Automated Electrophysiologic Hearing Testing Using a Threshold-Seeking Algorithm

Özcan Özdamar*
Rafael E. Delgado†
Rebecca E. Eilers‡
Richard C. Urbano§

Abstract

The efficacy of utilizing an automated algorithm to identify auditory brainstem responses (ABR) was studied. A microcomputer-based threshold-seeking algorithm utilizing click-evoked ABR was developed to determine evoked-response thresholds for automated hearing screening. The software consists of an evoked-response recognizer unit, which determines the presence or absence of a response, and a threshold-tracking unit, which controls the click intensity in order to track the threshold. The response recognizer is based upon correlation methods. Threshold tracking is accomplished using a Parameter Estimation by Sequential Testing (PEST) procedure, which is commonly used to study psychophysical properties of the auditory system. Sound level is automatically adjusted, based on the results of the recognizer and the threshold tracker. Test results were generally obtained in less than 15 minutes per ear. The results of the automated procedure correlate very highly with expert judgments of ABR threshold and show good test–retest reliability, suggesting that automated procedures are viable alternatives to traditional testing methods.

Key Words: Auditory brainstem response (ABR), electrophysiologic hearing testing, evoked potentials, hearing screening

Recently, hearing screening and testing of newborns and infants using electrophysiologic signals have gained widespread acceptance and have come to be preferred over other, less accurate techniques. The early portion of auditory evoked potentials called auditory brainstem response (ABR) has become the clinical standard of electrophysiologic hearing testing and has been the technique of choice for hearing screening in newborns (Hyde, Riko and Maliza, 1990; Jacobson, Jacobson, and Spahr 1990; Joint Committee on Infant Hearing, 1990). ABR reflects synchronous firings of the auditory nerve and brainstem pathways that are often associated with hearing sensitivity. ABR provides highly specific and sensitive test measures for purposes of hearing screening and testing (e.g., Stein, 1984; Hyde et al, 1990; Jacobson et al, 1990; Shimizu, Walters, Proctor, Kennedy, Allen, and Markowitz, 1990). Clinical use of ABR devices for both screening and threshold determination is limited, due to the high cost of programs both in terms of instrumentation and professional personnel necessary for implementation. Most current devices require clinical monitoring of patients during testing and customized expert parameter selection for acquisition of noise-free recordings.

Recently, the National Institutes of Health convened a consensus development conference on early identification of hearing impairment in infants and young children (NIH, 1993). The consensus panel recommended the adoption of universal hearing screening for newborns. Recognizing that universal screening would be expensive utilizing current technology, the

*Departments of Biomedical Engineering, Pediatrics, and Otolaryngology, University of Miami; †Intelligent Hearing Systems Corporation, Miami, Florida; ‡Departments of Psychology, Pediatrics, and Otolaryngology, University of Miami; §Departments of Pediatrics and Psychology, University of Miami, Coral Gables, Florida

Reprint requests: Özcan Özdamar, Department of Biomedical Engineering, University of Miami, P.O. Box 248294, Coral Gables, FL 33124

panel urged further research on new techniques for hearing assessment, including the development of automated procedures.

One method of facilitating the use of ABR for purposes of universal screening is to develop automated algorithms that require minimal human intervention. Automation can be achieved at various levels for different goals and applications. The simplest level of automation can be achieved by obtaining some requisite number of waveforms at specified levels followed by an analysis of pass/fail criteria for each stimulus level independently. Such a procedure can be used to classify responses into two or three categories: pass, fail, or could not test. A higher level of automation can be achieved if an algorithm is designed that determines the ABR threshold by properly changing stimulus level, detecting responses, and properly terminating the test when threshold is determined. This second level of automation requires the implementation of a tracking algorithm. Even higher levels of automation can be achieved by analyzing waveforms produced by different level stimuli in a context-sensitive manner, labeling waveforms, and extracting information such as latency-intensity functions for further diagnostic purposes.

Many techniques for achieving the first level of automation have been proposed (for a review, see Özdamar, Delgado, Eilers, and Widen, 1990; Dobie, 1993). Some of the methods that have been proposed are template matching/matched filtering (Thornton and Obenour, 1981; Woodworth, Reisman, and Fontaine, 1983), F-ratio/SNR calculation (e.g., Don, Elberling, and Waring, 1984; Elberling and Don, 1987), phase analysis (e.g., Beagley, Sayers, and Ross, 1979; Greenblatt, Zappulla, Kaye, and Fridman, 1985), signal power (variance) analysis (e.g., Arnold, 1985; Delgado and Özdamar, 1990b), cross-correlation analysis (e.g., Weber and Fletcher, 1980; Arnold, 1985; Özdamar et al, 1990), artificial neural networks (Alpsan and Özdamar, 1992; Özdamar and Alpsan, 1992), and coherence analysis (e.g., Dobie and Wilson, 1989). In spite of the multitude of proposed detection methods, only one device (Algo) has been commercialized for automated screening (Thornton and Obenour, 1981; Peters, 1986). This device was designed for hearing screening purposes only (classification of the test population into pass/refer categories) and has been used clinically (Kileny, 1987; Jacobson et al, 1990).

While the abundance of research has been directed toward the development of level one response detection, little research has been directed towards the development of level two automated threshold-finding algorithms. Recent studies show that machine detection of ABR thresholds can be nearly as accurate as human detection when off-line strategies are used (Don et al, 1984; Mason, 1984; Arnold, 1985; Elberling and Don, 1987; Özdamar, Delgado, Miskiel, Eilers, and Widen, 1987b; Özdamar et al, 1990). Although threshold seeking has seldom been a goal of screening, largely because threshold determination is typically a lengthy procedure, threshold information that could be obtained from a rapid, automated test would be preferable to information obtained from screening. In principle, a level two automation procedure could be valuable in screening, because it could be used to classify the test population into diagnostically significant risk categories such as those associated with normal, mild, moderate, or severe hearing loss. When effectively used, such information could reduce the number of clients who fall into the "fail" or "refer" groups by differentiating between potential losses that are likely due to middle ear function and those of greater immediate concern. If level three automation with an automated wave identification and latency-intensity determination algorithm (Delgado and Özdamar, 1990a) is added to threshold determination, then the procedure could be used for automated, rapid diagnostic testing as well.

In order to implement level two automation, a tracking procedure must be used. Adaptive threshold-tracking algorithms similar to those used in psychophysical research can be implemented to achieve reduced time for closed-loop operation and to obviate the need for professional intervention during testing. One adaptive tracking procedure, Parameter Estimation by Sequential Testing (PEST) (Taylor and Creelman, 1967), is widely used in auditory psychophysics. The PEST procedure generally starts at an intensity in the middle of the expected range and operates by halving of the step size upon reversal in the direction of the level of the stimulus. When the step size diminishes to a size smaller than a preset value, the procedure is terminated. PEST has been introduced to track evoked-response thresholds in humans (Özdamar et al, 1987b, 1990) and animals (Salvi, Ahroon, Saunders, and Arnold, 1987). These studies showed that PEST, when coupled with a response-detection algorithm, can be successfully used in electrophysiologic response determination.

In a previous study, we developed a computer simulation method to test various ABR recognition and threshold-tracking procedures for on-line use (Özdamar et al, 1990). Two response-analysis methods (amplitude variance ratio and cross-correlation) and three threshold-tracking methods (clinical, Bekesy, and PEST) were studied. Off-line simulations based on data obtained from 15 subjects showed that, on the average, both response-recognition methods gave acceptable results. Thresholds obtained with the variance ratio method, however, resulted in greater variability. The on-line use of both methods, however, resulted in significant differences in test efficiency, due to differences in tracking methods. The PEST method appeared to be the most efficient tracking method, with the clinical method second in efficiency. These results showed the feasibility of an on-line ABR threshold-seeking system for automated, electrophysiologic hearing testing.

This paper reports the development of a level two, automated, on-line algorithm implemented in a dedicated device for hearing testing with auditory brainstem responses. The device operates in real time to achieve a desired measurement goal and obviates the need for human intervention for: (1) determining stimulus levels by using a tracking procedure; (2) discarding responses of undesired origin; (3) recognizing the presence of a response; and (4) analyzing the overall results to predict hearing thresholds with accuracy and efficiency.

SYSTEM CONFIGURATION FOR THRESHOLD DETERMINATION

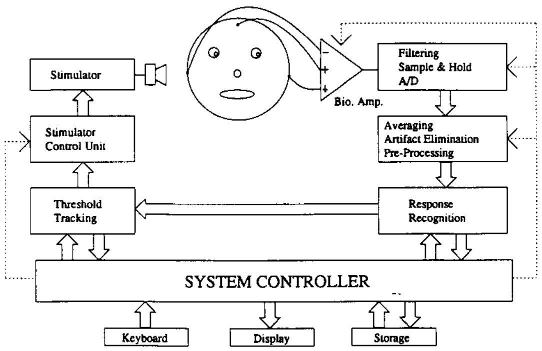

Figure 1 Functional system block diagram. Information flow during automated use is shown by thick arrows. Interrupted line arrows show information flow for parameter setting before initiating the automated operation.

METHOD

Hardware

The automated algorithm is implemented in a microcomputer (IBM PC AT) equipped with a digital signal processor (40 MHz TMS320C25 DSP chip) and a digital-sound stimulator. The system is controlled by an 80286 microprocessor using interactive, menu-based software designed for efficient use of dual processors (Özdamar, Kaplan, Miskiel, and Delgado, 1987a).

Brain waves were acquired with a bioamplifier (gain 100,000; filters 100–3000 Hz; 6 dB/oct) and digitized with a 12-bit A/D converter. Sampling rate was set at 40 kHz and 512 post-stimulus data points were recorded. Averaging was done by the DSP board by summing 1024 sweeps in a 32-bit buffer.

ABRs were generated by clicks, which were produced using a 16-bit D/A converter operating at a 10-kHz sampling rate. Rectangular clicks having widths of 100μsec were generated. Attenuation was accomplished using two digital attenuators, each with an 80-dB range.

Software

The functional configuration of the system is shown in Figure 1. The system software is organized into four subunits by a system controller, which provides the user interface. The controller directs the closed loop as a function of the outputs of the response-recognition and threshold-tracking units and directs the stimulator control programs.

Preprocessing and synchronized averaging were accomplished by the DSP program. ABRs were acquired by averaging 1024 sweeps containing 512 post-stimulus data points collected at a sampling rate of 25 μsec. Averaging is done by summing 12-bit data values in a 32-bit buffer. Prior to averaging, each single-sweep response was demeaned in order to prevent any unwanted overflow of the buffer, due to electrode or amplifier dc shifts. An artifact-rejection method based on amplitude level was implemented in the DSP to eliminate single-sweep responses with amplitudes over a pre-established criterion. This was accomplished in software by using a separate buffer area for single sweeps. Averaging was started after the complete sweep was collected and checked for artifacts. In this study, the artifact-rejection level was set at ±10 μV. After averaging, 512 post-stimulus ABR values were transferred to the microcomputer and converted to floating-point format for further processing.

Algorithms

Response Recognition

A method utilizing windowed cross-correlation of consecutive averages was implemented for response recognition (Özdamar et al, 1987b, 1990; Delgado, Özdamar, and Miskiel, 1988). In this method, two separate responses x(n) and y(n) obtained at the same level are cross-correlated within seven overlapping windows, with results R_{xy1}, R_{xy2}, ..., R_{xy7} as follows:

$$R_{xyl} = \frac{\sum_{n=a_i}^{b_i} [(x(n) - m_{xi})(y(n) - m_{yi})]}{\sqrt{\sum_{n=a_i}^{b_i} (x(n) - m_{xi})^2} \times \sqrt{\sum_{n=a_i}^{b_i}}}$$

$i = 1,...,7$ Eq. 1

where m_{xi} mean of x(n) within the ith window; m_{yi} mean of y(n) within the ith window; a_i beginning of the ith window; b_i ending of the ith window.

The window widths of 2 msec were used following optimization studies designed to minimize false positive and false negative errors. The windows were positioned to cover 5 to 10 msec ranges for the detection of wave V at all intensities. Again, a maximum correlation parameter R_{xymax} was selected from the seven outcomes (Eq. 2) and was compared with a level criterion C_{xy} for classification (Eq. 3).

$R_{xymax} = \text{Max}\ \{R_{xyl}\}$ $i = 1,...,7$ Eq.2

$T = +1$ if $R_{xymax} \geq C_{xy}$ for "Response"
$T = -1$ if $R_{xymax} < C_{xy}$ for "No Response" Eq. 3

The criterion C_{xy} was set to 0.7 after optimization studies. An example of response recognition with the windowed cross-correlation method is shown in Figure 2. A maximum cross-correlation outcome of 96.93 percent is obtained from the first window, which is above the criterion, and the recordings are recognized as a "Response."

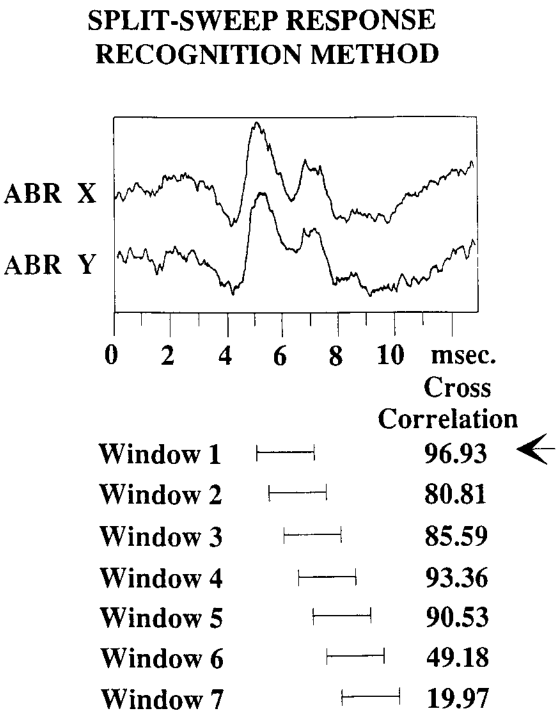

Figure 2 Windowed cross-correlation response-recognition method.

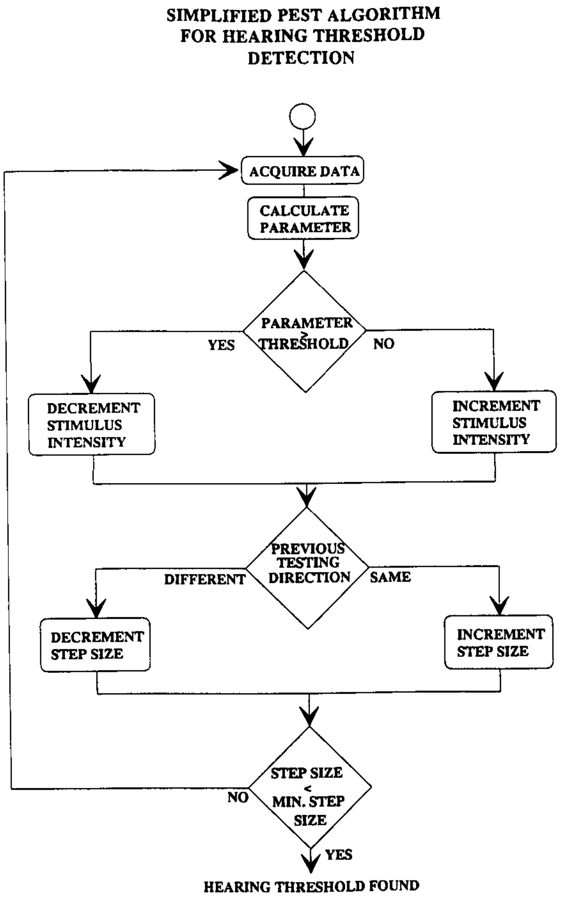

Figure 3 PEST algorithm for hearing threshold determination.

Threshold Tracking

As a result of its success in a previous simulation study (Özdamar et al, 1990), PEST procedure was utilized and adapted for electrophysiologic threshold tracking and detection (Fig. 3). The strength of the PEST procedure is that it rapidly converges on threshold, spending little time at high levels. The initial stimulus level was set to 50 dB HL and the initial step size utilized was 20 dB. A "Response" resulted in a reduced stimulus intensity, while a "No Response" resulted in an increased intensity based on the current step size. The step size was reduced by half, with each reversal in the staircase direction, and doubled, following a failure to detect a response in two consecutive level tests. The process was terminated when the step size became less than 5 dB and the threshold was specified as the level midway between the outcomes of the last two trials.

Figure 4 shows a testing path obtained from a normal subject (top) and the averaged responses obtained at the corresponding levels (bottom). Responses were obtained initially at 50 and 30 dB. No response was obtained at 10 dB. Accordingly, the step size was halved preceding the change in direction of the path. On the ascending series, a response was obtained at 20 dB, the step size was again halved, and the direction of the path changed to descending. A response was then obtained at 15 dB but not at 10 dB. Since step size was already 5 dB, the test was terminated and threshold determined halfway between the last two test levels (i.e., 13 dB HL).

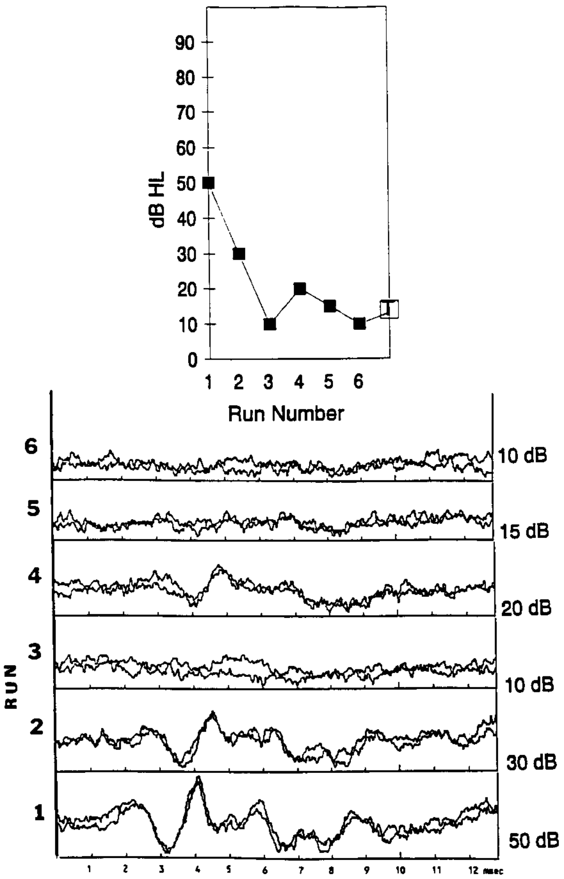

Figure 4 An example of a threshold-tracking path (top) and the corresponding ABR recordings (bottom) from a single test of a normal-hearing subject.

Data Acquisition

Subjects

In order to test the broad applicability of the level two automation approach to threshold seeking, a broad sample of subjects was desired. Accordingly, we tested infants, young children, adolescents, and a range of adults from young adulthood to old age. The procedure was tested in the laboratory with 105 subjects (54 males and 51 females). Of these subjects, 66 were adults (28 males and 38 females) with an age range of 13 to 85 years and a mean age of 34 years. Among these adult subjects, 31 ears had sensorineural hearing losses, while 2 ears had conductive and 2 had mixed losses.

The remaining 39 subjects were infants or young children (26 males and 13 females) with a mean age of 10 months and a range of 11 weeks to 3.6 years. Among the infants, 8 ears had sensorineural hearing loss and 6 had conductive losses.

Of the 105 subjects, 46 subjects (29 adults, 17 infants) were tested using both ears, 56 subjects (35 adults, 21 infants) were tested using one ear, and 3 subjects (2 adults and 1 infant) could not be tested, due to excessive muscle artifacts. In addition, 45 tests (28 adult and 17 infant) were repeated for the purpose of obtaining test–retest reliability. In order to simulate conductive hearing loss, 6 normal adult ears were plugged with foam earplugs and ABR tests were conducted, followed by assessment of behavioral threshold for the plugged ears. Four of these ears were retested, providing 10 tests with plugged ears.

One hundred and thirty-one tests were obtained from adults, including repeated and plugged ear tests. Six of these tests were eliminated, because ABR experts found them too noisy to be used for threshold determination.

Seventy-two tests were obtained from infants, including repeated tests. Four tests were eliminated, because they were too noisy for accurate threshold determination. Altogether, 193 tests were analyzed.

Procedures

Automated ABR testing was conducted in a sound-attenuated room by a technician who did not know how to interpret ABR tests. Adults were tested using either TDH-39 earphones coupled with MX/41 AR cushions or Etymotic ER-3A insert earphones with modified impedance tip adapters. Infants were tested with insert earphones. Alternating clicks were used as stimuli. The peak equivalent SPL for 0 dB HL (in reference to normal hearing level) was 32 dB (reference 3 kHz).

Recording was accomplished using gold-cup electrodes filled with bentonite paste attached to vertex (noninverting), ipsilateral mastoid (inverting), and contralateral mastoid (ground). Impedance of the electrodes was less than 10 kΩ at all times. Testing was accomplished without sedation. Infants were tested in natural sleep, often following feeding. All recordings, along with the duration of tests and the resulting automatically determined thresholds, were automatically saved for later analysis and evaluation by experts.

ABR recordings from all tests were separately evaluated by three experts who had at least 2 years of experience in interpreting ABRs. Two of these experts were auditory electrophysiologists and biomedical engineers (first two authors) and the third was a certified audiologist specializing in ABR testing. The experts were presented with all recordings (containing suprathreshold as well as subthreshold plots) on paper, ordered according to stimulus level. The experts were blind to the level of the automated threshold and did not have access to the nature or degree of the subject's hearing loss, if any.

The experts evaluated the recordings for each test independently and determined a response threshold. If an expert determined that the recordings were too noisy for reliable threshold determination, the test was marked as noisy. The average of the expert-derived thresholds was used to compare and evaluate the results of the automated method. Tests were identified as unreliable if one or more of the experts rated the recordings as too noisy to score and refused to assign a threshold. These tests were eliminated from further evaluation. Tests were also eliminated if experts differed by more than 10 dB on their judgments of threshold, since for these tests a "gold standard" could not be determined. Altogether, 10 tests were eliminated as too noisy or unreliable.

The hearing status of adults was evaluated primarily by pure-tone audiometry (air and bone conducted) obtained within a few days of ABR testing. Adult data consisted of audiograms with thresholds at 0.25, 0.5, 1, 2, 4, and 8 kHz. Impedance and speech audiometry were performed as needed for evaluation of hearing status. Hearing status of ears plugged with foam earplugs was separately evaluated using standard pure-tone audiometry with the earplugs inserted. For four ears, adequate audiograms could not be obtained.

The hearing status of the infants was evaluated when they reached about 6 months of age or, if they were older, within a few days of the administration of the ABR test, using an automated visual reinforcement audiometry device (Widen and Bull, 1984; Widen, 1990) in the sound field. Accordingly, threshold estimates were assumed to apply to both ears, and, in fact, in infants in whom both ears were tested with ABR, expert-derived thresholds for the two ears did not differ by more than 8 dB, with one exception. Hearing was evaluated with the two signals that were available in the device, low-pass filtered speech noise and a 4-kHz narrowband noise. Infants and young children were first screened with these stimuli, followed by a maximum likelihood sequence to determine threshold. For the purpose of this study, threshold information at 4 kHz for 55 tests was available and used for comparison with the ABR results, since the high-frequency information obtained from the behavioral test could be expected to correlate well with click ABR data.

RESULTS

Automated versus Expert ABR Threshold Errors

The automatic thresholds determined by the algorithm were evaluated by comparing them with the expert-derived thresholds. The mean automatic threshold for adults was 22.8 dB HL (SD = 14.8), compared to 23.3 dB HL (SD = 14.0) for experts. The mean automatic threshold for infants was 20.4 dB HL (SD = 20.9), compared to mean expert thresholds of 20.9 dB HL (SD = 17.1). Threshold errors were com-

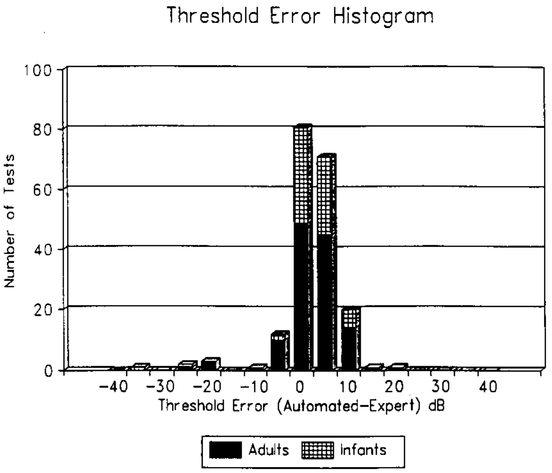

Figure 5 Threshold-error histogram of the automated system, including all tests. All errors are calculated by referring to the expert decisions.

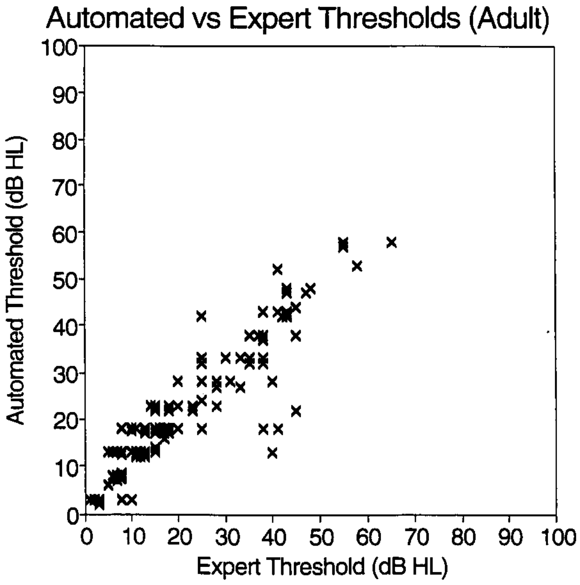

Figure 6 Scattergram of automated versus expert-derived thresholds for adults.

puted by subtracting the machine threshold from the expert threshold. The distribution of the errors is plotted for all subjects and tests in Figure 5. As expected, the distribution of deviation scores appears to be Gaussian. For adult subjects, the mean error threshold score (automated threshold–expert-derived threshold) is –0.5 dB, with a standard deviation of 5.9 dB. For infants, the mean is –0.5 dB, with a standard deviation of 6.4 dB. In 193 tests, 7 (3.6%) threshold scores fell outside of 2 standard deviations of the mean. Of these scores, most (five) were machine underestimates of threshold. One test produced an overestimate of threshold.

In order to evaluate the degree of agreement between automated and expert judgments, behavioral and electrophysiologic estimates, and multiple tests of the same ear, two statistical measures were used. Regression lines were fit to the data to assess the linear relationship between automated and expert threshold judgments. Regression provides a least-squares fit of the data where disagreements are weighted by their magnitude. Thus, it is possible to have high regression coefficients if error is symmetrically distributed around the regression line. Kappa, on the other hand, was used to assess the degree of agreement in a categorical manner (Kraemer, 1982). The data were sorted into two categories, agreement or nonagreement. In the following analysis, two scores were considered in agreement if they were within ±5 dB. The magnitude of the disagreement is not considered by Kappa. Thus, Kappa provides a view of absolute agreement and will typically be smaller than would be expected from the regression results. In evaluating the magnitude of Kappa, the following terms are applied: < .8 = almost perfect agreement; .6 – .8 = substantial agreement; .4 – .6 = moderate agreement; .2 – .4 = fair agreement; > .2 = slight agreement (Landis and Koch, 1977).

Figure 6 shows a scatterplot of the relationship in dB between expert-derived and automated thresholds for adult tests (n = 125). The scatterplot illustrates that the deviations around expert threshold level are not a function of absolute threshold level, though most of the deviations occur between 40 and 50 dB. A regression analysis of these data shows a regression coefficient of 0.87, with a standard error of 5.6 dB. The correlation coefficient (R^2) indicates that 84 percent of the variance is accounted for by the relationship between the two test scores. Kappa is also quite high (.78), indicating substantial agreement between machine and expert observers.

Figure 7 shows a scatterplot of the relationship in dB between expert-derived thresholds and automated thresholds for all infant tests (n = 68). As can be seen from the figure, the relationship between expert and automated thresholds is even more robust for the infant than for the adult data. Only two test scores (2.9%) fell beyond two standard deviations of the regression line. The regression coefficient for infants is 0.80, with a standard deviation of 5.0 dB. R^2 indicates that 91.6 percent of the variance is accounted for. Kappa (.85) is also

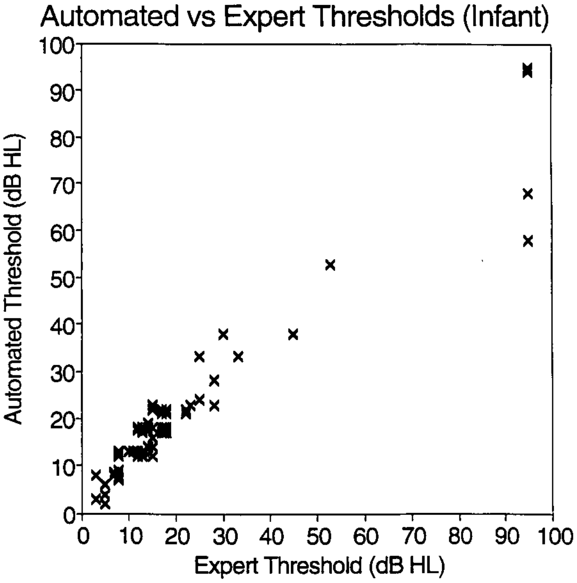

Figure 7 Scattergram of automated versus expert-derived thresholds for infants.

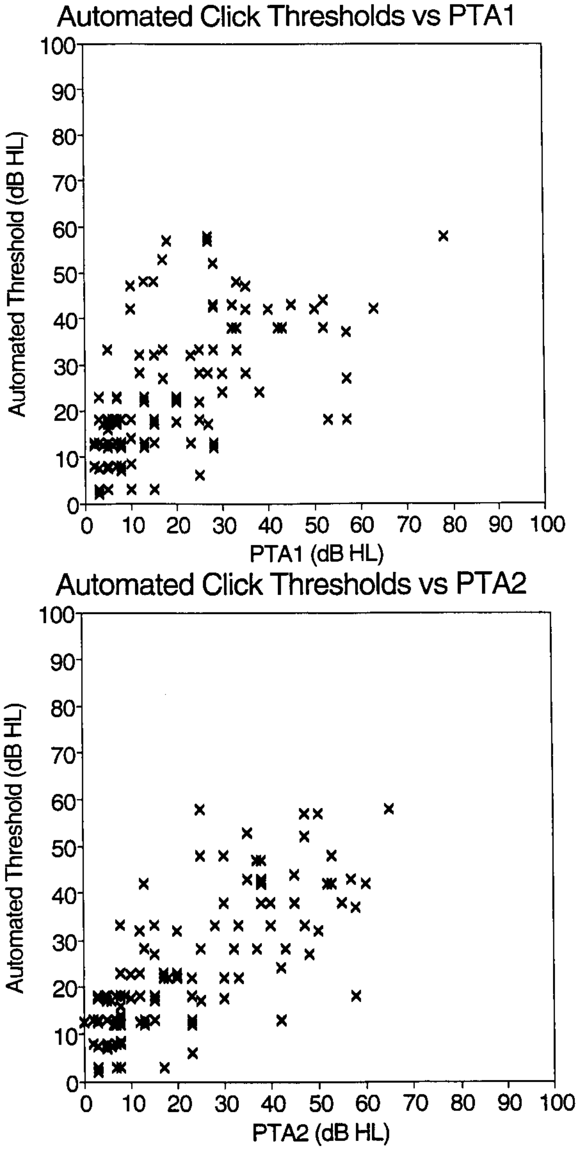

Figure 8 Comparison of automated thresholds and pure-tone averages in adults. *A*, PTA_1 (0.5, 1, and 2 kHz); *B*, PTA_2 (1, 2, and 4 kHz).

higher for infant tests than for adult tests. This highly significant value suggests almost perfect agreement between the expert and machine judgments. Overall, automated thresholds for infants seemed to be more accurate than for adults, probably because infant ABRs were acquired while infants slept, while adult data was collected from resting, but not usually sleeping, adults.

Automated ABR Thresholds versus Behavioral Hearing-Threshold Correlates

Pure-tone averages (PTA_1: mean threshold at 0.5, 1, and 2 kHz and PTA_2: mean threshold at 1, 2, and 4 kHz) have long been used to capture the essence of a hearing loss and to compare behavioral thresholds with ABR click data (Jerger and Mauldin, 1978). Pure-tone averages were compared with both automated thresholds and thresholds derived by the experts. Figure 8 illustrates the relationship between PTA_1 (*A*) and PTA_2 (*B*) and the automated threshold levels. As can be seen, the regression coefficient between PTA_1 and automated threshold (x = 0.53) is slightly poorer than the coefficient between PTA_2 and threshold (x = 0.58). The relationship between PTA_1 and automated threshold accounts for 35 percent of the variance, while the relationship between PTA_2 and automated threshold account for 52 percent of the variance. The standard deviations of PTA_1 and PTA_2, respectively, are 11.3 and 9.8 dB. Kappa suggests that the absolute agreement between PTA_1 and automated threshold (0.17) and PTA_2 and automated threshold (0.05) are slight, with only the agreement between PTA_1 and automated threshold reaching statistical significance ($p < .05$).

Figure 9 shows the relationship between PTA_1 and PTA_2 and expert-derived thresholds. Once again, the regression coefficient is higher for PTA_2 (x = 0.68, SD = 8.8) than for PTA_1 (x = 0.64, SD = 11.0), though both are somewhat improved versus the automated thresholds. The relationships between the expert-derived thresholds and PTA_1 and PTA_2 account for 46.1 percent and 65.2 percent of the variance, respectively. Kappa is somewhat higher for each pure-tone average versus expert-derived thresh-

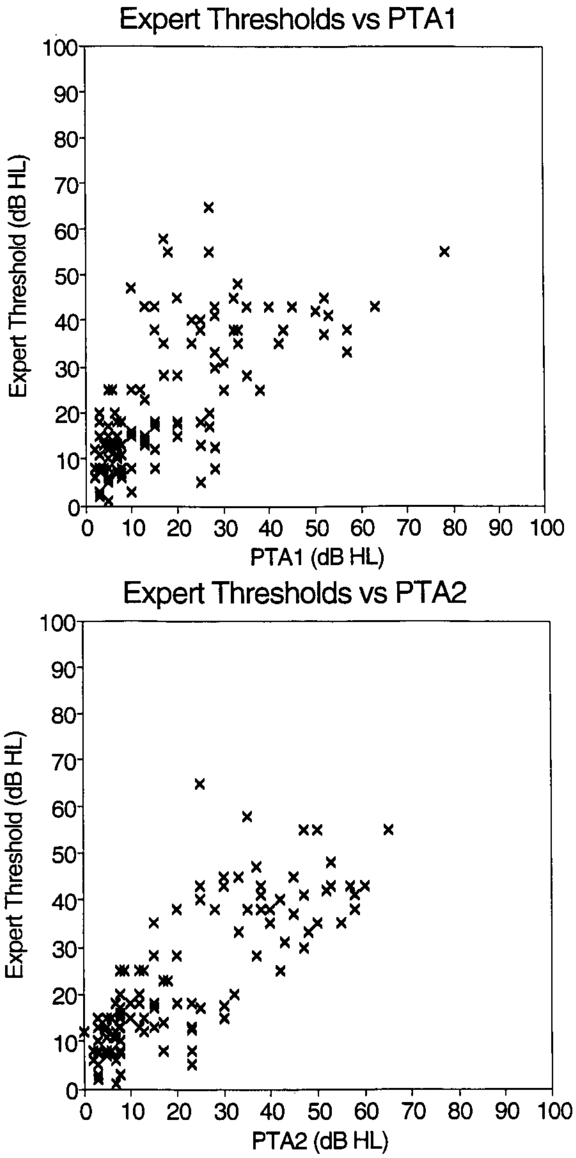

Figure 9 Comparison of expert-derived thresholds and pure-tone averages in adults. *A*, PTA_1 (0.5, 1, and 2 kHz); *B*, PTA_2 (1, 2, and 4 kHz).

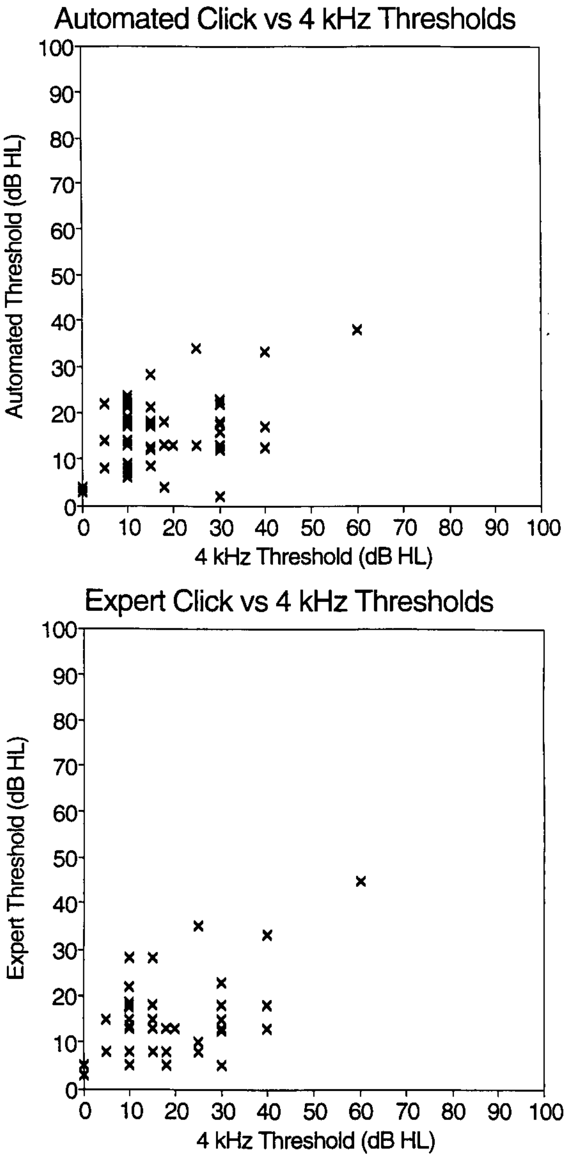

Figure 10 Comparison of automated thresholds and behavioral sound-field (4 kHz) thresholds for infants. *A*, automated thresholds; *B*, expert judgments.

olds (Kappa is .33 and .16 for PTA_1 and PTA_2, respectively) than for the same comparisons with automated thresholds. Nevertheless, absolute agreement between pure-tone averages and expert thresholds are only slight to fair. This is not surprising, given the fact that ABR click thresholds are highly influenced by audiogram configuration (Gorga, Worthington, Reiland, Beauchaine, and Goldgar, 1985; Keith and Greville, 1987).

Figure 10 shows the relationship between infant threshold at 4 kHz and expert (*A*) and automated (*B*) thresholds, respectively. The regression coefficients between each measure of ABR click threshold (automated: x = 0.24, SD = 6.9; expert: x = 0.31, SD = 7.0) and threshold at 4 kHz are poorer than the coefficients for the relationship between pure-tone averages and expert or automated scores for adults. The relationship between 4-kHz threshold and automated results account for only 14 percent of the variance, while the relationship between expert judgments and threshold accounts for 21 percent. Neither the expert judgments nor the automated threshold values are good predictors of high-frequency hearing in this group. This point is further made by examining Kappa, which shows poor agreement for both automated (.07) and expert (.11) judgments of threshold, based upon hearing levels at 4 kHz.

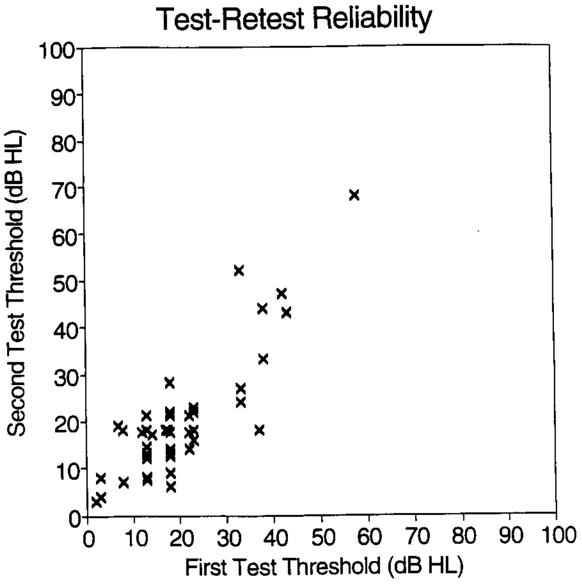

Figure 11 Test–retest reliability for 45 subjects.

Test–Retest Reliability

Figure 11 shows the outcome of 45 ears, each of which were tested twice. The regression coefficient between tests 1 and 2 is 0.95 (SD = 6.94), and 72 percent of the variance is accounted for by the relationship between the two tests. Since infant data were collected from sleeping infants and adult data from largely wakeful adults, we expected the infant data to be more reliable than the adult data. This was, in fact, the case. The infant data had smaller standard deviations and higher R^2 than the adult data (x = 1.13, SD = 8.6, R^2 = 0.84 for infants; x = 0.92, SD = 10.7, R^2 = 0.55 for adults). The overall Kappa was .77, indicating substantial agreement between tests.

Test Efficiency

The PEST algorithm determined the number of runs (2 recordings with 1024 sweeps each) per test. For 155 tests, the minimum number of runs was 4, while the maximum was 20 (mean = 7.4, SD = 2.8). Figure 12 shows histograms of the number of runs required to converge on an automated threshold. The figure shows the modal number of runs to be 6. In general, large numbers of runs were associated with nonoptimal subject states, resulting in multiple artifacts and complicated response-detection problems. Removal of these artifacts would significantly reduce run time. Since there was no appreciable time devoted to computational overhead, the average test time following placement of electrodes to determination of threshold was 12.6 minutes.

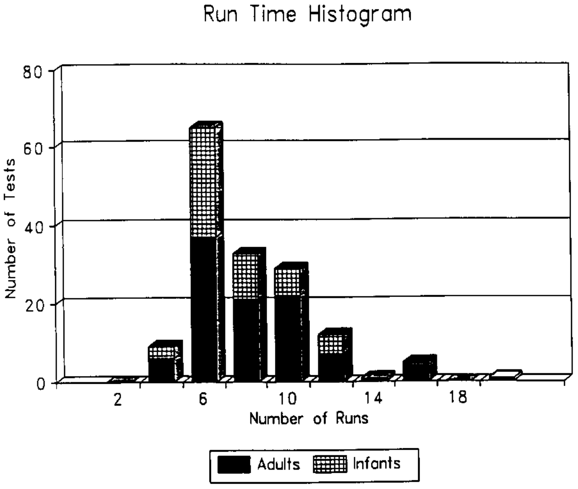

Figure 12 Run-time histogram for all tests for adults and infants.

DISCUSSION

This study describes the results of an automated, threshold-seeking method of ABR testing. The method employs a PEST procedure to acquire ABR recordings at appropriate levels and a sliding window cross-correlation to detect the presence of an ABR. The method shows promise in that it correlates nearly as well with behavioral measures of hearing as do thresholds determined by expert judges. In this study, the highest correlations were found between PTA$_2$ and click thresholds, but absolute agreement was greater for PTA$_1$ and click threshold.

A more rigorous test of the efficacy of automated threshold determination is the relationship between automated ABR thresholds and those derived from experts viewing the same recordings. Automated ABR thresholds derived from the sliding window cross-correlation agree with expert-derived thresholds (within ±10 dB) over 96 percent of the time. Of 193 tests, 186 agreed with experts.

Despite the extremely small number of discrepancies between judges and machine scores, it is worthwhile to examine the probable cause of instances of inaccurate automated thresholds. The sliding window correlation method operates with an empirical cut-off criterion. This criterion represents a fulcrum between false positives and false negatives. If the criterion is set high, false negatives will decrease. If the criterion is too low, false negatives will increase. The criterion for the present study was chosen based on off-line preliminary analysis of ABR recordings yielding a 95 percent confidence interval. Examination of the errors indicated that, as expected, the cross-correlation values for these recordings were near the cut-off point.

In general, human observers are not subject to the same kinds of errors as machines. Humans learn to use contextual information and ignore odd responses that do not fit into the general picture. Thus, a flat or noisy response is generally ignored if clear responses are found at intensities just above and below a questionable response. The automated decisionmaking, however, is largely based on cross-correlation. Cross-correlation increases for noisy recordings as a function of chance. Thus, machine errors tend to be underestimates of hearing loss.

The actual number of errors, however, was fewer than predicted by the confidence interval. Error reduction was achieved by using PEST, which has built-in self-correction features. When PEST fails to detect a response, step size is increased and a higher level is tested. As long as step size remains greater than 5 dB, the procedure will move between levels at which no response is detected and levels at which clear responses are detected. Thus, an inaccurate designation of a nonresponse will result in an immediate increase in level, followed by a subsequent decrease. An inaccurate designation of a response will generally self-correct when the next lowest level is tested.

Although PEST clearly has desirable self-correction features, it may not ultimately be the most effective method for threshold testing from a practical clinical point of view. First, an efficient PEST algorithm generally starts at a moderate level (50 dB), obviating the need for high-level testing. Unfortunately, responses to high-level signals provide valuable neurologic information concerning the integrity of the brainstem pathways. In particular, wave I can usually only be obtained at high levels of testing and is essential to determine the wave I–V interpeak interval. Second, since PEST does not test at regular level intervals, it does not provide all of the data needed for plotting latency-intensity functions. These functions are helpful in providing the clinician with information about the configuration of the audiogram. Third, PEST does not spend much test time at suprathreshold levels. Thus, it is difficult for clinicians to verify the integrity of the waveforms in their most robust forms.

On average, 12.6 minutes were sufficient to test one ear with within ±5 dB accuracy, excluding time lost to artifact-rejected sweeps. The PEST procedure generally finished in 6 runs, provided no spurious reversals were obtained as a function of the response-recognition unit. When erroneous reversals were obtained, they tended to self-correct, but in the process, the PEST procedure was prolonged (to as many as 20 runs in the current study). Currently, a variety of options are being studied to further reduce test time in an automated threshold-tracking system. For example, using a faster click rate (one that does not reduce response amplitude) and/or a variable sweep averager based on an adaptive stopping rule seem likely to make automated testing more time efficient.

Regardless of the clinical properties of PEST, it is clear that the sliding window cross-correlation method offers a viable alternative to manual testing of hearing in both adults and infants. The procedure demonstrates high test–retest reliability, especially for sleeping infants. The procedure can easily be applied to other threshold-tracking methods for clinically efficient automated hearing testing.

Acknowledgment. This study is supported by NINCDS SBIR Grants to Intelligent Hearing Systems Corporation (formerly Artificial Hearing Systems), Miami, Florida. The assistance of Edward Miskiel, Laurie Karbowski, and Ayse C. Özdamar in data collection and patient management are kindly appreciated.

REFERENCES

Alpsan D, Özdamar Ö. (1992). Auditory brainstem evoked potential classification for threshold detection by neural networks. I. Network design, similarities between human-expert and network classification, feasibility. *Automedica* 15:67–82.

Arnold SA. (1985). Objective versus visual detection of the auditory brain stem response. *Ear Hear* 6:144–150.

Beagley HA, Sayers BMcA, Ross A. (1979). Full objective ERA by phase spectral analysis. *Acta Otolaryngol* 87: 270–278.

Delgado RE, Özdamar Ö. (1990a). Automated ABR peak labelling using matched filters In: *IEEE Proc 12th EMBS Ann Int Conf.* Piscataway, NJ: IEEE Press, 870–871.

Delgado RE, Özdamar Ö. (1990b). ABR threshold determination using pre- and post-power spectral differences. Presented at the ASHA convention. *Asha* October 176.

Delgado RE, Özdamar Ö, Miskiel E. (1988). On-line system for automated auditory evoked response threshold determination. In: *IEEE Proc 10th EMBS Ann Int Conf.* Piscataway, NJ: IEEE Press, 1472–1473.

Dobie RA. (1993). Objective response detection. *Ear Hear* 14:31–35.

Dobie RA, Wilson MJ. (1989). Analysis of auditory evoked potentials by magnitude-squared coherence. *Ear Hear* 10:2–13.

Don M, Elberling C, Waring M. (1984). Objective detection of averaged auditory brainstem responses. *Scand Audiol* 13:219–228.

Elberling C, Don M. (1987). Threshold characteristics of the human auditory brainstem response. *J Acoust Soc Am* 81:115–121.

Gorga MP, Worthington DW, Reiland JK, Beauchaine KA, Goldgar DE. (1985). Some comparisons between auditory brainstem response thresholds, latencies and the pure-tone audiogram. *Ear Hear* 6:105–112.

Greenblatt E, Zappulla RA, Kaye S, Fridman J. (1985). Response threshold determination of the brainstem auditory evoked response: a comparison of the phase versus magnitude derived from the fast Fourier transform. *Audiology* 24:288–296.

Hyde ML, Riko K, Maliza K. (1990). Audiometric accuracy of the click ABR in infants at risk for hearing loss. *J Am Acad Audiol* 1:159–166.

Jacobson JT, Jacobson CA, Spahr RC. (1990). Automated and conventional ABR screening techniques in high-risk infants. *J Am Acad Audiol* 1:187–195.

Jerger J, Mauldin L. (1978). Prediction of sensorineural hearing level from brainstem evoked responses. *Arch Otolaryngol* 104:456–461.

Joint Committee on Infant Hearing. (1991). 1990 Position statement. *Asha* 33(Suppl. 5):3–6.

Keith WJ, Greville KA. (1987). Effect of audiometric configuration on the auditory brainstem response. *Ear Hear* 8:49–55.

Kileny PR. (1987). Algo-1 automated infant hearing screener: preliminary results. *Semin Hear* 8:125–131.

Kraemer HC. (1982). Kappa coefficient. In: Kotz S, Johnson NL, eds. *Encyclopedia of Statistical Sciences.* New York: J. Wiley & Sons, 352–354.

Landis JR, Koch GG. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33: 159–174.

Mason SM. (1984). On-line computer scoring of the auditory brainstem response for estimation of hearing threshold. *Audiology* 23:277–296.

National Institutes of Health. (1993). *Consensus Development Conference on Early Identification of Hearing Impairment in Infants and Young Children.* Bethesda, MD: NIH.

Özdamar Ö, Alpsan D. (1992). Neural-network classifier for auditory evoked potentials. In: Fishman MB, Robarts JL, eds. *Advances in Artificial Intelligence Research.* Vol. 2. Greenwich, CT: JAI Press, 165–175.

Özdamar Ö, Delgado RE, Eilers RE, Widen JE. (1990). Computer methods for on-line hearing testing with auditory brainstem responses. *Ear Hear* 11:417–429.

Özdamar Ö, Delgado RE, Miskiel E, Eilers R, Widen J. (1987b). *Comparison of Computer Methods for Automated Hearing Testing with Auditory Brainstem Responses.* Presented at the International Electric Response Audiometry Study Group, Tenth Biennial Symposium, Charlottesville, VA.

Özdamar Ö, Kaplan R, Miskiel E, Delgado RE. (1987a). Human-machine interface for an interactive evoked potential electrodiagnostic system. In: Asfour SS, ed. *Trends in Ergonomics / Human Factors IV.* Amsterdam: Elsevier, 1121–1129.

Peters JG. (1986). An automated infant screener using advanced evoked response technology. *Hear J* 39:25–30.

Salvi RJ, Ahroon W, Saunders SS, Arnold SA. (1987). Evoked potentials: computer-automated threshold tracking procedure using an objective detection criterion. *Ear Hear* 8:151–156.

Shimizu H, Walters RJ, Proctor LR, Kennedy DW, Allen MC, Markowitz RK. (1990). Identification of hearing impairment in the neonatal intensive care unit population: outcome of a five-year project at the Johns Hopkins hospital. *Semin Hear* 11:150–160.

Stein LK. (1984). Evaluating the efficiency of auditory brainstem response as a neonatal hearing screening test. *Semin Hear* 5:71–76.

Taylor MM, Creelman CD. (1967). PEST: efficient estimates on probability function. *J Acoust Soc Am* 41:782–787.

Thornton AR, Obenour J. (1981). *Auditory Response Detection Method and Apparatus.* US Patent #4,275,744.

Weber BA, Fletcher GL. (1980). A computerized scoring procedure for auditory brainstem response audiometry. *Ear Hear* 1:233–236.

Widen JE. (1990). Behavioral screening of high-risk infants using visual reinforcement audiometry. *Semin Hear* 11:342–356.

Widen JE, Bull DH. (1984). *An Automated Version of Visual Reinforcement Audiometry for Clinical Use.* Poster presentation at the American Speech and Hearing Association convention, San Francisco, November, 1984.

Woodworth W, Reisman S, Fontaine AB. (1983). The detection of auditory evoked responses using a matched filter. *IEEE Trans Biomed Engr BME* 30:369–376.