J Am Acad Audiol 7: 125–129 (1996)

Evaluation of a Maximum Likelihood Procedure for Measuring Pure-tone Thresholds Under Computer Control

Craig Formby*
LaGuinne P. Sherlock*
David M. Green†

Abstract

An adaptive, maximum likelihood (ML) procedure was assessed as an automated tool for estimating audiometric pure-tone thresholds in the clinic under computer control. Pure-tone air-conduction thresholds were measured from 101 workmen who received annual hearing rechecks as part of their employee hearing conservation program. A pure-tone threshold was measured bilaterally for each of the standard audiometric frequencies in a 15-trial block to yield 60 percent correct detection with the ML procedure. The workmen were tested on a modified "yes-no" task. On a trial, the signal was presented in a visually cued 200-msec observation interval. Each workman then had 1000 msec to make a "yes" response. If the workman did not respond during the 1000-msec response period, then the computer assumed a "no" response. After either the "yes" or "no" response, the computer adjusted the signal level for the next trial. The thresholds measured by ML procedure compared favorably with thresholds measured from the same listeners by conventional (CONV) audiometry. The efficiency of the ML procedure was also compared in terms of the time necessary for an experienced audiologist to instruct the listener and perform CONV audiometry. CONV audiometry (3–4 minutes per listener) required about half of the time needed for the ML procedure (6–7 minutes per listener). The relatively longer time associated with measuring an audiogram with the ML procedure was due primarily to more trials being used to estimate threshold.

Key Words: Computerized audiometry, maximum likelihood procedure

Conventional (CONV) audiometry is effectively performed as a yes-no task in which the audiologist adjusts the signal level adaptively to obtain a 50 percent correct detection response at each test frequency. The listener responds "yes" only if he or she believes that the test signal was presented by the audiologist. Otherwise, the listener makes no response. For CONV audiometry, the observation interval is undefined for the listener and the timing of signal presentation is random.

*Division of Otolaryngology — Head and Neck Surgery, Department of Surgery, University of Maryland School of Medicine, Baltimore, Maryland; †Psychoacoustic Laboratory, Department of Psychology, University of Florida, Gainesville, Florida

Reprint requests: C. Formby, Division of Otolaryngology — Head and Neck Surgery, Department of Surgery, University of Maryland School of Medicine, 419 W. Redwood Street, Suite 360, Baltimore, MD 21201

An alternative to CONV audiometry is an automated approach that uses the maximum likelihood (ML) principles described by Green (1993) to estimate a signal level corresponding to a 60 percent correct detection, sensitivity index (60%). The ML method is conceptually similar to CONV audiometry in that an adaptive psychophysical procedure is used with a single-interval, yes-no task to estimate pure-tone thresholds. The ML test, however, is controlled by computer rather than by an audiologist.

The ML procedure, as implemented in this report, estimated a detection threshold from the 60 percent correct value of the psychometric function yielding the maximum probability for a given signal frequency. On a block of trials, the ML procedure selected this threshold condition from one of 240 possible psychometric functions. The set of functions spanned a range of 60 dB, in 1-dB increments, for each of four false alarm ("guessing") probabilities: 0.0, 0.1, 0.2, and 0.3.

The principles of the ML procedure may be understood in terms of an example from Green (1993), which supposes a limited set of only four psychometric functions. The four functions shown in Figure 1 differ in sensitivity: curves A and C reflect a higher proportion of "yes" responses at lower intensity levels than do curves B and D. The false alarm rates also differ among the four functions. The false alarm rate associated with each function reflects the proportion of "yes" responses to signals at very low presentation levels and typically will be measured about 20 to 30 dB below the estimated threshold value for a given threshold function. In Figure 1, curves C and D have false alarm proportions of about 0.2, whereas curves A and B have negligible false alarm proportions.

Each of the four psychometric functions shown in Figure 1 is a logistic of the form

$$P(\text{yes}) = \alpha + (1-\alpha)\ \frac{1}{1 + e^{-k(x-m)}}, \quad (1)$$

where x is the signal level in decibels, m is the mean of the logistic (the point where it is equal to 0.5), and k is the slope of the logistic. The logistic is similar in form to a cumulative normal function and has the advantage of being easier to compute than the cumulative normal function. These four functions differ in the value of alpha (α), the false alarm parameter, and the threshold value, m. They all have the same value of k. Green (1993) has shown that the value of the constant k is not very critical in the ML procedure.

We can test whether or not one of the curves in Figure 1 represents the true psychometric function for a hypothetical listener with the sequence of stimulus levels from three successive trials shown in Table 1. For example, starting on trial 1, the listener responds (a "yes" response) to the stimulus level presented at +10 dB. The probability of this response is computed for each psychometric function and entered in the first row of Table 1. Function C yields the highest probability (**.9980**) on the first trial. The associated value of S(60%) for function C in Figure 1 is –2 dB, which then is the stimulus level presented on the second trial. On trial 2, the response is "no." The probability entries for the four functions are again recorded in Table 1 and are shown in the second row. The probability of a "no" response is simply 1 minus the probability of a "yes" response. The third row of Table 1, labeled PROD, reflects the product of the preceding probabilities associated with trials 1 and 2, which are assumed to be independent for

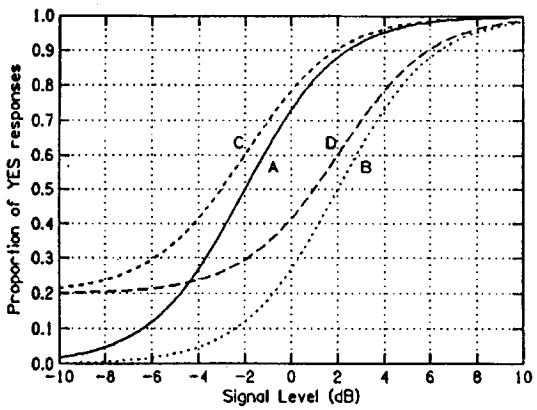

Figure 1 The main principles underlying the ML procedure for estimating pure-tone thresholds can be understood with reference to a limited set of four psychometric functions, A, B, C, and D. These four psychometric functions represent different proportions of "yes" responses to a signal as a function of the signal level. (From Green [1993]. Reprint with permission.)

each function. The maximum probability (.8650) is the PROD value for function B, which has a corresponding level for S(60%) = +3 dB in Figure 1. On the third trial, the signal level presented is therefore +3 dB. A "yes" response occurs on trial 3, and the probabilities for all four functions are again entered in Table 1. The probabilities correspond to the proportion of "yes" responses shown in Figure 1 for each function at +3 dB. The product is again computed, and the maximum probability (.5384) is still that associated with function B. If we had only the results from these three trials, then we would conclude that function B has the maximum likelihood of being the one used by our hypothetical listener at that point in the experiment. The final threshold estimate is therefore based on the signal level coinciding with the 60 percent value of the psychometric function that yielded the maximum probability. The 60 percent value is used because it is nearly the midpoint for the psychometric functions having false alarm rates over the range from α = 0.0 to 0.3. This range of α values is that considered previously by Green (1993) and corresponds to the range of α values evaluated in this study.

Thus, the ML procedure differs from CONV audiometry in that a computer applies rigorous statistical rules to select and control the signal presentation level on each trial based on the listener's response history. Moreover, an estimate of the listener's false alarm rate is also

Table 1 Hypothetical Sequence of Three Trials Leading to ML Estimate of Correct Psychometric Functions*

Trial	*Stimulus Level*	*Response*	*A*	*B*	*C*	*D*
1	+10 dB	Yes	.9975	.9820	**.9980**	.9856
2	-2 dB	No	.5000	.8808	.4000	.7046
	PROD		.4988	**.8650**	.3992	.6945
3	+3 dB	Yes	.9241	.6225	.9393	.6980
	PROD		.4609	**.5384**	.3750	.4847

*Adapted from Green (1993).

obtained at signal levels far below the listener's detection threshold and is available to the audiologist. In this study, we compared pure-tone thresholds measured by CONV audiometry and by the ML procedure. Our purpose was to assess the validity, efficiency, and reliability of the ML method as a clinical procedure with reference to CONV audiometry.

METHOD

Pure-tone air-conduction thresholds were measured as part of an annual employee hearing conservation program for 101 men (mean age = 43 years) with a history of occupational noise exposure. Thresholds were measured in two ways: (1) by an audiologist using a modified Hughson-Westlake (Hughson and Westlake, 1944) "up 5 dB — down 10 dB" CONV method (Carhart and Jerger, 1959) for the octave frequencies 250 to 8000 Hz, including 3000 and 6000 Hz per ASHA guidelines (1978), and (2) by a computer-controlled ML method, at the octave frequencies 250 to 8000 Hz. CONV thresholds were measured with a conventional audiometer (Madsen, model OB822) and headphones (Telephonics, type TDH-39) in a sound-attenuating booth. Pure-tone signals for the ML method were generated by a digital-to-analog converter (DAC) (TDT, model Quikki QDA1). The DAC output was low-pass filtered (TDT, model FLT2) below 7.5 kHz to prevent aliasing. The filtered signal was attenuated by programmable attenuators (TDT, model PA3) and presented through headphones (Telephonics, type TDH-39) in a sound-attenuating booth.

In the ML method, a threshold for each standard frequency was measured in a 15-trial block to yield 60 percent correct detection. On a trial, a 200-msec pure-tone signal was presented in a visually cued 200-msec observation interval. The signals included 10-msec rise-fall times as part of the nominal durations. Subjects had 1000 msec to make a "yes-only" response, which attenuated the signal level. If the subject did not respond during the 1000-msec response period, then the computer assumed a "no" response for the trial, and the signal level was increased according to the ML algorithm. This modified yes-no method differed from that reported by Green (1993), who required his listeners to make a keyboard entry to record a "no" response, and used 12 rather than 15 trials for each ML threshold estimate. The active "no" response was eliminated in this study to accelerate the threshold estimate process and to reduce confusion among the workmen.

To evaluate the efficiency of the CONV and ML methods, the time required to instruct each subject and complete the threshold measurement was recorded for each method. Thresholds were also measured by both methods from 20 subjects (n = 39 ears) on two occasions, separated by approximately 1 year, to assess test–retest reliability.

RESULTS AND DISCUSSION

In preliminary testing, thresholds were measured with the CONV and ML procedures from the right and left ears of 55 subjects. There were no significant differences at any test frequency between the right-ear and left-ear thresholds. Thus, the data to be reported here were combined for the right and left ears for all subsequent statistical analyses.

Figure 2 shows a comparison of the CONV and ML thresholds averaged for 202 ears as a function of test frequency. The threshold differences between the two methods were not significant statistically at any test frequency except 250 Hz ($t = 2.54$; $p < .01$), where the ML threshold was higher, but was within 3 dB of the CONV threshold. Thus, the ML thresholds agreed well with the "gold standard" CONV thresholds. These results are consistent with the pure-tone thresholds that Green (1993) measured with 12 ML trials and with a traditional yes-no procedure

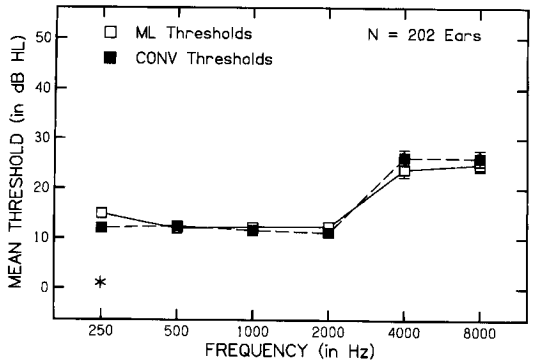

Figure 2 Comparison of CONV and ML group mean thresholds (and standard error bars) as a function of test frequency averaged for 202 ears of 101 listeners. Asterisk (*) indicates statistical difference at .05 significance level for 250 Hz.

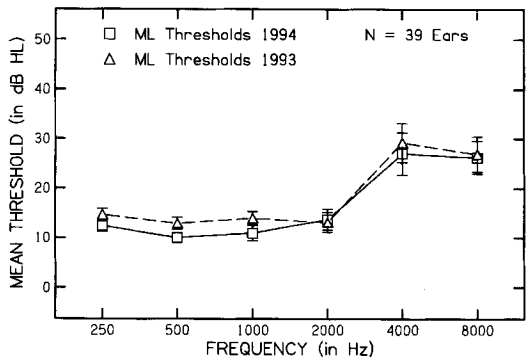

Figure 3 ML group mean thresholds (and standard error bars) as a function of test frequency for 39 ears of 20 listeners compared in 1993 and 1994.

for 13 "trained" listeners. The ML thresholds reported by Green (1993) were about 10 and 5 dB higher at 250 and 500 Hz, respectively, than comparable International Standards Organization (1964) recommended thresholds. We suspect that the higher thresholds measured here and by Green (1993) at the low audiometric frequencies are primarily due to the relatively brief 200-msec signal durations used for the ML procedure. Low-frequency sinusoids require a longer integration time than high-frequency sinusoids (Plomp and Bouman, 1959; Watson and Gengel, 1969), and 200-msec signal durations may not have been sufficiently long for complete temporal integration at the lower audiometric frequencies.

Figure 3 shows the test–retest ML thresholds as a function of test frequency averaged for 39 ears of 20 listeners. There were no significant test–retest differences at any test frequency. Thus, ML reliability was very good over the period of 1 year for these listeners. Figure 4 shows comparable thresholds for CONV audiometry. Test–retest reliability was equally good for CONV audiometry, and no significant test–retest differences were found at any test frequency.

The average test times, as timed with a stopwatch by a single clinician who administered both audiometric methods to all workmen, revealed that the CONV method was significantly quicker ($t = 30.54$; $p < .0001$), requiring 3 minutes and 46 seconds (SD = 55.2 sec) versus 6 minutes and 43 seconds (SD = 27.6 sec) for the ML method. This difference was primarily associated with the larger number of trials required to measure an audiogram with the ML procedure (15 trials × 12 test frequencies = 180 ML trials), compared with the CONV procedure (8–10 trials × 14 test frequencies ~ 126 CONV trials). A secondary time factor, perhaps adding 30 seconds more to the ML test, can be attributed to the extra time needed for instruction of the ML method.

Finally, an analysis of the ML false alarm proportions for 97 workmen revealed that 44 of these men (i.e., 45%) had α values of 0 at all 12 test frequencies. False alarm proportions for the other 53 workmen were inconsequential. Across the 1164 test conditions, which were produced by all 97 listeners for the 12 test frequencies, only 94 α values were nonzero. This represents nonzero α values for only 8 percent of the 1164 conditions, and only 15 of these α values were as large as 0.3. Thus, the false alarm rates were negligible in our measurements of pure-tone thresholds by the ML method. It is important to point out, however, that precise estimates of false alarm rates probably cannot be assured with only 15 ML trials per test frequency (Green, 1993). In this study, we did not use catch trials to estimate false alarm rates, but rather we estimated α from the asymptotic function at very low signal levels, which typically were 20 to 30 dB below the final threshold estimates. Gu and Green (1994) recently showed that, by adding catch trials with the signal trials, they could improve the estimate of α. Thus, it is possible to obtain a better estimate of the false alarm rate, but at the cost of more ML trials.

The virtues of the ML procedure are significant and highly desirable now for some audiometric applications. The ML procedure is probably best suited for large-scale, routine audiometric checks, which are common in the

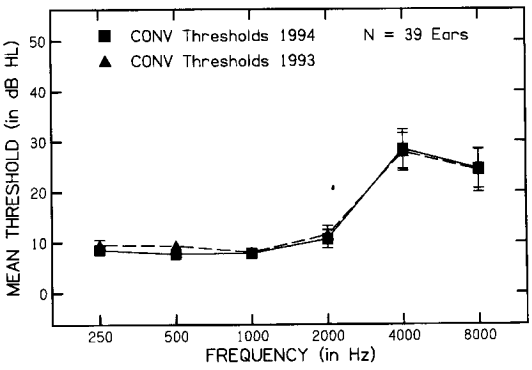

Figure 4 CONV group mean thresholds (and standard error bars) as a function of test frequency for 39 ears of 20 listeners compared in 1993 and 1994.

military and in industrial hearing conservation programs. For these applications, the advantages of the ML method may offset the longer test time. These advantages include (1) automated control of signal presentation level and frequency, and of response collection and analysis by computer; (2) visually defined observation intervals; (3) an estimate of the listener's false alarm rate; (4) a rigorous and consistent theoretical basis for selecting signal presentation level on each trial and for estimating detection threshold; and (5) active participation by the listener in a self-test task.

It is inevitable that as computers become more widely available in clinics, the ML procedure and related automated methodologies will find their place in audiology. The ML method, even in its present form, reflects an important step toward an efficient and reliable automated clinical audiometric procedure. We are confident that, with further study, we can optimize the efficiency of the ML procedure, either by reducing the number of trials per threshold estimate or the 1000-msec response period. Other refinements may include longer (>200 msec) signal interval presentations at low audiometric frequencies to assure complete temporal integration of these signals.

SUMMARY

Pure-tone air-conduction thresholds were measured for the standard audiometric frequencies by CONV audiometry and by a computer-controlled ML method. CONV and ML thresholds averaged across 202 ears differed by less than 3 dB at all test frequencies. Test–retest thresholds measured on two occasions separated by about 1 year were not significantly different at any test frequency for 20 listeners. CONV audiometry required about half of the time needed for the ML procedure, which used about twice as many trials as CONV audiometry to estimate each threshold.

Acknowledgments. These data were collected while C. Formby and L. P. Sherlock were on staff at The Johns Hopkins University School of Medicine, and the results were presented at the XXII International Congress of Audiology, Halifax, Nova Scotia, July, 1994. Partial support for this study and for preparation of this manuscript was provided by NIH award R01 DC00951. We gratefully acknowledge M. Johnson, who provided editorial assistance.

REFERENCES

American Speech and Hearing Association, Committee on Audiometric Evaluation. (1978). Guidelines for manual pure-tone threshold audiometry. *ASHA* 20:297–301.

Carhart R, Jerger J. (1959). Preferred method for clinical determination of pure-tone thresholds. *J Speech Hear Disord* 24:330–345.

Green DM. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *J Acoust Soc Am* 93:2096–2105.

Gu X, Green DM. (1994). Further studies of a maximum-likelihood yes-no procedure. *J Acoust Soc Am* 96:93–101.

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

International Standards Organization. (1964). *Standard Reference Zero for the Calibration of Pure-tone Audiometers. ISO Recommendation R389.* New York: American National Standards Institute.

Plomp R, Bowman MA. (1959). Relation between hearing threshold and duration for tone pulses. *J Acoust Soc Am* 31:749–758.

Watson CS, Gengel RW. (1969). Signal duration and signal frequency in relation to auditory sensitivity. *J Acoust Soc Am* 46:989–997.