Published in final edited form as:

Ear Hear. 2020 ; 41(6): 1692–1702. doi:10.1097/AUD.0000000000000891.

Dynamically Masked Audiograms with Machine Learning Audiometry

Katherine L. Heisey¹, Alexandra M. Walker^{1,2}, Kevin Xie^{1,3}, Jenna M. Abrams^{1,2}, Dennis L. Barbour¹

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: When one ear of an individual can hear significantly better than the other ear, evaluating the worse ear with loud probe tones may require delivering masking noise to the better ear in order to prevent the probe tones from inadvertently being heard by the better ear. Current masking protocols are confusing, laborious and time consuming. Adding a standardized masking protocol to an active machine learning audiogram procedure could potentially alleviate all of these drawbacks by dynamically adapting the masking as needed for each individual. The goal of this study is to determine the accuracy and efficiency of automated machine learning masking for obtaining true hearing thresholds.

Design: Dynamically masked automated audiograms were collected for 29 participants between the ages of 21 and 83 (mean 43, SD 20) with a wide range of hearing abilities. Normal hearing listeners were given unmasked and masked machine learning audiogram tests. Listeners with hearing loss were given a standard audiogram test by an audiologist, with masking stimuli added as clinically determined, followed by a masked machine learning audiogram test. The hearing thresholds estimated for each pair of techniques were compared at standard audiogram frequencies (i.e., 0.25, 0.5, 1, 2, 4, 8 kHz).

Results: Masked and unmasked machine learning audiogram threshold estimates matched each other well in normal hearing listeners, with a mean absolute difference between threshold estimates of 3.4 dB. Masked machine learning audiogram thresholds also matched well the thresholds determined by a conventional masking procedure, with a mean absolute difference between threshold estimates for listeners with low asymmetry and high asymmetry between the ears, respectively, of 4.9 dB and 2.6 dB. Notably, out of 6200 masked machine learning audiogram

tone deliveries for this study, no instances of tones detected by the non test ear were documented. The machine learning methods were also generally faster than the manual methods, and for some listeners, substantially so.

Conclusions: Dynamically masked audiograms achieve accurate true threshold estimates and reduce test time compared to current clinical masking procedures. Dynamic masking is a compelling alternative to the methods currently used to evaluate individuals with highly asymmetric hearing, yet can also be used effectively and efficiently for anyone.

Keywords

Audiogram; Audiology; Audiometry; Estimation; Machine learning; Psychoacoustics; Psychophysics

INTRODUCTION

Hearing naturally involves two ears, though clinicians and researchers typically evaluate one ear at a time because each can malfunction independently. Diagnostic sounds delivered over headphones can be manipulated to independently test each ear's hearing threshold. A threshold audiogram is a straightforward assessment of audibility that requires a patient to simply listen for and respond to the occurrence of pure tones presented to each ear. In most audiology practices today a clinician manually obtains pure-tone hearing thresholds following a procedure that was recommended as the standard for audiometric testing 60 years ago (Hughson & Westlake 1944; Carhart & Jerger 1959): the manually-delivered modified Hughson-Westlake audiogram (HWAG). This adaptive up-down staircase method continues to be emphasized in the most recent clinical guidelines (American Speech-Language-Hearing Association 2005) and is valued for being fast and reliable for many patients.

A particular case where manual HWAG is inadequate, however, is when cross hearing is likely to occur. Cross hearing arises when loud tones presented to the test ear cross over and are heard by the non-test ear via bone conduction through the skull (Martin & Blosser 1970). If a tone presented to the test ear is actually heard by the non-test ear, what was intended to be an independent assessment of the test ear's hearing ability is now confounded by the contralateral ear's response. In such cases, the estimated threshold of the test ear is artifactually lower than the true threshold.

A sound delivered to the test ear will lose some intensity as it travels to the contralateral ear, and it arrives at the non-test ear at a reduced sound level compared to its starting intensity. Interaural attenuation reflects the amount of sound energy that dissipates as the tone travels from the ipsilateral test ear to the contralateral non-test ear. Because interaural attenuation varies for each individual based on the dimensions of their skull, transducers used, frequency of the sound, and other testing factors, current compensatory testing methods rely on a conservative estimate of interaural attenuation for each transducer. For supra-aural and circumaural headphones, 40 dB is used across all frequencies (Smith 1968; Brännström & Lantz 2010). Having less contact with the skull, insert headphones have a higher estimated interaural attenuation of 50 dB – 75 dB depending on the frequency tested (Killion et al.

1985; Sklare & Denenberg 1987; Munro & Contractor 2010). Tones are conventionally considered at risk of cross-hearing only if their intensities are greater than the interaural attenuation estimate plus the hearing threshold of the non-test ear.

To offset the effects of cross-hearing, narrowband noise is introduced to the non-test ear to mask any potential cross tone detection in that ear (Denes & Naunton 1952; Hood 1960; Studebaker 1964). Current methods do not assess the need for masking until after initial unmasked threshold estimates are determined. Only frequencies with significantly asymmetric left and right ear thresholds after this testing are suspected to be incorrect due to contralateral ear responses. A masking procedure is then employed to re-analyze the thresholds at the identified frequencies. Because measuring the actual interaural attenuation at each frequency is impractical with conventional testing, a complex yet generic protocol must be used to re-evaluate the threshold estimates and achieve effective masking levels without overmasking (i.e., allowing the masker to cross over and affect tone detection in the test ear). One method to administer masking requires multiple iterations of re-establishing the threshold in the test ear while systematically adjusting the amount of masking in the non-test ear (Hood 1960). Optimized methods requiring fewer iterations have been proposed (Smith 1968; Turner 2004a; Turner 2004b) but are similarly constrained by the need to perform masking after initial unmasked audiograms are completed, thus substantially increasing true threshold estimation time.

Individuals with highly asymmetric hearing, defined here as differences in same-frequency thresholds between the two ears greater than the estimated interaural attenuation, are at risk of cross hearing. While only a subset of the general population, for them masking is an essential component of hearing assessment, aiding in differential diagnosis and hearing loss management decisions. Unfortunately, masking is a time-consuming process and is often cited as one of the most challenging procedures for audiologists to learn (Sanders & Rintelmann 1964; Ho et al. 2009; Valente 2009; Yacullo 2015; Gumus et al. 2016; Hamil 2016). No universally accepted masking standard or guideline exists. In the most recent surveys of audiologic practices conducted by the American Academy of Audiology (Martin et al. 1994; Martin et al. 1998), researchers noted that audiologists were using a broad range of masking methods and further determined that over half of the respondents were using inappropriate masking procedures.

Automated audiometry methods present the opportunity for standardization and uniformity of hearing assessments regardless of patient hearing status. While manual HWAG is considered the clinical standard for threshold estimation, automated and adaptive techniques have demonstrated similar accuracy and reliability to manual audiometry (Ho et al. 2009; Swanepoel et al. 2010; Mahomed et al. 2013; Shojaemend & Ayatollahi 2018). These methods have yet to see widespread clinical use, and few can accommodate highly asymmetric hearing individuals without additional modification.

We have developed a novel computational framework that uses active sampling of the most informative stimuli to estimate pure-tone audiogram thresholds. This active machine learning audiogram (AMLAG) has been shown to be as accurate as and more efficient than manual HWAG methods for normal and hearing loss populations (Song et al. 2015; Barbour

et al. 2019b). In order to create a comprehensive procedure that can be used clinically to further reduce testing time while retaining accuracy and improving informativeness, an automated masking procedure must be developed.

To this end, we have integrated a dynamic masking protocol into AMLAG to create the masked AMLAG. Unlike masking during manual audiometry, masked AMLAG presents suitable masking noise to the non-test ear throughout the entire audiogram test procedure. Every tone presented to the test ear is paired with masking noise in the non-test ear. Masking noise levels are derived from a combination of the interaural attenuation estimate and the intensity of the test ear tone. Every audiogram becomes a masked audiogram, and accurate thresholds are estimated directly because cross hearing is dynamically eliminated. AMLAG so rapidly homes in on hearing thresholds (Heisey et al. 2018) that individuals with fairly symmetric hearing should almost never be presented a suprathreshold masking noise, making masked and unmasked AMLAG procedurally equivalent for this large population. Here we show that masked AMLAG delivers true audiogram threshold estimates as accurately as manual HWAG regardless of hearing ability. Further, masked AMLAG is as efficient as manual HWAG and unmasked AMLAG for symmetric hearing individuals and shows significant efficiency gains in highly asymmetric hearing participants relative to HWAG.

MATERIALS AND METHODS

Participants

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

Equipment

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

Experimental Procedure

The 29 participants were split into two experimental groups according to their reported hearing ability. The first group consisted of nine participants with self-reported normal

hearing, designated as No Loss (NL). The remaining 20 participants reported some degree of hearing deficit and were designated as Hearing Loss (HL). The cohort with hearing loss exhibited a variety of etiologies based upon the relationships between their air-conduction and bone-conduction audiograms, including sensorineural loss, conductive loss and mixed losses.

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

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

Participants were instructed to click the left mouse button whenever they heard a tone, even if it was very soft. They were informed that the frequency, or pitch, would change between each tone sequence and that there could be relatively long periods of silence. Participants were instructed to ignore any wind or white noise they heard and were reminded to only click the mouse when they heard a pure tone. All participants were asked after each ear's test if they had heard any wind or white noise. Their responses to this question were recorded.

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

Masked AMLAG presented 1/3 octave narrowband noise to the contralateral non-test ear while simultaneously presenting a three-pulse sequence of tones to the test ear. This

procedure was performed for every tone presentation, even if a participant would not typically require masking. Masking noise began randomly in the 250 – 1500 ms interval before the onset of the pure-tone sequence and remained on for a total of 3.0 – 5.5 sec. The noise ramped on for 100 ms at the beginning of the intersequence interval and ramped off during the final 100 ms. All masking noise presentations began seamlessly at the conclusion of the preceding noise presentation, centered at the frequency of the test-ear tone and presented at 40 dB below the tone's presentation level. This masking presentation level is based on a conservative interaural attenuation level of 40 dB for supra-aural headphones (Yacullo 2015).

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

Masking for air conduction was performed when the air conduction threshold of the test ear was worse than the bone or air conduction threshold of the non-test ear by greater than or equal to 40 dB. To ensure the non-test ear was not responding to the tone, narrowband noise was presented at a suprathreshold level. Specifically, 10 dB was added to the air conduction threshold of the non-test ear and presented as narrowband noise. The true air conduction threshold of the test ear was then found using the plateau method (Hood 1960; Martin 1980; Yacullo 2015). A true threshold was determined when a participant responded to a tone after the noise was raised by 5 dB three times. In other words, when the participant heard the tone even after the noise was increased by a total of 15 dB.

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

Extended details on the masking procedure and other experimental details can be found in Supplemental Materials and at <https://osf.io/gj8s5>.

Statistical Analysis

AMLAG returns a continuous estimate of the probability of hearing any frequency-intensity pair in the stimulus domain. Hearing thresholds at octave frequencies were determined at the

0.707 detection probability to match the standard probability of detection for HWAG estimates. Any threshold estimate that was greater than 100 dB HL was designated as a “no response” at that frequency.

The unmasked and masked AMLAG thresholds were compared at the standard audiogram frequencies for Group NL. Accuracy and efficiency of the masked AMLAG in this case were evaluated via comparison to the unmasked AMLAG. Accuracy was assessed with Bland-Altman analysis. Bland-Altman plots reveal the difference between test methods (e.g., unmasked and masked AMLAG in this case) against the mean of the two test methods for each frequency in each ear for each participant. The goal of these plots is to demonstrate that bias does not change with the magnitude of hearing loss. If the plot shows little change in bias with hearing level, and constant variance in bias with hearing level, then a single measure of method agreement, called the Limits of Agreement (LOA) can be reported. The 90% LOA is the range of test differences that one might expect to find in 90% of participants tested. For example, a 90% LOA of 10 dB indicates that the difference between test methods would be expected to fall within 10 dB for 90% of individuals.

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

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

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

RESULTS

Group NL Analysis:

Masked AMLAG thresholds estimated in Group NL listeners were consistent with thresholds estimated by unmasked AMLAG, which has been previously validated as equivalent in accuracy to HWAG (Song et al. 2015; Heisey et al. 2018; Barbour et al. 2019b). The similarity of unmasked AMLAG and masked AMLAG threshold estimates at

the standard audiogram frequencies across all tests within Group NL is depicted in Bland-Altman plots in Figure 1 (Bland & Altman 1999). Differences are small on average and do not appear to be a function of threshold magnitude, indicating that bias and variability in bias do not appear to depend on degree of hearing loss. The 90% limits of agreement ($1.645 \times$ standard deviations) are therefore reasonable summaries of test equivalence, indicating that 90% of normal hearing participants would be expected to exhibit masked AMLAG thresholds within 5–9 dB of the unmasked AMLAG thresholds. Mean signed differences are close to 0, as would be expected if the two tests were evaluating the same underlying physiological process.

Additional numerical summaries are given for Group NL in Table S1. Previously published studies have demonstrated that variability in pure-tone manual HWAG thresholds obtained with supra-aural transducers in the age range studied here are considered clinically relevant only when exceeding 10 dB, and mean deviations within 5 dB are commonly cited as clinically acceptable (Stuart et al. 1991; Landry & Green 1999; Mello et al. 2015). The mean absolute difference between masked and unmasked AMLAG was under 5 dB at all frequencies with an overall mean of 3.4 dB. Collectively, these results indicate that masked AMLAG yields threshold estimates comparable in value to unmasked AMLAG in normal hearing individuals.

All AMLAG tests in this study were designed to deliver 100 tone presentations per ear in order to ensure confident final threshold estimates. Previous research has demonstrated that unmasked AMLAG often converges to a threshold estimate within 5 dB of the final threshold estimate in considerably fewer than 100 tone presentations per ear (Song et al. 2015; Heisey et al. 2018). For each participant in Group NL, the total number of tone presentations and average time for unmasked and masked AMLAG to converge to a threshold estimate within 5 dB of the final estimation in both ears were calculated (Table 1). Figure 2 shows the average absolute difference between the threshold estimate at each tone presentation and the final estimate after 100 tones averaged across all Group NL ears. It demonstrates a very similar convergence profile for both tests in this group.

At each tone presentation, masked AMLAG presented narrow band noise in the ear contralateral to the ear being tested. Because AMLAG so rapidly identifies putative thresholds and spends most of its sampling effort at nearby intensities, the paired masking noise level was almost always subthreshold and therefore expected to be undetectable by the non-test ear (Table 2 and Figure 3). To determine if dynamic masking subjectively altered the test experience, participants were asked following each AMLAG test (unmasked and masked) if they had heard any white noise and if so, in which ear they had heard it. Of the 36 automated audiogram assessments for Group NL, five tests were identified by participants as having presented detectable white noise in the non-test ear. Three of those five were actually unmasked AMLAG tests with no noise delivery at all, and the reported perception most likely was due to occlusion effects. The two masked AMLAG tests during which the participants noted hearing white noise were both tests in which suprathreshold masking levels were presented to the non-test ear. The participants commented that the masking noise was not distracting and described the noise as “soft.”

Group HL Analysis:

The accuracy of masked AMLAG was evaluated at standard audiogram frequencies relative to manual HWAG and averaged across all tests for Group HL. The similarity of masked AMLAG and HWAG threshold estimates at the standard audiogram frequencies across all tests is depicted in Bland-Altman plots in Figure 4 for group HL-LA and Figure 5 for Group HL-HA. Means and 90% limits of agreement are again depicted. Differences generally do not appear to be a function of threshold magnitude, though the variability in differences appears to be higher with higher thresholds for 4 kHz, Group HL-LA. Given that this trend was not found at adjacent frequencies or for 4 kHz in other groups, it seems likely to reflect participant sampling. The large outlier at the highest threshold for 1 kHz, Group HL-LA, may be attributable to this participant's self-reported tinnitus, and is a scenario worthy of further investigation.

Aside from this isolated behavior, the bias and variability in bias do not appear to depend on degree of hearing loss. The 90% limits of agreement therefore represent reasonable summaries of test equivalence, indicating that 90% of hearing loss participants would be expected to exhibit masked AMLAG thresholds within 9–14 dB of HWAG thresholds for Group HL-LA and within 5–9 dB of HWAG thresholds for Group HL-HA. Somewhat higher variability is apparent in individuals with low-asymmetry hearing loss with this analysis. Given that Group NL and Group HL-HA exhibit similar limits of agreement and that the few outliers inflating this estimate are evident in the measurements for Group HL-LA, it seems likely that a reasonable estimate for the 90% limit of agreement comparing masked AMLAG to conventional HWAG is around 10 dB. Because the Group NL evaluation was functionally implemented as a test-retest procedure, the agreement between masked AMLAG and conventional HWAG therefore appears to be similar to AMLAG's test-retest performance.

Group HL-LA numerical summaries are given in Table S2, and Group HL-HA numerical summaries are given in Table S3. Once again, mean signed differences near 0 imply that one test is not biased in its threshold estimates relative to the other. The small mean absolute differences between masked AMLAG and HWAG convey that the tests consistently deliver similar estimates. Group HL-LA had a mean absolute difference of 4.9 dB and Group HL-HA had a 2.6 dB difference. These results are within the published variability of 5-10 dB shown between traditional and other automated audiometry assessments (Shojaeemend & Ayatollahi 2018).

In addition to estimating accurate pure-tone thresholds, masked AMLAG was able to generate these thresholds with significantly fewer tone presentations ($p = 3.92 \times 10^{-3}$ for Group HL-LA, $p = 2.95 \times 10^{-4}$ for Group HL-HA, paired t -tests) and significantly more quickly than manual HWAG ($p = 5.48 \times 10^{-2}$ for Group HL-LA, $p = 5.66 \times 10^{-4}$ for Group HL-HA, paired t -tests). The efficiency of masked AMLAG was evaluated through a comparison of the average number of tone presentations and the average test time required to estimate thresholds within 5 dB of the final threshold estimates relative to manual HWAG's final threshold determinations. Because left and right threshold estimates are necessary to determine masking needs for HWAG, ears were analyzed as left and right pairs, giving overall results for each participant. The two Group HL-LA participants with a single excluded ear were removed from this analysis. Overall results are shown in Table 3. Masked

AMLAG estimated thresholds for both ears with, on average, 64 fewer tones per Group HL-LA participant and 136 fewer tones per Group HL-HA participant. For hearing losses where no masking was required during manual HWAG (Group HL-LA), the average masked AMLAG test time to estimate both ears for a single participant was 3.8 minutes faster than the average manual HWAG time. For hearing losses requiring masking during manual HWAG (Group HL-HA), the difference was much greater, with masked AMLAG estimating thresholds an average of 13.1 minutes faster than manual HWAG. Clinically, bone conduction is needed to determine a participant's masking needs in the presence of an air-bone gap. Accordingly, both air conduction and bone conduction tone counts were included in the total manual HWAG convergence analysis. No Group HL participants presented an air-bone gap that required additional air conduction masking. Therefore, the mean number of tone presentations and minutes required for all HWAGs with bone-conduction assessment removed from analysis are also summarized in Table 3.

Similar to the analysis for Group NL, each masking noise presentation was assessed to determine the effect of dynamic masking on Group HL tests. Group HL-LA participants did not clinically require masking, and we anticipated that much like Group NL, most masking levels would be presented at levels below the non-test ear threshold. On the other hand, Group HL-HA participants did require masking, and it was expected that masking noise would be heard in the test ear at suprathreshold levels. These results are shown in Table 4 and Figure 3. After each masked AMLAG test, participants were asked if they had heard any white noise and in which ear it had been heard. Listeners from Group HL-LA noted hearing masking noise in eight out of 22 masked AMLAG tests. Six of the eight were tests in which a fraction of the tones were paired with masking noise levels that would have been above the contralateral ear threshold for the Group HL-LA participants. One participant identified masking noise during left and right masked AMLAG, yet analysis shows that no suprathreshold masking noise was delivered during either test. It is suspected that occlusion effects or tinnitus might account for the perceived noise heard during both tests. Nevertheless, the noise perception did not appear to interfere with testing procedures or results. Two AMLAG tests in Group HL-LA had tone presentations paired with suprathreshold masking noise delivered to a non-test ear that were not identified by the participant. In these tests, masking noise levels may have been infrequent or quiet enough to be unremarkable. All eight Group HL-HA participants reported hearing masking noise in their better-hearing ear during the worse-hearing test ear assessment. No masking noise was discerned in the worse-hearing ear. No participant in Group HL reported the onset of the masking noise to be distracting or to inhibit their ability to perform the task.

Figure 6 shows the mean absolute difference between the final threshold estimate at 100 tone presentations and each increment iteration of masked AMLAG averaged across all of Group HL, Group HL-LA, and Group HL-HA participants. Test results converged faster for individuals with normal hearing, most likely because the initial seven fixed frequency/intensity combinations were particularly informative for this group and enabled active learning to select tone queries that rapidly reduced errors. Highly asymmetric hearing thresholds can also be estimated relatively rapidly, presumably for the complementary reason that extremely high thresholds near or beyond the maximum stimulus can also be identified relatively quickly in an active testing scenario. It is not surprising given this

consideration that individuals with thresholds in both ears near the middle of the testing range would require the most test tones to achieve comparable accuracy. Incidentally, these individuals are exactly the patient population for whom bilateral audiometry can most speed up testing (Heisey et al. 2018).

Figure S1 visually depicts the thresholds estimated for all ears with all air conduction tests for this study. Participants are sorted by group (NL, HL-LA and HL-HA), and within each group by pure tone average of the better hearing ear. This visualization demonstrates the variety of hearing profiles for the participants in this study, as well as the agreement between testing procedures. Most agreement is high, with occasional disparities at individual frequencies. Asymmetry alone is not associated with the disparities because Group HL-LA exhibited the least overall agreement between threshold estimates and not Group HL-HA. Figure S2 depicts the bone conduction thresholds that were estimated for this study population for reference. Participant order is the same as in Figure S1.

DISCUSSION

Masked AMLAG demonstrated similar accuracy and improved efficiency when compared to unmasked AMLAG and manual HWAG. These results were observed in normal hearing, symmetric loss and asymmetric loss participants. This finding is particularly important as it indicates that masked machine learning audiometry delivers accurate true threshold estimates even for patients with highly asymmetric hearing where substantial masking is required. Exploiting the relationships between interaural attenuation, intensity and frequency, dynamically masked AMLAG achieves its test time reduction for patients with asymmetric hearing by eliminating the need for a separate masking step. Notably, adding contralateral masking to every tone does not significantly increase test time for listeners with normal or symmetric hearing. For most of these participants, masking levels remained below hearing thresholds and were undetected throughout the test. A dynamically masked audiogram therefore allows for individual differences in masking needs to be addressed in real time without increasing test time.

It is important to consider that masked AMLAG was set to deliver 100 tone presentations per ear even if it was confident in the estimated thresholds at earlier tone counts in order to ensure the acquisition of complete audiogram models. Therefore, tone counts and test times were calculated at the point when masked AMLAG's estimation fell within 5 dB of its final estimation. Test stopping criteria, such as were used previously (Song et al. 2015), are the subject of ongoing research. A notable difference between HWAG and AMLAG is that the former must reach the end of its testing procedure before a complete threshold estimate is available, while the latter delivers a complete estimate for any length of test, though it converges closer to a more accurate model as more tones are delivered (Figure 1). AMLAG is therefore very flexible in its test length and can deliver useful results even in extremely short testing scenarios, such as with pediatric patients.

Manual HWAG test time and tone counts included the collection of both air conduction, bone conduction, and any masked thresholds. This procedure likely increased both measures significantly. This comparison is reasonable, however, because the manual masking protocol

used in this study requires bone conduction thresholds to determine if air conduction masking was needed due to an air-bone gap. None of the HL study participants had an air-bone gap requiring additional air conduction masking, so masked AMLAG has yet to be tested under those conditions. While masked AMLAG is currently limited to testing air conduction, it dynamically masks all tone presentations and, therefore, does not require bone conduction thresholds to effectually mask air conduction thresholds, making it more efficient than manual HWAG. To evaluate a more direct comparison, however, we have also tallied manual HWAG tone counts for this study population that only include air-conduction threshold and masking presentations, thereby excluding bone conduction counts from analysis (Table 3). Excluding bone conduction tone counts highlights that masked AMLAG is already more efficient than manual HWAG without masking and substantially outperforms manual HWAG when masking is required. For Group HL-LA, masked AMLAG estimated air conduction thresholds with fewer tone counts but in the same number of minutes as manual HWAG. For these participants, all of whom did not require contralateral air conduction masking, manual HWAG benefited from the proficiency and adaptability of an individual clinically trained to perform audiograms. The current implementation of masked AMLAG has a static response window of 1.5 seconds, regardless of when the participant responded to the tone. Future implementations could, for example, commence the inter-sequence wait time immediately after recording a heard response to more closely mimic the actions of skilled audiologists.

Three Group HL-HA participants had unilateral cochlear implants with no residual hearing in the implanted ear. For these participants, masked AMLAG for the implanted ear executed only Halton sampling because no heard tone was ever detected in that ear. Figure 7 shows the final left and right ear thresholds and each tone presented for one unilaterally deaf participant. While the tones presented in the better-hearing ear are almost all focused near the threshold estimate, the ‘dead’ ear samples canvas the entire frequency/intensity domain. Because no tone was heard and there is no initial threshold estimate, we declared masked AMLAG to be converged to “no response” at every frequency if there was no threshold estimate after 15 tones. This value accounts for a reasonable number of tones to adequately sample the domain and deduce a complete lack of hearing. In an eventual clinical version of AMLAG, Halton sampling will not be used, and a dead ear would be determinable rapidly by active sampling. The purpose of the extensive sampling in the current study was to determine if dynamic masking ever failed to properly mask a test tone. No examples of such failure were noted in 6200 tone deliveries. Ears with no residual hearing almost always elicit cross-hearing and require extensive masking when tested, as shown in the rightmost histogram of Figure 3. Masked AMLAG is able to effectively sample throughout the domain and cancel out all cross tones without requiring any additional procedure. Audiogram estimates of more nearly symmetric ears can be seen in Figure S3.

Because AMLAG frequency and intensity levels are selected by an active learning algorithm, subsequent masking levels rove across the entire frequency and intensity spectrum. It is possible for long periods without suprathreshold masking to be followed by a tone presentation paired with an audible contralateral masking level. While listener performance has been shown to be unaffected by tones roving between frequency, intensity, and ears (Song et al. 2015; Heisey et al. 2018; Barbour et al. 2019b), masking is unique in

that tones in a test ear are paired with masking noise in a non-test ear. For any test stimulus, both, either, or neither sound might be heard by the listener. The onset of sound, be it tone or noise, requires the listener to discern if a response is appropriate or should be inhibited. The consistent threshold estimation results in all groups demonstrate that masking noise did not disorient listeners or induce false positives. It was anticipated that participants requiring masking would have had more masking protocol exposure as a part of routine manual HWAG assessments, whereas participants to whom masking noise is a novel experience might have struggled to ignore masking noise. Eight of the 22 Group HL-LA tests, however, were presented with audible masking during masked AMLAG. Presumably, these participants had not previously experienced audiograms that included any masking protocol. These unfamiliar listeners successfully completed the assessment and had similar results as those without any audible masking.

Clinically, masked AMLAG offers several potential benefits compared to manual HWAG. As this study showed, masked AMLAG provides an opportunity for the standardization of masking, a challenging procedure with multiple variations that are frequently implemented incorrectly (Sanders & Rintelmann 1964; Yacullo 2015; Gumus et al. 2016; Hamil 2016; Valente 2009). Uniformity of clinical procedures is imperative in order to reduce inter-clinician variability and ensure that best practices are being achieved. Automation of these methods would also allow technicians to perform some routine testing, providing audiologists with more time for complex cases and to perform other clinical duties.

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

Second, previous research using AMLAG has included discussion of the unpredictable nature of the constituent tone sequences (Table S4) and corresponding difficulty for malingering patients to thwart the test (Song et al. 2015). The use of consistent relative contralateral noise levels that begin prior to tone delivery reinjects some predictability into the tone sequences for individuals with asymmetric hearing loss who definitely detect the maskers. The ultimate solution in this case may be to allow masker level to vary just as tone frequency and level do and explicitly estimate the frequency-dependent interaural attenuation at every test. This next-generation automated masking audiogram would then no longer rely on rules of thumb adopted from evaluating interaural attention in small numbers of individuals in the distant past. Additional bone conduction data may be required to properly select dynamic masking levels, but total maskers delivered would decrease over the method presented here, and the unpredictability of the masker presentations would cement a difficult if not impossible test procedure to thwart.

Bone conduction has been presented in this study as the source of a phenomenon that can confound accurate air conduction threshold estimation under some conditions. Bone conduction thresholds are useful in their own right, however, to aid in the differential diagnosis between sensorineural and conductive hearing loss. These thresholds for the subset of ears where clinically indicated in the current study can be seen in Figure S2. Adding pure-tone bone conduction threshold estimation to AMLAG would represent another advance toward efficient standardization. Given the demonstrated flexibility of AMLAG, doing so would be straightforward. It is possible that new transducer configurations may be needed in this case, though AMLAG may prove able to compensate for hardware limitations with an advanced software implementation.

Unmasked AMLAG includes the ability to estimate the hearing thresholds of both ears simultaneously through bilateral testing (Heisey et al. 2018; Barbour et al. 2019a). Adding air conduction masking to this procedure is straightforward. In fact, three normal hearing participants were recruited in this study to demonstrate the feasibility of masked bilateral AMLAG. All three participants were given an unmasked and a masked bilateral AMLAG. The average mean signed difference of 0.10 dB and average mean absolute difference of 3.0 dB between unmasked and masked bilateral AMLAG (Table S5) are similar to the differences seen between unmasked and masked unilateral AMLAG (Heisey et al. 2018). Additionally, the mean tone count and time to reach threshold estimates within 5 dB of the final estimate for both ears were 19 ± 25 tones and 2.1 ± 2.7 minutes for unmasked bilateral AMLAG, and 23 ± 20 tones and 2.5 ± 2.2 minutes for masked bilateral AMLAG (Table S6). Accurate and efficient bilateral estimation of air conduction thresholds for normal hearing individuals under conditions of dynamic masking suggests the successful extension of masked bilateral AMLAG to participants with symmetric or asymmetric hearing loss.

CONCLUSIONS

The incorporation of automatic dynamic masking into AMLAG demonstrates the versatility of active machine learning diagnostic procedures. AMLAG finds hearing thresholds so rapidly, most patients will never know they are taking a masked test because all the masking noise will fall below their detection thresholds. For the patients with asymmetric hearing, however, masked AMLAG delivers true thresholds much more quickly than conventional techniques and in about the same time as unmasked AMLAG would require to estimate thresholds potentially contaminated with cross hearing. Machine learning audiometry therefore has great potential to enhance patient care by simultaneously standardizing a challenging clinical procedure and optimizing both clinician and patient time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Jan-Willem Wasmann for comments on a previous version of this manuscript.

Funding for this project was provided by NIH grant UL1 TR002345 and NSF grant DGE-1745038. K. L. H. and A.M.W. wrote the article, K. L. H., A. M. W. and J. M. A. designed and conducted the experiments; K. L. H., K. X.,

and D. L. B. analyzed the data. All authors discussed the results and implications and commented on article revisions.

D. L. B. has a patent pending on technology described in this article and has equity ownership in Bonauria, LLC. The authors have no other disclosures.

References

- American Speech-Language-Hearing Association (2005). Guidelines for manual pure-tone threshold audiometry. Available at: [paper/Guidelines-for-manual-pure-tone-threshold/2f88a70daa6a64c70b71a072e46de8e0d2bbc4a1](https://www.speech-hearing.org/paper/Guidelines-for-manual-pure-tone-threshold/2f88a70daa6a64c70b71a072e46de8e0d2bbc4a1) [Accessed October 5, 2019].
- Barbour DL, DiLorenzo JC, Sukesan KA, et al. (2019a). Conjoint psychometric field estimation for bilateral audiometry. *Behav Res Methods*, 51, 1271–1285. [PubMed: 29949072]
- Barbour DL, Howard RT, Song XD, et al. (2019b). Online Machine Learning Audiometry. *Ear Hear*, 40, 918–926. [PubMed: 30358656]
- Bland JM, Altman DG (1999). Measuring agreement in method comparison studies. *Stat Methods Med Res*, 8, 135–160. [PubMed: 10501650]
- Brännström KJ, Lantz J (2010). Interaural attenuation for Sennheiser HDA 200 circumaural earphones. *Int J Audiol*, 49, 467–471. [PubMed: 20482294]
- Carhart R, Jerger J (1959). Preferred method for clinical determination of pure-tone thresholds. *Journal of Speech & Hearing Disorders*, 24, 330–345.
- Denes P, Naunton RF (1952). Masking in Pure-tone Audiometry. *Proc R Soc Med*, 45, 790–794. [PubMed: 13003999]
- Edgerton BJ, Klodd DA (1977). Occlusion effect in bone conduction pure tone and speech audiometry. *J Am Audiol Soc*, 2, 151–158. [PubMed: 845070]
- Gumus NM, Gumus M, Unsal S, et al. (2016). Examination of Insert Ear Interaural Attenuation (IA) Values in Audiological Evaluations. *Clin Invest Med*, 39, 27507. [PubMed: 27917798]
- Hamil TA (2016). Making Masking Manageable, North Charleston, SC: CreateSpace Independent Publishing Platform.
- Heisey KL, Buchbinder JM, Barbour DL (2018). Concurrent Bilateral Audiometric Inference. *Acta Acustica united with Acustica*, 104, 762–765.
- Ho ATP, Hildreth AJ, Lindsey L (2009). Computer-assisted audiometry versus manual audiometry. *Otol. Neurotol*, 30, 876–883. [PubMed: 20179426]
- Hood JD (1960). The principles and practice of bone conduction audiometry: A review of the present position. *The Laryngoscope*, 70, 1211–1228. [PubMed: 13715685]
- Hughson W, Westlake HD (1944). Manual for program outline for rehabilitation of aural casualties both military and civilian. *Transactions of the American Academy of Ophthalmology & Otolaryngology*, 48, 1–15.
- Killion M, Wilber L, Gudmundsen G (1985). Insert earphones for more interaural attenuation. *Hearing Instruments*, 36, 1–2.
- Landry JA, Green WB (1999). Pure-Tone Audiometric Threshold Test-Retest Variability in Young and Elderly Adults. *Journal of Speech-Language Pathology & Audiology*, 23, 74–80.
- Mahomed F, Swanepoel DW, Eikelboom RH, et al. (2013). Validity of automated threshold audiometry: a systematic review and meta-analysis. *Ear Hear*, 34, 745–752. [PubMed: 24165302]
- Martin FN (1980). The masking plateau revisited. *Ear Hear*, 1, 112–116. [PubMed: 7372016]
- Martin FN, Armstrong TW, Champlin CA (1994). A Survey of Audiological Practices in the United States. *Am J Audiol*, 3, 20–26. [PubMed: 26661604]
- Martin FN, Blosser D (1970). Cross hearing—air conduction or bone conduction. *Psychon Sci*, 20, 231–231.
- Martin FN, Champlin CA, Chambers JA (1998). Seventh survey of audiometric practices in the United States. *J Am Acad Audiol*, 9, 95–104. [PubMed: 9564671]
- Masterson EA, Tak S, Themann CL, et al. (2013). Prevalence of hearing loss in the United States by industry. *Am. J. Ind. Med*, 56, 670–681. [PubMed: 22767358]

- Mello L.A. de, Silva R.A.M. da, Gil D, et al. (2015). Test-retest variability in the pure tone audiometry: comparison between two transducers. *Audiology - Communication Research*, 20, 239–245.
- Munro KJ, Contractor A (2010). Inter-aural attenuation with insert earphones. *Int J Audiol*, 49, 799–801. [PubMed: 20735285]
- Sanders JW, Rintelmann WF (1964). Masking in audiometry. A clinical evaluation of three methods. *Arch Otolaryngol*, 80, 541–556. [PubMed: 14216015]
- Shojaemend H, Ayatollahi H (2018). Automated Audiometry: A Review of the Implementation and Evaluation Methods. *Healthc Inform Res*, 24, 263–275. [PubMed: 30443414]
- Sklare DA, Denenberg LJ (1987). Interaural attenuation for tubeophone insert earphones. *Ear Hear*, 8, 298–300. [PubMed: 3678645]
- Smith CR (1968). Clinical masking during pure tone audiometry. *Arch Otolaryngol*, 88, 169–170. [PubMed: 5662928]
- Song XD, Garnett R, Barbour DL (2017). Psychometric function estimation by probabilistic classification. *J. Acoust. Soc. Am*, 141, 2513. [PubMed: 28464646]
- Song XD, Wallace BM, Gardner JR, et al. (2015). Fast, Continuous Audiogram Estimation Using Machine Learning. *Ear Hear*, 36, e326–335. [PubMed: 26258575]
- Stuart A, Stenstrom R, Tompkins C, et al. (1991). Test-retest variability in audiometric threshold with supraaural and insert earphones among children and adults. *Audiology*, 30, 82–90. [PubMed: 1877901]
- Studebaker GA (1964). Clinical masking of air- and bone-conducted stimuli. *J Speech Hear Disord*, 29, 23–35. [PubMed: 14122665]
- Swanepoel DW, Mngemane S, Molemong S, et al. (2010). Hearing assessment-reliability, accuracy, and efficiency of automated audiometry. *Telemed J E Health*, 16, 557–563. [PubMed: 20575723]
- Turner RG (2004a). Masking redux. I: An optimized masking method. *J Am Acad Audiol*, 15, 17–28. [PubMed: 15030099]
- Turner RG (2004b). Masking redux. II: A recommended masking protocol. *J Am Acad Audiol*, 15, 29–46. [PubMed: 15030100]
- Valente M (2009). *Pure-Tone Audiometry and Masking*, Plural Publishing.
- Yacullo W (2015). Clinical masking In Katz J, ed. *Handbook of clinical audiology*. (pp. 77–111). Philadelphia, PA: Wolters Kluwer Health.

Figure 1:

Bland-Altman plots at the 6 frequencies of threshold comparison for unmasked AMLAG versus masked AMLAG in Group NL. Mean signed difference (MSD) in dB is indicated numerically and by a horizontal dashed line in each plot. Limits of agreement (LOA) in dB is indicated numerically and by 2 horizontal dotted lines in each plot. LOA is computed as $1.645 \times$ the standard deviation of the signed differences, reflecting the central 90% of the estimated distribution. Ordinate range is matched to the Group NL data at all frequencies and is identical for all plots; abscissa domains are matched to the Group NL data at each frequency.

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

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

Figure 4:

Bland-Altman plots at the 6 frequencies of threshold comparison for HWAG versus masked AMLAG in Group HL-LA. Ordinate range is matched to the Group HL-LA data at all frequencies and is identical for all plots; abscissa domains are matched to the Group HL-LA data at each frequency. Otherwise, plot details are identical to Figure 1.

Figure 5:

Bland-Altman plots at the 6 frequencies of threshold comparison for HWAG versus masked AMLAG in Group HL-HA. Ordinate range is matched to the Group HL-LA data at all frequencies and is identical for all plots; abscissa domains are matched to the Group HL-LA data at each frequency. Otherwise, plot details are identical to Figure 1.

Figure 6:

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

Figure 7:

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

Table 1:

Average number of tones and minutes required to achieve threshold estimates for each participant, Group NL (N = 9 participants)

	Mean Tone Count	Mean Number of Minutes
Group NL participants		
Unmasked AMLAG	37	4.0
Masked AMLAG	34	3.7

Table 2:

Masking noise above non-test ear threshold, Group NL (N = 18 ears)

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

Table 3:

Average number of tones and minutes required to achieve threshold estimates for each participant, Group HL

Group	Test	N	Mean Tone Count	Mean Minutes	Mean Tone Count: Air Conduction Only	Mean Minutes: Air Conduction Only
HL-LA	Manual HWAG	10	127	10.7	93	6.9
HL-LA	Masked AMLAG	10	63	6.9	63	6.9
HL-HA	Manual HWAG	8	186	18.5	114	9.9
HL-HA	Masked AMLAG	8	50	5.4	50	5.4

Table 4:

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

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