Detection-theoretic analysis of the observer-based psychophysical procedure (L)

Raman Arora

Department of Electrical and Computer Engineering and Auditory Behavioral Research Lab,
University of Wisconsin, Madison, Wisconsin 53706

Robert A. Lutfi^{a)}

Department of Communicative Disorders and Auditory Behavioral Research Lab, University of Wisconsin,
Madison, Wisconsin 53706

(Received 1 December 2006; revised 3 October 2007; accepted 25 October 2007)

The observer-based procedure, used in research applications to measure the hearing sensitivity of young infants, is analyzed within the framework of a two-stage (infant-judge) detection model involving a minimum number of assumptions. The model is taken to develop convergence theorems useful for estimating the number of judges and/or trials required to achieve a desired level of accuracy using the procedure. The model is also used to consider ways in which bias in estimates, known to be associated with the procedure, might be evaluated and reduced. Finally, a method is proposed by which the results of the analysis and different procedural variations designed to improve estimates can be evaluated empirically. © 2008 Acoustical Society of America.

[DOI: 10.1121/1.2816559]

PACS number(s): 43.66.Yw, 43.66.Fe, 43.66.Ba [JHG]

Pages: 1850–1853

I. INTRODUCTION

In the clinical evaluation of hearing, behavioral audiometry remains the gold standard. Yet, behavioral measures are not easily obtained from those who stand to benefit most from early diagnosis of a hearing loss: infants less than 5 months of age. Before 5 months of age the infant does not reliably turn their head in the direction of a sound; hence, hearing sensitivity is difficult to measure using standard visual reinforcement audiometry (cf. Moore and Wilson, 1978). Olsho *et al.* (1987) have proposed an alternative behavioral procedure for evaluating the hearing sensitivity of infants at this young age. In the observer-based procedure the sensitivity of the infant is estimated from a judge's ability to detect the presence or absence of the signal based exclusively on the infant's gross behavior. The procedure is essentially an adaptation to hearing of the forced-choice, preferential-looking procedure, developed some years earlier by Teller (1979) to measure visual acuity in infants. Though the procedure has been used primarily in research applications, it has a significant advantage of providing an objective means of evaluating the hearing sensitivity of infants. Specifically, there is a standard (the presence or absence of the signal) by which the judge's decision on each trial can be scored as correct or incorrect (see Olsho *et al.*, 1987 for a discussion of other advantages of the procedure). In this paper, we examine the observer-based procedure within the context of a simple two-stage (infant-judge) detection model to arrive at a few analytical results that might prove useful in evaluating and/or improving the accuracy of the procedure and its variants.

II. TWO-STAGE DETECTION MODEL

Our analysis is undertaken within the framework of the general two-stage detection model shown in Fig. 1. The first-stage detector is the child whose sensitivity to sound we are interested in measuring; the second stage detector is the judge who must decide on each trial whether or not a signal was presented based exclusively on the various aspects $r_1, r_2, \dots, r_N$ of the child's observable behavior. The signal in this case is represented by the binary variable $s \in \{0, 1\}$ indicating simple presence or absence; although s could just as easily represent two different sound waveforms or classes of waveforms that are to be discriminated by the child. All factors that would limit the child's ability to detect or discriminate s (neural noise, imperfections in auditory transduction and the like) are incorporated in the stochastic function f . Based on the outcome of $f(s)$ on each trial the child comes to a binary decision $x \in \{0, 1\}$ regarding the presence or absence of the signal. The sensitivity of the child is, therefore, completely described by the hit and false alarm probabilities

$$\beta_C = P(x = 1 | s = 1), \quad (1)$$

$$\alpha_C = P(x = 1 | s = 0). \quad (2)$$

Now we wish to measure β_C and α_C , but cannot do so directly because we do not have direct access to child's decision, x . Instead, in the observer-based procedure, we take as an estimate of β_C and α_C the hit and false alarm rate of the child-judge duo. The estimate of the hit and false alarm rate of the child is

$$\hat{\beta}_C = \beta_{JC} = P(z = 1 | s = 1). \quad (3)$$

$$\hat{\alpha}_C = \alpha_{JC} = P(z = 1 | s = 0), \quad (4)$$

where, β_{JC} and α_{JC} are taken to denote the hit and false alarm rate of the child-judge duo. These estimates, of course,

^{a)}Author to whom correspondence should be addressed. Electronic mail: ralutfi@wisc.edu

FIG. 1. Model of two-stage detection representing observer-based method for estimating child hearing sensitivity (see text for specific details).

are not perfect because the child does not generally convey their decision unambiguously through their behavior. Random movements of the child may be mistaken for a response to the sound, or no obvious response may be elicited even though the sound is clearly audible to the child, as might be the case, for example, when the child is not attending.[†] Here we represent the ambiguous relation between the child's decision and their behavior by the probabilistic function g . Other inaccuracies in the estimates of β_C and α_C result from the judge's perception of the child's response. For example, certain reliable indicators of the child's decision (possibly head movements) may go unnoticed by the judge, even while others (possibly facial expressions) are given much greater attention than they deserve. We represent all such imperfections in the judge's perception by the probabilistic function h . Taken together g and h represent all factors that limit the accuracy and reliability of the estimates of the child's sensitivity in the two-stage detector model. By allowing these functions to be quite general we encompass a wide variety of models of the possible relation between the child's decision and the judge's response.

The hit rate of the child-judge duo can now be expanded from the general expression of the two-stage detector model. Conditioning the probability β_{JC} on the child's decision x and using the law of total probability, we get

$$\beta_{JC} = P(z=1|s=1) = P(z=1|s=1, x=1)P(x=1|s=1) + P(z=1|s=1, x=0)P(x=0|s=1). \quad (5)$$

Now since $s \rightarrow x \rightarrow z$ forms a Markov chain, z is conditionally independent of s , given x . This means that $P(z=1|s=1, x=1) = P(z=1|x=1)$, which is the hit rate of second stage detector alone (i.e., the hit rate of the judge for the child's decision). Similarly, $P(z=1|s=1, x=0) = P(z=1|x=0)$ is the false alarm rate of the judge for the child's decision. Denoting these probabilities respectively as β_J and α_J , the hit rate of the child-judge duo from Eq. (5) is

$$\beta_{JC} = \beta_J \beta_C + \alpha_J \alpha_C. \quad (6)$$

The false alarm rate of the child-judge duo is obtained in similar fashion,

$$\begin{aligned} \alpha_{JC} &= P(z=1|s=0) = P(z=1|s=0, x=1)P(x=1|s=0) \\ &\quad + P(z=1|s=0, x=0)P(x=0|s=0) \\ &= \beta_J \alpha_C + \alpha_J (1 - \alpha_C). \end{aligned} \quad (7)$$

Now, Eqs. (6) and (7) lend themselves to mathematical analysis of various conditions in which the particular experiment or similar tests using the observer-based procedure may be performed. We consider three such conditions intended to improve estimates of the hit and false alarm rates of the child.

III. MULTIPLE JUDGES

Let us suppose that instead of a single judge, we have n judges, each of whom makes a binary decision $\{z_i\}_{i=1}^n$ regarding the presence or absence of the signal. The overall decision is taken to be some function of all the z_i . Here we give sufficient conditions on the statistical characterization of the judges to attain optimal detection performance. At the same time we develop expressions for the rate at which the estimation error decays with number of judges. This will give us the trade-off between the cost and accuracy of using multiple judges.

Consider then the following conditions for the second-stage detector: (1) The judges form independent decisions regarding s . (2) The statistical performance of individual judges differs, but for each judge the hit rate is at least slightly greater than the false alarm rate; i.e., $P(z_i=1|s=1) > P(z_i=1|s=0)$ for all $i=1 \dots n$. (3) The individual decisions from all judges are simply averaged to arrive at a group decision. Denoting $z^{(n)} = \sum z_i / n$ as the average, the group decision is

$$z = \begin{cases} 1, & z^{(n)} \geq \eta^{(n)} \\ 0, & z^{(n)} < \eta^{(n)}, \end{cases} \quad (8)$$

where $\eta^{(n)}$ is an arbitrary judgment criterion. Note here that we do not claim the simple averaging of the judges' responses to be optimal in any sense. We use the average as a means to derive the expressions for convergence which is mathematically tractable. Later we consider how more efficient combination rules may be implied by the results of experiments designed specifically to test the derived analytic expressions. Note too that condition (2) might otherwise be stated that the hit rate of a judge is *more likely* to be greater than their false alarm rate; the statement as given is simply more tractable mathematically.

Now, let $\beta_{JC}^{(n)}$ and $\alpha_{JC}^{(n)}$ denote the group hit and false alarm rates of judges, comparable to β_{JC} and α_{JC} for a single judge. We use a special case of the Hoeffding concentration inequality (Hoeffding, 1963) to show that $\beta_{JC}^{(n)}$ converges rapidly with n to β_C , while $\alpha_{JC}^{(n)}$ converges rapidly to α_C .

That is, for any $\varepsilon > 0$

$$P(z^{(n)} - E[z^{(n)}] \geq \varepsilon) \leq \exp(-2n\varepsilon^2). \quad (9)$$

where $z_i \in \{0, 1\}$ and $S_n = n z^{(n)}$. Now, consider the false alarm rate of the second-stage detector alone; that is, the probability that the judges' decide a signal is present ($z=1$) when the child, in fact, decides it is not ($x=0$). We can express this probability as

$$\begin{aligned}\alpha_j^{(n)} &= P(z^{(n)} \geq \eta^{(n)} | x=0) \\ &= P(z^{(n)} - E_0[z^{(n)}] \geq \eta^{(n)} - E_0[z^{(n)}]),\end{aligned}\quad (10)$$

where E_0 is used here to denote the expected value conditioned on $x=0$. The expression for convergence of $\alpha_j^{(n)}$ follows from Eq. (9) by conditioning on $x=0$ and by setting $\varepsilon = \eta^{(n)} - E_0[z^{(n)}] > 0$ according to condition (2),

$$\alpha_j^{(n)} \leq \exp[-2n\varepsilon^2]. \quad (11)$$

This shows that $\alpha_j^{(n)}$ converges to 0 exponentially as n becomes large. The convergence of the hit rate of the second-stage $\beta_j^{(n)} = P(z^{(n)} \geq \eta^{(n)} | x=1)$ is obtained in a similar fashion from the miss rate

$$1 - \beta_j^{(n)} = P(z^{(n)} - E_1[z^{(n)}] \geq E_1[z^{(n)}] - \eta^{(n)}). \quad (12)$$

Setting $\varepsilon = E_1[z^{(n)}] - \eta^{(n)} > 0$,

$$1 - \beta_j^{(n)} \leq \exp[-2n\varepsilon^2]. \quad (13)$$

This further shows that $\beta_j^{(n)}$ converges exponentially to 1 as n becomes large. Finally, from Eqs. (6) and (7) we determine that as $\alpha_j^{(n)} \rightarrow 0$ and $\beta_j^{(n)} \rightarrow 1$, $\beta_{JC}^{(n)}$ converges exponentially to β_C , while $\alpha_{JC}^{(n)}$ converges exponentially to α_C .

Equations (11) and (13) are useful for estimating the number of judges required to achieve a desired level of accuracy in the estimate of the child's hearing sensitivity. For example, if we wish to limit the false alarm rate of the second stage detector to a value α (Neyman-Pearson criterion) then the number of judges should be at least

$$n \geq \frac{1}{2\varepsilon^2} \ln\left(\frac{1}{\alpha}\right). \quad (14)$$

To get an idea of the rate of convergence implied by this inequality consider the following hypothetical case: The experimenter chooses an arbitrary criterion of $\eta^{(n)}=0.7$ for determining whether the response of the judges as a group is positive or negative. They wish the false alarm rate for the child's response ($x=0$) to be no greater than $\alpha=0.2$. Moreover, they estimate the average of judges' decisions when $x=0$ to be $E_0[z^{(n)}]=0.3$ (yielding a value of $\varepsilon = \eta^{(n)} - E_0[z^{(n)}] = 0.4$). For these conditions the inequality above indicates that the clinician will require $n \geq 5$ judges. Furthermore, by adding two more judges the experimenter can cut this false alarm rate in half (for $\alpha=0.1$, $n \geq 7$). The exponential rate of approximation is encouraging. It suggests that accurate measures of child sensitivity can be obtained in many fewer trials than when a single judge is used. Note in this regard that the conclusions derived here for multiple judges apply as well to multiple trials assuming that the decisions across trials are independent, as they were assumed for the individual judges.

IV. JUDGE'S TASK

Despite its many positives, the observer-based procedure is known to have at least one notable disadvantage: it generally underestimates the sensitivity of the child. The bias occurs because the expected performance of the judge can at best equal, but never exceed that of the child. In the context of the two-stage detector model, detection is limited by the second stage as an added noisy channel. Here we consider a change in the judge's task intended to reduce the bias in estimate.

Consider first that there is no bias when the second stage detector is transparent ($y=x$). Clearly, the sensitivity of the judge must equal that of the child when the judge's decisions exactly mimic those of the child. In theory, then, β_{JC} and α_{JC} are equivalently estimates of β_C and α_C when the judge is instructed to report the decision of the child, x , rather than the presence or absence of the signal. This, parenthetically, is the instruction given to the judge in what is known as behavioral observation audiometry (Thompson and Thompson, 1972). The advantage of this approach is that the judge may be given prior knowledge of the signal on each trial to both improve their performance and to provide an equal opportunity to overestimate as well as underestimate the sensitivity of the child.

Consider then that second stage detector conditions its decision on knowledge of the random variable s on each trial. The hit rate of the two-stage detector in this case is given by Eq. (5) as before. However, $s \rightarrow x \rightarrow z$ now no longer forms a Markov chain; z is now conditioned on s . Reflecting this dependency, the hit rate of the second-stage detector with prior signal knowledge is denoted $\tilde{\beta}_j = P(z=1 | s=1, x=1)$. The probability $\tilde{\beta}_j$ can be expressed as

$$\tilde{\beta}_j = P(z=1 | x=1) \cdot \frac{P(z=1 | s=1, x=1)}{P(z=1 | x=1)} = \beta_j \cdot b, \quad (15)$$

where $\beta_j = P(z=1 | x=1)$, as in the previous section, denotes the hit rate of the second-stage detector *without* prior knowledge of the signal. In a similar fashion, the false alarm rate of the second-stage detector with prior knowledge is given by

$$\tilde{\alpha}_j = P(z=1 | x=0) \cdot \frac{P(z=1 | s=0, x=0)}{P(z=1 | x=0)} = \alpha_j \cdot a. \quad (16)$$

Now, knowing that the signal was present, the judge can improve their performance by more often reporting that the child responded positively. That is, for $P(z=1 | s=1, x=1) > P(z=1 | x=1)$, $b > 1$ so that $\tilde{\beta}_j > \beta_j$. Similarly, knowing that the signal was not present, the judge can improve their performance by more often reporting that the child did not respond; in this case, $a < 1$ so that $\tilde{\alpha}_j < \alpha_j$.

Of course, whether or not judges will make appropriate use of the signal knowledge is ultimately an empirical question. Too much or too little reliance on the signal will either result in an overestimate or underestimate of the sensitivity of the child. However, by allowing the outcome to go in either direction, the average is expected to be a less biased estimator than when all judges underestimate sensitivity, as they can only do not knowing the signal.

Finally, it is important to note here that, while asking the judge to report the child's decision may reduce the problem of the bias in estimate, the cost is a loss of objectivity in the estimate. One can determine with certainty how accurately judges report the presence or absence of a signal, but not the decision of the child. What is needed is a cross check on the judge's performance against some other type of empirical estimate of accuracy. One such measure, which can be applied with or without signal knowledge, is proposed in the next section.

V. HIGH SIGNAL STRENGTH

We consider next a means by which the analytic results of the previous two sections might be put to an empirical test. The method involves presenting the signal at a high level so that, except in the case of a profound hearing loss, one can be reasonably sure that the signal is audible to the child. In this case we expect the hit rate of the child to be near one and the false alarm rate to be near zero. (Note here we take the position, consistent with other authors, that inattention would affect how the child responds to the signal, g in the model, not what they hear, f in the model; cf. [Bargones et al., 1995](#); [Werner and Boike, 2001](#)). Using an asterisk to denote the high signal level condition, we have $\beta_C^* \approx 1$ and $\alpha_C^* \approx 0$. The judge is not given information about the level of the signal; hence, it seems reasonable to assume that neither g nor h is changed by presenting the signal at a high level. If so, then β_C^* and α_C^* can be substituted for β_C and α_C in Eqs. (6) and (7) to yield $\beta_J \approx \beta_{JC}^*$ and $\alpha_J \approx \alpha_{JC}^*$. In other words, the observed hit and false alarm rates of the child-judge duo for the high signal level condition can be used to estimate the hit and false alarm rates of the judge alone. Substituting these empirically obtained values into Eqs. (6) and (7) and solving for β_C and α_C then gives an estimate of the hit and false alarm rate of the child when the signal is presented at a lower level. Specifically,

$$\beta_C \approx \frac{\beta_{JC} - \alpha_{JC}^* \alpha_C}{\beta_{JC}^*}, \quad (17)$$

$$\alpha_C \approx \frac{\alpha_{JC} - \alpha_{JC}^*}{\beta_{JC}^* - \alpha_{JC}^*}, \quad (18)$$

where, β_{JC} and α_{JC} denote, as before, the hit and false alarm rate of the child-judge duo obtained at the lower-level signal.

Note here that β_{JC}^* and α_{JC}^* represent the empirical equivalent of the asymptotic performance of judges as the number of judges increases. Given that the child detects the high-level signal with perfect accuracy, β_{JC}^* and α_{JC}^* are only limited by the random component of the child's response and the perception of the judge. This leads to two outcomes. First, it allows the analytic results of Secs. III and IV to be tested empirically. This can be done by measuring the rate at which β_{JC} and α_{JC} converge to β_{JC}^* and α_{JC}^* for various conditions of feedback and multiple judges, and then comparing the obtained rate in these cases to that predicted by the analytic expressions of Secs. III and IV. Second, it means that the high-level signal method can be used to improve procedures for estimating the child's response. This can be done, for example, by selecting only the best judges to participate in the observation; those previously found to produce the highest values of β_{JC}^* and the lowest values of α_{JC}^* . Or, it can be done by differentially weighting judges' responses based on these values for each judge. It can also be done by informing an appropriate selection of criterion $\eta^{(n)}$ for deciding the group or individual response of judges. Note, in this regard, that having obtained estimates of hit and false alarm rates of the child using the high-level signal method, we also have an estimate of the child's response bias. One may then use this information in subsequent observations to improve estimates by selecting a judgment criterion $\eta^{(n)}$ that more

closely matches that of the child. The benefit could be great as a mismatch of judgment criteria for the two-stage detector model has been shown to have a significantly adverse effect on estimates of sensitivity ([Sorkin et al., 2001](#)). The authors know of two cases in the literature where results from high signal level conditions have been used to calibrate judge's performance in a fashion similar to that proposed here ([Bargones et al., 1995](#); [Werner and Boike, 2001](#)). However, we know of no attempts so far to use such information to specifically select or give differential weighting to judges.

VI. CONCLUSIONS

The observer-based procedure offers an objective behavioral assessment of the hearing sensitivity of young infants who are otherwise incapable of performing visual reinforcement audiometry or other standard behavioral tests. In the present paper, we offer a theoretical analysis of the procedure within the context of a two-stage detection model to derive some analytic results that might be used improve estimates of the hearing sensitivity of infants. Clearly it remains for future studies to determine whether the analytic results presented here will survive empirical tests, and whether they can be used to any advantage in clinical and/or research settings.

ACKNOWLEDGMENTS

The authors wish to thank Dr. Lynne Werner and an anonymous reviewer for helpful comments on an early version of this manuscript. The work was supported by NIDCD Grant No. R01 DC01262-15.

¹Here we assume a model of attention, consistent with the tenets of detection theory, in which inattention is a source of noise that serves to weaken (more or less) the relation between signal and infant response on different trials. An alternative model is to assume that there is absolutely no information in the child's response when they are not attending (i.e., an all-or-none model of attention). While the latter has been used as a convenient vehicle for analyzing the effects of inattention, it has not been suggested in this work as a realistic model of attention ([Viemeister and Schlauch, 1992](#)). Notwithstanding, the analytic development given here would not apply if the all-or-none model is correct.

- Bargones, J. Y., Werner, L. A., and Marean, G. C. (1995). "Infant psychometric functions for detection: Mechanisms of immature sensitivity," *J. Acoust. Soc. Am.* **98**, 99–111.
- Hoeffding, W. (1963). "Probability inequalities for sums of bounded random variables," *J. Am. Stat. Assoc.* **58**, 13–30.
- Moore, J. M., and Wilson, W. (1978). "Visual reinforcement audiometry (VRA) with infants." In *Early Diagnosis of Hearing Loss*, edited by S. E. Gerber and G. T. Mencher (Grune and Stratton, New York), pp. 177–213.
- Olsho, L., Koch, E., Halpin, C., and Carter, E. (1987). "An observer-based psychoacoustic procedure for use with young infants," *Dev. Psychol.* **23**, 627–640.
- Sorkin, R. D., Hays, C. J., and West, R. (2001). "Signal detection analysis of group decision making," *Psychol. Rev.* **108**, 183–203.
- Teller, D. Y. (1979). "The forced-choice preferential looking procedure: A psychophysical technique for use with human infants," *Infant Behav. Dev.* **2**, 135–153.
- Thompson, M., and Thompson, G. (1972). "Responses of infants and young children as a function of auditory stimuli and test methods," *J. Speech Hear Disord.* **15**, 699–707.
- Viemeister, N. F., and Schlauch, R. S. (1992). "Issues in infant psychoacoustics," in *Developmental Psychoacoustics*, edited by L. A. Werner and E. W. Rubel (APA, Washington, DC), pp. 191–209.
- Werner, L. A., and Boike, K. (2001). "Infant's sensitivity to broadband noise," *J. Acoust. Soc. Am.* **109**, 2103–2111.