply the participant responses regarding tone detection, and the output is simply the tone parameters. This thin data transfer requirement makes MLAG well suited for separating the machine-learning computations into a back-end server implementation and the user interface into a front-end client implementation. Recent developments in real-time web audio (W3C, 2015) provide the opportunity to deliver machine learning audiometry, which can achieve arbitrary frequency and sound level resolution, online over the internet via a web browser interface. Online MLAG has been implemented in this fashion and is evaluated here for its ability to yield threshold estimates and test-retest reliability similar to online HWAG implemented on the same platform.

Materials & Methods

Stimulus synthesis and delivery

Pure-tone stimuli for this study were all generated in real-time using the WebAudio library (W3C, 2015) implemented within the Bonauria online audiometry platform. Pure tones were generated by a web browser on the client computer from stimulus parameters determined by the server computer running the active machine learning software.

Online machine learning algorithm

Details of this method are provided elsewhere (Song et al., 2015). Briefly, a participant's fully predictive unilateral air-conduction audiogram, defined here as the probability of tone detection for one ear as a function of frequency and sound level, is modeled as a Gaussian process (GP) observed through a sigmoidal link function to generate output values in the range [0,1]. This formulation is an implementation of a probabilistic classifier, where the estimator attempts to discern the boundary between categorical output states (Song et al., 2017; Song et al., 2018). In the case of pure-tone audiometry, Gaussian process classification (GPC) attempts to localize the boundary between undetected and detected tones as a function of frequency and sound level. This boundary is equivalent to the threshold audiogram. A probabilistic classifier delivers not only its estimate of the class boundary, but also a transition zone whose width (or spread) is inversely proportional to the confidence of that class boundary's location. Probabilistic machine-learning classification has been shown to be equivalent to classical psychometric curve estimation, though considerably more efficient under conditions of multiple input dimensions (Song et al., 2017).

GPC is a nonparametric Bayesian estimation method, where the output of the estimator is a posterior probability distribution at each point in the input domain of frequency and sound level. In this study, the posterior mean is used for comparison to Hughson-Westlake thresholds. A powerful feature of GPs is that they are completely characterized by their first two moment functions, a mean function (not to be confused with the posterior mean) and a covariance function. The covariance function is modeled as a linear function in sound level and a squared exponential function in frequency. When combined with a sigmoidal probit link function, these constraints model the audiogram as monotonically increasing in sound level and continuously smooth in frequency.

The probabilistic nature of GPC also provides information about the certainty of the posterior at any combination of frequency and sound level. Given participant responses from a particular set of previous tones delivered, the next best tone to deliver can then be optimally determined according to some criterion. After an initial random tone delivery, each new tone in this study was selected according to the principle of Bayesian active learning by disagreement (BALD) (Garnett, Osborne, & Hennig, 2013; Houlisby, Huszár, Ghahramani, & Lengyel, 2011). BALD ensures that each new tone delivered is the best one to differentiate among competing hypotheses of what the most accurate model structure should be.

Online Hughson Westlake Algorithm

Online HWAG complies with the ANSI standard for conventional threshold audiometry (American National Standards Institute, 2004; American Speech-Language-Hearing Association). Briefly, tone presentation begins at 1 kHz and 60 dB HL. Detected tones result in a 10 dB sound level decrement in the following tone. Undetected tones result in a 5 dB increment in the following tone. Tone presentation proceeds until at least 3 reversals in direction (i.e., from ascending to descending levels or from descending to ascending levels) have occurred and terminates when a majority of ascending queries result in a detection at a particular sound level. The threshold returned at the end of this procedure is the lowest sound level at which a majority of tone deliveries is detected. Estimation proceeds by octaves to 8 kHz, then by octaves from 250 Hz. Altogether, 6 thresholds are estimated at 6 frequencies.

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 19 and 79 years of age (mean 41, standard deviation 21), fluent English speakers and with no history of neurological disorder. Approval for completion of the study was received from Washington University in St. Louis's Human Research Protection Office (HRPO), and all participants provided informed consent before any testing protocol began. Hearing ability in these participants ranged from normal to profound loss.

Experimental procedure

Separately for each ear of each participant, 2 repetitions of online MLAG and 2 repetitions of online HWAG were conducted. Each acoustic stimulus consisted of a three-pulse sequence of 200-ms pure tones with inter-pulse intervals of 200 ms. Participants were seated within a sound isolation booth, and all auditory stimuli were delivered using a Dell XPS 13 5156 laptop computer running custom MatLab code and Etymotic Research 3A insert earphones connected to the native computer RealTek sound card operating at 16 bits and 44.1 ksamples/s. Computer audio output was calibrated to match the output of a GSI-61 two-channel clinical audiometer. The relative order for the online MLAG and online HWAG tests was randomized for each listener. Listeners were asked to remove any hearing-assist devices prior to data collection. Otoscopy was performed on all participants prior to testing. Each participant's right and left ears were examined for normal pinnae, external ear canals and tympanic membranes. Occluding cerumen was removed via handheld curette. The presence of atresia, otitis media, otitis externa, or other physical abnormalities or conditions were exclusion criteria for this study. Short periods of rest (~2 mins) were administered between each set of audiogram runs.

Online HWAG: A computerized version of conventional threshold audiometry programmed according to accepted standards as described above was initiated from the web site and then supervised unaltered by an audiologist. No intervention by the audiologist was required because the online HWAG procedure was completely automated. Each listener was

instructed to press a button upon detection of a presented pure-tone sequence of 3 identical tones as described above. Each tone sequence was separated from other stimuli by a randomized inter-trial interval of between 3 and 8 seconds to reduce listener prediction of stimulus presentation times. A response within 2000 ms following the onset of the tone sequence was logged as a detected sample; no response within 2000 ms was logged as an undetected sample. Hearing ability was assessed at standard audiogram frequencies presented in the order (1, 2, 4, 8, 0.25, 0.5 kHz), with the possible sound level ranging from -25 to 100 dB HL in a minimum of 5-dB increments. The conventional threshold audiogram was conducted separately for right and left ears.

Online MLAG: As with online HWAG, the online MLAG implementation asked listeners to respond with a button press upon detection of a three-pulse sequence of 200-ms pure tones. Each stimulus was identical to the HWAG tone sequence design, as were the inter-trial and response intervals. The range of possible sample points fell within 250 to 8000 Hz in semitone increments including 1000 Hz along the frequency dimension, and within -25 to 100 dB HL in 1-dB increments including 0 dB HL along the sound level dimension. The first sample was randomly delivered near 0 dB HL and between 1 and 2 kHz. After the response was recorded to this initial sample point, the algorithm followed the iteration cycle of hyperparameter learning (i.e., covariance function tuning), posterior estimation and BALD sampling of next stimulus as previously described (Song et al., 2018). This cycle was iterated for an overall total of 49 presentations. “Detected” responses for which no tone presentations occurred within 2000 ms (i.e., false detections) were not used in evaluating MLAG or in learning the hyperparameters. The automated audiogram was conducted separately for right and left ears. To accommodate user comfort, delivered tone sound levels never exceeded 10 dB higher than the maximum sound level delivered up to that point in the test or 60 dB HL, whichever was greater. This rule combined with the acquisition function and initial sampling strategy ensured that only 1.2% of the more than 4000 MLAG tones in this study were delivered more than 20 dB greater than the eventually estimated threshold value at each frequency and 0% of the tones were delivered more than 40 dB greater.

A flow chart of the online MLAG procedure is depicted in Figure 1.

Analysis

Following completion of online MLAG data collection, each GP posterior mean was contoured at a detection probability of 0.707, the standard probability of a detection at convergence for a transformed 2-up, 1-down method like HWAG (Levitt, 1971). These threshold values therefore become a continuous (in frequency) estimate of the listener’s threshold audiogram.

HWAG and MLAG threshold audiograms were compared at the standard audiogram frequencies because HWAG generates no direct threshold estimates at other frequencies. Accuracy of the MLAG procedure was assessed via comparison to the results of HWAG by calculating 1) the mean signed difference (MLAG – HWAG) and standard deviation; 2) the mean absolute difference and standard deviation; and 3) the root mean square difference between the estimated MLAG and HWAG audiogram thresholds (Mahomed et al., 2013;

Swanepoel, Mngemane, Molemong, Mkwanazi, & Tutshini, 2010). Test-retest reliabilities of MLAG and HWAG were assessed by calculating 1) the mean signed difference (second – first) and standard deviation; 2) the mean absolute difference and standard deviation; and 3) the root mean square difference between the estimated thresholds produced by the 2 runs of each algorithm (Mahomed et al., 2013).

Results

A screen shot of the web page control panel for initiating a new test can be seen in Figure 2A. The clinician or researcher conducting the test has control over multiple testing parameters, and these options periodically grow as new features are added to the website. Plots of online HWAG and online MLAG threshold estimates along with tones delivered and participant responses for one ear can be seen in Figure 2B,C. Notable is the dispersion in frequency of the tones delivered for MLAG estimation relative to HWAG. It is this dispersion that enables MLAG to provide continuous estimates of hearing threshold as a function of frequency.

Active sampling for MLAG provides the efficiency gains delivered by this novel estimation method (Song et al., 2018). An example of the active sampling procedure can be seen in Figure 3. After 5 tones have been delivered, the posterior probability surface corresponding to the best estimate of hearing given prior beliefs about the hearing function (uninformative in this case) coupled with the participant's responses can be seen in Figure 3A. Note that this surface estimates the probability that this participant can hear a tone with any combination of frequency and sound level within the input domain evaluated. Estimator assumptions are fairly weak in these experiments, which is why this surface does not resemble known hearing functions with such few samples. It is a feature of MLAG that either weak or strong assumptions can be attached to the method, placing more or less emphasis on the observations vs. the prior beliefs when forming an estimate.

Each subsequent tone is selected by the acquisition function, which quantifies the uncertainty regarding the probability surface (Figure 3B). The most uncertainty is present at frequencies where few observations have been made and at frequencies where observations of only one type (i.e., either detected or undetected) have been made. More sophisticated tone-selection criteria could be formulated, but the current strategy is to sample where uncertainty would be reduced the most with a new observation, indicated by the pink star. The participant detected this new tone (Figure 3C), which results in revision of the probability surface at all nearby frequencies. Active sampling allows a rough hearing function to be estimated rapidly, with greater refinement upon accumulation of more observations.

Improvement in threshold estimation decelerates with many more tone delivery iterations, however. Figure 3D shows the probability surface formed after 40 tones, and Figure 3E shows where the maximum of the acquisition function lies. The participant detected the tone delivered at this point, which updated the estimate in the local vicinity, though much less substantially than for earlier tones. Later MLAG tone delivery turns out to be more valuable

for estimating psychometric spread (i.e., the extent of the transition from undetected to detected as level is increased) than threshold (Song et al., 2018).

Each online MLAG run was terminated after delivering exactly 49 tone presentations. The median number of online HWAG tone presentations across all participants was 67. Online MLAG can be programmed to terminate upon reaching a preset criterion, as was done previously (Song et al., 2015). Because MLAG computes the best stimulus to present given all the stimuli that have already been delivered, analyzing estimate quality for the first n stimuli provides meaningful information about general estimator performance with that number of stimuli.

Both testing methods were evaluated under identical conditions in order to ascertain their comparability in estimating hearing thresholds. Table 1 shows the results of evaluating the similarity of online MLAG threshold estimates at standard audiogram frequencies to online HWAG averaged across all listeners and estimation runs. For the 6 standard audiogram frequencies, the mean signed estimated threshold difference was -0.969 ± 6.02 dB, the mean absolute estimated threshold difference was 3.24 ± 5.15 dB, and the root mean square estimated threshold difference was 5.58 dB. These values compare favorably with historical differences in audiogram estimation methodologies (Gosztonyi Jr., Vassallo, & Sataloff, 1971; Ishak et al., 2011; Mahomed et al., 2013; Schmuziger, Probst, & Smurzynski, 2004) as well as previous offline MLAG procedures (Song et al., 2015), indicating that both HWAG and MLAG appear to be measuring similar quantities. This observation is supported by the signed difference 95% confidence interval of $[-9, 8]$, which includes 0.

It should be noted that while HWAG is a standardized procedure and is implemented consistently from study to study, MLAG is still a method in active development. As a result, variability between the current MLAG implementation and previous MLAG implementations may include design modifications as well as traditional sources of experimental variability, such as natural variation across different study populations.

An informative test would be expected to provide similar estimates when the underlying system is not changing, such as repeated tests closely spaced together in time. Tables 2 and 3 show the results of evaluating the test-retest reliability of online MLAG and online HWAG, respectively, at standard audiogram frequencies averaged across all listeners and estimation runs. For the 6 standard audiogram frequencies, the mean signed difference (second minus first) between the online MLAG runs was -0.486 ± 7.15 dB, the mean absolute difference was 2.85 ± 6.57 dB, and the root mean square difference was 6.32 dB. Similarly, for the 6 standard audiogram frequencies the mean signed difference (second minus first) between the online HWAG runs was -0.208 ± 4.02 dB, the mean absolute difference was 1.88 ± 3.56 dB, and the root mean square difference was 3.69 dB. These values imply somewhat less reliability for online MLAG compared to online HWAG. Nevertheless, both the HWAG signed difference 95% confidence interval of $[-5, 5]$ and the MLAG signed difference 95% confidence interval of $[-7.8, 5]$ imply that both methods are consistently estimating similar thresholds upon repeat testing.

More detail about the behavior of online HWAG and MLAG can be observed by examining individual threshold audiogram estimates. Figure 4 includes estimates generated from 2 repeats of HWAG and 2 repeats of MLAG for all 42 ears of this study sorted by pure-tone average. The major observable trend is the high concordance between HWAG and MLAG estimates for most ears, as reflected in Table 1. The next most obvious trend is the high reliability of both methods upon repeat testing, summarized in Table 2 and Table 3. Some interesting exceptions are also notable. In particular, the largest deviations between test type and between repeats of MLAG occur mostly at the highest frequencies and occasionally at the lowest frequencies. It has been observed previously that MLAG estimation tends to be the most variable at the highest and lowest frequencies, most likely because of its inability to sample on both sides of the edge frequencies to improve estimates (Song et al., 2015). The total tone count for this study was capped at 49 tones per ear, which had the consequence of leading to occasional undersampling of the highest or lowest frequencies. In about 10% of the ears, 49 tones were apparently too few for the current version of MLAG to achieve reliable estimates at extreme frequencies. Effective methods to circumvent this issue include rigorous but variable termination criteria, as has been employed in previous studies, sampling beyond the edge frequencies, and density-based acquisition criteria. Undersampling, should it occur in future ears, is readily detectable and therefore useful as a quality metric for the algorithm or audiologist to either update the estimate or reject it altogether.

Because MLAG produces threshold estimates across all frequencies at any stage of acquisition, its estimate quality can be compared against the final HWAG estimate continuously as testing progresses. Figure 5 shows the similarity of the online MLAG estimate to the final online HWAG estimate as a function of algorithm iteration (i.e., the number of tones delivered). This post-hoc analysis was performed by constructing an online MLAG threshold audiogram estimate from the posterior distribution after each iteration of the GPC algorithm and then evaluating the absolute difference from the final online HWAG threshold audiogram at the six standard audiogram frequencies. As found previously for offline MLAG, online MLAG estimates tend to achieve near-final absolute difference values after about 20 samples on average (Gardner, Song, Cunningham, Barbour, & Weinberger, 2015; Song et al., 2018). Additional online MLAG tones beyond 20 do continue to improve threshold estimates for all frequencies, though at a considerably lower rate than the earlier tones. As noted previously, most of the MLAG estimation errors after delivery of the full tone complement can be traced to undersampling near edge frequencies for a few outlier participants, which is resolvable by the presentation of additional informative tones for those participants.

Discussion

Online machine learning audiometry represents a novel approach for testing hearing function. The method described here consists of a nonparametric Bayesian estimator of tone detection probability as a function of frequency and sound level. Included in this complete audiogram estimate are a continuous threshold audiogram estimate and a continuous psychometric spread estimate as a function of frequency. Each new tone delivered to a participant is selected via an active sampling procedure to provide the most information

given the previous tones delivered and any prior beliefs about the audiogram. The result is an estimator that delivers a near-optimal audiogram estimate for any given number of sample tones. In this and all MLAG studies to date, only uninformative prior beliefs have been employed in order to quantify how this Bayesian estimator performs naively with the least amount of prior information possible.

Accuracy of online MLAG was assessed by comparing against online HWAG administered on the same web platform. HWAG was explicitly designed to adapt tone delivery to each participant and to provide reliable estimates of hearing threshold at discrete frequencies (Carhart & Jerger, 1959; Hughson & Westlake, 1944). It does so by delivering multiple repeated tones at each frequency of evaluation—over 10 tones per frequency on average in this particular study. The current version of MLAG, on the other hand, was not explicitly designed for reliability. In fact, not only were multiple MLAG tones unlikely to have been delivered at any of the octave HWAG frequencies, but typically no tones at all were delivered at those exact frequencies. The evaluation as it was structured in this study was therefore intended to be most favorable to HWAG in order to provide a lower bound on estimated MLAG accuracy, reliability and efficiency given uninformative prior beliefs.

The mean signed difference between threshold estimates across all frequencies was less than 1 dB, just as with the offline MLAG compared against a manually delivered HWAG (Song et al., 2015). This result provides further confidence that both methods are estimating similar quantities and are therefore either unbiased or biased in the same ways. The signed threshold difference data revealed considerably more variability at 8 kHz (the highest frequency tested) than the other frequencies. The second greatest variability can be seen at 250 Hz, the lowest frequency tested. This increased variability arises from a few participants and from both HWAG and MLAG test results, as can be seen in Figure 4.

The mean absolute difference between threshold estimates across all frequencies was 3.24 dB. This value compares favorably to the average correspondence of 4.16 dB between offline MLAG and manual HWAG (Song et al., 2015) and also compares favorably to the average correspondence of 4.2 dB between manual and automated audiometry generally (Mahomed et al., 2013). These results provide further evidence that online MLAG and HWAG are estimating similar quantities.

Reliability assessment indicates that online HWAG may be somewhat more reliable at octave threshold estimates than the current online MLAG version, at 1.88 dB mean absolute difference between repeat tests for HWAG versus 2.85 dB for MLAG. The latter value is an improvement over first-generation MLAG reliability, estimated to be 4.51 dB (Song et al., 2015). Furthermore, these test-retest values also compare favorably to the average reliabilities of 3.2 dB and 2.9 dB for manual and automated audiometry, respectively, observed with a variety of estimation techniques (Mahomed et al., 2013). Both the online methods evaluated here therefore exhibit reliability within the range expected for practical threshold audiogram estimators.

The greater variability in repeat threshold estimates at 8 kHz for both test types in this cohort can be seen in Tables 2 and 3. Because higher variation appears in both MLAG and HWAG

measures and did not appear to this extent in previous studies of MLAG, the possibility exists that at least part of its source resides within the study population, instructions to participants, stimulus delivery system, or some similar considerations. The extra variability is confined to a few participants, making the possibility of a systematic issue less likely. Regardless, this unexplained variance will be a topic of future research.

The greater variability at high and low edge frequencies for MLAG can be seen in Tables 1 and 2. This variability arises because MLAG uses observations at adjacent frequencies to improve estimates, but at either edge of the frequency domain such observations are limited. Probing at frequencies lower than the lowest target frequency and higher than the highest target frequency by about one-third or one-half an octave successfully mitigates these edge effects, as does de-tuning acquisition samples slightly from absolute optimality (Song et al., 2017; Song et al., 2018). These adjustments were not performed in the present study both to allow a clearer comparison to previous research on offline MLAG where such adjustments were not made (Song et al., 2015) and to provide a clear lower bound on MLAG performance as a baseline against which to compare additional procedural advances added in future versions to improve estimation quality. Allowing acquisition to continue by delivering additional tones would also mitigate these effects.

The present implementation of the online MLAG algorithm assumes uninformative prior beliefs, meaning that no probabilistically useful information about a participant is incorporated into the estimator before acquiring the first sample from that participant. The audiogram is only assumed to be continuous and smooth in the frequency dimension and monotonic in the sound level dimension. MLAG has the capability of incorporating more meaningful prior beliefs, such as the distribution of audiograms in the human population or in past measurements of a particular participant. Indeed, some of the best prior beliefs may be the audiogram of the contralateral ear or of the most recent audiogram on record (Barbour et al., 2018). In the present study, however, initial beliefs about the nature of a particular participant's audiogram were primed only by sampling with a single initial random tone. Active sampling then proceeded to select the most informative tones for further estimation.

It is worth keeping in mind that by definition HWAG can only deliver threshold estimates in multiples of 5 and 10 dB, yielding a theoretical resolution of ± 3.75 dB around the 0.707 probability threshold. MLAG can deliver threshold estimates at arbitrary sound level values, though for this study the nearest whole dB was returned, yielding a theoretical resolution of ± 0.5 dB. The lower threshold resolution of HWAG could lead it to having an inflated reliability measure because many repeated measures would likely end up with thresholds falling within the same 5 dB bin and thus producing an error of exactly 0. Furthermore, because the 0.707 probability point was selected for MLAG threshold to compare with HWAG threshold and because the threshold values matched so closely, the MLAG method therefore must also be estimating psychometric spread accurately as well. This capability has indeed been observed in extensive simulations (Song et al., 2017; Song et al., 2018). Delivering a fully predictive audiogram in the current context is equivalent to saying that MLAG accurately estimates a complete psychometric curve at every frequency. The utility of this extra information regarding hearing is indeterminate, as acquiring enough data for such inference using traditional techniques would require so much time as to be impractical

for clinical and most research purposes. For these reasons, relatively few examples of this more complete audiogram information have been collected previously.

The current study was intended to compare a conventional threshold audiogram estimation method and an advanced method using the same online platform and identical testing conditions. Previous work has compared offline MLAG versus manual HWAG and found similar results (Song et al., 2015). Future work needs to determine the sensitivity of this method for determining threshold audiograms under different conditions, such as in different sound booths and with different audiologists. Fortunately, the online nature of this implementation makes such a comparison easy for audiologists to undertake. Access to the Bonauria online audiometry portal used to collect the data presented here is available to any clinician or researcher in order to conduct this or any other related study.

Conclusion

Online machine learning audiometry delivers hearing threshold estimates similar to those of online Hughson-Westlake audiometry for a variety of audiogram shapes. Notable improvements provided by MLAG that are not possible with HWAG, however, include arbitrarily short acquisition times, continuous-frequency threshold estimates, and continuous-frequency psychometric spread estimates. Furthermore, the machine-learning algorithm can readily be extended to other types of hearing tests, leading to efficient, unified tests of overall hearing ability. The online nature of MLAG using modern web audio standards frees clinicians and researchers from specific hardware requirements, as web browsers implementing updated web audio standards exist on computer, tablet and smartphone platforms. The online HWAG procedure used for comparison in this study itself appears to be a novel development that will deliver discrete thresholds from a web browser consistent with international standards familiar to audiologists. Overall, online MLAG provides numerous procedural and inferential advantages that cannot be matched by other extant methods, making it an attractive choice for the basis of a next-generation audiometric standard.

Acknowledgements

Funding for this project was provided by NIH grants UL1 TR002345, T35 DC008765, T32 NS073547, NSF grant DGE-1745038, and the Center for Integration of Medicine and Innovative Technology (CIMIT).

DLB has a patent pending on technology described in this manuscript and has equity ownership in Bonauria, LLC.

DLB wrote the manuscript; RTH designed and conducted the experiments; NM helped design and conduct the experiments; XDS, KAS, JCD and BRDS developed and implemented the algorithm; JYC, EAD, JMB, KLH and DLB analyzed the data. All authors discussed the results and implications and commented on manuscript revisions.

Financial disclosures:

Funding for this project was provided by NIH grants UL1 TR002345, T35 DC008765, T32 NS073547, NSF grant DGE-1745038, and the Center for Integration of Medicine and Innovative Technology (CIMIT). DLB has a patent pending on technology described in this manuscript and has equity ownership in Bonauria, LLC.

References

- American National Standards Institute. (2004). Methods for manual pure-tone threshold audiometry *ANSI 3.21* (Vol. 3). New York.
- American Speech-Language-Hearing Association. (2005). Guidelines for manual pure-tone threshold audiometry. Retrieved from <http://www.asha.org/policy>
- Barbour DL, DiLorenzo J, Sukesan KA, Song XD, Chen JY, Degen EA, ... Garnett R (2018). Conjoint psychometric field estimation for bilateral audiometry. *Behav Res Meth*. doi:10.3758/s13428-018-1062-3
- Carhart R, & Jerger J (1959). Preferred method for clinical determination of pure-tone thresholds. *J Speech Hear Disord*, 24, 330–345.
- Fechner GT (1860). *Elements of Psychophysics* (Adler HE, Howes Trans. D. H. & Boring EC Eds.). New York: Holt, Rhinehart & Winston.
- Franks J (2001). Hearing measurement In Goelzer B, Hansen CH, & Sehrndt GA (Eds.), *Occupational exposure to noise: evaluation, prevention and control*. (pp. 183–232). Dortmund: World Health Organization.
- Gardner JM, Song XD, Cunningham JP, Barbour DL, & Weinberger KQ (2015). Psychophysical testing with Bayesian active learning. Paper presented at the Uncertain Artif Intell.
- Garnett R, Osborne MA, & Hennig P (2013). Active learning of linear embeddings for Gaussian processes. *arXiv*, 1012.2599.
- Gosztonyi RE Jr., Vassallo LA, & Sataloff J (1971). Audiometric reliability in industry. *Archives of Environmental Health: An International Journal*, 22(1), 113–118.
- Houlsby N, Huszár F, Ghahramani Z, & Lengyel M (2011). Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Hughson W, & Westlake H (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Trans Am Acad Ophthalmol Otolaryngol*, 48(Suppl), 1–15.
- Ishak WS, Zhao F, Stephens D, Culling J, Bai Z, & Meyer-Bisch C (2011). Test-retest reliability and validity of Audioscan and Békésy compared with pure tone audiometry. *Audiological Medicine*, 9(1), 40–46.
- Levitt H (1971). Transformed up-down methods in psychoacoustics. *J Acoust Soc Am*, 49(2B), 467–477.
- Mahomed F, Eikelboom RH, & Soer M (2013). Validity of automated threshold audiometry: A systematic review and meta-analysis. *Ear and hearing*, 34(6), 745–752. [PubMed: 24165302]
- Schmuziger N, Probst R, & Smurzynski J (2004). Test-retest reliability of pure-tone thresholds from 0.5 to 16 kHz using Sennheiser HDA 200 and Etymotic Research ER-2 earphones. *Ear and hearing*, 25(2), 127–132. [PubMed: 15064657]
- Song XD, Garnett R, & Barbour DL (2017). Psychometric function estimation by probabilistic classification. *J Acoust Soc Am*, 141(4), 2513–2525. [PubMed: 28464646]
- Song XD, Sukesan KA, & Barbour DL (2018). Bayesian active probabilistic classification for psychometric field estimation. *Atten Percept Psychophys*, 80(3), 798–812. doi:10.3758/s13414-017-1460-0 [PubMed: 29256098]
- Song XD, Wallace BM, Gardner JR, Ledbetter NM, Weinberger KQ, & Barbour DL (2015). Fast, continuous audiogram estimation using machine learning. *Ear and hearing*, 36 (6), e326–e335. [PubMed: 26258575]
- Swanepoel DW, Mngemane S, Molemong S, Mkwanazi H, & Tutshini S (2010). Hearing assessment—reliability, accuracy, and efficiency of automated audiometry. *Telemed J E Health*, 16(5), 557–563. [PubMed: 20575723]
- W3C. (2015). Web Audio API. Retrieved from <https://www.w3.org/TR/webaudio/>

Figure 1:

ML Procedure: Stimuli were presented through insert earphones and participants were instructed to press a button upon detection of any stimulus. A response within 2 s following the onset of the tone stimulus was considered a detected sample; no response was counted as an undetected sample; a response after 2 s was recorded as a false detection and not counted toward the audiogram estimate.

Figure 2:

Screen shots of online implementations of HW and ML audiometry. Control panel for the audiologist (A) demonstrates some of the parameters that can be adjusted for a test. Thresholds estimated from the HW procedure (B) compare favorably to thresholds estimated from the ML procedure (C). Tone frequencies are much more dispersed in the latter, enabling a continuous threshold estimate as a function of frequency. Thresholds at octave frequencies are not special relative to other frequencies in ML audiometry.

Figure 3:

Posterior mean probability function (A) for one ear is computed using 5 sampled points. Red diamonds indicate the tone was inaudible; black pluses, audible. Acquisition function (B) is computed and identifies the point of maximum uncertainty between models (pink star). This point where models differ most is then queried for participant audibility (C, pink arrow). The participant detected this tone, so the updated set of points is used to recompute the posterior probabilities with a lower threshold near the frequency of that tone. After 40 sampled points, the updated posterior probability (D) has greatly reduced the uncertainty in the acquisition function (E) to the area we consider threshold. As such, the posterior probabilities after 40 (D) and 41 (F) samples are almost identical due to the small amount of uncertainty. Each tone response therefore updates the entire audiogram estimate across all frequencies, though much less as the test progresses.

Figure 4:

Threshold audiogram estimates for all 42 ears of this study sorted by pure-tone average. Two repetitions each of HWAG (dashed purple line) and MLAG (solid orange line) are shown. Concordance of all estimates is generally high. Any discordance of MLAG performance is largely confined to extreme frequencies in a few ears.

Figure 5: Cumulative agreement between HW and ML thresholds as a function of algorithm iteration. Plotted are the mean absolute difference and 1 standard deviation. The automated HW algorithm presented a median of 67 tones to find thresholds at the standard 6 frequencies. The ML algorithm could have terminated before the 49-sample criterion without sacrificing average accuracy.

Table 1:

Differences between the online ML and online HW threshold estimates.

Frequency (kHz)	0.25	0.5	1	2	4	8	All
Signed differences (ML – HW)							
Mean signed difference (dB)	−0.867	−0.253	−0.160	−0.162	−0.297	−4.18	−0.969
Standard deviation (dB)	8.19	4.98	2.62	3.96	4.51	8.51	6.02
Absolute differences							
Mean absolute difference (dB HL)	3.49	2.51	1.65	2.26	3.24	6.42	3.24
Standard deviation (dB HL)	7.46	4.30	2.04	3.25	3.13	6.96	5.15
Root mean square differences							
RMS difference (dB HL)	8.08	4.68	2.84	4.32	4.76	8.79	5.58

Table 2:

Test-retest reliability of online ML audiogram.

Frequency (kHz)	0.25	0.5	1	2	4	8	All
Signed differences (second – first)							
Mean signed difference (dB)	0.556	–1.33	0.139	–0.342	–0.400	–0.412	–0.486
Standard deviation (dB)	13.1	6.78	1.78	2.98	3.55	8.32	7.15
Absolute differences							
Mean absolute difference (dB)	5.11	2.61	1.08	1.66	2.00	4.71	2.85
Standard deviation (dB)	12.06	6.39	1.40	2.48	2.94	6.82	6.57
Root mean square differences							
RMS difference (dB HL)	12.07	6.32	1.86	4.02	5.90	7.74	6.32

Table 3:

Test-retest reliability of online HW audiogram.

Frequency (kHz)	0.25	0.5	1	2	4	8	All
Signed differences (second – first)							
Mean signed difference (dB)	–0.156	0.00	0.00	–0.313	0.469	–1.25	–0.208
Standard deviation (dB)	2.97	3.11	2.20	3.09	3.45	7.30	4.02
Absolute differences							
Mean absolute difference (dB)	1.41	1.56	0.94	1.56	1.41	4.38	1.88
Standard deviation (dB)	2.61	2.68	1.98	2.67	3.17	5.93	3.56
Root mean square differences							
RMS difference (dB HL)	3.12	3.12	2.67	3.09	3.36	6.64	3.69