British Journal of Audiology, 2001, **35**, 43–52

Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking

R.J. Baker[1] and S. Rosen[2]

[1]*Centre for Human Communication and Deafness, The University of Manchester, UK and*
[2]*Department of Phonetics and Linguistics, University College London, UK*

(Received 2 May 2000; accepted 26 July 2000)

Abstract
There has been much recent interest in the use of adaptive psychophysical procedures based on maximum-likelihood estimation (MLE) in order to minimize testing time. The speed and accuracy of MLE was compared to a standard transformed up–down algorithm in a two-interval forced-choice task. Thresholds for detecting a 2 kHz tone in either a broadband or a notched-noise were estimated in three normal-hearing listeners. The transformed up–down algorithm tracked 79% correct with either two, four, six or eight final turnarounds, whereas the MLE procedure tracked 70%, 80% or 90% correct. MLE was always quickest, but with a penalty in increased variability. Use of the MLE procedure to track 70% or 80% correct also resulted in a tendency to overestimate listeners' sensitivity. Reducing the number of turnarounds in the up–down procedure from eight to two reduced the number of trials required by nearly half and resulted in thresholds with similar magnitude and variability to those obtained using MLE to track 90% correct.

Key words: adaptive procedure, maximum likelihood estimation (MLE), threshold measurement

Introduction

One of the major concerns when 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 the step-size used and by the response to the current stimulus. Correct responses lead to the task being made 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 essential idea here is that a parametric form for the psychometric function is assumed, whose parameters are set so that the probability of the set of obtained responses is maximized given the set of stimuli presented so far. This function is then used to determine the level of the next presentation – the level most likely to lead to the desired probability of a correct response (typically 70–80%). Several studies have compared MLE techniques with other techniques in either computer simulations or empirical measurements (Pentland, 1980; Hall, 1981; Shelton et al., 1982; Shelton and Scarrow, 1984; Madigan and Williams, 1987; Green, 1990; Gu and Green, 1994; Saberi and Green, 1997).

Address for correspondence: R.J. Baker, Centre for Human Communication and Deafness, Faculty of Education, The University of Manchester, Oxford Road, Manchester M13 9PL (E-mail: richard.baker@man.ac.uk).

Shelton et al. (1982) showed that MLE and Levitt procedures produced similar thresholds and variability, but that MLE was able to converge on the threshold in fewer trials. This suggests that MLE may be more useful in situations where time is an important constraint. Also, Green (1990) suggested that it was possible, at least theoretically, to minimize the variability of the obtained thresholds by the appropriate choice of the level of performance to track in MLE. For a logistic model, the level which results in minimum variability (the so-called 'sweetpoint') occurs for a probability of 0.809 (close to the 0.794 given by a three-down/one-up Levitt procedure). Thus Green (1990) recommends that the next stimulus value chosen equates to a probability of 0.809 on the current best estimate of the psychometric function.

The motivation behind the present study was to evaluate the MLE procedure in a notched-noise masking task and, in particular, to ascertain whether significant benefits could be gained by tracking performance levels near the sweetpoint. The aim of the masking 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 (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. In the present study, this Levitt procedure, as used by Rosen et al. (1998), was compared with three implementations of the MLE procedure, placing the stimulus at the 70%, 80% or 90% correct point on the psychometric function.

Method

Notched-noise masked thresholds were measured in three normal hearing listeners (<20 dB HL, 125 Hz to 8 kHz). The notched-noise conditions were chosen to be representative of the studies of Rosen and coworkers (Rosen and Baker, 1994; Rosen et al., 1998). In all cases a probe-tone frequency of 2000 Hz was used. The masker noise consisted of either a broadband noise (400–3600 Hz), or the same noise with a spectral notch (1200–2800 Hz). In keeping with the previous literature on notched-noise masking, these are referred to as normalized notch widths of 0.0 and 0.4, respectively. 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 is the procedure that has typically been used, whereas Rosen et al. (1998) argue that the latter is more appropriate given the nature of the auditory filter non-linearity.

For each of the four conditions (two notches x two levels) thresholds were measured in a two-interval two-alternative forced choice task using four tracking procedures (see below for details). 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 converter (40 kHz sampling frequency), anti-aliasing filter (Kemo, 48 dB/oct, 10 kHz), PA4 attenuators, SM3 mixer, headphone amplifier and Etymotic ER-2 insert earphone monaurally to subjects' right ears.

Adaptive techniques

Transformed up–down adaptive staircase

The 'baseline' threshold estimates were made using the procedure described by Levitt (1971) in which the task is made more difficult after three correct responses, and made 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) and was chosen to be representative of those used in similar psychoacoustic tasks.

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 levels at the following eight turnarounds was taken as the threshold. From each adaptive track, the average level at the first two, first four and first six turnarounds was also calculated. Of course, the parameters used to control the Levitt algorithm may

not be optimal in terms of speed or accuracy and their adjustment could speed up the procedure.

MLE

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:

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

this should read: (1 + e...)

where $p(x)$ is the probability of a correct response given a stimulus value x in dB. The parameter m represents level at the mid-point of the psychometric function. In this case m is the level giving a probability of 0.75 or 75% correct, since the false-alarm rate (f) is 0.5 for a two-alternative forced-choice task. The parameter s represents the slope of the psychometric function.

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 mid-point of value m gives 75% correct so it is a trivial matter to calculate the stimulus level x to give an 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 mid-points was chosen so that the upper end of the range was 20–30 dB above the estimated masked threshold and the total range of possible mid-points was 60 dB. The spacing between possible mid-points was 1 dB.

After each stimulus presentation and response the likelihood was calculated for each mid-point within the above range based on all the responses obtained thus far.* The mid-point was then chosen that gave the greatest likelihood of fitting the data thus far, and from this mid-point the next stimulus level was calculated as required to satisfy the desired performance criterion based on this best fitting function.

This procedure was repeated until a predefined stopping criterion was 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. If this criterion was not reached within a maximum of 50 trials the procedure was halted, and the result from that run was discarded. To avoid large changes in stimulus level, levels were not permitted to change by more than 10 dB from one trial to the next. This was especially important for the fixed-probe conditions. Without a limitation on the size of level changes, the masking noise would change from quiet to very loud from the first trial to the second, possibly startling the listener.

Estimation of psychometric function slope

Although it is possible to use the MLE procedure to estimate the slope of the psychometric function as well as its mid-point, 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, a fixed slope was used, based on estimates from previous notched-noise masking experiments (16 different notch conditions at a range of levels in three normal hearing subjects; subjects JD, RJB and WC in Rosen et al., 1998). These thresholds were obtained by use of the three-down, one-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 1) in order to estimate the slope. The mean fitted slopes (standard deviation; number of thresholds) for the three subjects were 0.98 (0.55; 216), 1.01 (0.53; 266) and 0.83 (0.53; 219). To approximate these, a value of 1.0 was chosen for use in the maximum-likelihood procedure. A slope value of 1.0 equates to a psychometric function that covers a range of 4.4 dB between 55% and 95% correct. Alternatively, a 2-dB step (as used here in the Levitt procedure) centred around the mid-point would move between 63.4% and 86.6% correct.

*After each trial the probability (p) of a correct response for that stimulus level is calculated from equation 1. The likelihood of a correct response is given by p, and of an incorrect response by (1–p). The appropriate value for the stimulus/response pair is calculated and stored for each of the possible psychometric functions (60 in this study). After each trial, a cumulative product of the likelihoods for each psychometric function is stored and the function with the maximum value used to determine the next stimulus level.

Results

Comparison of measured thresholds

In the present implementation of the MLE procedure, the stimulus is always presented at the current best estimate of 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 assumption 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. For a fixed slope of 1, the largest adjustment required, for the 90% MLE procedure was 1.4 dB. Thus comparisons of estimated thresholds were made for seven 'procedures': three MLE procedures and the Levitt procedure with thresholds estimated after two, four, six or eight final turnarounds.

Figure 1 shows box-plots of the thresholds adjusted as above for each of the three listeners and four notch conditions (16 repeated measurements at each condition). The key points to note from Figure 1 are:

- The results are consistent across the three listeners.
- As expected, the mean thresholds vary considerably between the four different masker conditions.
- The largest interquartile ranges result from use of the MLE procedure to track 70% and 80% correct.
- 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.

Analysis of variance of estimated thresholds

The data represented in Figure 1 were first analysed using repeated measures analysis of variance (ANOVA) with the two factors of masker condition (four levels) and tracking procedure (seven levels). There was a significant interaction between notch and tracking procedure ($p<0.001$), showing that the adaptive procedures produce different trends depending on the configuration of the masking task. This interaction is evident from Figure 1 where the median thresholds 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 those for the fixed masker conditions (also evident in Table 1).

To investigate whether the different adaptive procedures resulted in different thresholds, the data were partitioned into the four separate masker conditions and re-analysed using four one-factor repeated measures ANOVA.

For each of the four masker conditions the estimated thresholds showed significant differences across the seven adaptive procedures ($p = 0.011$ for masker-fixed and no-notch condition; $p<0.001$ for all others). A pairwise comparison (Tukey HSD; $\alpha = 0.05$) revealed that the mean threshold estimated by use of the MLE 70% procedure always differed significantly from all thresholds obtained by use of the Levitt procedure, and that there was never a significant difference between thresholds for the Levitt procedure and the 90% MLE procedure. Strikingly, the threshold estimate from the Levitt procedure did not depend upon the number of turnarounds. Mean thresholds and the groups revealed by the pairwise comparison are shown in Table 1.

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 Figure 1 that variability is not the same for all seven measurement procedures. In order to compare the variability of threshold estimates from the seven measurement procedures the spread of the 16 repeated threshold estimates was quantified (standard deviation and interquartile range) and pooled across notch conditions. A single factor repeated measures ANOVA was used for the comparison.

Using standard deviations as the measure of variability, there was a significant difference between the seven measurement procedures ($p<0.001$). The resulting groups derived from pairwise comparisons for the different adaptive procedures are shown in Table 2. It is clear from this that all the Levitt procedures produce the smallest variability followed by the MLE procedure tracking the 90% correct point. The MLE at 80% and 70% correct produce the most variable estimates of masked threshold. It is worth noting that reducing the number of final turnarounds in the Levitt procedure does not produce a significantly greater variability in the threshold

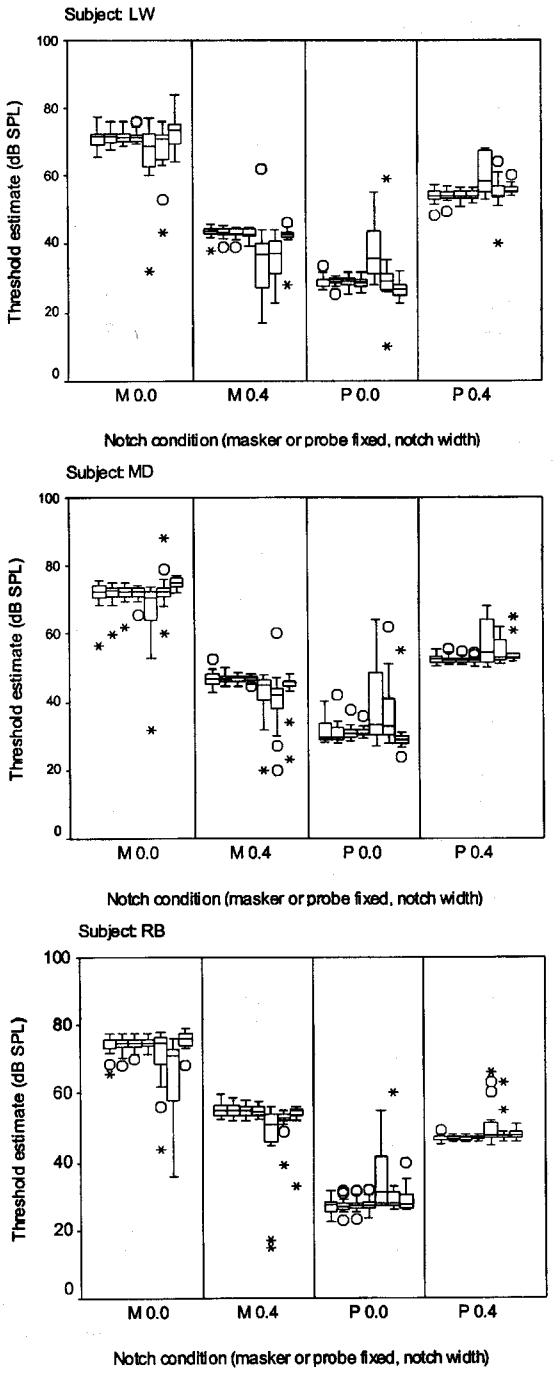

Fig. 1. Box-plots of notched-noise masked thresholds. Each plot shows the median, interquartile range (box), outliers ($1.5 < x < 3.0$ times interquartile range from box edge), extremes ($* > 3.0$ times interquartile range from box edge) and range excluding outliers and extremes (whiskers). For each notch condition the data are arranged in the adaptive procedure order: Levitt two-turns, Levitt four-turns, Levitt six-turns, Levitt eight-turns, MLE 70%, MLE 80% and MLE 90%.

Table 1. Pairwise comparison of mean thresholds for each masker condition*

Notch	Procedure	Pairwise comparison of means (significance level 0.05) Mean threshold		
		I	II	III
M 0.0	MLE 70%	67.69		
	MLE 80%	68.25		
	Levitt (2 turns)		72.14	
	Levitt (4 turns)		72.48	
	Levitt (6 turns)		72.66	
	Levitt (8 turns)		72.78	
	MLE 90%		74.45	
M 0.4	MLE 70%	41.26		
	MLE 80%	42.52		
	MLE 90%		46.24	
	Levitt (8 turns)		48.09	
	Levitt (6 turns)		48.20	
	Levitt (4 turns)		48.32	
	Levitt (2 turns)		48.46	
P 0.0	MLE 90%	28.69		
	Levitt (8 turns)	29.05	29.05	
	Levitt (6 turns)	29.09	29.09	
	Levitt (4 turns)	29.13	29.13	
	Levitt (2 turns)	29.21	29.21	
	MLE 80%		32.36	
	MLE 70%			37.10
P 0.4	Levitt (2 turns)	51.08		
	Levitt (4 turns)	51.11		
	Levitt (6 turns)	51.12		
	Levitt (8 turns)	51.13		
	MLE 80%	52 38		
	MLE 90%	52.72		
	MLE 70%		56.27	

*Measurement procedures within the same group do not produce significantly different thresholds from each other.

measurements, although there does appear to be an increase in variability as the number of turnarounds decreases.

Reducing the influence of outlying data by use of the interquartile range as the measure of variability reveals much the same picture as the standard deviations, that is, a significant effect of procedure on the variability of threshold measurement ($p<0.001$). The grouping of the seven procedures, obtained from pairwise comparisons for the adaptive procedures (Table 3), show that the MLE 70% correct produces significantly more variability than the MLE 80% and both these are significantly more variable than the others.

Speed of threshold estimation

As well as the masked threshold, the number of trials needed to obtain each threshold was also

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

Pairwise comparison of standard deviatiations (significance level 0.05)

Procedure	Average standard deviation		
	I	II	III
Levitt 8 turns	1.48		
Levitt 6 turns	1.67	1.67	
Levitt 4 turns	2.00	2.00	
Levitt 2 turns	2.46	2.46	
MLE 90%		3.72	
MLE 80%			7.28
MLE 70%			9.13

Table 3. Pairwise comparison of average interquartile ranges, each calculated from the 16 repeated measures within each cell and pooled across masker condition and subjects

Pairwise comparison of interquartile ranges (significance level 0.05)

Procedure	Average interquartile ranges		
	I	II	III
Levitt 8 turns	1.91		
Levitt 6 turns	2.07		
Levitt 4 turns	2.27		
MLE 90%	2.88		
Levitt 2 turns	2.94		
MLE 80%		6.52	
MLE 70%			11.38

recorded. By use of 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 little interaction between these two factors. Pairwise comparison of the means (Table 4, pooled across-notch conditions) showed that the Levitt procedure with eight turnarounds took approximately 2.5 times more trials to estimate thresholds than the three implementations of the MLE procedure used in the present study. However, restricting the Levitt procedure to two turnarounds rather than eight almost halved the number of trials required without, as shown above, resulting in a significantly different or more variable threshold estimate.

As there are many different possible rules for stopping the adaptive procedure, they are always somewhat arbitrary. The criterion of eight turnarounds for the Levitt procedure was the same as that 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

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

Pairwise comparison of number of trials (significance level 0.05)

Procedure	Average number of trials per threshold					
	I	II	III	IV	V	VI
MLE 70%	18.7					
MLE 80%	19.8	19.8				
MLE 90%		20.6				
Levitt 2 turns			28.2			
Levitt 4 turns				35.9		
Levitt 6 turns					43.4	
Levitt 8 turns						50.8

procedure requiring an increased number of trials before the criterion was met.

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 most widely used. As an alternative, the MLE procedure of Pentland (1980), and variations thereof, have recently been more widely utilized 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, placing the stimuli at the 80% correct point (near the 80.9% sweetpoint for the logistic function) should result in a smaller variability than at 70% or 90% correct. The present study shows that for a tone-in-noise task this was not the case. This suggests that either the chosen logistic model of the psychometric function is incorrect for this task or that other factors are 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 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 by use of the three MLE procedures. Specifically, tracking 70% correct results in thresholds that on average are 5.9 dB better than when tracking 90% correct. Given that the thresholds are adjusted to correspond to 75% correct for comparison across procedures, such a large difference is surprising. From Figure 1 it is clear that this discrepancy is related to the asymmetry in the distribution of threshold measurements obtained by use of the MLE procedure to track 70% correct.* Several consecutive correct guesses when the tone is inaudible result 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 90% correct allows the MLE procedure to get back to the 'true' threshold much more reliably than using the 70% correct point. Related behaviour was also noted in com-

*Calculating the median threshold for each listener and condition to reduce the effects of 'outliers' produces a mean difference of 3.6 dB.

puter simulations by Green (1990; his Figure 8) in which tracking 94% correct resulted in estimates converging to within 1 dB of threshold in 20 trials, whereas tracking 70.7% correct 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. First, the bias that results from estimating a low percentage correct tends to overestimate listeners' sensitivity in that the procedure seems to be abnormally influenced by correct guesses when the signal is inaudible. Related to this, a large number of consecutive presentations of inaudible stimuli may result in listeners 'forgetting' what they are listening to. This may result in an 'overshoot' when the stimulus becomes audible again, thus adding to the variability of the threshold estimate. Second, the stopping criterion used in the present study relies on the standard deviation of 10 successive trials becoming less than 1 dB (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 correct guessing.

It is also clear from Figure 1 that the spread of measurements for the 70% correct MLE procedure (and 80% in some conditions) is not only larger than for the 90% correct, but it is asymmetric in that the 70% correct procedure tends to overestimate listeners' ability to detect the tone in the masking noise. Analysis of the standard deviations confirms this increased variability when using the MLE procedure to track 70% or 80% correct when compared to tracking 90% correct or use of a Levitt procedure. Interestingly, reducing the number of turnarounds from which the threshold is estimated using the Levitt procedure only increases the variability of the estimated thresholds by about 1 dB.* It is also worth noting that even when tracking 90% correct with the MLE procedure the standard deviations of the threshold estimates are still greater than that obtained when using the Levitt procedure with two turnarounds.

A further point worth noting about the use of MLE to track high performance levels such as 90% correct is that it makes the procedure very unforgiving of lapses in attention, especially at the beginning of the run. Because the procedure only requires one wrong response out of every 10 trials, the estimated threshold is essentially limited to the easiest level at which a single error is made. Thus, a lapse of attention early in the run can lead to estimated thresholds which are much higher than they should be. Such runs are characterized by a single wrong response, as the MLE procedure is then never able to make the task difficult enough. Because the up–down procedures are only affected by a short stretch of trials, lapses of attention are of much less consequence.

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 variability, and a 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. However, both would be at the expense of the speed of threshold estimation. If the main requirement is to take several repeated measurements in as short a time as possible then use of either a MLE procedure to track a high performance level, or a Levitt procedure with few turnrounds, may be more useful than a Levitt procedure with a large number of turnarounds.

A further point worth considering is the use of a three-alternative forced-choice paradigm (Shelton and Scarrow, 1984). This would reduce the chance of correct guessing and thus perhaps the bias evident in tracking 70% or 80% correct with the MLE procedure. Finally, a hybrid of the two procedures may be worth investigating. The efficiency penalty of the up–down procedures appears attributable primarily to the initial run of trials from an easy level to near threshold, which MLE executes quickly. On the other hand, the MLE can be unusually sensitive to single lapses of attention, whereas the up–down procedures

*Excluding the MLE measurements from a pairwise comparison shows that the only significant difference (at $p = 0.05$) occurs between use of eight versus two turnarounds to estimate the thresholds. This is true both for the standard deviations and the interquartile ranges.

are relatively robust in this regard. It may therefore prove productive to use MLE initially to get near the threshold, followed by up–down tracking for 2–4 turnarounds to ensure that the true threshold has been reached. Even more simply, we have recently implemented a variation of the Levitt procedure in which a one-down/one-up rule is used up to the first turnaround, at which point the three-down/one-up rule is implemented.* On the basis of the data reported here, we estimate that nearly seven trials can be eliminated, on average, from every test session on the basis of this rule. It thus appears that a two-stage Levitt procedure, with two final turnarounds, would lead to thresholds at least as consistent as those from the best MLE, with equivalent numbers of trials, but which is more robust against lapses of attention.

Acknowledgements

Supported by the Wellcome Trust (grant reference 046823/Z/96). The authors wish to thank professors Mark Lutman and Brian Moore for their comments on an earlier draft.

References

Green DM. Stimulus selection in adaptive psychophysical procedures. J Acoust Soc Am 1990; 87: 2662–74.

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

Hall JL. Maximum-likelihood procedure for estimation of psychometric functions. J Acoust Soc Am 1968; 44: 370.

Hall JL. Hybrid adaptive procedure for estimation of psychometric functions. J Acoust Soc Am 1981; 69: 1763–9.

Levitt H. Transformed up–down methods in psychoacoustics. J Acoust Soc Am 1971; 49: 467–77.

Madigan R, Williams D. Maximum-likelihood procedures in two-alternative forced-choice: Evaluation and recomendations. Percept Psychophys 1987; 42: 240–9.

Patterson RD. Auditory filter shapes derived with noise stimuli. J Acoust Soc Am 1976; 59: 640–5.

Patterson RD, Moore BCJ. Auditory filters and excitation patterns as representations of frequency resolution. In: BCJ Moore, ed. Frequency Selectivity in Hearing. London: Academic Press, 1986; 123–77.

Pentland A. Maximum likelihood estimation: the best PEST. Percept Psychophys 1980; 28: 377–9.

Rosen S, Baker RJ. Characterising auditory filter nonlinearity. Hear Res 1994; 73: 231–43.

Rosen S, Baker RJ, Darling AM. Auditory filter nonlinearity at 2 kHz in normal listeners. J Acoust Soc Am 1998; 103: 2539–50.

Saberi K, Green DM. Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics. Percept Psychophys 1997; 59: 867–76.

Shelton BR, Picardi MC, Green DM. Comparison of three adaptive psychophysical procedures. J Acoust Soc Am 1982; 71: 1527–33.

Shelton BR, Scarrow I. Two-alternative versus three-alternative procedures for threshold estimation. Percept Psychophys 1984; 35: 385–92.

*The changes in step-size go on independently of this modification. Hence the changes in level which occur at the start of the session after a single correct response are the largest that are possible. Thus the step-size and modified rule both conspire to move the stimulus level quickly into the region of the threshold.