Speech, Hearing and Language: work in progress

Volume 11

Minimising boredom by maximising likelihood - an efficient estimation of masked thresholds.

Richard J. Baker and Stuart ROSEN

Department of Phonetics and Linguistics
UNIVERSITY COLLEGE LONDON

Minimising boredom by maximising likelihood - an efficient estimation of masked thresholds.

Richard J. BAKER[1] and Stuart ROSEN

Abstract

One of the main problems in carrying out psychoacoustic experiments is the time required to measure a single threshold. In this study we compare the accuracy of threshold estimation in a 2I2AFC task for detecting a 2kHz tone in either a broadband noise or a notched-noise. Tone thresholds were estimated in three normal-hearing listeners using either a Levitt procedure to track 79% correct, or a maximum-likelihood estimation (MLE) procedure to track 70, 80 or 90% correct. Given the chosen parameters for the different procedures, the MLE procedure proved to be approximately 2.5 times faster at estimating masked thresholds than the Levitt procedure. Only thresholds using the 70% MLE procedure were significantly different in magnitude from those obtained using the Levitt procedure. To test the repeatability of the measurements the standard deviations (SD) of the threshold were calculated. Statistical analyses show smallest SDs for the Levitt and 90% MLE procedures, with significantly larger SDs for the 70% and 80% MLE.

Introduction

Of major concern in designing psychoacoustic experiments is not just the issue to be investigated, but also the time available for the experimental tests to be carried out. The desire for efficient threshold estimation has led to the adoption of several different procedures typically using one, two or three alternative forced choice techniques.

Probably the most widely used procedure in psychoacoustics is the adaptive technique based on the transformed up-down procedure described by Levitt (1971). In this technique the initial stimulus is set “above” threshold and subsequent presentation levels are governed by (a) the step-size used and (b) the response to the current stimulus. Correct responses make the task harder by the given step-size and incorrect responses make the task easier. The choice of step-size (fixed or variable), and the patterns of correct/incorrect responses leading to a reversal are described in detail by Levitt (1971).

More recently, considerable interest has been shown in other threshold estimation techniques. In particular, with the advent of increased computing power in laboratories and clinics, the maximum-likelihood estimation (MLE) procedure (Hall, 1968) has become more widely used. The basic premise behind this procedure is that the experimenter assumes a parametric form for the psychometric function, and after each new response the values of the parameters are computed that “…maximise the probability of the set of responses that have been obtained, given the set of stimuli that have been presented”.

Several studies have compared maximum-likelihood threshold estimation techniques with other techniques in either computer simulations, or empirical measurements (e.g.

[1] Centre for Human Communication & Deafness, Faculty of Education, University of Manchester, Oxford Road, Manchester M13 9PL

Pentland, 1980; Hall, 1981; Shelton et al., 1982; Shelton and Scarrow, 1984; Madigan and Williams, 1987; Green, 1990; Gu and Green, 1994 and Saberi and Green, 1997).

The motivation behind the present study was to evaluate the MLE procedure in a notched-noise masking task, where the aim of the task is to obtain the threshold of a tone presented in a broadband noise (with or without a spectral notch around the tone frequency). This notched-noise masking procedure has been widely used in estimates of auditory frequency selectivity (e.g. Patterson, 1976; Patterson and Moore, 1986) and typically requires 10-16 thresholds to be measured at differing notch widths to obtain an accurate description of the auditory filter shape at one level and frequency. Recent systematic attempts to describe how auditory filtering changes across level have required as many as 160 threshold measurements at one frequency (Rosen and Baker, 1994; Rosen et al., 1998). The benefits of an efficient technique in such studies are obvious, especially if they are to be applied clinically.

This study set out to empirically compare implementations of these two adaptive threshold estimation procedures, and in particular to compare how the choice of performance level affects the variability of threshold estimation in the MLE procedure. When using the MLE procedure, Green (1990) suggested tracking a performance level so as to minimise the variability of the estimated threshold on each successive trial. The variance of the estimate (eq. 2 of Green, 1990) is given by:

$$\sigma^2 = \frac{[p(1-p)]}{\left(\frac{dF}{dx}\right)^2} \qquad \text{equation 1}$$

where *dF/dx* is the slope of the psychometric function and *p*(1-*p*) is an estimate of the variance of the estimated probability *p*. Combining this with a logistic model of the psychometric function (see below) leads to a probability of 0.809 at which the variance is a theoretical minimum, the so-called *sweetpoint*. This would suggest that, if the logistic function is a realistic model for the psychometric function for a tone-in-noise masking task, the least variability in threshold measurement would be achieved by placing the stimulus at the 81% point on the psychometric function (a point close to the 79.4% given by a three-down/one-up Levitt type procedure). Thus we chose to compare this Levitt procedure with 3 implementations of the MLE procedure, placing the stimulus at the 70%, 80% or 90% point on the psychometric function.

Method

Notched-noise masked thresholds were measured in three normal hearing listeners (<20dB HL). The notched-noise conditions were chosen to be representative of the studies of Rosen et al. (Rosen and Baker, 1994; Rosen et al., 1998). The masker noise consisted of either a broadband noise (400-3600 Hz), or the same noise with a spectral notch (1200-2800 Hz). The probe-tone frequency was 2000 Hz in all cases. For each of the notched-noise conditions, the masked threshold was measured for both a fixed-masker spectrum level of 50 dB SPL (probe-tone level adjusted to find threshold) and a fixed probe-tone level of 50 dB SPL (masker spectrum level adjusted). The former of these two is the procedure that has been typically used, while Rosen et al. (1998) argue that the latter is more appropriate given the nature of the auditory filter nonlinearity.

For each of the 4 conditions (2 notches x 2 levels) thresholds were measured using 4 tracking procedures (see below for details) in a two-interval two-alternative forced choice task. For each of these 16 conditions, the threshold estimates were repeated 16 times to give an estimate of the repeatability of each procedure. Thus a total of 256 thresholds were measured per subject.

All the stimuli were software generated and presented via Tucker-Davis AP1/DD1 D-A converters (40kHz sampling frequency), anti-aliasing filters (Kemo, 48 dB/oct, 10 kHz), PA4 attenuators, SM3 mixer, headphone amplifier and Etymotic ER-2 insert earphone monaurally to the subject's right ear.

Adaptive techniques

Transformed up-down adaptive staircase

The "base-line" threshold estimates were made using the procedure described by Levitt (1971) in which the subject must respond correctly three times before the task is made more difficult, and easier after one wrong response. This procedure tracks the 79.4% point on the psychometric function, and is the same as that used by Rosen et al. (1998). An initial step size of 5 dB was used, which was decreased by 1 dB after each turnaround until a final step size of 2 dB was reached. Once this final step size was reached the average of the following 8 turnarounds was taken as the threshold.

Maximum-likelihood estimation

The maximum likelihood procedure used was similar to that described by Green (1990). A logistic function was chosen to represent the form of the psychometric function. This function can be written as:

$$p(x) = f + \frac{f}{(1 + e^{s(m-x)})} \qquad \text{equation 2}$$

where *p(x)* is the probability of a correct response given a stimulus value *x* in dB and *m* represents the mid-point of the psychometric function. In this case *m* equates to a probability of 0.75 or 75% percent correct, since the false-alarm rate (*f*) is 0.5 for a two-alternative forced choice task.

As discussed above, the sweetpoint for this logistic function occurs at a probability of 0.809 (80.9% correct). To present the stimuli at the sweetpoint, a midpoint of value *m* gives 75% correct so it is a trivial matter to calculate the stimulus level *x* to give a 80.9 % rate of correct responses. Similarly, the presentation level of the stimulus is adjusted in this study to estimate the 70, 80 or 90% correct points on the psychometric function.

Given the psychometric function described above (and a fixed value of the slope – see below) a range of possible midpoints was chosen such that the upper end of the range was 10-20 dB above the estimated masked threshold, and the total range of possible midpoints was 60 dB. The spacing between possible midpoints was 1 dB.

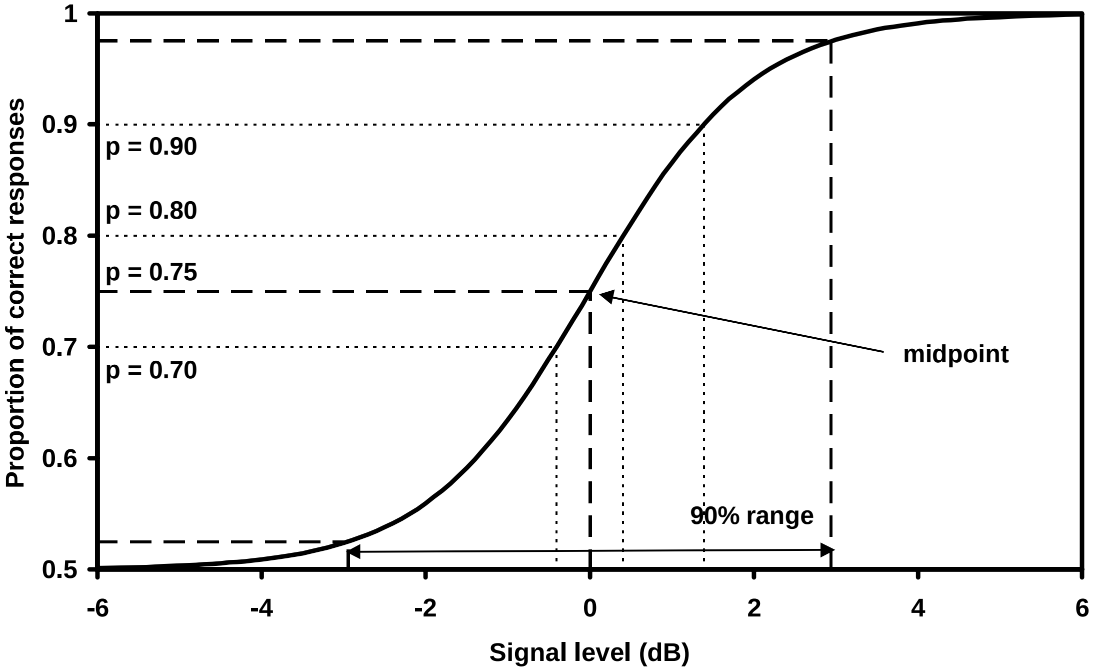

Figure 1. *Psychometric function used in the maximum-likelihood estimation of threshold. The slope of the function is fixed to a value of 1.0 (see text for details)*

After each stimulus presentation and response the likelihood was calculated for each midpoint within the above range based on all the responses obtained thus far [for a correct response probability = *p(x)*, for incorrect response probability = 1-*p(x)*]. The midpoint is then chosen that gives the greatest likelihood of fitting the data thus far. From this midpoint the next stimulus level is calculated as required to satisfy the desired performance criterion based on this best fitting function.

This procedure is repeated until a predefined stopping criterion is achieved. The procedure was successfully halted if, after a minimum of 15 trials, the standard deviation of the last 10 presentations was below 1 dB. The final threshold was then calculated based on all the presentations used. To avoid large changes in stimulus level, levels were not permitted to change by more than 10dB from one trial to the next. If this criterion was not reached within a maximum of 50 trials the procedure was halted, and the result from that run was discarded.

Estimation of psychometric function slope.

While it is possible to use the MLE procedure to estimate the slope of the psychometric function as well as its midpoint, Green (1990) showed that even a relatively large mismatch between the slope used in the MLE procedure and that of the underlying psychometric function had little effect on the measured thresholds. Here we use a fixed slope based on estimates from previous notched-noise masking experiments. Notched-noise masked thresholds were obtained using a 2kHz tone over a range of fixed masker and fixed probe levels and 16 different notch conditions in three normal hearing subjects (subjects JD, RJB and WC in Rosen et al., 1998). These thresholds were obtained using the 3-down, 1-up transformed adaptive procedure to track 79.4% correct (Levitt, 1971). For each threshold measurement, a logistic regression was used to fit the above psychometric function (equation 2) in order to estimate the slope. The distributions of the fitted slopes are shown in figure 2. The mean fitted slopes for the three subjects are 0.98, 1.01 and 0.83. To approximate these, a value of 1.0 was chosen for use in the maximum likelihood procedure.

It should be noted that the minimum step size of 2 dB used by Rosen et al. (1998) restricts the range of meaningful values of the fitted slope. For example, a slope of 3 means that 90% of the transition region of the psychometric function lies within a 1.96 dB range. Thus a fitted slope value of much greater than 3 is somewhat meaningless when the step size is 2 dB.

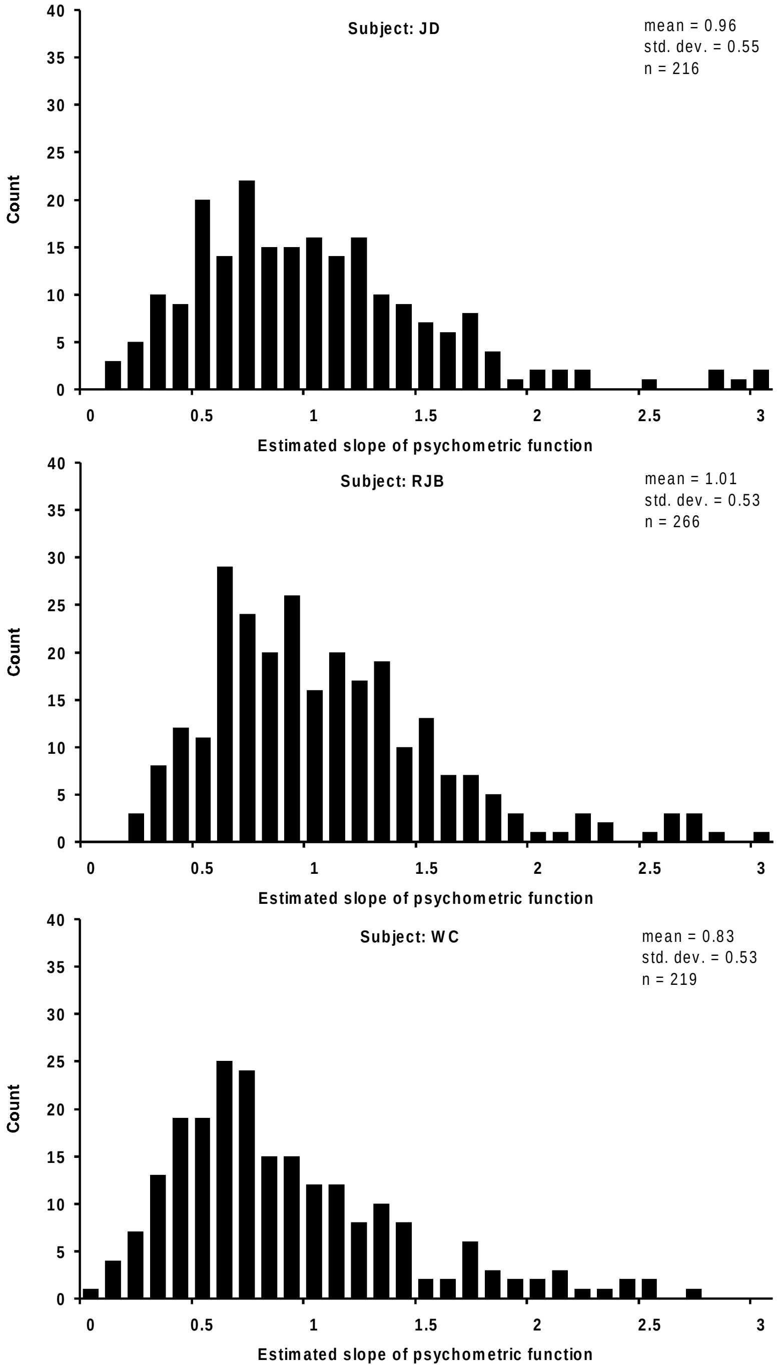

Figure 2. *Estimated slopes of psychometric functions calculated from notched-noise masking experiments using a transformed up-down procedure to estimate the 79% correct point on the psychometric function.*

Results

Comparison of measured thresholds.

In the present implementation of the MLE procedure, the stimulus is always presented at the chosen point on the psychometric function (70, 80 or 90%). To allow comparison of the different procedures, the estimated thresholds were adjusted to the midpoint of the psychometric function (the 75% correct point). This adjustment was made easy by the use of a fixed psychometric function slope in the MLE procedure, and the assumption that this psychometric function shape was also appropriate for the Levitt procedure. The adjustment is simply a different additive constant depending on which point on the psychometric function was being estimated.

Figure 3 shows box-plots of the thresholds adjusted as above for each of the three listeners, 4 tracking procedures and 4 notch conditions (16 repeated measurements at each condition). The key points to note from figure 3 are:

(a) The results are consistent across the three listeners

(b) As expected, the mean thresholds vary considerably between the 4 different masker conditions

(c) The smallest inter-quartile ranges result from using the Levitt procedure and the MLE to estimate the 90% point.

(d) Estimating the 70% point with the MLE procedure produces the largest inter-quartile range.

(e) The MLE procedure results in an asymmetric distribution of thresholds: It is less likely that the procedure will return from an extreme stimulus value when the probe is inaudible than when it is audible. This is less evident when tracking a higher proportion correct.

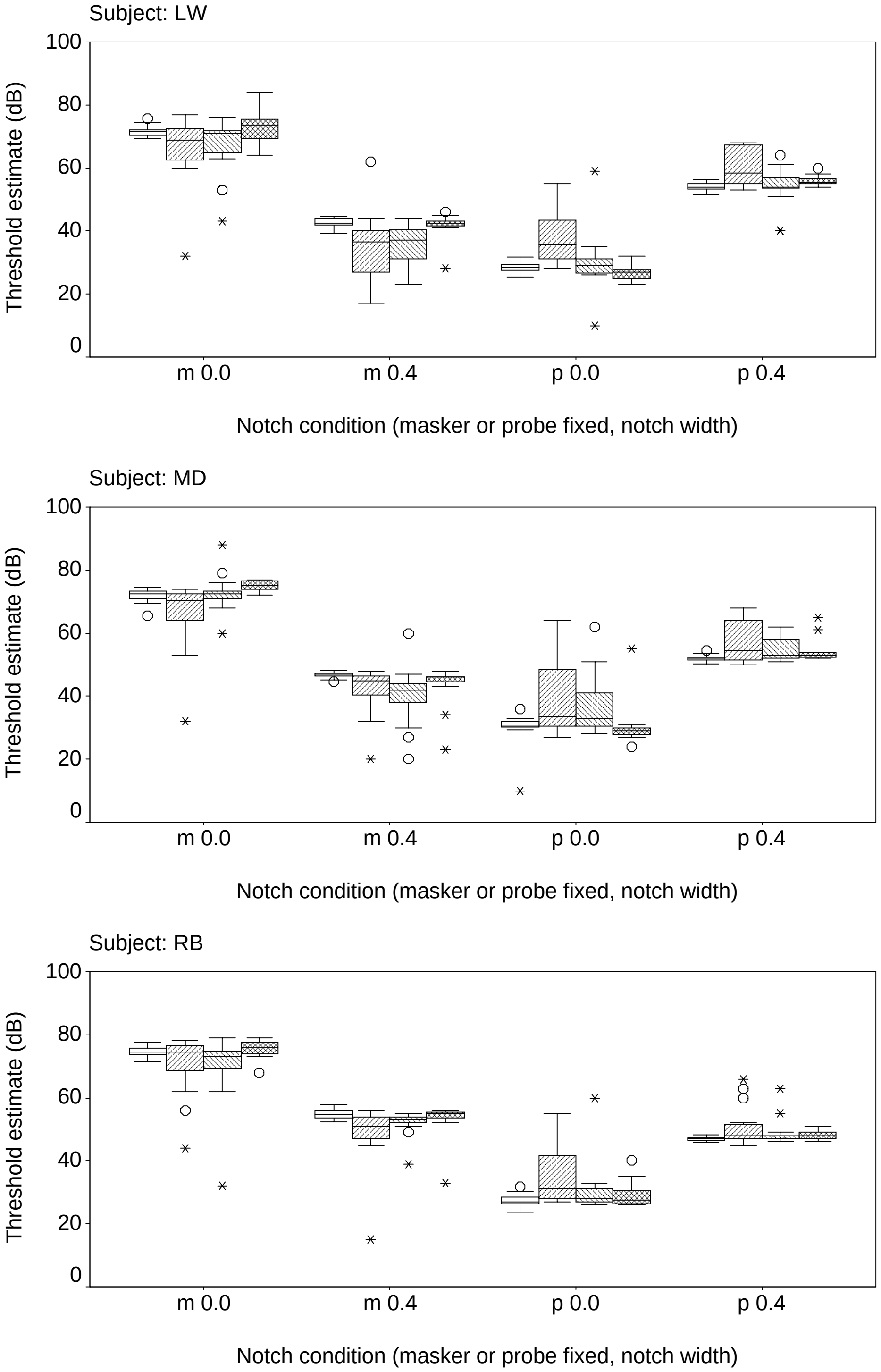

Figure 3. *Box-plots of notched-noise masked thresholds. Each plot shows the median, inter-quartile range (box), outliers (1.5 < o < 3.0 times inter-quartile range from box edge), extremes (* > 3.0 times inter-quartile range from box edge) and range excluding outliers and extremes (whiskers). For each notch condition the data are arranged in the adaptive procedure order: Levitt, 70%, 80% and 90%.*

Analysis of variance of estimated thresholds.
The data represented in figure 3 were first analysed using repeated measures analysis of variance with two factors of masker condition and tracking procedure (each with 4 levels; Howell, 1996). There was a significant interaction between notch and tracking procedure (p<0.001). This interaction, the fact that introducing a spectral notch into a broadband noise reduces the masked threshold, and that different levels of masker result in different masked thresholds makes statistical analysis of the effect of the notch condition somewhat meaningless. However, the existence of a significant interaction term in the analysis is important as it shows that the adaptive procedures produce different trends depending on the configuration of the masking task. This interaction is evident from figure 3 where the median threshold for the 70% and 80% MLE procedures tend to be above those for the Levitt and 90% MLE procedures for the fixed probe conditions, and below for the fixed masker conditions.

To investigate whether the different adaptive procedures resulted in different threshold measurements, the data were partitioned into the 4 separate masker conditions and re-analysed using a one-factor repeated measures ANOVA, with adaptive procedure type being the 4 level factor.

For each of the 4 masker conditions the estimated threshold showed significant variation between the four adaptive procedures (p < 0.05). A pairwise comparison revealed that the mean threshold estimated using the MLE 70% procedure always gave an estimate significantly different from the Levitt procedure, and that there was never a significant difference between the Levitt procedure and the 80 and 90% MLE procedures. The mean thresholds and the groups revealed by the pairwise comparison are shown in table 1.

Pairwise comparison of means (sig. level 0.05)				
notch condition	group I	group II	mean	std. error.
M 0.0	70%		67.693	1.438
	80%	80%	69.787	1.732
		Levitt	72.753	0.987
		90%	74.456	0.733
M 0.4	70%		41.880	4.078
	80%	80%	42.516	4.925
	90%	90%	46.243	3.578
		Levitt	48.066	3.603
P 0.0	90%		28.694	1.057
	Levitt		28.597	0.709
	80%	80%	32.359	2.269
		70%	37.099	1.201
P 0.4	Levitt		51.101	2.130
	90%		52.715	2.407
	80%	80%	52.380	1.897
		70%	56.266	2.914

Table 1. *Pairwise comparison of mean thresholds for each masker condition. Measurement procedures within the same group do not produce significantly different thresholds from each other.*

Variability of threshold estimates.

Along with the average threshold values measured, it is important to take into account the repeatability of the threshold estimates produced by a particular procedure. It is clear from fig. 3 that the variability is not the same for all 4 measurement procedures. In order to compare the variability of threshold estimates from the four measurement procedures (and for across the four masker conditions) the spread of the 16 repeated threshold estimates was quantified (standard deviation and inter-quartile range) and a two factor repeated measures ANOVA used for the comparison.

Using the standard deviations as the measure of variability, there was a significant difference between the four measurement procedures (p < 0.0001), and a borderline effect of notch condition (p = 0.05). The resulting groups derived from a pairwise comparison for the different adaptive procedures are shown in table 2. It is clear from this that the Levitt procedure produces the smallest variability of threshold measurement followed by the MLE procedure tracking the 90% correct point, while the MLE at 80% and 70% produce the most variable estimates of masked threshold.

Pairwise comparison of means (sig. level 0.05)				
Group			Mean	Std. Error
I	II	III		
Levitt			1.808	0.306
	90%		3.717	0.429
		80%	7.181	0.347
		70%	8.889	0.178

Table 2. *Pairwise comparison of average standard deviations, each calculated from the 16 repeated measures within each cell and pooled across masker condition and subjects.*

Using the inter-quartile range as the measure of variability reveals much the same picture as the standard deviations i.e. significant effect of procedure on the variability of threshold measurement (p = 0.001), with no significant effect of notch condition. The grouping of the four procedures, obtained from a pairwise comparison for the adaptive procedures (table 3), show that the MLE 70% procedure produces significantly more variability than the other three procedures which don’t differ significantly from each other at the p = 0.05 level.

Pairwise comparison of means (sig. level 0.05)			
Group		Mean	Std. Error
I	II		
Levitt		1.913	0.159
90%		2.875	0.407
80%		5.708	1.198
	70%	11.250	1.375

Table 3. *Pairwise comparison of average inter-quartile ranges, each calculated from the 16 repeated measurements within each cell and pooled across masker condition and subjects.*

Speed of threshold estimation.

While the rules for stopping the adaptive tracking are, more often than not, somewhat arbitrary it is clear that there are an infinite number of permutations that could be used. The criteria for the Levitt procedure were the same as those used by Rosen et al. (1998). For the MLE procedure the aim was to achieve a similar level of performance with as few trials as possible. Thus, rather than using a fixed number of trials, a stopping criterion as described previously was used in an attempt to obtain a stable threshold measurement as quickly as possible. Clearly, tightening this criterion would result in the procedure requiring an increased number of trials before the criterion was met.

As well as the masked threshold, the number of trials needed to obtain each threshold was also recorded. Using a two factor repeated measures ANOVA, there was a significant effect of the type of adaptive procedure on the number of trials needed to measure the threshold ($p < 0.001$), but no significant effect of notch condition and no significant interaction. Pairwise comparison of the means (table 4, pooled across notch conditions) showed that the Levitt procedure took approximately 2.5 times more trials to estimate thresholds than the three implementations of the MLE procedure used in this study.

Pairwise comparison of means (sig. level 0.05)			
Group		Mean	Std. Error
I	II		
70%		18.66	0.06
80%		19.75	0.44
90%		20.58	1.11
	Levitt	50.82	2.04

Table 4. *Pairwise comparison of average number of trials to estimate each threshold, calculated from the 16 repeated measurements within each cell and pooled across masker condition and subjects.*

Discussion

Adaptive tracking procedures of various forms have widely been used in psychoacoustic experiments to estimate different types of thresholds. For tone-in-noise masking experiments the transformed up-down adaptive procedures described by Levitt (1971) have been the procedure of choice. These have typically used either the two-down/one-up method to estimate the 70.7% point on the psychometric function, or the three-down/one-up method to estimate the 79.4% (other rules have also been put forward, e.g. Levitt, 1971, Brown, 1996). As an alternative, the maximum likelihood estimation procedure Pentland (1980), and variations thereof, have recently been more widely utilised in attempts to find a more efficient (i.e. quicker) estimation of threshold.

Green (1990) showed that, for a given number of trials, stimulus presentation at the sweetpoint resulted in lower variability than stimulus presentation at other performance levels. Following this argument, estimating thresholds by placing the stimuli at $p=0.809$ (the sweetpoint for the logistic model of the psychometric function used here) should result in the smallest variability of threshold estimates. That is, estimation of the 80% correct point in the present study should result in a smaller variability than at 70% or 90%. That this is not the case suggests that either the chosen logistic model of the psychometric function is incorrect for this task, or that other factors are also coming into play. Taking the first point, Green (1990) concluded that the variability of the threshold estimates (in simulations) is “not strongly affected by enormous mismatch between the observer’s psychometric function and that used in the maximum-likelihood analysis”. Furthermore, the form of psychometric function used in the MLE procedure in the present study was not arbitrary, but was based on previous threshold estimations in the same type of masking task and is thus unlikely to be very different from the true psychometric functions of the listeners. Thus it seems unlikely that inaccuracy in the model chosen for the listeners psychometric function is responsible for the improvement in measurement variability at 90% correct over that of 80%.

Comparison of the mean thresholds (table 1) shows a relatively large difference between the thresholds measured using the three MLE procedures. Specifically, tracking 70% correct results in thresholds that on average are 7.6 dB better than when tracking 90%. A difference in this direction is to be expected. However the magnitude of the difference is far greater than the theoretical 1.8 dB that would be expected given the psychometric function used in the MLE procedure. It is also clear from figure 3 that the spread of measurements for the 70% MLE is not only larger than for the 90%, but that it is asymmetric in that the 70% procedure tends to overestimate the listeners ability to detect the tone in the masking noise. Indeed, in some estimates of the threshold, this overestimation shows up as extremes in the box-plots of fig.3. Analysis of the standard deviations of the estimated thresholds shows that the Levitt procedure results in significantly less variability than when tracking 90% correct using the MLE procedure which in turn is significantly less variable than using MLE to track 70% or 80% correct. However, when the inter-quartile ranges, rather than standard deviations, the differences are much less significant.

In terms of running the experiment, several correct guesses when the tone is inaudible results in the tone level being decreased to well below threshold (or masker level increased if the tone level is fixed). When such a large “deviation” occurs, tracking the 90% correct allows the MLE procedure to get back to the “true” threshold much more reliably than using the 70% point. Related behaviour was also noted in computer simulations by Green, 1990; his fig. 8) in which tracking 94% percent correct resulted in estimates converging to within 1dB of threshold in 20 trials, while tracking 70.7% took about 100 trials to reach the same level of accuracy. Green (1990) also suggested that such behaviour would be evident as a bias in the estimate if insufficient trials were used in the measurement.

Such a bias is clearly evident in the present results, with two key differences. Firstly, the bias that results from estimating a low percentage correct tends to overestimate the sensitivity in that the procedure seems to be abnormally influenced by correct guesses when the signal is inaudible (false alarms). Secondly, the stopping criterion used in the present study relies on the standard deviation of 10 successive trials becoming less than 1dB (after a minimum of 15 trials). Clearly, reducing the limiting standard deviation or increasing the minimum number of trials would result in a greater degree of “success” when tracking 70% correct, as both would result in a greater number of trials over which the MLE procedure estimates the threshold, thus reducing the effects of false alarms.

It is clear from these and other results that an MLE procedure can offer significant speed advantages over the more traditional transformed up-down procedures. However, such an advantage may be offset by an increased tendency for the estimated thresholds to follow a somewhat skewed distribution. Such a skewing of estimated thresholds is particularly evident when the MLE procedure is used to estimate relatively low performance levels (e.g. 70% correct in a 2I2AFC task). These difficulties may be overcome by increasing the number of trials or tightening the stopping criteria. Both however would be at the expense of the speed of threshold estimation.

Acknowledgements

This work was supported by the Wellcome Trust (Grant No.046823/Z/96) and Defeating Deafness: The Hearing Research Trust (Grant No.161:UCL:SR).

References

Brown, L. (1996). “Additional rules for the transformed up-down method in psychophysics.” Perception and Psychophysics **58**(6): 959-962.

Green, D. M. (1990). “Stimulus selection in adaptive psychophysical procedures.” J. Acoust. Soc. Am. **87**(6): 2662-2674.

Gu, X. and Green, D. (1994). “Further studies of a maximum-likelihood yes-no procedure.” J. Acoust. Soc. Am. **96**(1): 93-101.

Hall, J. L. (1968). “Maximum-likelihood procedure for estimation of psychometric functions.” J. Acoust. Soc. Am. **44**(1): 370.

Hall, J. L. (1981). “Hybrid adaptive procedure for estimation of psychometric functions.” J. Acoust. Soc. Am. **69**(6): 1763-1769.

Howell, D. C. (1996). Statistical Methods for Psychology. Belmont, CA, Wadsworth Inc.

Levitt, H. (1971). “Transformed up-down methods in psychoacoustics.” J. Acoust. Soc. Am. **49**(2): 467-477.

Madigan, R. and Williams, D. (1987). “Maximum-likelihood procedures in two-alternative forced-choice: Evaluation and recomendations.” Perception and Psychophysics. **42**(3): 240-249.

Patterson (1976). “Auditory filter shapes derived with noise stimuli.” J. Acoust. Soc. Am. **59**: 640-645.

Patterson, R. D. and Moore, B. C. J. (1986). Auditory filters and excitation patterns as representations of frequency resolution. In: Frequency Selectivity in Hearing. B. C. J. Moore. London, Academic Press**:** 123-177.

Pentland, A. (1980). “Maximum likelihood estimation: The best PEST.” Perception and Psychophysics **28**: 377-379.

Rosen, S. and Baker, R. J. (1994). “Characterising auditory filter nonlinearity.” Hear. Res. **73**: 231-243.

Rosen, S., et al. (1998). “Auditory filter nonlinearity at 2kHz in normal listeners.” J. Acoust. Soc. Am. **103**(5): 2539-2550.

Saberi, K. and Green, D. M. (1997). “Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics.” Perception and Psychophysics **59**(6): 867-876.

Shelton, B. R., et al. (1982). “Comparison of three adaptive psychophysical procedures.” J. Acoust. Soc. Am. **71**(6): 1527-1533.

Shelton, B. R. and Scarrow, I. (1984). “Two-alternative versus three-alternative procedures for threshold estimation.” Perception and Psychophysics **35**(4): 385-392.