MLP: A MATLAB toolbox for rapid and reliable auditory threshold estimation

Massimo Grassi
Università di Padova, Padua, Italy

and

Alessandro Soranzo
University of Teesside, Middlesbrough, England

In this article, we present MLP, a MATLAB toolbox enabling auditory thresholds estimation via the adaptive maximum likelihood procedure proposed by David Green (1990, 1993). This adaptive procedure is particularly appealing for those psychologists who need to estimate thresholds with a good degree of accuracy and in a short time. Together with a description of the toolbox, the present text provides an introduction to the threshold estimation theory and a theoretical explanation of the maximum likelihood adaptive procedure. MLP comes with a graphical interface, and it is provided with several built-in, classic psychoacoustics experiments ready to use at a mouse click.

In this article, we present MLP, a MATLAB toolbox enabling auditory threshold estimation via the adaptive maximum likelihood (ML) procedure proposed by David Green (1990, 1993). This procedure is particularly suitable for estimating thresholds with an optimal compromise between accuracy and rapidity. For this reason, the ML procedure has been used successfully in clinical contexts (e.g., Florentine, Buus, & Geng, 2000), in studies with children (e.g., Wright et al., 1997), and in studies with a large number of subjects (e.g., Amitay, Irwin, & Moore, 2006). For the same reason, it is suitable for those studies in which subjects perform various tasks—that is, when each task has to consume only a portion of the subject's time. The ML procedure is largely known, used, and appreciated by the auditory community; it has collected more than 120 citations, and the majority of these citations have come from journals specializing in auditory research.[1] Thus, the user of this procedure can benefit from a large background literature to optimize his/her own threshold estimation. As far as we know, MLP is the first software implementing an adaptive psychophysical procedure with a graphical interface in a freely downloadable version, and it is provided with several built-in, classic psychoacoustics experiments ready to use at a mouse click.

In the next section, we will give a short introduction to threshold estimation theory. The reader familiar with these concepts may wish to skip this section. The ML procedure and the MLP toolbox will be illustrated after this section.

Sensory Threshold Estimation Theory

Sensation moves within and across two types of thresholds: *detection* and *discrimination*. The detection threshold is the minimum detectable stimulus *level*,[2] in the absence of any other stimuli of the same sort. In other words, the detection threshold marks the beginning of the sensation of a given stimulus. The discrimination threshold is the minimum detectable *difference* between two stimuli levels. Therefore, for a given sensory continuum, the discrimination threshold *cuts* the sensory continuum into the steps into which it is divided.

The *detection* threshold can be estimated either via yes/no tasks or via multiple-alternative forced choice tasks (in brief, *n*AFC, with *n* being the number of alternatives). The *discrimination* threshold, on the contrary, must be estimated exclusively via multiple *n*AFC tasks. In yes/no tasks, the subject is presented with a succession of different stimulus levels (spanning from below to above the subject's detection threshold) and is asked to report whether he or she has detected the stimulus (*yes*) or not (*no*). In an *n*AFC task, the subject is presented with a series of *n* stimuli differing in level. In audition, because the various stimuli have to be presented in temporal succession, the tasks are often multiple-interval tasks (i.e., mI0-*n*AFC). In an *n*AFC task, one stimulus (the *variable*) changes its level across the trials, whereas the level of the others (the *standards*) is fixed. The difference between standard and variable ranges from below to above the subject's detection (or discrimination) threshold. After each trial, the subject is asked to report which was the variable stimulus.

Figure 1 shows the hypothetical results of a yes/no task. The graph shows the relation between the stimulus level and the subject's performance, together with one function fitting the hypothetical data. This function is referred
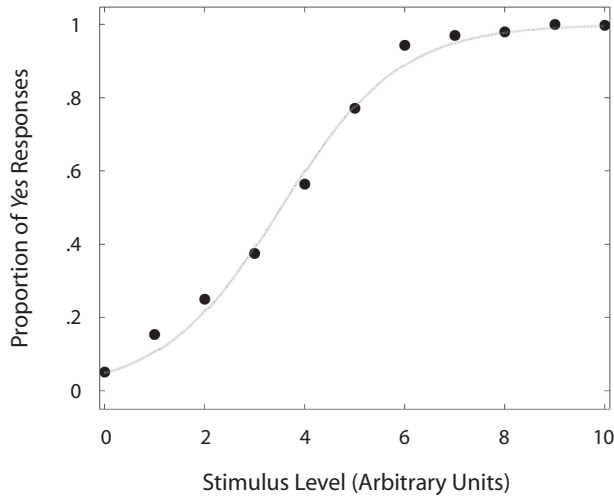
M. Grassi, massimo.grassi@unipd.it

Figure 1. Results of a hypothetical yes/no task. The subject's data are fitted with a logistic function (dashed curve). Note that this subject committed some false alarms (see later in the text) because, at the zero stimulus level (i.e., no stimulus actually presented), we can still observe a certain number of *yes* responses.

to as the *psychometric function.* Independently from the task type and from the type of threshold being measured, behavioral data are fitted with a sigmoid function such as that represented in Figure 1. Different types of psychometric functions can be adopted to fit experimental data—for example, the logistic, the Weibull, and the cumulative Gaussian.

The general equation of a psychometric function is the following (adapted from Wichmann & Hill, 2001), representing a subject's performance as a function of the stimulus level x:

$$\Psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)f(x; \alpha, \beta), \qquad (1)$$

where $f(x)$ is the sigmoid function chosen by the experimenter (i.e., logistic, Weibull, cumulative Gaussian); β determines the function's slope, whereas α determines the displacement of the function along the abscissa (see Figure 2); and γ and λ are "psychological" parameters that will be discussed shortly.

Among the sigmoid functions, the logistic is the most widely used, because of its computational simplicity. Its formula is the following:

$$f(x) = \frac{1}{1 + e^{\beta(\alpha-x)}}. \qquad (2)$$

Therefore, the corresponding psychometric function Ψ (in the toolbox, the logistic psychometric function is the Logistic.m function) is

$$\Psi = \gamma + (1 - \gamma - \lambda)\left[\frac{1}{1 + e^{\beta(\alpha-x)}}\right]. \qquad (3)$$

In the logistic psychometric function, α (often referred to as the *midpoint*) enables the displacement of the function along the stimulus-level axis. It corresponds to the average between γ and λ [*i.e.,* $\Psi(\alpha) = (\gamma + \lambda)/2$]. β is the function slope—that is, the rate of change in the subject's performance with changes in stimulus level. The greater the absolute value of β, the steeper the psychometric function will be. Moreover, for positive values of β the function increases, whereas for negative values it decreases (see Figure 2). When the function is adapted to psychological needs, γ and λ come into play. Depending on the task type (i.e., yes/no or nAFC), γ assumes a different meaning. There is, in fact, a major difference between these tasks: In yes/no tasks, the subject's response criterion is not under the control of the experimenter; on the contrary, it is in nAFC tasks (Green & Swets, 1966; Stanislaw & Todorov, 1999). The reason for this difference is that in a yes/no task, when a *yes* response is collected for a very low stimulus level, it is difficult at-
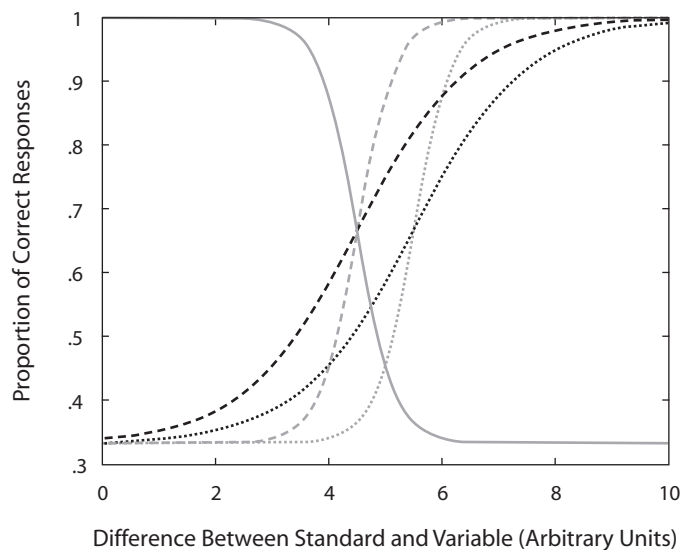

Figure 2. Five logistic psychometric functions. The gray (or black) functions have identical slopes. The dashed (or dotted) functions have identical midpoints. The gray solid function has a negative slope.

tributing this response to a high subject sensitivity or to a bias toward the *yes* response. Biased responses are called *false alarms*, and they affect the lower limit of the psychometric function, which can assume values greater than zero (see Figure 1). In other words, the probability of getting a *yes* response in the absence of the stimulus is greater than zero (Green, 1993). Hence, in yes/no tasks, γ corresponds to the subject's false alarm rate. False alarms are absent in *n*AFC tasks (Green & Swets, 1966; Stanislaw & Todorov, 1999). In *n*AFC tasks, the level of the standard(s) is always different from the level of the variable, and trials have, therefore, correct and incorrect answers. When the difference in level between standard and variable is below the subject's threshold (i.e., "when the stimulus level is low") the probability that the subject will return a correct answer is determined by chance, and chance will depend on the number of alternatives. For this reason, in an *n*AFC task, γ corresponds to chance level—that is, the reciprocal of the number of alternatives (e.g., 50% for 2AFC, 33% for 3AFC, 25% for 4AFC, and so on).

The meaning of λ, on the contrary, is independent of the task and refers to another error. In both yes/no and *n*AFC tasks, subjects can commit errors independently from the stimulus level; these are lapses of attention. Lapses of attention are estimated to be a small percentage of the subject's responses (i.e., 1%–5%; Saberi & Green, 1997; Wichmann & Hill, 2001), and they can affect the psychometric function fitting by decreasing the upper limit of the function (see Wichmann & Hill, 2001, for an extended discussion). Attentional lapses are particularly problematic at high stimulus levels (in yes/no tasks) or high differences in level between the standard and the variable(s) (in *n*AFC tasks; Saberi & Green, 1997; Wichmann & Hill, 2001).

Researchers are often interested in estimating a single point of the psychometric function, which is a subject's threshold. In probabilistic terms, the threshold corresponds to an arbitrary point of the psychometric function p (hereafter referred to as the *p*-target) included between the lower and the upper limits of the function (i.e., γ and λ). In other words, when we estimate a threshold, we search for the stimulus level eliciting the *p*-target proportion of *yes* (or correct) responses. Treutwein (1995) proposed that the *p*-target should be the middle of the psychometric function (e.g., 50% for yes/no tasks, 75% for 2AFC, 66% for 3AFC, etc.). However, other authors have suggested that higher values should be targeted (e.g., Amitay, Irwin, Hawkey, Cowan, & Moore, 2006; Baker & Rosen, 1998, 2001; Green, 1990).

Thresholds can be estimated by means of two classes of procedures: adaptive and nonadaptive. In nonadaptive procedures—for example, the constant stimuli method—the stimulus levels (or differences between standard and variable levels) are preset before the beginning of the experiment. The stimuli should span from below to above a subject's threshold. During the experiment, the stimuli are presented to the subject in random order, and the proportion of *yes* (or correct) responses is calculated for each stimulus. In other words, the subject's threshold will be interpolated from a fully sampled psychometric function, making

the measurement of the threshold expensive in terms of an experiment's time. This represents the major drawback of this class of procedures when the experimenter needs to estimate the subject's threshold only. For the reason above, when they need to estimate a threshold, psychophysicists prefer adaptive over nonadaptive procedures. In adaptive procedures, the stimulus levels are selected at the same time as the experiment is running, depending on the subject's answers. Adaptive procedures maximize the ratio between the number of stimuli presented at/near threshold and the number of stimuli presented far from threshold. Adaptive procedures can be grossly divided into two types: nonparametric (also known as staircases) and parametric. The only assumption made by nonparametric procedures is that the psychometric function is monotonic. Parametric procedures, on the contrary, make more assumptions. For example, they assume the shape of the psychometric function. Examples of nonparametric procedures are the method of limits by Fechner (1889), the simple up–down by von Békésy (1947), and the transformed up–down by Levitt (1971). Examples of parametric procedures are the PEST by Taylor and Creelman, (1967), the "best" PEST by Pentland (1980), and the QUEST by Watson and Pelli, (1983). The ML procedure studied by Green (1990, 1993) is a parametric procedure. Although the paternity of ML threshold estimation cannot be attributed directly to Green (Hall, 1968, and Pentland, 1980, for example, suggested this same approach), there is no doubt that he is the author who studied this approach in the most detail (Green, 1990, 1993, 1995; Gu & Green, 1994; Saberi & Green, 1997).

Nonparametric procedures are generally used more than parametric ones, even if they involve some disadvantage. The major one is that they tend to be more time consuming (e.g., Amitay, Irwin, Hawkey, et al., 2006; Leek, 2001).[3] Nonetheless, nonparametric procedures are used more than parametric ones, because they are theoretically simpler and can be easily implemented via conventional software (e.g., MEL, E-Prime), whereas parametric procedures are theoretically more complex and require more advanced programming skills. Given the current state of the art, we are aware of just one parametric procedure implemented in a freely downloadable version—namely, the QUEST procedure (Brainard, 1997; Pelli, 1997; Watson & Pelli, 1983).

The Maximum Likelihood Procedure

The ML procedure is composed of two independent processes: the *maximum likelihood estimation* and the *stimulus selection policy* (in the toolbox, these two processes are synthesized in the FindThreshold.m function). The algorithm used by the ML procedure differs slightly for yes/no and *n*AFC tasks. Because the *n*AFC task can be used to estimate all types of thresholds, in the following paragraphs we will explain the ML procedure as applied to this task only. Readers interested in how the procedure works in the yes/no task can refer to the original work by Green (1993).

Maximum likelihood estimation. Before the beginning of the experiment, the experimenter hypothesizes several psychometric functions, called *hypotheses*. The

hypotheses have the same slope β, attentional lapse rate λ, and chance level γ but differ in midpoint α so as to cover the range of stimulus levels at which a subject's threshold is supposed to be (see Figure 3).

An experiment might begin by providing the subject with a stimulus level that is above threshold. The subject's response is then collected and utilized to calculate the likelihood of each hypothesis. Likelihood is calculated by means of the following function:

$$L\left(H_j\right) = \prod_{i=1}^{n} H\left(x_i\right)^C \left[1 - H\left(x_i\right)\right]^W, \qquad (4)$$

where $L(H_j)$ is the likelihood of the jth hypothesized function and i is the trial number. The exponents C and W are equal to 1 and 0, respectively, when the response is correct, and to 0 and 1 otherwise. The product above can be simplified into a sum by means of a logarithmic transformation as follows (CalculateLikelihood.m in the toolbox):

$$L\left(H_j\right) = \sum_{i=1}^{n} C \log H\left(x_i\right) + W \log\left[1 - H\left(x_i\right)\right]. \qquad (5)$$

Once the likelihood of each hypothesis has been calculated, the ML procedure selects the highest likelihood hypothesis. This hypothesis is the one having the highest likelihood of resembling the actual subject's psychometric function. The highest likelihood hypothesis will be identified by its midpoint α.

The likelihood of the hypotheses is calculated after each trial. Hence, even after the very first trial, an ML estimate is returned by the procedure, although it may be highly inaccurate. However, the greater the number of trials, the more the estimate becomes accurate. Therefore, the best estimate is that returned by the last trial.

Stimulus selection policy. Once the most likely hypothesis has been found, at which level does the next stimulus have to be presented? The common response is to assert that we should set the stimulus level at threshold (Simpson, 1989)—that is, at the p-target. Even after the very fist trial, the ML procedure has enough information to select the threshold level that is used as the stimulus level for the successive trial. The most likely hypothesis contains also the most likely subject's threshold. The threshold will be the inverse function of the most likely hypothesis at the p-target (InvLogistic.m in the toolbox); therefore,

$$\Psi^{-1}\left(p_{\mathrm{t}}\right) = \alpha_j - \frac{1}{\beta} \ln\left(\frac{1 - \gamma - \lambda}{p_{\mathrm{t}} - \gamma} - 1\right), \qquad (6)$$

where p_{t} is the p-target.

Green (1990, 1993) showed analytically that there is an optimal p-target that researchers should track (as well as many p-targets that experimenters should avoid; see the ML Procedure Guidelines section). This particular p-target (often referred to as the *sweet point*) optimizes the estimate of the subject's threshold, because the variance associated with the estimate of this particular p-target is smaller than the variance associated with any other possible p-target. The variance of the threshold estimate associated with any p-target is equal to the binomial variance, $\Psi(1 - \Psi)$, divided by the slope of the psychometric function squared; therefore,

$$\sigma^2 = \frac{\Psi(1 - \Psi)}{\Psi'^2}, \qquad (7)$$
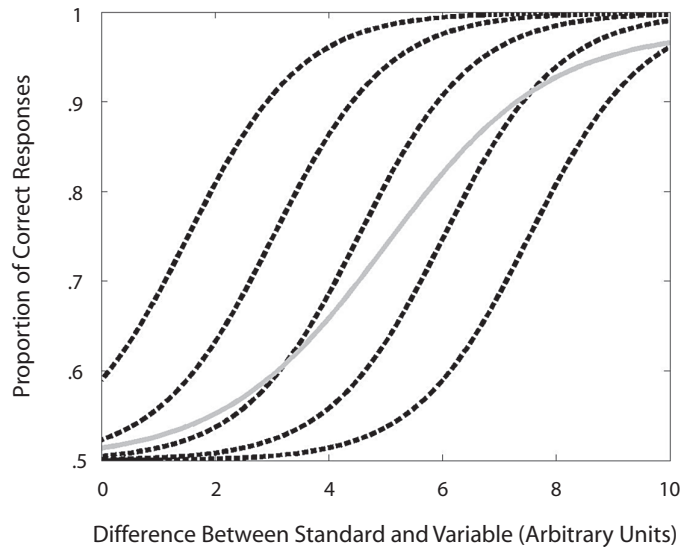

Figure 3. The solid gray function is the actual subject's psychometric function. The black dotted functions are the hypotheses we set for running the maximum likelihood procedure and estimating the subject's threshold. Note that the actual subject's function is different from all of the hypothesized functions. However, within the hypothesized functions, the middle one is most similar to the subject's psychometric function.

Table 1
Comparison Between Sweet Point and the Arithmetic Mean
Between γ and λ for the Logistic Function

γ (%)	Average(γ, λ) (%)	Sweet Point (%)
0	50.0	50.0
25 (i.e., 4AFC)	62.5	68.3
33 (i.e., 3AFC)	66.6	72.9
50 (i.e., 2AFC)	75.0	80.9

where Ψ'^2 is the derivative of the psychometric function slope squared. The best possible *p*-target is that minimizing the above ratio. In the case of the logistic function, the sweet point (p_{sw}) can be calculated analytically as follows (Green, 1993):

$$p_{sw} = \frac{2\gamma + 1 + \sqrt{1 + 8\gamma}}{3 + \sqrt{1 + 8\gamma}}. \qquad (8)$$

If we assume that the subject does not produce any attentional lapse (and this assumption is convenient; see the ML Procedure Guidelines section), the sweet point depends exclusively on γ. Table 1 compares sweet points and *p*-targets calculated as the arithmetic means between γ and λ, for some γ values. When γ increases, the sweet point becomes correspondingly greater than the average between γ and λ.

Altogether, the ML procedure can be understood better with an example (see Figure 3 and Table 2). Let us suppose that we want to estimate a detection threshold by means of a 2AFC task. We select five hypotheses whose midpoints range from ~1.5 level units to ~7.5 level units, so that the step between each midpoint is equal to 1.5 level units (see Figure 3). We choose to track the sweet point that, in 2AFC tasks, is the stimulus eliciting 80.9% correct responses (as reported in Table 1). We arbitrarily set the first variable stimulus level to 11 and present the stimuli to the subject. Let us suppose that the subject's answer is *correct* (Table 2, row 1, column 3). Each hypothesis expects, for that level, a certain proportion *p* of correct responses (Table 2, row 1,

columns 4–8). The likelihood of each hypothesis is calculated by means of Equation 5 and is reported in Table 2, row 1, columns 9–13. As is shown in Table 2, H1 results in the highest likelihood hypothesis. By means of Equation 6 we calculate the subject's threshold: We take H1 and calculate the stimulus level that corresponds to 80.9% correct responses. The subject's threshold result is equal to 2. Therefore, the stimulus level for the second trial will be set at 2. On the second trial, the subject gives a wrong response. Each hypothesis expects, for this stimulus level, a certain proportion $1 - p$ of wrong responses. This proportion, together with that for the first trial, can be passed to Equation 5. After the likelihood calculation, the hypothesis that most likely resembles the subjects' psychometric function is now H5. The stimulus level corresponding to the sweet point in H5 equals 8. The stimulus level for the third trial will be set at 8. The process just described is iterated until the subject has run the number of trials that we set at the beginning of the experiment.

The major benefit of the ML procedure is that it makes maximal use of the available data: The data from all the trials are used to estimate the subject's threshold. If the experimenter has set the procedure appropriately, the ML procedure will arrive at the threshold more accurately than will a staircase procedure. A further advantage of the ML procedure is its rapidity. Green (1993) claimed that 12 trials of ML are sufficient for a reliable threshold estimate. Although recent evidence suggests that this initial claim was too optimistic (e.g., Amitay, Irwin, Hawkey, et al., 2006; Leek, Dubno, He, & Ahlstrom, 2000), the procedure still remains a rapid one. An additional advantage of the procedure is the possibility of tracking any point of the psychometric function—in other words, the possibility of tracking any *p*-target. The reader, however, should know that this characteristic is shared also by some nonparametric procedures (e.g., Kaernbach, 1991). This characteristic is particularly appealing, for example, when the experimenter needs to replicate the results of a study targeting a specific point of the psychometric function.

Table 2
The Possible Development of an ML Procedure

Trial	Level	Answer	H1	H2	H3	H4	H5	L(H1)	L(H2)	L(H3)	L(H4)	L(H5)	Th
1	11	c	1.00	1.00	1.00	1.00	.99	*0.00*	0.00	0.00	0.00	−0.02	2
2	2	w	.19	.37	.46	.49	.50	−1.66	−1.00	−0.77	−0.71	*−0.71*	8
3	8	c	1.00	1.00	.99	.94	.81	−1.66	−1.00	−0.79	*−0.78*	−0.92	6.5
4	6.5	c	1.00	.99	.94	.81	.63	−1.66	−1.02	*−0.85*	−0.99	−1.38	5
5	5	c	.99	.94	.81	.63	.54	−1.67	−1.08	*−1.06*	−1.45	−2.00	5
6	5	c	.01	.06	.19	.37	.46	−1.69	*−1.15*	−1.27	−1.90	−2.62	3.5
7	3.5	c	.94	.81	.63	.54	.51	−1.75	*−1.36*	−1.73	−2.53	−3.30	3.5
8	3.5	w	.06	.19	.37	.46	.49	−4.56	−3.01	*−2.73*	−3.30	−4.01	5
9	5	c	.99	.94	.81	.63	.54	−4.57	−3.08	*−2.94*	−3.75	−4.63	5
10	5	w	.01	.06	.19	.37	.46	−8.78	−5.88	*−4.60*	−4.76	−5.40	5
11	5	c	.99	.94	.81	.63	.54	−8.79	−5.94	*−4.81*	−5.21	−6.02	5
12	5	c	.99	.94	.81	.63	.54	−8.81	−6.00	*−5.02*	−5.67	−6.65	5

Note—The table represents numerically the example reported in Figure 3. In the table, "trial" is the trial number, "level" is the difference in level between standard and variable in a detection task, and "answer" is the subject's answer ("c," correct response; "w," wrong response). Columns H1–H5 show the probability of correct/wrong response predicted by each of the five hypotheses set for the maximum likelihood procedure. The successive columns [i.e., from L(H1) to L(H5)] are the likelihood of each hypothesis given the subject's answers to the stimulus levels presented thus far. The highest likelihood hypothesis is given in italics. The last column (i.e., "Th") is the estimate of the subject's threshold calculated at the end of each trial. In the present example, we track a *p*-target that is the sweet point of a logistic function for a 2AFC task (i.e., 80.9%).

Besides its advantages, the ML procedure also involves some disadvantages. The major is that both the shape and the slope of the hypothesized psychometric functions are set by the experimenter before running the experiment. These parameters could be unknown, and the selected ones might not coincide with the "actual" ones (see Figure 3). This disadvantage, however, does not seem to affect experiment reliability significantly. By means of several simulations, Green evaluated whether the mismatch between the "actual" subject's psychometric function and the psychometric function hypothesized by the procedure affects the threshold estimate. From these simulations, it emerged that (1) a shape mismatch (Green [1990] compared logistic and Gaussian) does not have an evident effect on the threshold estimate and (2) a slope mismatch can affect the threshold estimate by increasing estimate variance (Green, 1990, 1993). By the same token, the experimenter may not know in advance the rate of attentional lapses (λ) or the false alarm tendency (γ) of the subject. In three articles, Green and colleagues (Green, 1993, 1995; Gu & Green, 1994) investigated whether the lack of knowledge of these two parameters affects the threshold estimate and whether these two parameters can be estimated by the ML procedure together with the subject's threshold.[4] Lapses of attention affect the threshold estimation when they are numerous and occur in the very first trials (Green, 1995). In particular, the bias produced by lapses of attention is large if they occur within the first five trials but becomes almost negligible in the later ones (Gu & Green, 1994). Also, false alarms can bias the threshold estimate (Green, 1993; Gu & Green, 1994). However, the ML procedure can be used to make a rough estimate of the subject's false alarm rate. Solutions to the problems just presented will be given in the ML Procedure Guidelines section.

The ML Toolbox

MLP has been developed to work with MATLAB 7.0 or higher[5] and can be downloaded from the following Web page: www.psy.unipd.it/~grassi/mlp.html. In the Web page, the user will find the complete list of the toolbox's functions and the most updated list of the available experiments, together with a brief description of them. At the moment we are writing, MLP is provided with 24 built-in experiments. MLP works with any operative system and does not require additional MATLAB toolboxes. All MLP functions are compressed in a zip archive that the user needs to expand and copy into the MATLAB "toolbox" folder. The user needs also to "add with subfolder" all the toolbox contents in the MATLAB path. All functions have a command line help. The help can be seen by typing "help" followed by the function name at the MATLAB prompt.

To start an experiment, type "mlp" (i.e., the main function) at the MATLAB prompt. This calls up a graphical interface (see Figure 4). Now the user has to select the experiment he or she wants to run. The majority of the built-in experiments are classic psychoacoustics experiments. Some are "translations" for the ML procedure of a subset of experiments performed by Kidd, Watson, and Gygi (2007). These authors ran 19 classic psychoacous-
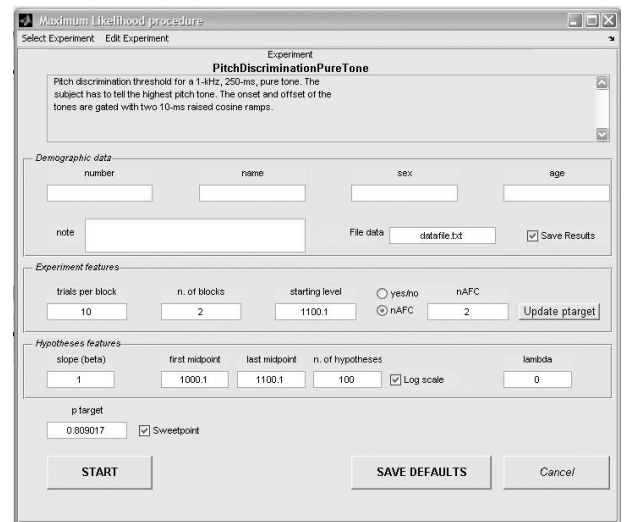

Figure 4. The MLP graphical interface.

tics experiments on a large ($N = 340$) number of adult subjects. The MLP user running these experiments can thus compare his/her own results with those reported in that study.[6] Built-in experiments come with default parameters. The user can, however, change parameters at will (see the ML Procedure Guidelines section). If changes are made and saved, they will be kept until the next "SAVE DEFAULTS" command is called. When the user presses "START" the experiment begins, and the stimuli are presented to the subject. In all the built-in experiments, the subject responds by pressing the key numbers of the computer keyboard. In nI-nAFC experiments, the subject reports the temporal position of the variable stimulus. For example, in a 4AFC task, if the subject thinks that the variable stimulus was the third stimulus presented, he or she has to press "3." In yes/no tasks, the "1" number corresponds to the answer "yes, I perceive/detect," and any other number (e.g., "0") corresponds to the "no, I did not perceive/detect" answer. Keypresses must be followed by the "Return" key. After each block of trials, the subject's threshold is echoed on the computer screen. MLP saves two data files (tab-delimited, flat format, text files) in the MATLAB current directory. The first is an extended data file that contains all the experiments' events—that is, subject number, name, sex, age and note, block number, trial number, level of the stimulus presented, subject's response, threshold estimated after each trial, and estimated false alarm rate γ (please note that this last estimate is for yes/no tasks only). In this file, the subject's responses are coded as "1" ("yes," or correct) and "0" ("no," or wrong). The name of this file can be set by the user through the graphical interface. The second data file contains only the subject's threshold, and it is saved after each subject. By default, the name of this file is the subject's name. If the user does not input the subject's name, the file is called "untitled.txt."

In the case in which the specifics of the built-in experiments do not match the experimenter's needs, he or she can edit them and adapt them to his/her own needs. The

characteristics of the sounds of each experiment are written at the beginning of the experiment.m files and can be easily changed. More advanced MATLAB users can write their own experiments by taking as an example any of the built-in experiments. The MLP Web page provides detailed instructions on how to write custom experiments.

ML Procedure Guidelines

In this section, we will provide the reader with a set of guidelines for a fruitful use of the ML procedure by means of MLP. Users who want to edit existing experiments (or create their own) can use these guidelines for a fruitful optimization of the threshold search.

The first decision the experimenter has to make before starting a detection threshold estimation is about the kind of task the subjects will perform (we remind the reader that discrimination thresholds must be estimated via *n*AFC tasks only). The choice of the specific task depends on two factors: the desired experiment duration and the desired robustness of the threshold estimation (yes/no experiments are usually shorter). However, as was previously mentioned, in yes/no tasks, the subject's bias is not under the control of the experimenter. For this reason, thresholds gathered with yes/no tasks can be less robust than those gathered with *n*AFC tasks.[7] In brief, if duration is essential, we suggest using the yes/no task, whereas if robustness is essential, the *n*AFC task should be preferred. A corollary question is about the number of alternatives in an *n*AFC task. Once again, there is a trade-off between duration and robustness. An increase in the number of alternatives leads to an augmented robustness of the threshold estimation (Schlauch & Rose, 1990), but it augments the experiment duration also. For the ML procedure, Amitay, Irwin, Hawkey, et al. (2006) suggest using three alternatives.

Once the task has been chosen, the experimenter has to set the parameters of the ML procedure, such as slope, range, and number of hypothesis. As we wrote previously, the ML procedure uses, for the threshold estimation, a set of hypotheses (i.e., a set of psychometric functions) with identical slope, but differing in the position along the abscissa over a certain range. Ideally, the slope of the hypotheses should be identical to the subject's psychometric function slope, but occasionally this parameter could be unknown. The choice of the wrong slope can affect the threshold estimate—in particular, when the chosen slope is (much) less steep than the actual one (Green, 1990, 1993). When the actual slope is unknown, the experimenter might need to estimate it (at least roughly) before starting the experiment. The classic method for performing this estimation is to run a constant stimuli experiment (e.g., Saberi & Green, 1997) and successively interpolate the subject's performance with the logistic psychometric function.

The selection of range and number of hypotheses is less problematic than the selection of the hypotheses' slope. The range of the hypotheses must cover abundantly the range of stimuli levels where we expect the subject's threshold is. In practice, the midpoint of the first hypothesis should be well below the subject's threshold, whereas that of the last hypothesis should be well above the subject's thresh-

old. Green (1993) showed that the number of hypotheses set to fill this range does not affect the standard deviation of the threshold estimate. The number of hypotheses affects, however, how close the final threshold estimate will be to the actual threshold. For this reason, our suggestion is to use the highest possible number of hypotheses compatible with computer potentiality. The computational load required by the procedure is, in fact, proportional to the number of hypotheses. Finally, the range of the hypotheses can be spaced either linearly or logarithmically. We recommend the use of logarithmic spacing for physical quantities expressed on a linear scale (e.g., frequency, duration, etc.) and linear spacing for units that are already expressed logarithmically (e.g., sound pressure level when expressed in decibels).

The next hypothesis parameter one should set pertains to the yes/no task only. It is the false alarm rate γ. In the current version of the toolbox, this parameter ranges between five possible fixed values: 0%, 10%, 20%, 30%, and 40%. The most likely of these γ values is returned after each trial in the MLP data file. If the subject has a high false alarm rate, the threshold could be underestimated. In practice, the subject reports being able to detect very low-level stimuli. Consequently, the ML procedure will look for the subject's threshold also in the very low-level range. However, the ML procedure returns only a rough estimate of the subject's false alarm rate. In particular, Green (1993) showed that the ML procedure underestimates the subject's false alarm rate. The underestimation of the false alarm rate can be reduced by introducing a number of catch trials during threshold estimation. In catch trials, the stimulus level is set either to zero or to the minimum level of the range in which one is looking for the subject's threshold (Gu & Green, 1994; Leek et al., 2000). The answer to catch trials (which is expected to be *no* unless the subject is producing a biased response) is included in the calculation of the hypotheses' likelihood. With MLP, the user has the option to include catch trials in the ML procedure to reduce the underestimation of the subject's false alarm rate. Catch trials will be presented at any moment of the threshold estimation, excluding the very first trial. The occurrence of catch trials is determined probabilistically by a proportion that the user can set in the graphical interface. During the ML procedure, when a catch trial will occur, the stimulus level will be set to the minimum level of the range in which one is looking for the subject's threshold (i.e., the first midpoint). We recommend keeping the catch trial rate to about 20% of the total number of trials (Leek et al., 2000).

After the selection of the parameters for the ML procedure, the parameter for the stimulus selection policy (i.e., the *p*-target) has to be chosen. The criterion for selecting a *p*-target differs slightly between yes/no and *n*AFC tasks. In *n*AFC tasks, Treutwein (1995) suggested tracking the middle of the psychometric function (i.e., 75% for 2AFC, 66% for 3AFC, etc.). As we wrote previously, Green (1990, 1993) suggested tracking the sweet point, which is generally higher than the middle of the psychometric function (see Table 1). Baker and Rosen (1998, 2001) and Amitay, Irwin, Hawkey, et al. (2006) tracked *p*-targets even higher
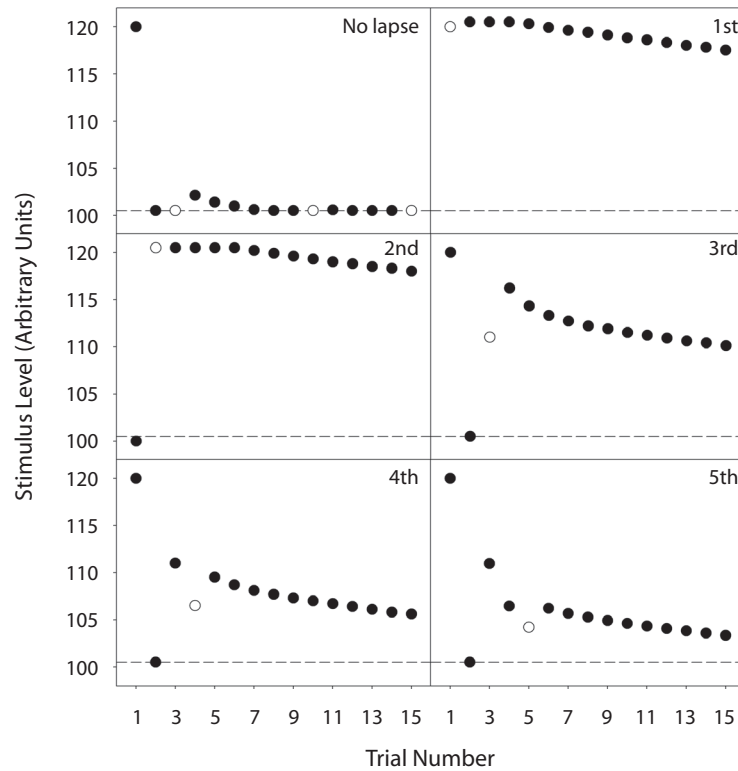
Figure 5. Stimulus level presented by the maximum likelihood procedure as a function of trial number, showing the effect of attentional lapses on Trials 1–5. The top left graph shows threshold estimation with no lapse. In the graphs, black symbols show *yes* (or correct) responses and white symbols show *no* (or wrong) responses. The dashed line shows the subject threshold.

than the sweet point. As a rule of thumb, we suggest tracking a *p*-target not lower than the sweet point.

The selection of the *p*-target for the yes/no task is slightly more complex. The general tendency is to track the middle of the psychometric function (i.e., 50%). However, this *p*-target could be too low, especially when working with subjects that are known to have a high false alarm rate (e.g., children). Unlike with *n*AFC tasks, with yes/no tasks the calculation of the sweet point is not straightforward, because the sweet point depends on the subject's false alarm rate (see Equation 8), which is unknown. In summation, if we presume that subjects are reliable, we can track the middle of the psychometric function (i.e., 50%). If, on the contrary, we presume that subjects are unreliable, we have to track a higher *p*-target. Green (1993) tracked the average between the minimum and the maximum sweet points of the false alarm range expected to be observed. In other words, if one expects a false alarm rate ranging from 0% to 40%, one will track 63.1%—that is, the average between the sweet point for a 0% false alarm rate (i.e., 50%) and the sweet point for a 40% false alarm rate (i.e., 76.2%). This is the option we implemented in the toolbox.

Now that all the parameters of the ML procedure have been set, one has to decide the length of the experiment— that is, the number of trials. As we wrote previously, one of the major advantages of ML method is its rapidity. Re-

liable threshold estimates can be obtained with very few trials. Green (1993) showed that 12 trials may be enough. That initial claim was perhaps too optimistic. Recent studies have suggested that an optimal threshold estimate should require about 24 (Leek et al., 2000) or about 30 (Amitay, Irwin, Hawkey, et al., 2006) trials; that still remains a small figure.[8]

The very last thing the experimenter has to decide before starting the experiment is the level of the first stimulus that the procedure will present to the subject. This level is the only one set by the experimenter. Green (1993) demonstrated that the starting level value has no effect on the threshold estimate. We suggest setting this level relatively high, in order to offer the subject an easy first trial.

After the ML procedure has terminated, there is one last thing the experimenter can do to further control the goodness of the threshold estimates: controlling for attentional lapses. If the subject produces attentional lapses, the threshold is likely to be overestimated. The reason is simple. The subject states that he or she is not able to detect (or to discriminate) the very high-level stimulus. Therefore, the ML procedure will look for the subject's threshold also in a very high-level range. However, attentional lapses affect the threshold estimate only if they occur within the first five trials (Gu & Green, 1994). Blocks of trials characterized by attentional lapses are easy to spot (and remove) in the data analysis. Figure 5 compares six

blocks of trials. In five of these blocks, the ML procedure is "affected" by an attentional lapse that occurred in the first, second, third, fourth, or fifth trial of the block.

REFERENCES

AMITAY, S., IRWIN, A., HAWKEY, D. J. C., COWAN, J. A., & MOORE, D. R. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *Journal of the Acoustical Society of America*, **119**, 1616-1625.

AMITAY, S., IRWIN, A., & MOORE D. R. (2006). Discrimination learning induced by training with identical stimuli. *Nature Neuroscience*, **9**, 1446-1448. doi:10.1038/nn1787

BAKER, R. J., & ROSEN, S. (1998). Minimizing the boredom by maximising likelihood: Efficient estimation of masked threshold. *British Journal of Audiology*, **32**, 104-105.

BAKER, R. J., & ROSEN, S. (2001). Evaluation of maximum likelihood threshold estimation with tone in noise masking. *British Journal of Audiology*, **35**, 43-52.

BRAINARD, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, **10**, 433-436.

FECHNER, G. T. (1889). *Elemente der Psychophysik* (2nd ed.). Leipzig: Breitkopf & Härtel.

FLORENTINE, M., BUUS, S., & GENG, W. (2000). Toward a clinical procedure for narrowband gap detection: I. A psychophysical procedure. *Audiology*, **39**, 161-167.

GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2005). Sampling plans for fitting psychometric function. *Spanish Journal of Psychology*, **8**, 256-289.

GREEN, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **87**, 2662-2674. doi:10.1121/1.399058

GREEN, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. *Journal of the Acoustical Society of America*, **93**, 2096-2105. doi:10.1121/1.406696

GREEN, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America*, **97**, 3749-3760. doi:10.1121/1.412390

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

GU, X., & GREEN, D. M. (1994). Further studies of a maximum likelihood yes–no procedure. *Journal of the Acoustical Society of America*, **96**, 93-101. doi:10.1121/1.410378

HALL, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **44**, 370.

KAERNBACH, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics*, **49**, 227-229.

KIDD, G. R., WATSON, C. S., & GYGI, B. (2007). Individual differences in auditory abilities. *Journal of the Acoustical Society of America*, **122**, 418-435.

LEEK, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, **63**, 1279-1292.

LEEK, M. R., DUBNO, J. R., HE, N.-J., & AHLSTROM, J. B. (2000). Experience with a yes–no single-interval maximum-likelihood procedure. *Journal of the Acoustical Society of America*, **107**, 2674-2684.

LEVITT, H. (1971). Transformed up–down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467-477.

PELLI, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437-442.

PENTLAND, A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics*, **28**, 377-379.

SABERI, K., & GREEN, D. M. (1997). Evaluation of maximum-likelihood estimators in nonintensive auditory psychophysics. *Perception & Psychophysics*, **59**, 867-876.

SCHLAUCH, R. S., & ROSE, R. M. (1990). Two-, three-, and four-interval forced-choice staircase procedures: Estimator bias and efficiency. *Journal of the Acoustical Society of America*, **88**, 732-740. doi:10.1121/1.399776

SIMPSON, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics*, **45**, 572-576.

STANISLAW, H., & TODOROV, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, **31**, 137-149.

STEVENS, S. S. (1957). On the psychophysical law. *Psychological Review*, **64**, 153-181. doi:10.1037/h0046162

STEVENS, S. S., & GALANTER, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, **54**, 377-411. doi:10.1037/h0043680

TAYLOR, M. M., & CREELMAN, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**, 782-787.

TREUTWEIN, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**, 2503-2522. doi:10.1016/0042-6989%2895%2900016-S

VON BEKESY, G. (1947). A new audiometer. *Acta Otolaryngology*, **35**, 411-422.

WATSON, A. B., & PELLI, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.

WICHMANN, F. A., & HILL, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, **63**, 1293-1313.

WRIGHT, B. A., LOMBARDINO, L. J., KING, W. M., PURANIK, C. S., LEONARD, L. M., & MERZENICH, M. M. (1997). Deficits in auditory temporal and spectral resolution in language-impaired children. *Nature*, **387**, 176-178. doi:10.1038/387176a0

NOTES

1. Source: ISI Thompson Web of Science.

2. In this article, the term *level* refers to both intensive (e.g., luminance, sound pressure) and nonintensive (e.g., light wavelengths, acoustic frequency) physical quantities. These two classes of physical quantities are perceptually mapped into the so-called prothetic sensations (e.g., brightness, loudness), as opposed to metathetic sensations (e.g., perceived color and pitch) (Stevens, 1957; Stevens & Galanter, 1957).

3. This is the case for the most widely used nonparametric procedure (i.e., transformed up–down; Levitt, 1971), which requires about twice as many trials as the ML procedure (Leek et al., 2000).

4. The ML algorithm can be used to estimate all psychometric function parameters (i.e., β, γ, and λ)—thus, the subject's whole psychometric function. However, the estimation of the whole psychometric function requires thousands of trials (García-Pérez & Alcalá-Quintana, 2005), and it is therefore performed exclusively in Monte Carlo simulations.

5. Unfortunately, some of the graphical characteristics of the toolbox are incompatible with older versions of MATLAB.

6. Readers interested in an identical replicate of the experiments run by Kidd, Watson, and Gygi (2007)—and thus in a direct comparison with the published results—should refer to the *Test of Basic Auditory Capabilities* by the same authors (Communication Disorders Technologies Inc.).

7. There are some exceptions to this. Green (1995) found that, in some conditions, the threshold estimates gathered with yes/no tasks were more robust than those gathered with nAFC tasks.

8. The literature reports a way to further shorten the length of experiments, which is to halt the ML procedure when the variance of the threshold estimation drops below a certain preset value (Leek et al., 2000). As we wrote previously, the ML procedure returns a threshold estimate after each trial. For Leek et al., the ML procedure was halted when the variance of the last 10 threshold estimates was smaller than a certain preset value. This option is not implemented in the toolbox.