Classification of Audiograms by Sequential Testing: Reliability and Validity of an Automated Behavioral Hearing Screening Algorithm

Rebecca E. Eilers*
Ozcan Ozdamar†
Michele L. Steffens‡

Abstract
In 1990, CAST (classification of audiograms by sequential testing) was proposed and developed as an automated, innovative approach to screening infant hearing using a modified Bayesian method. The method generated a four-frequency audiogram in a minimal number of test trials using VRA (visual reinforcement audiometry) techniques. Computer simulations were used to explore the properties (efficiency and accuracy) of the paradigm. The current work is designed to further test the utility of the paradigm with human infants and young children. Accordingly, infants and children between 6 months and 2 years of age were screened for hearing loss. The algorithm’s efficacy was studied with respect to validity and reliability. Validity was evaluated by comparing CAST results with tympanometric data and outcomes of staircase-based testing. Test-retest reliability was also assessed. Results indicate that CAST is a valid, efficient, reliable, and potentially cost-effective screening method.

Key Words: Audiogram, auditory brainstem response (ABR), classification of audiograms by sequential testing (CAST), hearing testing, tympanogram, visual reinforcement audiometry (VRA)

The evaluation of adult hearing is usually accomplished using standard pure-tone audiometric procedures (ANSI, 1978; ASHA, 1978), which yield a multi-frequency audiogram consisting of threshold values at octave or half-octave intervals from 250 Hz to 8 kHz. Standard clinical procedures test one frequency at a time by probing for threshold using an up/down sequence of trials to find the lowest presentation level at which a response is obtained three times (Hughson and Westlake, 1944; Carhart and Jerger, 1959). For cooperative adults and older children, the procedure is both accurate and efficient. For infants, the demands of the test situation exceed their response and attentional capabilities. Visual reinforcement audiometry (VRA) was developed to remedy those aspects of testing associated with response capability of infants and other developmentally young populations (Suzuki and Ogiba, 1960, 1961; Liden and Kankkunen, 1969; Moore et al, 1977). VRA is based on operant conditioning techniques, which are used to teach infants to respond with head turns to audible signals in order to view visual reinforcers. VRA procedures, however, usually do not generate enough trials to obtain a minimally adequate four-frequency audiogram without compromising standard methods of probing for threshold.

As a consequence, infant hearing testing and screening is often reduced to either tests of suprathreshold signals, usually wide-band in nature (Widen, 1990) or tests of fewer than four frequencies. The need to obtain a four-frequency
audiogram is usually met by repeated testing over several visits.

Currently, a nationally based initiative, in the form of an infant hearing screening bill (The Hearing Loss Testing Act, HR2089) is being deliberated by the United States House of Representatives' Health and Environment Subcommittee. The bill reflects the fact that 14 states currently have some form of mandated infant hearing screening in the belief that early detection can prevent the sequelae associated with undetected hearing loss. While much of this screening is now done during the neonatal period, a significant amount is delayed to times when VRA-based behavioral screening is possible. Furthermore, follow-up protocols for neonatal failures typically include behavioral testing during the first year of life. Traditional VRA techniques are not well suited to the task of follow-up of large groups of infants due to the costs involved in the multiple visits typically needed for audiogram determination.

In anticipation of the need for rapidly determined infant audiograms for purposes of infant hearing screening, Özdamar et al (1990) proposed a new algorithm called CAST (classification of audiograms by sequential testing) and tested it using computer simulation and an infant response model. (For details of the theory and algorithm, see Özdamar et al, 1990.) CAST differs from traditional audiometric procedures in that it selects a best-fitting audiogram from a set of predetermined audiogram patterns. These patterns represent the range of typical audiograms found in a given population of a given age (in this case, the population of infants). Each of the audiogram patterns has an estimated frequency of occurrence derived from known population parameters. CAST utilizes a modified Bayesian method of diagnosis based on the dynamically changing information collected during the course of a screening test. The traditional Bayesian method of diagnosis uses the conditional probabilities of symptoms to identify the occurrence of a specific disease. CAST treats audiometric configurations as "diseases" and trial-by-trial information as "symptoms." During the test, the probability associated with the occurrence of each audiogram pattern is updated based on trial-by-trial information provided during the test. Trial parameters (level and frequency) are determined automatically based upon the results of previous trials during a test. The outcome of a CAST test is one of the audiogram patterns from the audiogram template set. CAST converges on a pattern as determined by a preset stopping criterion (e.g., a probability of greater than 95% associated with one pattern in the set).

The algorithm as tested by computer simulation proves to be efficient and accurate. The 13 to 20 trials typical of simulated tests translates into about 3 to 5 minutes of infant test time. Simulations suggest that the procedure is well suited to screening since it does not pass pseudosubjects (computer-simulated subjects) with hearing loss. The added advantage of the procedure is that it offers more than a pass or fail classification. The procedure provides an estimate of hearing configuration for individuals who do not pass screening.

In the current work, the CAST algorithm was evaluated with human infants and young children in a typical large urban area clinical population in order to investigate the clinical biostatistical properties of the algorithm for use as a universal screening tool. CAST outcome was compared to tympanometric findings and staircase-based four-frequency audiograms. In addition, CAST test-retest reliability was investigated.

METHOD

Subjects

In order to subject CAST to rigorous clinical evaluation, 403 infants, approximately half of whom (232) appeared on a high-risk register, were recruited to the Mailman Center for Child Development. Infants were recruited from two main sources, a "high-risk" follow-up program for graduates of the neonatal intensive care units at the University of Miami/Jackson Memorial Hospital and mail solicitation of families with young infants within the catchment area of the Mailman Center. At the time of the initial evaluation, the infants had a mean chronologic age of 62.5 weeks with a range of 27 to 134 weeks. All preterm infants had corrected ages of 6 months or greater prior to testing.

The testing carried out for the present study was accomplished in several stages. Some infants were tested and retested with CAST, some received CAST followed at a second visit by Optimized Hearing Test Algorithm (OHTA), a psychophysically based diagnostic staircase procedure, and some received OHTA first followed by CAST at a second visit. The data were used to assess reliability and validity in a number of ways.

Tympanometry was used to separate test results based upon probable hearing status of
each subject. During the course of the study, infants may have received up to four separate tests (usually administered over two visits) including multiple CAST tests and/or CAST and OHTA. Tympanometry may have been performed at each test visit, one of the visits or not at all. During the study, 215 infants had normal tympanograms bilaterally, 79 infants showed abnormal tympanograms bilaterally, 72 were mixed (one normal ear, one abnormal), and tympanograms were either not attempted or unattainable for 146 infants on at least one visit.

Description of the Algorithms

Classification of Audiograms by Sequential Testing

CAST is based upon a dynamic implementation of the Bayesian model of medical diagnosis. Essentially, the procedure is designed to select, through testing, an audiogram pattern (from a data base of templates) that best approximates that of the infant. Each template consists of a four-frequency (0.5, 1, 2, and 4 kHz) pattern of threshold levels. Initially, CAST was implemented with the seven prototype patterns evaluated by Özdamar et al (1990). Pilot studies suggested that additional template types were needed and the template set was expanded to include the nine prototype patterns shown in Figure 1. This template set is referred to as the Original Set. The prototype patterns were chosen to represent a broad range of hearing impairment based upon examination of hundreds of partial infant audiograms in our clinical database. The first template of the set in Figure 1 is consistent with a “normal” pattern of hearing and the remaining templates represent various patterns of hearing loss that differ with respect to configuration and magnitude. During the course of the study, a second template set was derived from a modification of the first. The second template set or “Revised Set” is shown in Figure 2 and differs from the Original Set only in templates 1 and 2. The Revised Set was developed to help overcome deficiencies noted in the Original Set and referred to below.

The CAST algorithm is iterative and trial parameter selection (i.e., signal level and frequency) is under computer control. The algorithm is summarized in outline form in Table 1. In step 1, the optimum test frequency is selected by calculating the mean weighted level of each template frequency for the threshold values of all templates. The weight is determined by the estimated frequency of occurrence of each template. Once the frequency with greatest template variance is selected, the mean level of the nine templates is calculated and a trial is presented at the frequency with greatest variance at the mean weighted level for that frequency. The infant’s response is then used to recalculate each of the pattern weights. For instance, if an infant failed to respond to the mean weighted level at 1000 Hz, then the probability that the infant’s audiogram corresponds to template 1 is decreased and the probability that the infant’s audiogram corresponds to one of the templates indicating hearing loss is increased. After each trial, the program checks to see if the probability associated with any template exceeds a
Classification of Audiograms by Sequential Testing/Eilers et al

The CAST procedure has a built-in front end that serves to train infants, if necessary. In order to proceed beyond the first frequency and signal level, the infant must turn correctly to two consecutive trials (out of six), which are presented at the initial frequency and level selected by CAST. If the infant fails to achieve two consecutive correct trials, the level is increased by 20 dB. The two-trial criterion continues to apply. Failure at the increased level is followed by automatic activation of the reinforcers in concert with signal presentation in the absence of headturns. Two consecutive correct headturns without reinforcer prompting result in a 10 dB decrease in level and a new test frequency. Success at this level leads back to the stochastic procedure. Failure to reach this level leads to discontinuance of testing. Training failure is indicative of either severe to profound hearing impairment (rarely) or developmental delay and/or behavioral problems.

**Optimized Hearing Test Algorithm**

The Optimized Hearing Test Algorithm was developed through the use of computer simulation and subsequent testing with infants (Eilers et al, 1991a, b). The algorithm is modeled on an up/down staircase procedure utilizing a 5-dB-step size following a QUIR (quickly in range) (Trehub et al, 1986) introductory and training phase. This phase, as the name suggests was designed to move the majority of testing quickly to near threshold values.

During testing, OHTA intermingles the frequency of test signals and works on the assumption that in normally hearing individuals, there will be little difference in minimal response levels or threshold for different frequencies. Thus, OHTA assumes that a response at 40 dB HL at one frequency (e.g., 1 kHz) indicates the likelihood of a similar response to the same level signal at another frequency. OHTA does not initially test all levels at all frequencies but rather a range of levels at some frequencies. Thus, for most normally hearing infants, time is saved by not testing all suprathreshold signals at each signal level. During OHTA, QUIR continues until the infant fails to turn to a signal trial after which the algorithm moves to the 5-dB-step size and up/down procedure and proceeds to probe threshold at each frequency. Each threshold probe is terminated after three reversals in the direction of up/down movements of the staircase (see Eilers et al, 1991a, b). If an infant has a sloping hearing loss or another irregular pattern of loss, the algorithm automatically finds minimum response levels at each test frequency, beginning the threshold search at the level of the first missed trial at a given frequency. The OHTA procedure is most efficient for normally hearing infants; time of test for an infant with an asymmetric pattern of loss across test frequencies approaches that of a standard staircase procedure.

OHTA contains a fixed percentage of randomly dispersed control trials which can be used to assess the infant's readiness to turn and the degree of behavioral control. Probe trials (15 dB above the current test threshold estimate) are also used to keep infants on task during the near threshold trials and as a measure of internal consistency of infant test-taking behavior.

**Procedure**

Infants were tested in a sound-attenuated booth containing a sound field speaker system, a control box, a small table, two chairs placed on adjacent sides of the table, one for the examiner and one for the parent. The infant was seated on the parent's lap during testing. The infant was required to make a 90 degree headturn to view one of four reinforcers, housed on top of the speaker, when he or she detected a signal. The
test and signals were controlled by an IBM XT microcomputer and special hardware and software (IVRA model 100, Intelligent Hearing Systems). The IVRA system and amplifier were housed in a control room adjoining the booth.

For any test, CAST and OHTA were administered by a single examiner, trained in VRA, who was largely blind to the frequency and intensity of signals. The examiner wore sound attenuating ear plugs and tightly fitting headphones through which music was played during testing. While the examiner was able to hear very loud signals, mid-range and near-threshold signals were effectively masked.

The examiner's role was to keep the child visually interested at midline and to initiate trials by pressing a button on the response box. The examiner also judged whether or not an infant turned toward a visual reinforcer during a test interval (control or signal). If the infant's headturn occurred within the 4-sec response interval, one of four visual reinforcers was randomly activated. CAST required from 10 to 20 trials (including control and training trials) to select the best fitting audiogram template. The average administration time was from 3 to 5 minutes per child. The CAST program kept track of trial-by-trial data and, at the end of the test, produced an audiogram pattern corresponding to one of the templates.

The output of OHTA was a four-frequency audiogram and trial-by-trial data. The test took approximately 15 to 20 minutes and required an average of 45 trials.

Tympanometry was accomplished using a Micro Audiometrics Earscan. On the basis of tympanometric pattern, infants were classified into three middle-ear status groups: normal bilaterally (Type A pattern with peak pressure between -150 and +50 daPa), abnormal bilaterally (Type B, no measurable peak, or Type C pattern with negative pressure peak greater than -150 daPa), and abnormal unilaterally (Jerger, 1970). Those infants with unilateral losses were excluded from some test comparisons (e.g., middle ear status compared to CAST results) because clear cut predictions could not be made about their hearing status; they were not excluded from other comparisons (e.g., those that entailed test-retest reliability performed on the day of tympanometry).

RESULTS

Phase 1: Results of CAST Testing Using the Original Templates

Results of tympanometry were used to accrue a group of infants and children with an increased probability of hearing loss. It was reasoned that CAST should be expected to identify more infants and children as hearing impaired who exhibit abnormal tympanometric findings than similar children with normal tympanograms. Accordingly, infants tested using the Original CAST template set were divided into four groups on the basis of middle ear status as defined by tympanometry: 32 infants were normal bilaterally (Normal Tymp), 25 had bilaterally abnormal findings (Abnormal Tymp), 83 infants did not have tympanometry data available (No Tymp), and 11 had mixed results (Mixed Tymp). The focus of this analysis is on the Normal and Abnormal Tymp categories. Data on infants without tympanometric findings are presented as well with the assumption that the No Tymp infants represent an average unscreened population.

CAST categorized the audiograms of 84.4 percent of infants in the Normal Tymp group as pattern 1, the normal pattern. Patterns 4, 5, and 6 were each assigned to one infant and pattern 7 to two infants. CAST categorized 76 percent of the infants with Abnormal Tymp as having normal hearing. The remaining Abnormal Tymp infants were categorized as having pattern 4 audiograms (8%) or pattern 6 audiograms (16%). Finally, infants in the No Tymp group who received abnormal pattern designations fell into one of the following categories: pattern 4 (8%), pattern 5 (4%), and pattern 6 (16%).
predominantly into patterns 5 and 6. These data are summarized in Figure 3. Inspection of the data reveals that CAST did quite well in assigning normal outcomes to infants who had normal middle ear function but missed some infants who presumably had hearing affected by abnormal middle ear conditions. Our best estimate from traditional VRA procedures in our clinic is that about half of infants who show abnormal middle ear function also show elevated hearing thresholds (see also Widen, 1990).

CAST, using the original templates, failed to detect about half of the children we would expect to have had abnormal hearing.

**Phase 2: Results of CAST Testing with Revised Templates**

Accordingly, a new template set was developed (see Figure 2), which in effect made the CAST passing criteria more stringent. Pattern 1 was adjusted downward 5 dB at each frequency and pattern 2 was adjusted downward at .5 and 1 kHz to represent a hearing loss with a shallower slope, one often associated with middle ear pathology.

A second set of CAST tests was conducted with the Revised Template set. Following tympanometry, infants were classified into four categories: Normal Tymps (183), Abnormal Tymps (54), Mixed Tymps (61), and No Tymps (103). Figure 4 shows the percentages of patterns selected by CAST for the Normal, Abnormal, and No Tymp conditions. CAST classified 75.9 percent of tests of infants with Normal Tymps as normal pattern 1 with the next largest group assigned to pattern 2. Tests of infants with Abnormal Tymps were scored as abnormal (patterns 2–9) 50 percent of the time. For the No Tymp (unselected) condition, 70.8 percent of tests were normal.
The New CAST templates seemed to detect appropriate numbers of hearing-impaired infants based on tympanometric criteria. In the process of improving the hit rate for CAST, the estimate of false positive rate was also increased somewhat from about 15 percent to about 25 percent to 30 percent in Normal Tymp and No Tymp groups. The empirical assessment of false positives with the new templates follows in phase 3.

Phase 3: Results of CAST Testing Compared to Conventional Tests of Threshold

Tympanometry provides at best only a rough validation procedure for evaluating a screening test. Although we know that infants with abnormal tympanograms are much more likely to have hearing abnormalities than infants with normal middle ear function (especially those with Type B tympanograms), there is only a moderate correlation between infant hearing tests and results of tympanometry. Accordingly, we began phase 3 testing which consisted of comparisons of CAST outcomes with results of tests designed to probe infant thresholds. Infants involved in phase 3 were tested with both CAST and OHTA at separate visits. The two tests were presented in counter-balanced order about 1 week apart. Infants were considered to have passed OHTA if their mean minimum response level across the four test frequencies was less than 12 dB (the mean threshold estimate for CAST for pattern 1).

Eighty-five infants participated in phase 3. Of these, 42 had OHTA tests that were deemed sufficiently reliable to serve as a "gold standard" for computation of clinical biostatistics. Reliable tests were defined as those with greater than 75 percent correct controls and 75 percent correct probe trials. CAST testing revealed that of 33 infants who received pattern 1, 30 had normal OHTA thresholds. Two of the OHTA failures had minimum response levels at three frequencies in the normal range and highly elevated thresholds in the last frequency probed. These abnormal levels at one frequency did not fit coherently with these infants' normal findings at the other three frequencies. Given the facts that in each case the last tested frequency was elevated and that these infants had normal tympanometric findings, it was judged that the abnormal findings were fatigue-related and not indicative of hearing loss. This judgment was later confirmed by retesting at a subsequent visit with the affected frequency. Following retesting, these infants were judged to be normally hearing. One infant had several elevated thresholds yielding a mean threshold of 32 dB HL with no tympanograms available for that date. This infant could not be rescheduled. The percentage of infants with both congruent and incongruent scores on the two tests is shown in Figure 5A.

Figure 5B shows the comparable data for the nine infants who failed CAST (other than pattern 1). Of these, three also failed OHTA. For these failures, the pattern chosen by CAST was the best possible fit to the available templates in the set. The other six infants had minimal response levels in the normal range on OHTA.

Of 27 infants who passed CAST but had OHTA tests with less than 75 percent correct controls (infants who do a great deal of head turning), all passed OHTA. Infants who had fewer than 75 percent correct trials but more than 75 percent correct control trials (i.e., infants who do not often turn) showed elevated thresholds on OHTA in 33 percent of tests. Two out of three infants who failed CAST in this group also showed elevated thresholds.

The results of infants with reliable OHTA tests were used to calculate sensitivity and specificity of CAST by assuming OHTA as a gold standard. Sensitivity and specificity calculations are based on the assumption that diagnoses are already confirmed at the time statistics are compiled. They are, however, often applied to predictions of diagnosis in individuals whose condition is unknown. Thus, sensitivity is the capacity for correct diagnosis in confirmed positive cases. Specificity, however, is the capacity for correct diagnosis in confirmed
negative cases. For CAST, sensitivity is 75 percent and specificity is 84 percent. Although these measures are often used to quantify characteristics of screening tests, they really do not reflect information that is crucial to the patient and examiner. Ideally, what we wish to know is the likelihood that a positive (i.e., an infant identified through screening as hearing impaired) is really hearing impaired and conversely, the likelihood that a pass is an infant who has normal hearing sensitivity. Positive accuracy and negative accuracy (see Feinstein, 1976), respectively, provide this information. The positive accuracy of CAST measured in this study is 33 percent. Two-thirds of the 21 percent of infants who failed screening had normal hearing, therefore yielding a false positive rate of 14 percent. The failure rate generated by the CAST-OHTA comparison (21%) compares favorably to ABR failure rates in screening programs of general NICU populations tested in hospital.1 The rates range from 12 percent (Fifer and Gonzales, 1991) to 4 to 20 percent (Cevette, 1984) to 20 percent (Salamy and Weyland, 1986). Stein (1984) suggests that, as a general rule, screening at lower click intensities (30 or 40 dB HL) yields a 15 to 20 percent failure rate. CAST screens for hearing thresholds considerably lower than the 30 to 40 dB level of most ABR screens and screens at four frequencies as well. Accordingly, a 20 percent failure rate seems in line with current state-of-the-art neonatal screening tests.

The negative accuracy of CAST as measured in this study is 97 percent. Virtually all of the infants who pass CAST have normal hearing as judged by a psychophysically based staircase procedure.

Phase 4: Evaluating CAST Test-Retest Reliability

CAST test-retest reliability was obtained by testing infants twice during a single visit or twice on visits separated by about 1 week. Infants whose visits were separated by 1 week were required to maintain the same Tympanometry status from week 1 to week 2 in order to be included in the study. One hundred and thirty-seven infants participated. Figure 6 shows the results of the reliability analyses. Normal hearing results were obtained for 68.6 percent of infants tested on both test 1 and test 2 while 15.3 percent of infants tested in abnormal categories on both tests. Of infants who changed outcome on the second test, 10.3 percent changed from pattern 1 to an abnormal category and 5.8 percent changed from abnormal to normal. The majority of the latter category were training failures in test 1 who conditioned during test 2. Many of the subjects who were abnormal in test 2 only, may have tired during the second test as reflected by failure to turn to signals they had previously responded to. Overall test-retest reliability was 83.9 percent.

DISCUSSION

One goal of our work over the past few years has been to develop an automated hearing screening algorithm for infants to provide audiologists with information more nearly comparable to that obtainable for adults. We sought to develop a methodology that could be implemented on a personal computer allowing for all the modern conveniences of data storage, retrieval, automatic audiogram printouts, etc. More importantly, we wanted to provide reliable audiometric information that could be easily interpreted by a professional and could be obtained without the necessity of utilizing two examiners.

We accomplished these goals by developing and testing automated algorithms through both computer simulation and with large numbers of infants with differing hearing status. We found the algorithm behaved much as it had during earlier reported simulation tests (Ozdamar et al, 1990). In particular, CAST testing required the same number of trials as simulated tests (10–20) despite the fact that the current tests with infants were done with an expanded template set allowing for a larger number of differ-
ential diagnoses. In addition, the outcome of CAST procedures evaluated with infants included very few misses (the negative predictive accuracy was 97%). The presumption that infants who passed CAST had normal hearing was highly justified. Thus, CAST has many of the properties necessary for a screening test. It is quick, inexpensive in terms of examiner time, and does not produce many false negatives. Retesting of failures (21% of infants tested) was, in our experience, often accomplished during the initial visit. Since many of the false positives resulted from failure to condition early in the test, retesting eliminated the need for a subsequent visit to confirm normal hearing sensitivity.

Recently, in an independent evaluation of CAST, Widen et al (1991) compared CAST with a threshold-finding procedure with 43 children in the same age range as the present study. Widen found negative predictive accuracy of 93 percent using a staircase as the "gold standard." Widen and colleagues add, however, that two children identified as false negatives on the basis of the staircase results were most probably normally hearing since there was no history of risk factors, no concern about hearing status, normal middle ear function, and no history of otitis media. Thus, the negative predictive accuracy would have been 100 percent.

Widen and co-workers found fewer false positives than the current study as reflected in significantly higher positive predictive accuracy (79%). The difference in false positive rate might have been the result of the difficult-to-test high-risk infants included in the present sample but not in the Widen sample. In addition, many of the high-risk infants in our study were tested following medical screening and interventions while the Widen et al subjects were scheduled at their own convenience solely for the purpose of hearing testing (Widen, 1992). Thus, Widen's procedure would have resulted in infants being in more optimal states for hearing testing. Despite these differences, CAST functioned similarly in the two settings with respect to test time and average number of test trials. These results offer further encouragement that an automated procedure can function effectively in a number of clinical environments.

In summary, a 5-minute automated test procedure does quite well in identifying hearing impairment in infants and young children considered at risk for such deficits. The procedure provides an estimate of a four-frequency audiogram in addition to pass or fail information. The paradigm seems to be sensitive to detection of mild to moderate losses. It overestimates the incidence of hearing loss at about the same rate as ABR in the infant nursery but, to our knowledge, has not missed hearing-impaired subjects. As such it holds promise as a cost-effective screening tool for infants and young children.

Acknowledgment. Support for this work was provided by NIH SBIR grant #2R44 HD23057 to Intelligent Hearing Systems.

 Portions of this paper were presented at the American Speech-Language-Hearing Association Annual Convention in Atlanta in November 1991.

REFERENCES


