Mechanisms of auditory perceptual learning

PETER RICHARD JONES, BA, MSc.

Thesis submitted to the University of Nottingham
for the degree of *Doctor of Philosophy*

JULY 2013

Thesis Abstract

Practice improves performance on many basic auditory tasks. However, while the phenomenon of auditory perceptual learning is well established, little is known about the mechanisms underlying such improvements. What is learned during auditory perceptual learning? Here I attempt to address this question by applying models of performance to behavioural response data, and examining which parameters change with practice.

In Chapter 1, the phenomenon of auditory perceptual learning is described and a theory of decision-making introduced. Within this framework, a number of potential learning mechanisms are operationalised: encoding efficiency, internal noise, bias and inattention. Previous research concerning these constructs is presented.

Chapters 2 and 3 examine the possible sensory mechanisms of learning. In Chapter 2, four models are developed, each of which is applied to the responses of listeners given a pure tone frequency discrimination task. Learning is shown to primarily represent a reduction in internal noise, with encoding efficiency, attentiveness and bias appearing invariant. That auditory learning primarily represents a reduction in internal noise conflicts with a prominent claim in the visual literature that signal not noise changes with learning, and possible reasons for this conflict are discussed. In Chapter 3 training data are reported for tone detection in unpredictable noise. This is a more complex auditory task, which requires information to be appropriately integrated across channels. To examine whether in such conditions learning is subserved by factors other than internal noise, reverse correlation was used to calculate the relative weight that listeners attributed to each spectral region. The optimisation of these weights is shown to be the primary mechanism of learning on this task, and their inclusion obviates the need to invoke internal noise as an explanation of improved performance.

In Chapter 4, a series of experiments are reported that investigated the extent to which learning and development share common mechanisms. Developmental differences in masked tone-detection are evidenced. Using methods extended from Chapter 3, these are shown to be partly explained by differences in encoding efficiency, with younger listeners inappropriately integrating information over a wider spectral range. This result recapitulates the practice-induced differences in encoding efficiency observed in adult listeners (Chapter 3).

Task performance is potentially constrained not just by the strength of the sensory evidence, but also by the efficiency of the wider decision process that the sensory evidence informs. Accordingly, Chapters 5 and 6 examine the role of non-sensory factors in learning. In Chapter 5, the role of interval bias in yes/no task learning is evaluated. Naïve listeners are shown to be overly liberal in their responses. This tendency is shown to be eradicated through practice, leading to improved detection limens on a modulation detection task. In Chapter 6, the role of sequential response dependencies in forced-choice task learning is evaluated. Naïve listeners are shown to respond in a manner conditional on their previous responses. This nonstationary bias is shown to be militated through practice. Simulations are used to relate the observed changes in performance and bias.

In Chapter 7 the experimental results are reviewed and implications discussed. I conclude that auditory perceptual learning is subserved by multiple mechanisms that: operate in parallel, vary in importance depending on the task demands, and incorporate both sensory and non-sensory processes. The methods of analysis described are shown to effectively partition components of perception in normal hearing children and adults, and may help to understand learning processes needed for the rehabilitation of listening difficulties.

Copyright Permissions

- Chapter 2 is adapted from the paper:

  Jones, P.R., Moore, D.R., Shub, D.E., and Amitay, S. (2013). Reduction of internal noise in auditory perceptual learning. *J. Acoust. Soc. Am.*, **133**(2), 970-981

- Data from Chapter 2 were presented in:

  Jones, P.R., Moore, D.R., and Amitay, S. (2010). Auditory perceptual learning: Signal enhancement or noise reduction? *Poster at BSA, Manchester*

- Data from Chapter 3 were presented in:

  Jones, P.R., Moore, D.R., and Amitay, S. (2011). Practice reduces informational masking by improving decision strategy and reducing lapse rates. *Poster at BSA, Nottingham*.

  Jones, P.R., Moore, D.R., and Amitay, S. (2012). Practice reduces informational masking by improving decision strategy and reducing lapse rates. *Poster at ARO Midwinter Meeting, San Diego*.

- Data from Chapter 5 were presented in:

  Ratcliffe, N., Jones, P.R., Moore, D.R., and Amitay, S. (2012). The role of response bias when learning a yes/no task. *Poster at BSA, Nottingham*.

- Data from Chapter 6 were presented in:

  Jones, P.R., Moore, D.R., and Amitay, S. (2012). The role of response bias when learning a forced-choice task. *Poster at BSA, Nottingham*.

- The LaTeX source-code used to typeset this thesis is freely available at:

  https://code.google.com/p/ihr-thesis/.

"Philosophy and science, and the springs
Of wonder, and the wisdom of the world,
I have essay'd, and in my mind there is
A power to make these subject to itself"

– George Gordon Noel

Nomenclature

γ	*Asymptotic psychometric performance*	The extrema of the psychometric function. Upper and lower asymptotes are indicated by the subscripts γ_{up} and γ_{lo}, respectively. For functions relating inputs to $\mathrm{P}(`yes')$, the ideal observer would exhibit $\gamma_{up} = 1$ and $\gamma_{lo} = 0$. Deviations from the ideal are assumed to reflect inattention
	Bias	$\lambda - \lambda_{ideal}$ A systematic tendency to favour ones response alternative independent of the sensory of evidence. The *a priori* decision factor. Modelled as the deviation of the listener's criterion from the ideal
c	*Bias metric*	$\lambda - \frac{1}{2}d'$ For functions relating inputs to $\mathrm{P}(`yes')$: $c < 0$ indicates a bias towards the 'yes' response; $c > 0$ indicates a bias towards the 'no' response; $c = 0$ indicates no bias. *Also known as:* λ_{centre}.
c_T	*Bias metric (Total)*	$\lambda_{obs} - \arg\max_\lambda \sum \mathrm{P}(hit \mid \lambda) + \mathrm{P}(Cor\ Rej \mid \lambda)$ Generalised form of c, appropriate for situations where N distributions ≥ 2
ϱ	*Bias under guessing*	$0 \leq \varrho \leq 1$ A systematic tendency to favour a response alternative when responding inattentively. When $\varrho = 0.5$, $(1 - \gamma_{up}) = \gamma_{lo}$. When $\varrho = 1$, $(1 - \gamma_{up}) = K$ and $\gamma_{lo} = 0$
CE	*Constant error*	$PSE - PPE$ A psychophysical metric of bias, derived from the lateral shift of the psychometric function relative to the ideal. For functions relating inputs to $\mathrm{P}(`yes')$: $CE < 0$ indicates a bias toward the 'no' response; $CE > 0$ indicates a bias toward the 'yes' response; $CE = 0$ indicates no bias (n.b. this is the opposite sign to the SDT metrics, c and c_T)

λ	*Decision criterion*	$-\Phi^{-1}(FA)$ An SDT metric of the cut-off point in a listener's decision rule. It is assumed that this listener will predicate their responses on whether or not the decision variable exceeds this value
DV	*Decision variable*	$\sum_{i=1}^{n} \omega_i Input_i$ A scalar value which, when compared to a criterion value, determines the listener's decision. In the tasks considered here, the DV is assumed to be linear sum of the weighted inputs
DL	*Difference limen*	The psychophysical threshold of perception. The smallest stimulus difference that can be reliably detected on some specified proportion of trials. The proportion is often denoted in a subscript. For example, DL_{79} is the smallest physical condition at which the listener responds correctly on 79% of trials. Limens are sometimes referred to simply as *thresholds*. In yes/no tasks *DL* may be analogously referred to as the *detection limen*: the minimum detectable difference from 0
η_{enc}	*Encoding efficiency*	$0 \leq \eta_{enc} \leq 1$ The goodness with which the information contained in multiple channels (e.g., spectral regions) is integrated. The ideal observer will weight each channel proportionate to the reliability of the task-relevant information. In contrast to internal noise, improved performance through increased encoding efficiency is a deterministic process. Sometimes referred to without the subscript, simply as η
ω	*Encoding weight*	$-1 \leq \omega \leq 1$ The (relative) degree to which a corresponding information channel informs the decision process. As such, a weight may be considered a metric of how much *attention* the listener pays to that aspect of the physical input. Ideally each information channel should be weighted proportional to its relative signal-to-noise ratio. The complete set of weights constitutes the listener's encoding strategy, and thus determines their encoding efficiency
A_f	*External noise exclusion*	A scalar value that acts multiplicatively on external noise magnitude to determine the *effective* external noise magnitude. Such a process is typically considered to represent the tuning of one or more perceptual filters, and so is conceptually related to the concept of encoding weights/efficiency
σ_{ext}	*External noise magnitude*	$0 \leq \sigma_{ext}$ The standard deviation of the additive Gaussian noise distribution, arising from sources extrinsic to the observer
λ_{ideal}	*Ideal decision criterion*	The value of λ that maximises the objective function, such as overall P_C

K	*Inattentiveness*	$K = 1 - \gamma_{up} + \gamma_{lo}$ The proportion of trials in which the listener responds randomly, independent of the stimulus. A non-sensory limitation on performance
σ_{int}	*Internal noise magnitude*	$0 \leq \sigma_{int}$ The standard deviation of the additive Gaussian noise distribution, arising from sources intrinsic to the observer
P_C	*Percent correct*	$0 \leq P_C \leq 100$ The proportion of responses that matched the target response ($\times 100$). Sometimes presented with a subscript denoting a particular subset of trials, e.g., $P_{C<SN>}$ denotes percent correct on signal-noise (i.e., 'Interval 1') trials
PPE	*Point of physical equality*	The abscissa value at which the ideal psychometric function is at chance on the ordinate. The point at which the stimulus magnitudes are equal (e.g., identical tones), and responses should be randomly distributed
POE	*Point of subjective equality*	The abscissa value at which the observed psychometric function is at chance on the ordinate. A metric of when the observer is equally likely to choose between each response alternative, sometimes taken to indicate the perceived equality between stimulus magnitudes
d'	*Sensitivity*	yes/no: $\Phi^{-1}(H) + \Phi^{-1}(FA)$ 2I2AFC: $\left[\Phi^{-1}(P_{C<SN>}) + \Phi^{-1}(P_{C<NS>})\right]/\sqrt{2}$ An SDT metric of sensitivity. The distance between the means of two Gaussian, internal response distributions, in z-score units. When bias = 0 d' is directly proportional to P_C. See Wickens (2002) for alternative formulations
β	*Signal gain*	A scalar value that acts multiplicatively on the signal magnitude (and, in some models, the external noise magnitude also). Such improvements can occur through more efficient tuning of filters to task-relevant stimulus components. This term is therefore largely analogous to the encoding efficiency term, η_{enc}. However, β has its own etymology within the visual literature and, unlike η_{enc}, is not constrained to values within zero and one
σ_{all}	*Total noise magnitude*	$\sigma_{all} = \sqrt{\sigma_{int}^2 + \sigma_{ext}^2}$ The standard deviation of the sum total noise distribution, including both internal and external sources

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Background

In this chapter the phenomenon of auditory perceptual learning is detailed and the research framework introduced. A model of decision-making is described, within which four potential limitations on performance are operationalised. Two of these – ***internal noise magnitude*** *and* ***encoding efficiency*** *– pertain to the efficiency with which sensory information is extracted. The other two constructs –* ***bias*** *and* ***inattentiveness*** *– are primarily non-sensory, and relate to higher order behaviour and decision processes. The thesis of the present work is that changes in one or more of these factors underlie perceptual learning. Previous research pertaining to these constructs and their role in learning is evaluated.*

1.1 Auditory perceptual learning

HUMAN sensory systems possess a striking capacity to acclimatise to their environment. This was demonstrated over a century ago, when subjects wearing prismatic glasses for prolonged periods were found to 're-invert' their distorted visual input (Stratton, 1897). Perceptual learning refers to a specific form of such acclimatisation, whereby performance on a sensory task improves as a function of practice (Goldstone, 1998; Fine and Jacobs, 2002). In contrast to the phenomena of sensitisation, habituation and priming, these improvements tend to be acquired over a protracted periods, and are often retained for many days (Karni and Sagi, 1993), weeks (Molloy et al., 2012), or months (Karni and Sagi, 1993). And, unlike more general arousal effects, these improvements tend to be relatively specific to the particular materials used during training (though see Fahle, 2005; Ahissar and Hochstein, 1997; Irvine et al., 2000; Wright and Zhang, 2009). However, whilst the phenomenon of auditory perceptual learning has been well characterised, comparatively little is known about

the functional changes underlying such improvements. How is it that perceptual judgements are improved through experience? In this thesis I attempted to address this question using models of behavioural data.

1.1.1 Perceptual learning tasks

Learning has been shown to occur across a wide range of auditory tasks, including basic judgements of sound spectrum, level, and timing, as well as more complex tasks involving the recognition of tonal patterns or speech stimuli (Table 1.1). Additional cases of learning have been reported anecdotally (e.g., Neff and Callaghan, 1988), or are suggested by the use of prolonged practice sessions prior to testing. This wealth of evidence suggests that learning is fundamental to the auditory decision making process, such that performance on any auditory judgement task may improve with practice.

1.1.2 Paradigms & Methods

In a psychoacoustic task the listener is presented with a stimulus (or set of stimuli), about which they must make a judgement and respond accordingly. More specifically, and in keeping with the overwhelming majority of the perceptual learning literature, the present work is restricted to tasks where:

1. Each trial is composed of a sequence of discrete events (e.g., cue, observation$_i$, observation$_{i+1}$, ..., observation$_m$, response, reward).
2. Responses are unspeeded. Listeners have an unlimited time to respond, and are encouraged to maximise accuracy rather than speed.
3. Listeners respond from a finite set of discrete alternatives (e.g., 'yes' or 'no').
4. The ideal strategy is to make decisions contingent purely upon the sensory evidence (i.e., trial variables are independent of those preceding/following).

The precise configuration of trial events often differs between studies. For example, some studies use 3+ observation intervals and ask the listener to 'pick the odd one out' (Amitay et al., 2005), while other studies ask listeners to compare two observations and 'select the greater' (e.g., in duration Wright et al., 1997), or make one observation and either identify the sound (Watson et al., 2008) or judge whether or not a particular sound was present (Wenger and Rasche, 2006). Amounts and rates of learning appear largely invariant across trial configurations (Amitay et al., 2006). The exception to this is the use of feedback, which when given after each

Domain	Task	Example Reference(s)
Frequency	tone discrimination	Campbell and Small, 1963 Demany, 1985 Micheyl et al., 2006
	f0 discrimination	Grimault et al., 2002
	harmonic complex discrimination	Carcagno and Plack, 2011
	tone identification	Hartman, 1954 Meyer, 1899 Cuddy, 1970
Intensity	tone detection in quiet	Zwislocki et al., 1958
	tone detection in broadband noise	Tucker et al., 1968 Gundy, 1961
	tone detection in a tonal sequence	Leek and Watson, 1984
	tone discrimination	Wright and Fitzgerald, 2005
	tone discrimination in noise	Buss, 2008
Contrast	amplitude modulation detection	Fitzgerald and Wright, 2011
	AM rate discrimination	Grimault et al., 2002 Fitzgerald and Wright, 2005
Timing	interval duration discrimination	Wright et al., 1997 Karmarkar and Buonomano, 2003
	position identification	Tanner and Rivette, 1963
	asynchrony detection	Virsu et al., 2008
	duration order discrimination	Mossbridge et al., 2006 Mossbridge et al., 2008
Spatial	time difference [ITD] discrimination	Rowan and Lutman, 2007 Wright and Zhang, 2006
	level difference [ILD] discrimination	Wright and Fitzgerald, 2001
	monaural cue localisation	Van Wanrooij and Van Opstal, 2005
Patterns	tone-complex identification	Green, 1992
	sequential tone identification	Leek and Watson, 1988 Warren, 1974
	sequential tone discrimination	Barsz, 1996
Speech	syllable-identification	Tremblay et al., 1998
	sentence recognition	Millward et al., 2011 Hagerman, 1982

Table 1.1: Auditory tasks for which perceptual learning effects have been evidenced. As discussed in §1.1.3, not every study uses a consistent criterion of learning.

trial or block tends to enhance the amount (Ball and Sekuler, 1987) and rate (Fahle and Edelman, 1993) of learning, though does not appear to be a prerequisite of learning (Campbell and Small, 1963; though see Herzog and Fahle, 1997).

1.1.3 Measures & Analysis

Dependent Variable The dependent variable in most perceptual learning studies is the difference limen, DL: an estimate of the smallest physical difference that can be correctly discriminated on some arbitrary proportion (e.g., 70.7%) of trials. Early learning studies derived this value from psychometric functions that were fitted to data acquired using the Method of Constant Stimuli (e.g., Campbell and Small, 1963). However, most contemporary studies track a particular performance level directly, using algorithms that adaptively vary the relevant stimulus parameter (Amitay et al., 2006). The move towards adaptive algorithms was partly motivated by their greater efficiency – adaptive tracks require an order of magnitude fewer trials to yield DL estimates (thereby minimising intra-test training effects). Adaptive tracking is also particularly well suited to situations where the magnitude of individual variability makes it difficult to specify the range of stimulus values in advance, as is often the case with naïve listeners (e.g., Amitay et al., 2005, observed frequency discrimination DLs that differed by over 2 orders of magnitude). Finally, adaptive tracks may also have the advantage of enhancing learning by ensuring that task difficulty is manipulated appropriate to the listener's abilities (Ahissar et al., 2009; Linkenhoker and Knudsen, 2002).

Changes in the signal detection theory sensitivity metric, d', are also commonly used as an index of learning. The principle advantage of d' is that it is a 'pure' measure of perceptual sensitivity. That is, unlike DL or P_C (percent correct), d' is considered to be independent of response bias (though see Chapter 6 for exceptions). The principle disadvantage of d' is that, unlike DL, it indexes sensitivity with regard to a specific physical stimulus value. Thus, the use of d' is problematic in situations where the appropriate physical values cannot be specified in advance.

Analysis Learning can be intuitively expressed in terms of the quantity of listeners that exhibit improved performance (e.g., lower DLs) after training (e.g., Campbell and Small, 1963; Demany, 1985). More formally, learning is usually evidenced either by a significant difference between mean session scores (e.g., Grimault et al., 2003), and/or a significant negative slope when performance is regressed against a unit of practice, such as block, session, or day (e.g., Wright et al., 1997; Buss, 2008). These

Study	$\sim\Delta DL$ (%)	DL Units	Discrim. task
Demany (1985)	48	Hz [$\Delta F_{70.7}$]	Frequency
Buss (2008)	37	dB [10log($\Delta I_{79}/I$)]	Intensity
Wright and Fitzgerald (2001)[a]	53	dB [10log(ΔI_{79})]	ILD
Wright and Fitzgerald (2001)[a]	54	msec [Δt_{79}]	ITD
Wright et al. (1997)	56	msec [$\Delta t_{79}/t$]	Temp. interval

[a] For overview see Wright and Fitzgerald (2003)

Table 1.2: Approximate auditory learning effect sizes on five example discrimination tasks, expressed in terms of percent improvement on initial threshold performance (DL). The units in which the DL was measured are shown in column three, with a subscript indicating the percent correct level at which the threshold was evaluated (in each case determined by the number of reversals in a Levitt (1971) staircase). In some cases the DL was normalised by the standard to derive a Weber fraction, and this fact is included for completeness.

analyses may be carried out within individuals, or at the group level using repeated-measures analyses. To demonstrate that a particular intervention was responsible for any observed learning it is also necessary to establish that any improvements are significantly greater than those in a suitable control population (e.g., Zhang and Wright, 2009).

The precise criterion of learning often differs between studies. For example, Wright and Fitzgerald (2001) employed a stringent requirement that mean DL scores must both significantly differ across, and negatively regress against, session number. Conversely, Demany (1985) required only that scores are improved in magnitude after practice. In the studies reported here, a significant group-mean difference is typically used as the criterion of learning, though regression statistics are also considered where appropriate.

Effect size Asymptotic learning effect sizes differ substantially between tasks (Wright and Zhang, 2005; Fine and Jacobs, 2002). However, in many basic auditory tasks it is common for group-mean starting performance (DL) to improve by 30 – 60% (Table 1.2).

Even greater variability is often observed between individuals, within a single study. Indeed, it is common for a proportion of listeners (e.g., 38%, Zhang and Wright, 2009) to exhibit no significant learning effect (Tremblay et al., 1998)[1]. Some of this variance may be explained by initial performance (Campbell and Small, 1963; Amitay et al., 2005), while additional individual variability has been associated with differences in

[1]Though it is not always possible to disambiguate an absence of improvement from very rapid (asymptotic) improvement within the first session.

non-verbal IQ (Amitay et al., 2010). A unified framework for explaining such individual differences remains lacking, however.

1.1.4 Temporal profile

Timescale The amount of practice required to reach asymptotic performance varies considerably between studies. The most important distinction appears to be the complexity of the task (Robinson and Summerfield, 1996). With simple judgements, such as those concerning the relative frequencies of two pure tones, the majority of learning occurs within 1000 trials, though further learning often continues for several thousand further trials (Demany, 1985; Hawkey et al., 2004; Molloy et al., 2012). In contrast, with judgements regarding more complex stimuli, the period of learning is substantially elongated. For example, the majority of improvement in non-native phonetic contrast discrimination is observed only after 20,000 trials, with further learning occurring even after 60,000 trials (Lively et al., 1993). Notably, it remains at present unclear whether this increase in learning duration reflects a corresponding increase in the sum-total amount of information being extracted from the stimulus, or whether in more complex situations the rate of learning is diminished, such that listeners require more trials to extract an equivalent amount of information.

Related to complexity is the notion of variability, which also appears to affect the period over which learning occurs. For example, the number of trials required to achieve asymptotic performance on a frequency-discrimination task has been found to increase when the frequency of the standard is randomly jittered across trials(Amitay et al., 2005). Numerous other latent factors may also contribute to the learning timescale, and are poorly understood at present. Some of these pertain to properties of the stimulus (e.g., the duration of the stimulus), while others relate to individual differences between listeners, for example in terms of motivation or IQ (Amitay et al., 2010).

Dynamic The shape of the auditory perceptual learning curve is typically characterised by a relatively short period of very rapid learning followed by a protracted period of more gradual learning[2]. Accordingly, it is common to model learning data with an exponential or power function (e.g., Molloy et al., 2012). The dynamic of auditory learning appears more linearised than in motor learning, where improvements commonly follow a double

[2]Furthermore, some authors have argued that the grossly continuous profile is actually constituted by discrete phases of within-session learning and between-session and/or over-night consolidation (Molloy et al., 2012; Donovan and Radosevich, 1999). However, such effects were not apparent in the present work.

exponential trajectory (Krakauer et al., 1999; Braun et al., 2009), but often appears less linear than in vision, where relative rates of learning often decrease more gradually across trials/sessions (e.g., Wenger and Rasche, 2006; Fahle, 2005; Lu and Dosher, 2004; Ball and Sekuler, 1987). These differences are quantitative rather than qualitative, however, and visual learning studies also exhibit (sometimes marked: Poggio et al., 1992) non-linearities. It is unclear at present whether these differences are modality-specific *per se*, or relate to more general differences in task complexity.

This non-linear learning profile has been proposed to indicate two distinct learning processes: a procedural phase in which learning pertains to the paradigm and procedure, followed by a perceptual phase, in which learning concerns the task-relevant sensory distinction. However, such a duality of processes need not necessarily obtain. Non-linear outputs may readily arise from linear changes in a single underlying parameter (Elman, 1997). Moreover, a strong mapping of early/late to procedural/perceptual learning is inconsistent with the double-dissociation by which brief exposure to test procedure does *not* elicit improved performance, whereas brief exposure to a task-relevant perceptual distinction *does* improve performance, even given a novel test procedure (Hawkey et al., 2004).

As well as the slow/rapid phases of learning, many experiments yield a third phase in which performance actually begins to decline towards the end of the regimen (e.g., Huyck and Wright, 2011). The mechanisms of this decline remain uncertain, but may relate to boredom or fatigue. Consistent with this, abrupt restorations of performance have been observed when a novel manipulation is introduced to the experiment (Neff and Dethlefs, 1995). Analogous deterioration effects have also been observed within sessions, and have been attributed to fatigue of primary sensory neural circuits (Mednick et al., 2008),

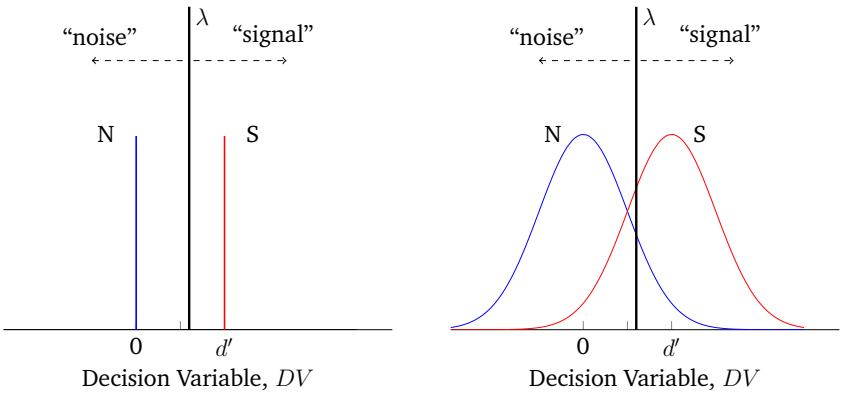
1.2 Model of decision making

In ∮1.2.1, the core tenets of detection theory are briefly described, with particular regard to the key requisite assumptions. This provides the framework within which listeners are assumed to make decisions, and introduces the essential concepts of the decision variable, DV, and decision criterion, λ. In ∮1.2.2 this process is elaborated upon, and the principles described by which a sensory observation is mapped to the DV, and the location of λ is determined. In doing so four factors are introduced that may potentially limit the efficiency of the decision process. Two of these, **internal noise magnitude** and **encoding efficiency**, pertain to the

efficiency with which sensory input is mapped to the DV. The other two, **bias** and **inattentiveness**, are primarily non-sensory, and relate to the listener's propensity to respond in a manner independent of the input. Section 1.3 further elucidates the nature of these mechanisms and how they may be measured.

1.2.1 Basic detection theory

In the types of tasks that shall be considered here (cf. ∮1.1.2), the listener's role is to judge to which category a given stimulus belongs. For example, on each trial the listener may be asked whether they heard a noise, N, or a signal, S. Signal Detection Theory (Peterson et al., 1954; Swets et al., 1961; Green and Swets, 1974; Macmillan and Creelman, 2005) provides a theoretical framework for how the listener makes this decision. As illustrated in Fig 1.1a, the process involves two steps.



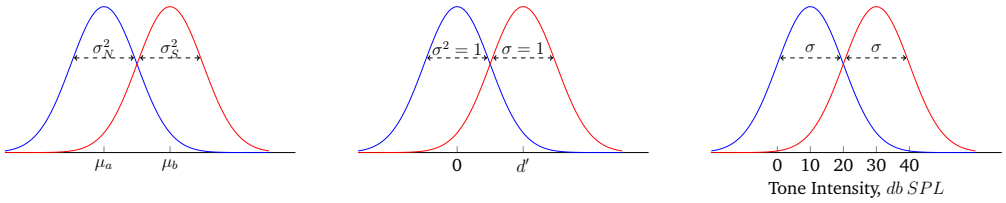
(a) **(b)**

Fig. 1.1: **(a)** DV values for a single observation of a noise, N, and a signal, S, stimulus. The vertical line represents the decision criterion. **(b)** DV probability density functions from which individual observations are drawn. Since the curves overlap, the DV for a noise stimulus will exceed that induced by a signal on some proportion of trials. N.B. While a detection task is described here, the same principles readily generalise to discrimination tasks also (see Macmillan and Creelman, 2005).

Firstly, a scalar decision variable, DV, is constructed from the sensory observation. Secondly, a hypothesis is selected by comparing the DV to a decision criterion, λ, thus:

$$response = \begin{cases} \text{“}signal\text{”}, & \text{if } DV > \lambda. \\ \text{“}noise\text{”}, & \text{otherwise.} \end{cases} \tag{1.1}$$

When practicable, the decision dimension shall be assumed to correspond directly to some physical parameter of the sensory input, such as frequency



(a) **(b)** **(c)**

Fig. 1.2: Gaussian SDT models when **(a)** The variance of each distribution is unconstrained **(b)** Each distribution is constrained to have equal (unit) variance (and the mean of the first distribution is arbitrarily set to zero) **(c)** Each distribution is constrained to have equal variance (and decision dimension is assumed to correspond to some physical scale)

in Hz, or intensity in dB SPL. Thus, the DV on a tone detection task may take either the value 0 or T dB (where T is the level of the target). Crucially though, while the expected value of DV is determined by the actual stimulus value, the trial-by-trial value of DV is subject to random variation, or 'noise'. Therefore, as shown in Fig 1.1b, the DV value associated with each stimulus category is a random variable, continuously distributed over the decision dimension. On each trial a sample is drawn from one of the distributions, and the listener must determine from which distribution the observation belongs.

As is common, the present work assumes that the random variability associated with each stimulus class is Gaussian and additive. These assumptions serve to limit the number of parameters in the underlying model, thereby allowing their estimates to be adequately constrained by the relatively small datasets afforded by perceptual learning studies. By assuming normality, each distribution can be fully specified by two parameters: mean, μ, and standard deviation, σ (Fig 1.2a). By assuming additivity, the standard deviation of both distributions can be specified by a common parameter $\sigma_a = \sigma_b = \sigma$. When one is only interested in the relative amount of overlap between the distributions (i.e., *sensitivity*), one can arbitrarily set $\sigma = 1$ and $\mu_N = 0$. Sensitivity is thereby fully determined by the distance between the expected values of the two distributions, d' (Fig 1.2b), the value of which is constrained by just two behavioural estimates – the listener's hit and false alarm rate (assuming a single, fixed criterion). Alternatively, if the amount of internal noise is of primary interest then μ_N and μ_S may be set to their physical equivalents (if such equivalence may be assumed), and σ becomes the single unknown parameter in the model, which, together with λ, fully predicts performance.

While mathematically convenient, the assumptions of Gaussian additivity are unlikely to hold universally. Thus, converging evidence from psychophysics (McGill and Goldberg, 1968; Swets, 1959; Watson et al., 1964) and neurophysiology (Young and Barta, 1986; Teich and Khanna, 1985; Siebert, 1970; Winter and Palmer, 1991) indicates that, over wide perceptual ranges, listeners are grossly limited by a form of noise, the amplitude of which is proportional to the magnitude of the stimulus ('multiplicative' noise). For example, Weber's law (that the 'just noticeable difference' between two stimuli is proportional to the magnitude of the stimuli) is often taken to indicate the presence of a limiting noise source that increases with stimulus strength. Similarly, auditory neurons are often shown to exhibit approximately Poisson processes with spike-rate variability increasing as a function of mean firing rate (e.g., Fig 6a of Young and Barta, 1986). Moreover, recent evidence suggests that distributions of internal noise, even at the gross level, may constitute a near-miss to normality (Neri, 2013). However, I contend that Gaussian additivity constitutes an acceptable approximation for the tasks considered here, where the stimulus magnitudes are narrowly distributed about the listener's threshold. Thus, near-threshold discriminations of simple (e.g., pure tone) auditory stimuli have been shown to produce *receiver operating characteristic*[3] [ROC] curves that are linear in Gaussian-transformed coordinates, implying that the underlying variables are Gaussian. Moreover, they have been shown to have unit slopes, implying that variance is equal across the two distributions (Viemeister, 1970; Talwar and Gerstein, 1999). Normality is also theoretically justified by reference to the Central Limit Theorem (i.e., since the DV is likely to be instantiated by large *populations* of roughly, mutually independent random variables).

In practice, when computing a given parameter estimate, it shall be further assumed that the observations from each category are independently, identically distributed [i.i.d.]. The assumption of i.i.d. is necessary, since it is necessary to aggregate over multiple trials when estimating hit or false alarm rates. However, the i.i.d. assumption is problematic in the context of perceptual learning since, *ex hypothesi*, the underlying decision parameters are changing throughout the course of the experiment. This makes it necessary to carefully select the number of trials in order to maximise statistical power. Too few trials will result in large deviations in sample-mean estimates. Conversely, too many trials will result in genuine shifts in the population-mean, thereby reducing the magnitude of the improvement (see Appendix A). In practice, any compromise will result

[3] plots of hit rates against false alarm rates, obtained by holding the stimuli constant and varying the criterion

in the size of the learning effect being underestimated. However, certain steps may be taken to maximise the acceptability of the i.i.d., and thus the effect sizes. For example, adaptive algorithms may be used to minimise the number of trials per measure, or the complexity or variability of the task may be increased, thereby minimising the learning rate (see ∮1.1.3).

1.2.2 The decision process

So far a simple model has been presented in which decisions are made by applying a decision criterion, λ, to the trial-by-trial value of a decision variable, DV, drawn from a random distribution. However, to see how the goodness of these decisions may be limited (and by implication, improved through practice), it is necessary to expatiate upon how sensory observations are mapped to DV values, and how the location of λ is determined. To wit, the assumed process by which a listener makes a decision given a sensory observation, and the sequence of internal transformations required therein, are depicted in Fig 1.3.

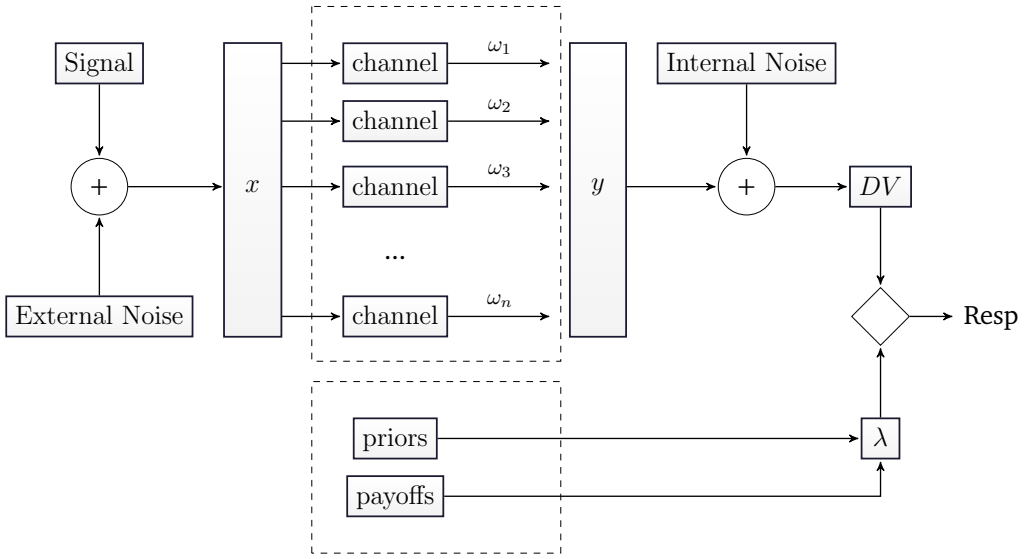


Fig. 1.3: The present model of decision making. The observer receives a noisy input, x, which is reduced to a scalar internal response, y, via the linear weighted sum of each information channel. The internal response is also corrupted by an (additive) noise source that is invariant of the stimulus magnitude. The observer makes a decision by comparing the magnitude of the resultant decision variable, DV, to a criterion, and responds accordingly. The criterion is informed by the observer's perception of relative utilities and prior expectations (see body text).

This model is conceptually similar to a number of previous expositions in both the visual (Gold et al., 2004; Pelli, 1991) and auditory (Richards and

Zhu, 1994) literatures. The input is a physical stimulus, x, comprised of a signal corrupted by an additive external noise. Often such external noise arises from energy distributed concomitantly with the signal in terms of their fundamental physical characteristics (e.g., frequency, amplitude, and temporalspatial location). However, external noise may also occur that has no obvious or direct physical connection with the signal, so long as it ultimately increases uncertainty at the level of the decision variable, DV. When normally distributed, the magnitude of external noise is denoted, σ_{ext}. Its *effective* magnitude is determined by the resultant standard deviation in the DV, and may be attenuated relative to σ_{ext} (e.g., if the listener is able to ignore the channels containing the noise). Sources of external noise include imperfect signal synthesis (e.g., sampling, quantization), transducer imperfections and incidental background sounds – none of which are typically assumed to have an impact in well-controlled laboratory experiments.

This stimulus sample, x is transformed into an internal response (or: *representation*), y, by summing over the weighted outputs of *n* independent information channels. How these channels are conceived depends on the task, the level of analysis, and the question of interest. For example, in a yes/no detection task each channel may be a frequency tuned filter, in which case y corresponds to activity summed across spectral regions. Alternatively, each channel may represent temporal bins, such as observation intervals in a mAFC or multiple-looks task. Each ω, indicates the relative degree to which the corresponding channel informs the decision process. As such, $|\omega|$ may be considered a metric of how much *attention* the listener pays to that aspect of the physical input. To the extent that the weights approximate their ideal values, a listener shall be deemed more or less efficient in their encoding of the stimulus information. The potential effects of greater **encoding efficiency** are twofold. Firstly, it may serve to deterministically increase the amount of task-relevant information the listener is able to extract from the physical stimulus[4]. And secondly, it may serve to reduce the *effective* level of the external noise by filtering out task-irrelevant information. For example, by ignoring the non-signal-frequency components in a notched-noise masking experiment.

ENCODING EFFICIENCY

The DV is formed by combining y with an additive **internal noise** of magnitude σ_{int}, analogous to how signal and external noise were combined previously. In contrast with *external* noise, internal noise is random variation arising from sources intrinsic to the listener. Its possible

INTERNAL NOISE

[4]Relating this concept back to the Signal Detection Theory schemas shown previously in Fig 1.2, increased encoding efficiency here implies a greater signal magnitude, as represented by increased separation between the means of the two DV distributions

sources include: non-deterministic transduction (e.g., due to Brownian motion of hair cells; Denk et al., 1989), stochastic neural encoding and transmission both in the auditory periphery (Javel and Viemeister, 2000) and more centrally (e.g., Vogels et al., 1989), physiological maskers such as heartbeats and blood flow (Soderquist and Lindsey, 1971; Shaw and Piercy, 1962), as well as random fluctuations in attention, motivation, memory, and other factors relating to the decision process. Some authors have further subdivided the internal causes of noise (e.g., sensation versus central: Durlach and Braida, 1969; Oxenham and Buus, 2000; Shinn-Cunningham, 2000), but I do not do so here. It is important to note that internal noise represents the sum of all of the observer's variable limitations. Thus, while it is often tempting to think of internal noise exclusively as originating from the statistical properties of sensory neurons, such an assumption is unnecessary for our purposes. Internal noise is here modelled as being 'late', or 'central'. However, in some scenarios it may be more appropriately thought of as being introduced within each encoding channel. In both cases its effects are equivalent, however early internal noise arising early may also be affected by the listener's encoding strategy.

The listener makes a decision by comparing the DV value to a criterion, λ, as per Eq 1.1. In most perceptual learning tasks the goal is to maximise overall percent correct, and the criterion should be placed accordingly. Furthermore, in virtually all learning tasks the prior probability of each stimulus category occurring is uniformly distributed. As such, the ideal criterion location in a two category (equal-variance) task is almost invariably at the point equidistant between the means of the two DV distributions (Wickens, 2002). Any systematic deviations from the ideal location will decrease performance, and shall be termed **bias**. Bias may occur if, for example, the listener perceives the relative utility (or: *payoff*) of each response outcome to be asymmetric (Maddox and Bohil, 1998) (e.g., if they believe the penalty of missing an *A* outweighs the benefit of spotting a *B*), or if they perceive the relative probability of each trial-type occurring to be asymmetric (Craig, 1976; Tanner et al., 1967; Parducci and Sandusky, 1965; Schulman and Greenberg, 1970; Parks and Kellicutt, 1968). Suboptimal performance will also occur if the listener adduces the correct criterion placement, but is unable to maintain a fixed criterion. However, any such instability in the decision criterion shall here be assumed random and subsumed under internal noise.

BIAS

Once the listener has thresholded the sampled sensory evidence the decision process is complete and the listener is assumed to respond accordingly. Decision and response are therefore identical in the model. This is not necessarily the case in every perceptual judgement task. For

example, motor errors may result in the response deviating from that intended, or the listener may correctly identify the response but forget which key to press. These inefficiencies are assumed largely negligible in the types of task considered here (i.e., where the sets of responses are small and clearly labelled, and where total emphasis is placed on the accuracy of responses, rather than speed). Any such errors that do occur are assumed to constitute random processes and are modelled as internal noise.

INATTENTION A fourth limitation on performance is **inattentiveness** – lapses in concentration resulting in the listener not perceiving the stimulus, misperceiving it, or otherwise rendering the listener incapable of making an informed judgement. Such lapses are not represented within Fig 1.3 since they are modelled as a complete departure from the decision process described above. Instead, it is assumed that in the event of a lapse, the listener's response is driven by a separate process in which on a proportion of trials, K, a response is selected at random based on a uniform random distribution (an unbiased guess).

To sum up, a simple model of perceptual decision making has been presented in which the accuracy of listeners responses are limited by four main forms of potential inefficiency. Two of the limitations, internal noise magnitude and encoding efficiency, relate to how sensory evidence is mapped to the DV. These factors determine the listener's *sensitivity*, in that improvements in either will increase the separability between the underlying DV distributions associated with each stimulus class. A third limitation, bias, relates to the listener's placement of their decision criterion, which may be more or less ideal. This is typically thought of as a higher order and/or non-sensory factor. The final component, inattentiveness, is similarly non-sensory, and relates to how well listeners are able to sustain concentration on the task. Improvements in any of these factors will increase performance, and so could in principle explain the perceptual learning phenomenon, either individually or in combination. The experimental work presented in subsequent chapters aimed to contrast empirically the relative importance of each factor as a mechanism of learning.

1.3 Mechanisms of Learning

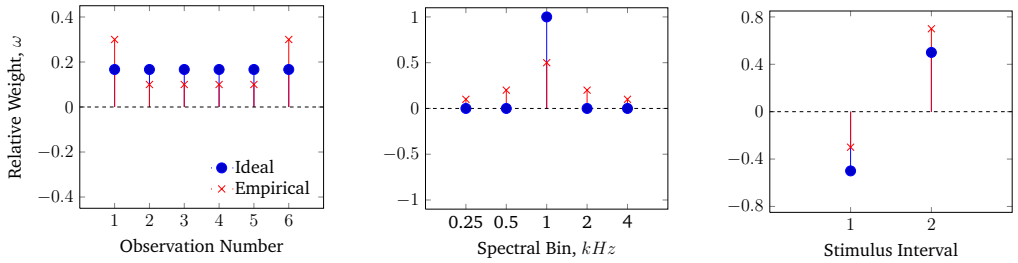
In ∮1.2 four potential mechanisms of learning were introduced in the context of a general decision making framework. Here each of these factors is elucidated in turn, with particular reference to how these concepts may be operationalised. Further information will be provided within

the relevant experimental chapter(s), as indicated at the end of each subsection.

1.3.1 Encoding efficiency

Encoding efficiency essentially describes how well a listener understands the structure of, and is able to extract task-relevant information from, a physical stimulus. In the decision model presented in ∮1.2.2, stimulus information is encoded within a number of independent, discrete channels. Each channel has an associated weight, ω, the relative magnitude of which may be considered a metric of how much *attention* the listener pays to that aspect of the stimulus. The values of $[\omega_1 \ldots \omega_n]$ can be estimated by correlating the trial-by-trial stimulus values in each channel with the listener's response, and by normalising the coefficients so that their magnitudes sum to one. The resultant values may be positive or negative, and their relative magnitudes indicate the degree to which the corresponding channel informs the decision process.

The complete vector of normalised weights, $[\omega_1 \ldots \omega_n]$, constitutes a listener's *encoding strategy*. In general, the ideal encoding strategy is to give non-zero weight only to task-relevant channels, and to scale each weight proportional to the relative reliability of the information contained in the associated channel. As illustrated in Fig 1.4, the ideal configuration of weights will depend on the task. For example, in a multiple-looks task the ideal strategy would be to assign the same weight to each observation, such that the output of each channel informs the decision variable equally. Conversely, in a detection task where each channel represents a spectral region, the ideal strategy would be to give weight only to those channels containing energy from the target stimulus, and to do so proportionally to the signal-to-noise ratio at each bin. In a two alternative forced-choice discrimination [2AFC] task, the best strategy is to predicate responses on the difference between the two intervals. This can be achieved by linearly summing over weights that are equal in magnitude but opposite in sign.



(a) **(b)** **(c)**

Fig. 1.4: Example ideal (blue) and hypothetical empirical (red) decision weights for a **(a)** multiple-looks **(b)** tone-detection **(c)** 2I2AFC discrimination task.

Encoding efficiency, η_{enc}, may be improved (increased) by the listener learning which information channels contain the greatest signal to noise ratio, and directing their attention accordingly. The role of encoding efficiency in learning is considered in Chapter 2, and forms the main focus of Chapter 3 and Chapter 4.

1.3.2 Internal noise magnitude

Noise refers to random variation in the listener's internal response to a signal, which is, *qua definitione*, unwanted and liable to interfere with the listener's judgements. In many contexts 'noise' is therefore synonymous with 'variability', and the two terms are often used interchangeably. That noise is present in sensory decision processes can be inferred from the fact that each presentation of a putatively identical input stimulus will, given a sensitive enough measure, produce a different output response on each occasion[5]. Internal noise is distinguished from external noise by its point of origin.

Over the last 50 years a number of measures of internal noise magnitude, σ_{int}, have been developed. These include external noise titration (Lu and Dosher, 2008), model-fitting (Jesteadt et al., 2003), *n*-pass consistency (Green, 1964), multiple-looks (Swets, 1959), and direct variability estimates derived from distributions of errors (e.g., Buss et al., 2009). Details of these techniques will be discussed within the relevant experimental chapters. Notably, however, each of the first three methods share the common principle of referencing an unknown internal noise magnitude, σ_{int}, to a known external noise magnitude, σ_{ext}.

It is therefore imperative to be able to parametrically control σ_{ext}. The classic method of introducing variability to a stimulus signal is through

[5]Indeed, as discussed in Chapter 2, the consistency of output responses can be used to index magnitude of internal noise

masking. This technique is very common in the visual literature (see Lu and Dosher, 2008), where the masker is typically wide-band (e.g., white noise) and presented simultaneously with the signal. However, simultaneous masking is problematic for auditory tasks, since the relatively complex interactions between masker and target (Gifford and Bacon, 2000; Moore and Vickers, 1997) often result in masking effects that are nuanced, and often unpredictable (e.g., Turner et al., 1992; Yost et al., 1976). Even for the relative simple case of tone-detection in bandpass noise, the interactions are not trivial, and not always fully understood (Richards et al., 1991; Humes and Jesteadt, 1989).

Alternatively, intra-trial variability can be introduced without masking by modulating some property of the signal, in a more (e.g., iterated rippled noise; Patterson et al., 1996) or less (e.g., pulse-train jitter; Rosenberg, 1966; Cardozo and Ritsma, 1968) elaborate manner. Unfortunately, as with masking, such *jitter* can often have confounding side-effects. For example, one relatively simple manipulation is to perturb the pitch of a sinusoidal tone by varying the repetition rate (i.e., *vibrato*). However, modulating the frequency in this way will also introduce amplitude modulations, since the signal will be moving across tuned filters in the cochlea (Plack et al., 2005, pp. 40), and there is no simple relationship between the manipulation to the standard deviation of the listener's responses (Shonle and Horan, 1980).

Finally then, variability can be introduced more exactly through jittering parameters across trials (e.g., Jesteadt et al., 2003). In this method, an i.i.d. random value is added to the task-relevant parameter of the stimulus prior to every presentation. As long as the mean value of the DV is proportional to the physical stimulus value, this affords a relatively direct method of manipulation that is ensured to be Gaussian, additive, and of known magnitude.

Internal noise magnitude, σ_{int}, may be improved (reduced) by attenuating any of the random limitations in the decision process. For example, σ_{int} could equally be reduced through a reduction in the variability in cell firing rates, or through the observer learning to sit quietly in their chair. The role of internal noise in learning is primarily considered in Chapter 2 and Chapter 3.

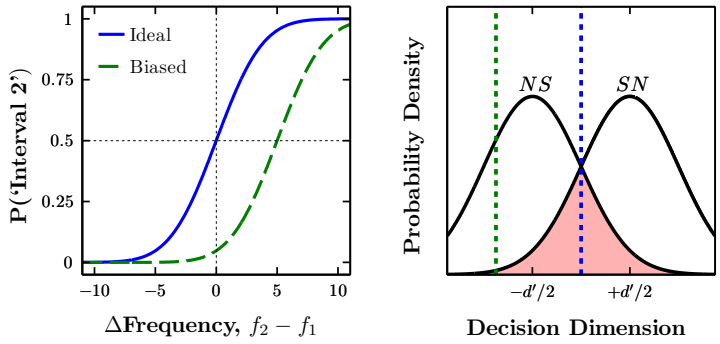
1.3.3 Bias

Bias refers to any predilection for or against making a particular response that exists independent of the sensory evidence. This is a relatively broad definition, and includes both consistent preferences and ones that vary over

time. By definition bias is non-sensory. There therefore exists a natural distinction between bias on the one hand, and constructs such as internal noise magnitude and encoding efficiency that fundamentally relate to the listener's sensitivity to sensory information.

Formally, bias, is defined as the difference between the listener's criterion, λ, and the ideal, λ_{ideal}. Given the assumptions of $\oint$1.2.1, the value of λ can be estimated from the listener's false alarm rate, F, by finding the point on the standard normal distribution above which $F\%$ of the probability lies (Green and Swets, 1974). The value of λ_{ideal} is that which maximises P_C, given (1) the listener's sensitivity to each stimulus, d', (2) the *a priori* probabilities of each stimulus class being observed, and (3) the rewards/penalties associated with each outcome (see Chapter 5 for a more detailed exposition). When each of these factors are distributed symmetrically about the point of physical equality [PPE], bias can also be intuitively derived from the psychometric function by observing how the point of subjective equality deviates from the PPE (cf. Fig 1.5a). However, if any of these assumptions fail, then the PPE no longer indexes the ideal, and the ideal criterion location must be explicitly calculated as per the above. This may occur, for example, if the observer is more sensitive to a particular stimulus/stimulus-configuration (e.g., ascending versus descending pitch differences), in which case the ideal strategy would be to favour the less sensitive category at the PPE.

Bias may be improved (reduced) through better approximation of λ_{ideal}. In practice, this may require listeners learning the statistics of the stimulus probabilities and response outcomes (both of which are generally uniform distributions), or generally minimising the extent to which responses are contingent on anything other than the current sensory information. The role of bias in learning is considered in Chapter 2, and forms the main focus of Chapter 5 and Chapter 6.



(a) **(b)**

Fig. 1.5: Psychometric (left) and SDT (right) schematic illustrations of an unbiased observer (blue) and an observer with an 'Interval 1' response bias (greens). Bias is evidenced by lateral shifts (in psychometric function or criterion).

1.3.4 Inattentiveness

Attention can be operationalised in various ways. For example, in the decision framework of ∮1.2.2, changes in selective attention are manifest as systematic changes in channel weights. Inattentiveness is distinct from this, and concerns a more general disregard for the stimulus. More formally, inattentiveness can be modelled as the listener responding randomly on a certain proportion of trials. Given the simplifying assumption that such guesses are uniformly randomly distributed across trials, observed performance can be thought to represent a probability mix of two processes. Firstly, a basic sensory process, $\Phi(x)$, which causes the probability of responding 'Interval 2' to vary sigmoidally between zero and unity as a function of stimulus magnitude. And secondly, a guessing process which occurs with probability K. Following Green (1995), the probability of an 'Interval 2' response in a two-alternative response design is thus:[6]:

$$P(\textit{'Interval 2'}) = K\varrho + (1 - K)\Phi(x). \tag{1.2}$$

Where K is the probability of inattention, and ϱ is a bias factor which controls the relative likelihood of 'Interval 1' or 'Interval 2' responses when guessing. If $\varrho = 0$ then the listener will always respond 'Interval 1' when guessing, while $\varrho = 1$ means "always guess interval 2", and $\varrho = 0.5$ is unbiased.

Such a model has been favoured here and elsewhere (e.g., Green, 1995; Viemeister and Schlauch, 1992; Wightman and Allen, 1992) due to the relatively simple predictions it makes regarding the shape of the psychometric function. Namely, as the proportion of attentional lapses

[6] See Green (1995) for an analogous formulation when P_C is the dependent variable

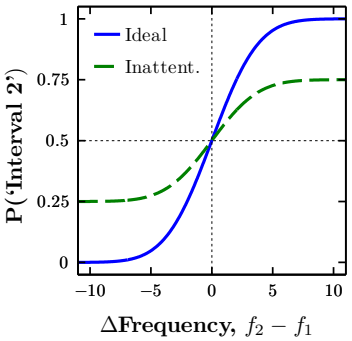


Fig. 1.6: Effects of inattention on the psychometric function. Notice that the inattentive observer (green, dashed) asymptotes above chance and below unity when probability of responding 'Interval 2' is plotted as a function of signal magnitude.

increases, performance in all conditions declines and the slope of the psychometric function shallows. Thus, while inattentiveness is assumed to be independent of sensory process, its effects are not. Increased inattention results in DL estimates being greater and less variable (Green, 1995), and may cause internal noise magnitude to be overestimated (i.e., in instances such as Buss et al., 2009, where σ_{int} is inferred from the reciprocal of the slope).

Since inattentiveness is assumed to be invariant across stimulus conditions, the proportion of inattentive trials can be inferred from asymptotic performance in the most difficult/easy conditions, where it would be reasonable to expect performance to otherwise be at chance/unity. For example, in 1.2, the upper asymptote, γ_{up}, corresponds to $K\varrho + (1 - K)$, whereas the lower asymptote, γ_{lo}, is $K\varrho$. The rate of inattention, K, may therefore be derived from the estimated asymptote values, thus:

$$K = (1 - \gamma_{up}) + \gamma_{lo}, \tag{1.3}$$

while response bias under guessing is given by:

$$\varrho = \frac{\left(\frac{K-1+\gamma_{up}}{K} + \frac{\gamma_{lo}}{K}\right)}{2}. \tag{1.4}$$

Notably, while the assumption that inattention is invariant across stimulus conditions is mathematically convenient, it is likely a crude approximation. It may be that inattentiveness increases at very low levels, if the listener perceives the task to be impossible, or at very high levels if the listener perceives the task to be trivially easy. Moreover, in some instances extremely large stimulus levels may be virtually impossible to ignore (e.g., if they cause pain). To some extent these considerations may have been mitigated in the present work by the use of stimuli predominantly

focused around the listener's threshold. Nevertheless, the derived values of inattention should be treated with caution.

Inattentiveness, K, may be improved (reduced) by listeners minimising the proportional of trials on which they undergo lapses in concentration. The role of inattentiveness in learning is considered in Chapter 2.

1.4 Previous data concerning mechanisms of learning

1.4.1 Sensory factors

The majority of previous behavioural research into perceptual learning has concentrated on the conditions that maximise learning outcomes, either in terms of the amount/rate of learning (Molloy et al., 2012; Seitz and Dinse, 2007; Herzog and Fahle, 1997), or the generalisability of what is learned (Ahissar et al., 2009; Brady and Kersten, 2003; Kurt and Ehret, 2010). In comparison, there has been little research into *what* is improving during learning. The most concerted efforts are found in the visual psychophysical work of Dosher, Lu and colleagues (Lu and Dosher, 2008), together with that of Gold, Bennett and Sekuler (Gold et al., 1999). These authors have assumed that learning represents an increase in sensory sensitivity, and have used external noise techniques to accredit learning to various sensitivity limiting constructs. The precise interpretation of these constructs depends on the underlying model. These models are grossly similar across research groups. In each case, the observer's sensitivity is limited by the signal-to-noise ratio, which is determined by the efficiency of the weights (or: *template*) through which the stimulus is filtered, and by the presence of any internal noise. However, subtle differences between approaches have lead to different interpretations of similar observations.

The linear amplification model Gold et al. (1999) modelled learning using a linear amplification model [LAM]. Herein, enhancement of the signal occurs via a task-specific filter which responds with a gain of β to a signal stimulus of unit magnitude. Improved performance can be achieved through more efficient tuning of the filter to the signal component(s) of the stimulus, resulting in increased signal gain. Alternatively, improved performance can also be achieved by a reduction in an additive internal noise, which is conceptualised as the aggregate of all the observer's intrinsic additive noise sources. As such, the primary distinction is between

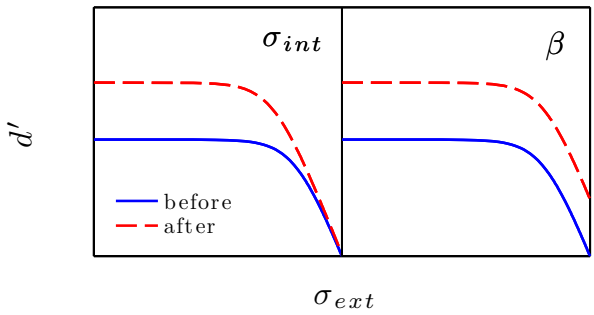


Fig. 1.7: Schematic learning profiles given the observer model of Gold et al. (1999). Training-induced improvements in internal noise (left) and signal gain (right) are shown by red dashed lines, relative to the solid blue baseline.

changes in deterministic *signal gain* process versus a random *internal noise* component. This model is formalised as:

$$d' = \frac{S}{\sigma_{all}} = \frac{\beta\Delta}{\sqrt{{\sigma_{int}}^2 + {\sigma_{ext}}^2}}, \tag{1.5}$$

where β is the gain of perceptual template to the signal stimulus, Δ is the signal magnitude, and σ_{int} and σ_{ext} are the standard deviations of the additive internal and external noise distributions, respectively.

Changes in the β and σ_{int} parameters produce distinct behavioural effects. As shown in Fig 1.7, when performance is plotted as a function of external noise, changes in the internal noise magnitude, σ_{int}, are characterised by a lateral shift in the point of inflection, with improvements in performance occurring only at relatively low levels of external noise. Conversely, changes in signal gain, β, cause the performance curves to be uniformly shifted vertically, with improvements occurring at both low and high external noise levels.

Based on empirical observations of uniform improvement across multiple external noise levels, Gold et al. (1999) concluded that *signal gain enhancement* (i.e., via filter retuning) underlies perceptual learning on various visual identification tasks (see also Gold et al., 2004).

The perceptual template model Lu and Dosher's observer perceptual template model (PTM; Lu and Dosher, 1999) differs from the LAM in three key respects. Firstly, they introduce a *multiplicative noise* term to account for behaviour resembling Weber's law. The magnitude of such noise, σ_{mul}, is proportional to magnitude of the DV. Secondly, they argue that any gain applied to the signal will apply to the external noise also, with an increase in β causing an increase in the total output of the perceptual filter. If this is the case then the processes of signal gain and internal noise reduction are

formally equivalent (Lu and Dosher, 1998). As such, Lu and Dosher (2009) do not attempt to draw a distinction, and instead term both processes *stimulus enhancement*. Thirdly, they note that the effective level of external noise can be reduced by tuning the perceptual filter to the 'appropriate time, spatial region, and/or content characteristics of the signal stimulus' (Dosher and Lu, 1998). This *external noise exclusion* mechanism, A_f, can essentially be thought of as representing the efficiency of a perceptual filter. For a narrowband signal embedded in a broadband noise this is proportional to the *width* of the filter around the region of interest. Finally, the authors also introduce a non-linear transduction constant, ζ. This is primarily designed to account for various phenomena not directly related to the question of which mechanisms underlie learning (e.g., see Gold et al., 2004), and is only included here for completeness. Performance given this model is formalised thus:

$$d' = \frac{(\beta\Delta)^{\zeta}}{\sqrt{{\sigma_{int}}^2 + (\beta\sigma_{ext}/A_f)^{2\zeta} + \sigma_{mul}^2}}, \tag{1.6}$$

where ζ is the non-linear transduction constant, A_f represents the amount of (effective) external noise attenuation, and σ_{mul} is the standard deviation of a multiplicative internal noise distribution. The value of σ_{mul} is given by the effective magnitude of the input, thus:

$$\sigma_{mul}^2 = N_{mul}^2\left[(\beta\sigma_{ext}/A_f)^{2\zeta} + (\beta\Delta)^{2\zeta}\right], \tag{1.7}$$

where N is a coefficient determining the standard deviation of the multiplicative noise, proportional to the magnitude of the stimulus (cf. Lu et al., 2000). Since in this model β is applied to σ_{ext}, an increase in β is equivalent to a reduction in σ_{int} – both serve to reduce the relative contribution of internal noise to the decision variable. The behavioural signatures of these various processes are shown in Fig 1.8. Enhancement of the signal, either via a reduction in additive internal noise magnitude or an increase in signal gain, improves performance at low external noise levels only. Improved external noise filtering produces the converse pattern of improvements at high external noise only (i.e., having no effect in quiet conditions). Finally, a uniform enhancement in sensitivity is here attributed – in contrast with Gold et al. (1999) – to a reduction in multiplicative internal noise.

The empirical observations from Lu, Dosher and colleagues (e.g., Lu et al., 2006; Lu and Dosher, 2004; Duncan, 2001) have tended to closely resemble those of Gold et al. (1999), described above. However, due

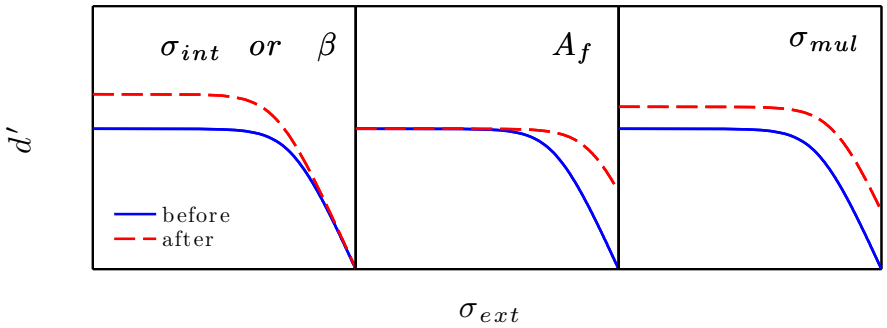


Fig. 1.8: Schematic learning profiles given the observer model of Lu and Dosher (1999). Improvements in internal noise magnitude or stimulus gain (left), external noise filtering (middle) and multiplicative internal noise magnitude (right) are shown by red dashed lines, relative to the solid blue baseline. Same format as Fig 1.7.

to the differences in the underlying model these have been consistently attributed to reductions in additive σ_{int}, together with increases in external noise exclusion. This dual-mechanism account is supported by double dissociations, with internal noise reduction (Dosher and Lu, 2006) and external noise exclusion (Lu and Dosher, 2004) having been observed independently and in isolation. It is also supported by asymmetric transfer patterns (Dosher and Lu, 2005), with improvements from training in quiet transfering to performance in noise, but not vice versa, indicating the presence of independent learning mechanisms.

Comparison with the present decision model The decision process described earlier in ∮1.2.2 incorporates several of the features from these visual learning models. Internal noise reduction is similarly modelled as a reduction in the magnitude of an additive Gaussian random variable, σ_{int}, and will similarly manifest as improvements in low external noise conditions. Improvements in sensitivity may also occur through the optimisation of channel weights, ω. However, as will be seen in Chapters 2 and 3, the precise characterisation of this process will depend on the task demands and/or how the channels are conceived. In simultaneous masking situations, channel reweighting may manifest as external noise exclusion, via a decrease in the weight given to noisy channels. By contrast, in Chapter 2 a situation is presented wherein the distribution of internal noise is independent of the channel weights, and where channel reweighting is thus equivalent to the concept of *signal gain* proposed by Gold et al. (1999). A multiplicative internal noise reduction component would be required to explain performance more generally, but was omitted here due to the narrow range of stimulus magnitudes in the tasks under consideration (see

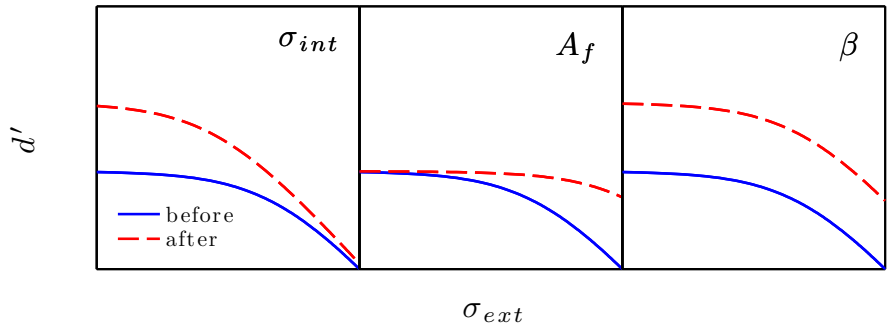


Fig. 1.9: Schematic learning profiles given the present observer model (as depicted in Fig 1.3). Training-induced improvements in internal noise (left), external noise filtering (middle) and signal enhancement (right) are shown by red dashed lines, relative to the solid blue baseline.

§1.2.2). The formalisation of this framework is therefore as follows:

$$d' = \frac{\beta\Delta}{\sqrt{{\sigma_{int}}^2 + (\sigma_{ext}/A_f)^2}} \tag{1.8}$$

where β and A_f may or may not be present, depending on the task demands and how the input channels are conceived. For completeness, the various learning profiles are shown for comparison in Fig 1.9.

From the visual perceptual learning research, detailed above, one might predict that learned improvements in auditory sensitivity would represent reductions in internal noise magnitude, σ_{int}, when the ratio of internal-to-external noise is high, and reductions in effective external noise, σ_{ext}/A_f, when $\sigma_{int}/\sigma_{ext}$ is low.

1.4.2 Non-sensory factors

The perceptual learning studies detailed in the previous section focused exclusively on factors relating to perceptual sensitivity. In contrast, there has been comparatively little research into the potential role of inefficiencies in the wider decision process, such as those relating listeners' placement of their decision criterion.

The reasons for this disparity appear to stem predominantly from four considerations. First and most directly, studies have tended to indicate that non-sensory factors such as response bias are low in naive listeners, and invariant across practice sessions (e.g., Schoups et al., 1995). Secondly, the specificity of learning to the trained stimulus (Wright et al., 1997; Fiorentini and Berardi, 1980), is sometimes taken to indicate that learning somehow represents a purely 'low-level' or 'bottom-up' process. Thirdly, multi-unit animal experiments have indicated physiological changes in

primary auditory cortex after learning (for a review see Irvine, 2007). For example, Recanzone et al. (1993) observed neural recruitment and tuning curve sharpening in the A1 of adult Owl Monkeys trained on a frequency discrimination task. Again, such data are often taken to indicate that what is learned is relatively 'low-level' or 'non-cognitive'. Fourth and finally, practical issues relating to the manner in which learning data are collected often prevent reliable indices of these constructs being derived (see Chapter 6).

However, none of these lines of evidence convincingly rule out a role for non-sensory changes during learning. As will be demonstrated, findings of low/invariant bias are neither unanimous (Chapter 5), nor necessarily veridical (Chapter 6). Similarly, observations concerning transfer and low level physiological changes are at best incomplete, and at worst misleading. They are incomplete in that partial transfer of learning has been observed on several tasks (e.g., Demany, 1985; Jeter et al., 2010, for review see Wright and Zhang, 2009), while the physiological changes have not always been consistently replicated (Brown et al., 2004) and are often small in comparison to the behavioural changes (see Gilbert et al., 2001; Petrov et al., 2005). They are potentially misleading, since neither observations of learning-specificity (see Mollon and Danilova, 1996; Petrov et al., 2005), nor physiological changes in more peripheral regions (e.g., Nakamoto et al., 2008), preclude the possibility that the underlying changes occur more centrally and/or at the level of decoding. Accordingly, some authors have postulated that refinements in the wider decision making process may have a causal role in explaining learning. This evidence is reviewed here.

Bias There exists considerable work showing how an observer's criterion may be manipulated through variations in feedback (Maddox and Bohil, 1998), by task instructions (Donaldson, 1996), or by making the relative stimulus frequencies unequal (Creelman, 1965). Similarly, it has long been known that observers are prone to shift their criterion as a function of time or experience (e.g., Kubovy and Healy, 1977). More generally, there is extensive research indicating that observers are liable to condition their responses on non-sensory factors, such as recently occurring events (Lindman and Edwards, 1961; Edwards, 1961) or their own previous responses (Green, 1964). This work has consistently indicated that naïve observers are prone to bias, and that such bias may be reduced through practice.

It is perhaps surprising then that there exists such a paucity of work relating these findings to the phenomenon of perceptual learning. One of the few studies to expressly examine the role of bias in perceptual

learning is that of Wenger and Rasche (2006). Therein, changes in response bias were measured on a visual yes/no (detection) task, as a function of practice. Across those observers who exhibited reliable improvements in performance, levels of bias were found to differ significantly across sessions (see Chapter 5 for discussion). The effect was consistent across a number of subjects/manipulations, indicating that shifts in criterion reliably occur as part of learning (see also Wenger et al., 2008). However, it remains unclear from these data what the direction of the effect is (in Wenger and Rasche, 2006, bias actually appeared to *increase* with practice), whether it also holds in auditory learning, and whether changes in bias are observed in more commonly used task paradigms, such as m-alternative forced-choice.

Inattentiveness If there has been little research into the relationship between bias and learning, then there has been virtually none into the role of inattentiveness. This is surprising for four reasons. Firstly, because there is a tendency to invoke changes in 'attentional capacity' as a possible explanation of perceptual learning (Green et al., 2003). Secondly, because IQ, which has often been shown to correlate with sustained attention, has also been found to be a predictor of learning (e.g., Schweizer and Moosbrugger, 2004, though see Finomore et al., 2009 for discussion). Thirdly, since inattentiveness often *is* thought to play a role in learning, in the form of decrements in performance observed during the latter stages of a regimen (i.e., and so could equally contribute to the increments in performance observed in the initial stages). And fourthly, because inattentiveness *has* been shown to play a role in differentiating performance amongst children (Moore et al., 2008; McArthur and Hogben, 2012; Wightman and Allen, 1992) and some clinical populations (Witton et al., 2002).

As with bias, this lack of discussion is likely to stem in part from the predominant use of experimental procedures that do not allow accurate measurements of inattentiveness to be derived (i.e., adaptive tracks, which do not typically afford reliable estimates of asymptotic performance), as well as from inertia in the underlying assumptions regarding listeners' decision making, many of which were originally developed exclusively from observations of highly practiced, and often self-selecting, individuals. The present work investigates whether these assumptions are acceptable in the case of perceptual learning, and will conclude that they are acceptable in the case of inattentiveness, but are almost certainly not in the case of bias.

1.4.3 Research Plan

To summarise, in this chapter a theoretical framework has been presented by which performance on psychoacoustical tasks may be seen to be limited by a variety of potential factors: internal noise magnitude, encoding efficiency, bias, and inattentiveness). Changes in any of these factors, either singularly or in combination, may underlie the improvements in performance observed in auditory perceptual learning. In the succeeding chapters behavioural models are used to evaluate the importance of each factor as a function of experience, given a variety of basic psychoacoustical tasks. Chapters 2, 3 and 4 examine in particular the relative importance of sensory factors, and complement recent similar work in the visual literature. Chapters 5 and 6 consider in more detail the possible impact of non-sensory on learning, and make a novel case for a substantive role of bias-reduction in perceptual learning. The overall conclusions and implications are discussed in Chapter 7.

CHAPTER 2

Pure tone discrimination learning

This chapter examines what mechanisms underlie auditory perceptual learning. Fifteen normal hearing adults performed two-alternative, forced choice, pure tone frequency discrimination for four sessions. External variability was introduced by adding a zero-mean Gaussian random variable to the frequency of each tone. Measures of internal noise, decision efficiency, bias, and inattentiveness were derived using four methods (model fit, classification boundary, psychometric function, and double-pass consistency). The four methods gave convergent estimates of internal noise, which was found to decrease from 4.52 *Hz to* 2.93 *Hz (S.D. of decision variable) with practice. No group-mean changes in encoding efficiency, bias or inattentiveness were observed. It is concluded that learned improvements in frequency discrimination primarily reflect a reduction in internal noise. Data from highly experienced listeners and neural networks performing the same task are also reported. These results also indicated that auditory learning represents internal noise reduction, potentially through the re-weighting of frequency-specific channels.*

2.1 Introduction

PERCEPTUAL learning is improved performance on a perceptual judgment task as a result of practice. While the phenomenon is well established, little is known about the mechanisms underlying such improvements. In the visual literature it has been variously suggested that reductions in internal noise (Dosher and Lu, 1998) or improvements in encoding efficiency (Gold et al., 1999) may underlie learning. In this paper we examine whether either of these factors change during auditory (frequency discrimination) learning. We also examine two further potential limiting factors that have not previously been considered: response bias and attentiveness.

Internal noise is uncertainty in the internal response to an input signal which, in contrast with external noise, is generated by sources intrinsic to the observer. Internal noise is therefore synonymous with intrinsic variability, and the two terms may be used interchangeably. Internal noise is fundamental to Signal Detection Theory (SDT; Green and Swets, 1974; Macmillan and Creelman, 2005). It is also a prominent concept in psychophysics (Gescheider, 1997; Klein, 2001), where the ogival psychometric function is theoretically justified as the cumulative form of a random variable with a bell-shaped distribution. Potential sources of internal noise include non-deterministic transduction (e.g., due to Brownian motion of hair cells; Denk et al., 1989), stochastic neural encoding and transmission both in the auditory periphery (Javel and Viemeister, 2000) and more centrally (e.g., Vogels et al., 1989), and physiological maskers such as heartbeats and blood flow (Soderquist and Lindsey, 1971).

Over the last 50 years a number of measures of internal noise have been developed. These include external noise titration (Lu and Dosher, 2008), model-fitting (Jesteadt et al., 2003), n-pass consistency (Green, 1964), multiple-looks (Swets, 1959), and direct variability estimates derived from distributions of errors (e.g., Buss et al., 2009). Following related work in the visual literature (e.g., Gold et al., 1999), we here utilised the model-fitting and double-pass consistency techniques. In addition, we also considered two direct variability estimates which were derived using the same data.

In contrast with internal noise, encoding efficiency constitutes a systematic rather than random limitation on performance (cf. Berg, 2004; Berg and Green, 1990). In sensory tasks, encoding efficiency primarily describes how well the listener is able to selectively integrate information across channels. How these channels are conceived depends on the task. For example, in spectral profile analysis, listeners must detect when the levels of one or more components of a multitone stimulus are changed. In such a task, if the frequency components are widely spaced then every frequency component in the complex can be considered a channel, and a good strategy would be to attend predominantly to the difference between the signal channel and the average of the non-signal channels. In the present study, each interval in a two-interval, forced-choice paradigm is considered to be a channel, with similar quantities of internal noise in both channels. In this case a good strategy would be to attend equally to both intervals. Encoding efficiency can either be inferred by comparing observed sensitivity to the ideal (e.g., Berg and Green, 1990; Tanner and Birdsall, 1958), or by comparing a listener's estimated strategy to the ideal (e.g., Dai and

Berg, 1992; Alexander and Lutfi, 2004). Here we used variations on both these approaches. Signal detection theory was used to derive a model containing an encoding efficiency parameter which was fitted to observed performance, while a novel classification boundary approach was used to estimate listeners' encoding strategies.

Response bias (*hereafter*: bias) is the tendency to favor one response over another, irrespective of the stimulus features. Thus, a listener who is biased towards one alternative may select it even when the sensory evidence makes it more likely that the other is true. Psychometric thresholds are liable to be negatively affected by such bias, unless either explicit corrections are made or metrics such as d' used that are designed to partial out these effects. Indices of response bias can be derived from lateral shifts in psychometric functions (Gescheider, 1997), or by using SDT to calculate the distance of the listener's criterion from the ideal (Macmillan and Creelman, 2005).

Inattentiveness is the complement of sustained attention. It expresses the fact that on a proportion of trials listeners appear to respond independently of the sensory information, possibly reflecting a lapse in concentration. For simplicity, it is common to assume that inattention is a binary process that occurs independently of the stimulus level or trial number (cf. Viemeister and Schlauch, 1992). Historically, inattentiveness has been little studied relative to the other limitations described here. This may in part be because inattentiveness is specifically selected against in many psychophysical experiments (which tend to be populated by highly experienced, reliable and well-motivated observers). Nonetheless, a number of behaviours have been identified from which metrics of inattention may be derived, such as the amount and/or profile of excursions from threshold in an adaptive track (Moore et al., 2008), or asymptotic performance on the psychometric function (Green, 1995).

In this study, we investigated the extent to which each of these mechanisms (internal noise; encoding efficiency; response bias; inattentiveness) contribute to auditory perceptual learning. The task was two-interval, two-alternative, forced-choice (2I2AFC) frequency discrimination in which the frequency of both tones was jittered by adding Gaussian noise. Frequency discrimination was selected due to both its prevalence in the learning literature (e.g., Hawkey et al., 2004; Demany, 1985) and its robust tendency to improve with practice relative to other psychoacoustic tasks (cf. Wright and Zhang, 2009). Jitter was used to introduce an external noise component as a reference for internal noise magnitude. On simple auditory tasks requiring judgements based on pure tone stimuli, the limiting factor in performance is often suggested to be internal noise (e.g., Houtsma,

1995; Durlach and Braida, 1969). If this is the case during learning, then the magnitude of internal noise should decrease as a function of practice, concomitant with improved discriminability. Conversely though, there has been a tendency in the visual literature to conclude that changes in encoding efficiency underlie learning (e.g., Gold et al., 1999, 2004; Chung et al., 2005, though see Lu and Dosher, 2009). If auditory perceptual learning is analogous to visual perceptual learning then we might expect predominant changes in encoding efficiency. There has been comparatively little research into response bias and inattentiveness during learning. We therefore made no predictions as to their prevalence or whether they would change with practice.

2.2 General Methods

2.2.1 Stimuli & Apparatus

The stimuli in all conditions were 300 ms (including ramp) sinusoids, gated on/off by 10 ms cos^2 ramps and presented at 70 dB SPL. Stimuli were digitally synthesised in Matlab v7.4 (2007a, The MathWorks, Natick, MA) using a sampling rate of 44.1 kHz and 24-bit quantisation. Digital-to-analog conversion was carried out by a PCI sound card (Darla Echo; Echo Digital Audio Corporation, Carpinteria, CA), interfaced via the Psychophysics Toolbox v3 (Brainard, 1997) ASIO wrapper (Steinberg Media Technologies, Hamburg). Stimuli were presented diotically via Sennheiser HD 25-I headphones. Participants were tested individually in a double-walled sound-attenuating booth. They responded by pressing one of two buttons on a button box. Visual fixation cues and feedback were presented on an LCD monitor.

2.2.2 Procedure

The task was 2I2AFC frequency discrimination, for which participants were asked to "pick the higher-pitched tone". Each trial commenced with a 400 ms warning interval during which a visual fixation cross was displayed, followed by two 300 ms tones separated by a 400 ms interstimulus interval. On each trial a pair of tones was sampled in random order from a pair of Gaussian distributions[1] with a common standard deviation of σ_{Hz} and randomly ordered means of $1000 \pm \Delta_{Hz}$ (Fig. 2.1). Thus in each trial samples were drawn in random order from two distributions symmetric about 1 kHz, and the target sample was that drawn from the distribution of mean $1000 + \sigma_{Hz}$. Participants were given an unlimited time to respond, after which visual feedback was presented for 400 ms prior to the next trial onset.

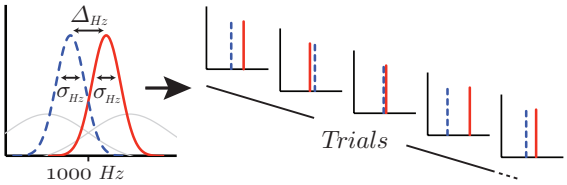


Fig. 2.1: Stimulus schema for a single external noise condition. The dashed and solid distributions are the jittered 'low' and 'high' tone distributions, respectively. On each trial a tone was independently drawn from each distribution in random order (example values for the first five trials are shown on the right). The difference in hertz between the means of the two normal distributions, Δ_{Hz}, was determined by the frequency-difference condition, which was fixed within each block. The common standard deviation of the two distributions, σ_{Hz}, was set so as: $\sigma_{Hz} = \Delta_{Hz}/2$. The frequency-difference condition was fixed within a block. An example pair of distributions corresponding to a greater Δ_{Hz} condition is shown in light gray hairlines.

The standard deviation of the jitter, σ_{Hz}, took on the values 0.5, 1.5, 2.5, 3.5, 4.5 and 5.5 Hz. This range of values was chosen to accommodate the most likely magnitude of internal noise based on pilot data. In keeping with Jesteadt et al. (2003), the separation between distributions, Δ_{Hz}, was co-varied along with the amount of jitter, σ_{Hz}, such that $\Delta_{Hz} = 2\sigma_{Hz}$. The overlap between distributions was therefore constant across all six conditions and resulted in an invariant d'_{ideal} of 2.0 (i.e., the ideal listener would be expected to score $\sim 92\%$ correct in all conditions).

Feedback was given, in the form of a visually presented 'happy' or a 'sad' smiley face. For the purposes of determining feedback (though not for scoring), the actual frequencies presented were used to calculate if the listener's response was correct. This was done in order to reinforce the optimal response behaviour of responding to the higher frequency tone, and to discourage the use of non-stimulus driven strategies. Additional feedback was presented at the end of each block in the form of a percentage score, again based on the frequencies of sounds presented (tones + noise) rather than on their values prior to jittering.

Each test block consisted of 50 trials drawn from one of the six frequency difference conditions, Δ_{Hz}. Each session consisted of 32 test blocks, presented in pseudorandom order. Thus conditions were fixed within each block, and randomly ordered between blocks. The number of trials per session (1600) was large given typical frequency-discrimination learning rates (e.g., Molloy et al., 2012), but is consistent with the slower learning observed when the training stimuli are randomly varied (Amitay et al., 2005).

The test blocks in the first session were preceded by two short practice blocks consisting of 10 'easy' (150 Hz difference) and 10 'difficult' (8 Hz

difference) trials, intended to familiarise participants with the procedure. In blocks one to 24, each frequency difference was tested four times in pseudorandom order. These 1200 trials were used in the model fit analysis (see below). In the final eight blocks, all the previous blocks from the narrowest ($\sigma_{Hz} = 0.5; \Delta_{Hz} = 1$) and broadest ($\sigma_{Hz} = 5.5; \Delta_{Hz} = 11$) frequency differences were repeated in pseudorandom order. These last 400 trials were used in the double-pass consistency analysis. They were identical to the trials heard earlier in the experiment, although the order of the trials within each block was randomised in order to avoid the potential confound of response dependencies on consistency (for discussion see Levi et al., 2005; Spiegel and Green, 1981). None of the listeners reported, when questioned, being aware of the fact that the last eight blocks consisted of repetitions of earlier trials. All 1600 trials were used to carry out the psychometric function and classification boundary analyses. Sessions lasted approximately 80-90 minutes in total, including two rest breaks. All listeners took part in one session per day for four consecutive days.

2.2.3 Analysing Learning

Learning was assessed by examining sensitivity as a function of session. For each stimulus condition, successive pairs of test blocks were concatenated to yield blocks of 100 trials. Each analysis block was then used independently to derive estimates of sensitivity, d', and response criterion, λ, as per Wickens (2002). In two blocks participants responded 100% correctly to one interval. In these two cases, the number of correct responses was adjusted by 0.5 to yield a defined d' value (Macmillan and Creelman, 2005).

2.2.4 modelling behaviour

Measures of internal noise, encoding efficiency, bias and inattentiveness were derived using four methods of analysis: model fit, classification boundary, psychometric function, double-pass consistency. Although all related, each method differs in terms of its precise derivations, assumptions, and how it partitions performance into various limiting parameters. The use of multiple methods allowed for constructs common across methods (e.g., internal noise) to be cross-validated, and for a greater range of constructs to be examined. Example individual data for a single listener derived using each method are shown in Fig. 2.2 (n.b. there is no graphical analogue to the double-pass method). Each panel is discussed in the context of its associated methodology.

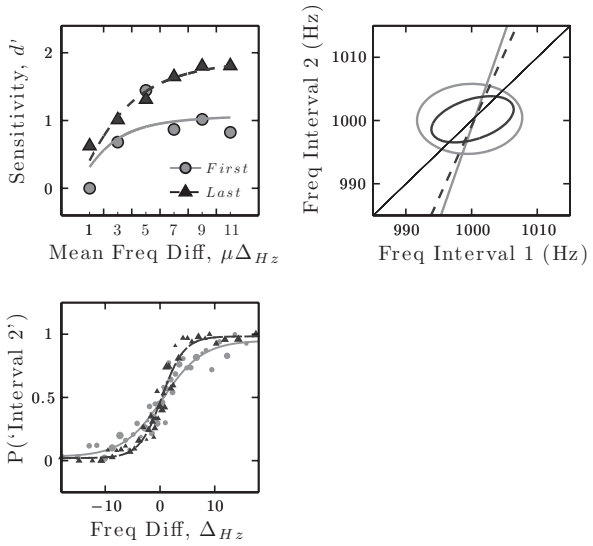


Fig. 2.2: Individual model fits for a single listener; first and last session only. (Top-left) Model-fits to observed sensitivities. Curves represent least-square fits to Eq. (2.1), from which internal noise and encoding efficiency parameters are derived. (Top-right) Estimated classification boundaries (solid lines) and standard deviations of errors with respect to their boundaries (ellipses). Smaller ellipses indicate less internal noise, while a classification boundary closer to the identity function indicates a more efficient encoding strategy. (Bottom-left) Cumulative Gaussian psychometric fits to Eq. (2.2). The proportion of 'Interval 2' responses are given as a function of frequency difference ($Freq_2 - Freq_1$), post-jittering. A steeper slope indicates less internal noise, while asymptotic performance closer to the upper/lower bounds (0 and 1) indicates more attentiveness. (See body text for further details.)

Model fit

Encoding efficiency, η (cf. Berg, 2004), and the standard deviation of a zero-mean Gaussian internal noise, σ_{Int}, were calculated by fitting observed sensitivities to the model:

$$d' = \frac{\eta \cdot \Delta_{Hz}}{\sqrt{\sigma_{Int}^2 + \sigma_{Hz}^2}}, \tag{2.1}$$

where Δ_{Hz} and σ_{Hz} represent the mean separation and the common deviations of the stimulus distributions, respectively. This model represents a version of that described previously by Jesteadt et al. (2003), extended to include an additional encoding efficiency parameter that reflects any deterministic limitations on performance arising from the listener's encoding strategy. The derivation of Eq. (2.1) is given in Appendix 2.A.

As shown in Fig. 2.2 (top-left), least-squares fits to Eq. (2.1) were made to observed sensitivities. These fits were constrained by transformation to yield finite and positive parameter values. Fits were made independently to each set of 600 trials (two blocks from each condition), yielding two

estimates of internal noise and encoding efficiency per listener, per session. These estimates were averaged to provide a single value for comparison with the other three measures.

Classification boundary

The listener's task in 2I2AFC frequency discrimination can be conceptualised as a binary classification problem. As shown in Fig. 2.2 (top-right), the decision space is two-dimensional, with each axis corresponding to the frequency in a given interval. The target variable is the interval containing the higher tone (either 'Interval 1' or 'Interval 2'). When Interval 1 is plotted on the abscissa, the data points belonging to class 'Interval 1' will be below the identity function, while class 'Interval 2' points will be above the identity function. Since the stimulus distributions are arranged symmetrically around 1 kHz, the ideal classification boundary will have a slope of one and pass through the origin. Alternatively, less optimal strategies may be employed. For example, the listener shown in Fig. 2.2 gives disproportionate weight to interval 1 in both session one and (to a lesser extent) in session four.

Each listener's classification boundary was estimated by finding the linear function that best predicts their responses given the presented frequencies (i.e., after the addition of external noise). The angle from the observed slope to the ideal was taken as an index of encoding efficiency, η. The spread of misclassifications given this boundary was interpreted as an index of internal noise magnitude, σ_{Int}. Spread was computed as the standard deviations of 2-D Gaussians fitted to errors (shown by the ellipses in Fig. 2.2). The Euclidean distance of the classification boundary from the point of physical equality $\{1000, 1000\}$ was interpreted as interval response bias, CE.

Linear discriminant analysis was used to fit classification boundaries to the data from each session[2] (1600 trials per fit). This yielded one estimate of internal noise, encoding efficiency and bias per listener, per session.

Psychometric function

Psychometric functions were estimated by maximum likelihood fits to the function:

$$P(\text{'Int 2'}) = \gamma_{lo} + (\gamma_{up} - \gamma_{lo})\Phi(x; \mu, \sigma), \tag{2.2}$$

where $P(\text{'Int 2'})$ is the proportion of 'Interval 2' responses, γ_{lo} and γ_{up} are lower and upper asymptotes, and $\Phi(x; \mu, \sigma)$ is the Gaussian cumulative distribution function with mean μ and standard deviation σ, evaluated at

the values x. In our task, x is the linear difference in frequency between the two intervals, with a positive value representing a higher frequency in the second interval. When fitting psychometric functions, some authors additionally include a variable exponent term, which introduces a potential non-linearity to the slope of the sigmoid (e.g., Dai and Micheyl, 2011; Dai and Richards, 2011). Such a term did not substantively effect the present findings, and so was omitted (see Appendix 2.B).

The fitted value of σ was taken as a measure of internal noise. The psychometric function was also used to derive two additional measures: response bias and inattentiveness. Response bias was indexed by constant error (CE): the estimated point of subjective equality, $\hat{\mu}$, minus the point of physical equality on the psychometric function. Inattention was modelled as a stationary, stochastic process by which listeners, on some proportion of trials K, respond independently of the sensory evidence. Following (Green, 1995, see also Wightman and Allen, 1992), K was derived from the estimated asymptote values, thus:

$$K = 1 - \gamma_{up} + \gamma_{lo}. \tag{2.3}$$

The main caveat with this approach as a measure of internal noise is that the psychometric function confounds random and deterministic limitations on performance, the latter of which are inconsistent with the notion of noise as random variability (Green, 1964). In the limit, a listener who attends only to uninformative channels will have a slope of zero. Changes in the gradient of the psychophysical slope are therefore ambiguous. They may reflect either more variability in the decision variable, or a less efficient strategy, or a mixture of both. This ambiguity can be resolved either by assuming that the encoding strategy is ideal (e.g., Glasberg et al., 2001; Tanner, 1958), or by estimating the listener's encoding strategy and making fits to the actual, trial-by-trial decision variable, thereby partialling out any systematic performance limitations (e.g., Berg, 2004). In the present work we assumed that the encoding strategy is ideal. However, in doing so we acknowledge that the resultant value will be an upperbound on internal noise magnitude. The extent that this value approximates the true value will depend on the efficiency of the encoding strategy. This will be indicated both by the model-fit analysis and the classification boundary analysis.

Psychometric functions were fitted using the 'psignifit' Matlab toolbox (v2.5.6), which implements the maximum-likelihood method described by Wichmann and Hill (2001). As shown in Fig. 2.2 (bottom-left), fits were made independently for each session, using all 1600 trials. This yielded

one estimate of internal noise, inattentiveness and bias per listener, per session.

Double pass consistency

The central tenet of the n-pass consistency technique (Green, 1964; Spiegel and Green, 1981) is that when the same stimulus is presented multiple times, the probability of agreement between each of the listener's responses is determined by the ratio of internal to external noise (n.b. the magnitude of external noise should not affect response consistency, since the noise sample is frozen across repetitions. Thus even in high external noise conditions, the response consistency of an ideal observer would be 100%). The mathematics of this is expounded by (Lu and Dosher, 2008, see also Burgess and Colborne, 1988), who show that, assuming a normally distributed internal noise drawn independently on each observation, the probability of two answers agreeing, P_A, is determined solely by the ratio of internal-to-external noise, α, together with the stimulus-determined parameters (Δ_{Hz}, σ_{Hz}):

$$P_A = \int \phi(x - \Delta_{Hz}; 0, \sqrt{2}\sigma_{Hz}) \Big\{ \Phi^2(x; 0, \sqrt{2}\alpha\sigma_{Hz}) \left[1 - \Phi(x; 0, \sqrt{2}\alpha\sigma_{Hz})\right]^2 \Big\} \, \mathrm{d}x, \tag{2.4}$$

where $\phi(x; 0, \sigma)$ is a Gaussian random variable with mean 0 and standard deviation σ, and $\Phi(x; 0, \sigma)$ is its cumulative distribution function. This equation states that the probability of agreement can be computed from the probability of the same response occurring twice for a given signal, weighted by the probability of that a signal of at least that magnitude occurring. In turn, the probability of the same response occurring twice is the probability of a greater 'Interval 1' internal response occurring on the first pass (cf. Fig 1.3), multiplied by the probability of a greater 'Interval 1' internal response occurring on the second pass (which, assuming independent, identically distributed noise, is the square of either probability considered singularly), additively combined with the analogous product of the corresponding 'Interval 2' probabilities.

Consistency was examined independently for each session, and separately for the low and high external noise conditions. Specifically, a subset of the trials were presented in a two-pass manner to allow for double pass consistency (DPC) to be estimated. Response consistency was calculated as the proportion of trials where the listener responded the same way across both presentations, irrespective of whether the response was correct. The consistency score was then used to derive estimates of internal noise

by numerically solving Eq. (2.4). This yielded two estimates of internal noise and encoding efficiency per listener, per session (i.e., one each for the lowest and highest external noise conditions). However, performance was so low in the hardest condition ($\sigma_{Hz} = 0.5; \Delta_{Hz} = 1$) that it appeared that some listeners were not able to maintain a stable criterion. Thus, only the internal noise estimates from the high frequency-difference condition ($\sigma_{Hz} = 5.5; \Delta_{Hz} = 11$) are reported here.

2.3 Experiment I: Learning in naïve listeners

2.3.1 Listeners

Sixteen listeners participated, none of whom had any prior experience of auditory psychophysics. Eleven were female (mean age 22.3), five were male (mean age 25.3). All had normal hearing, as assessed by audiometric screening administered in accordance with the BSA standard procedure ($\leq$ 20 dB HL or less bilaterally at 0.5 kHz to 4 kHz octaves; British Society of Audiology, 2004). Listeners were not screened based on initial task performance, were recruited through advertisements placed around Nottingham University campus, and received an inconvenience allowance for their time. The study was conducted in accordance with Nottingham University Hospitals Research Ethics Committee approval and informed written consent was obtained from all participants.

One listener was excluded from all analyses due to performing at chance in all conditions throughout all four sessions. Two additional listeners were not included in the double-pass analysis due to a technical error.

2.3.2 Results

Learning

Grand mean performance (d') for listeners across sessions is shown for each stimulus condition in Fig. 2.3. Sensitivity increased as a function of session [$F(3, 42) = 16.7, p < 0.001, \eta_p^2 = 0.54$], indicating improvement with practice. There was no significant interaction between session and condition [$F(15, 210) = 1.3, p = 0.21$], indicating that learning occurred irrespective of external noise condition. Response criterion (λ) did not change across sessions [$F(3, 42) = 1.3, p = 0.30$]. There was substantial variability in performance between listeners, with d' ranging by approximately one unit within each session. There was also a large degree of variability in learning, with changes in mean sensitivity, $\Delta d'$, varying from -0.04 to 0.92 across listeners.

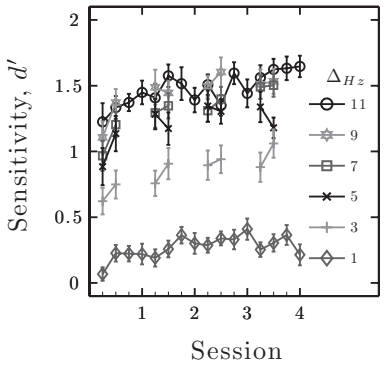


Fig. 2.3: Frequency discrimination learning. Each point shows group-mean sensitivity, d', as a function of session, averaged over all 15 listeners. Error bars represent $\pm$ 1 standard error of the mean (S.E.M.), both here and in all subsequent figures. Each stimulus condition is shown separately. The breaks between data points in conditions $\Delta_{Hz} = \{3 - 9\}$ are due to the fact that blocks from these conditions were not repeated at the end of each session (i.e., when assessing consistency).

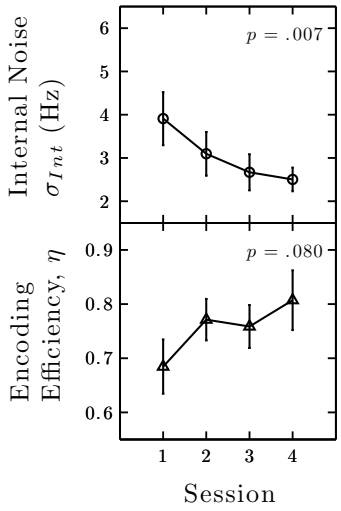


Fig. 2.4: Changes in model fit parameter estimates with practice. (Top) Group mean internal noise, σ_{Int}, as a function of session. (Bottom) Group mean encoding efficiency, η, as a function of session. In each panel, the main effect p value from the associated repeated measures ANOVAs are shown top-right; see body text for details.

Model fit

Best fits were made to the model given in Eq. (2.1). Figure 2.4 shows the group mean values of internal noise (σ_{Int}) and encoding efficiency (η). Internal noise estimates decreased significantly across sessions [$F(3, 42) = 4.7, p = 0.007, \eta_p^2 = 0.25$]. There was a non-significant trend towards an improvement in encoding efficiency, with improvements observed in 11 of 15 listeners [$F(3, 42) = 2.4, p = 0.08$]. Goodness-of-fit improved throughout the study, with median $r^2 = 0.53$ in session one increasing to $r^2 = 0.63$ in session four.

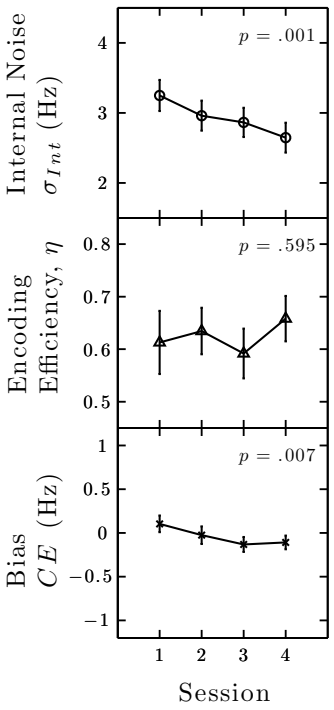


Fig. 2.5: Changes in classification-boundary parameter estimates with practice. Panels show the following group mean values as a function of session: (Top) Standard deviation of errors (given an estimated classification boundary) as a measure of internal noise; (Middle) Distance of the boundary slope from the ideal, as a measure of encoding efficiency; and (Bottom) CE as a measure of bias (a negative CE value indicates an'Interval 1' response preference). This figure follows the same format as Fig. 2.4, with which the internal noise estimates are directly comparable.

Classification boundary

Group mean values of internal noise (σ_{Int}), encoding efficiency(η) and bias (CE), as derived using the classification boundary technique, are given as a function of session in Fig. 2.5. Internal noise estimates decreased significantly across sessions [$F(3, 42) = 6.9, p < 0.001, \eta_p^2 = 0.33$]. No change in encoding efficiency was observed [$F(3, 42) = 0.6, p = 0.60$]. Bias did significantly change over sessions [$F(3, 42) = 4.6, p = 0.007, \eta_p^2 = 0.25$], with listeners tending to favour Interval 2 in session one ($CE = 0.10$), and Interval 1 in session four ($CE = -0.11$), though none of the session means significantly differed from 0 (no bias) [Hotelling's $T^2; T^2(4, 11) = 13.2, p = 0.10$].

Psychometric function

Psychometric function fits were made to Eq. (2.2). [Mean goodness-of-fit: $r^2 = 0.87$]. The slope of the function (internal noise) became steeper in 87% of listeners. There was little change in lower or upper asymptote (inattention) or in constant error (bias). Group mean values of internal

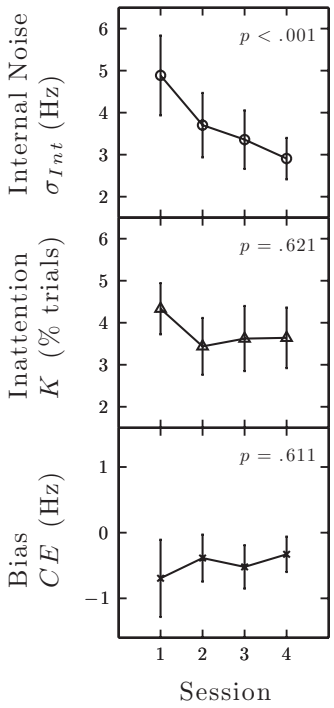


Fig. 2.6: Changes in psychometric function parameter estimates with practice. Panels show the following group mean values as a function of session: (Top) Fitted Gaussian standard deviation as a measure of internal noise, σ_{Int}; (Middle) Inattentiveness (derived from asymptotic performance), K, as a measure of sustained attention; (Bottom) CE as a measure of bias.

noise (σ_{Int}), inattention (K) and bias (CE) are given as a function of session in Fig. 2.6. Internal noise estimates decreased significantly across sessions [$F(3, 42) = 8.2, p < 0.001, \eta_p^2 = .37$]. No changes in inattention [$F(3, 42) = 0.60, p = 0.62$] or bias [$F(3, 42) = 0.68, p = 0.57$] were observed, with mean bias remaining indistinguishable from 0 throughout [$T^2(4, 11) = 2.9, p = 0.69$].

Double pass consistency

Group mean values of internal noise (σ_{Int}) as derived using the DPC technique are given as a function of session in Fig. 2.7. Internal noise estimates decreased significantly across sessions [$F(3, 36) = 9.9, p < 0.001, \eta_p^2 = .45$].

Comparison of metrics

As shown in Table 2.1, correlations between the four sets of internal noise estimates were strong [$r \geq 0.69$; all $p < .001$]. Positive correlations were also observed between the bias estimates from the classification boundary and psychometric fit approaches [$r = 0.63; p < 0.001$], and between the encoding efficiency estimates from the model fit and classification boundary

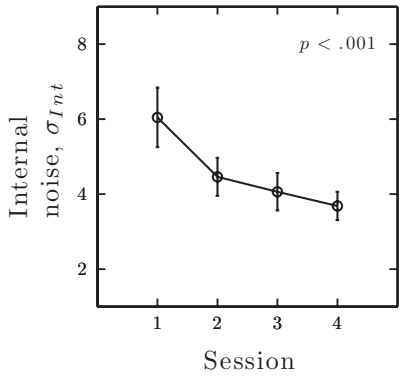


Fig. 2.7: Changes in double-pass internal noise magnitude estimates with practice. Each point shows group mean internal noise, σ_{Int}, as a function of session, estimated using the double-pass consistency method.

	MF	CB	PF
DPC	0.68	0.81	0.82
PF	0.80	0.82	-
CB	0.62	-	-

Table 2.1: Correlation coefficients, r, between internal noise estimates, σ_{Int}, from the model fit (MF), classification boundary (CB), psychometric function (PF) and double-pass consistency (DPC) methods.

measures [$r = 0.37; p = 0.004$]. Individual internal noise estimates for the first and last sessions are given for each test in Table 2.2. The double-pass consistency method tended to produce the somewhat larger estimates, being the greatest of the four in 88% of cases. Conversely, the model fit and classification boundary methods tended to produce the smallest noise estimates.

2.3.3 Discussion

Frequency discrimination sensitivity improved significantly with practice, although there was substantial individual variability in both performance and learning. Improvements in sensitivity were accompanied by a significant decrease in internal noise with little change in encoding efficiency, bias and inattentiveness. The results show that practice-induced improvements in frequency discrimination sensitivity primarily represent a reduction in internal noise. Averaged over the four methods, mean internal noise values ranged from 3.2 to 6.0 Hz in session one, and 2.5 to 2.9 Hz in session four.

The four methods yielded highly correlated estimates of internal noise. Notably, since encoding efficiency was less than ideal, the internal noise

	Session 1				Session 4			
Listener	MF	CB	PF	DPC	MF	CB	PF	DPC
L1	3.0	3.1	3.2	5.3	1.3	2.0	2.0	2.9
L2	4.6	3.2	4.5	6.1	4.8	2.9	4.0	4.5
L3	2.3	2.5	3.3	5.2	2.5	2.2	2.4	3.3
L4	1.8	2.2	2.2	3.2	2.0	2.1	1.9	3.2
L5	1.4	3.4	3.7	6.1	2.1	2.3	2.0	3.2
L6	5.0	4.1	6.1	11.6	4.5	3.7	3.9	7.4
L7	1.8	2.1	2.0	2.9	1.4	1.1	1.5	1.8
L8	2.7	2.3	2.0	n.a.	2.7	3.5	2.6	n.a.
L9	10.3	5.1	15.8	n.a.	2.0	4.8	9.4	n.a.
L10	5.1	2.9	4.1	5.0	3.3	2.0	2.7	3.0
L11	2.3	2.3	2.6	3.6	1.5	2.1	2.0	2.6
L12	5.3	4.8	10.2	11.9	1.6	3.0	1.9	4.5
L13	2.6	2.7	3.0	4.0	2.8	2.0	2.6	4.4
L14	6.8	3.5	5.5	6.9	3.1	2.8	2.6	3.2
L15	3.5	3.2	5.1	6.7	2.0	1.5	2.1	4.0

Table 2.2: Summary of internal noise results, σ_{Int}, for individual listeners during the first and last session. Initialisms follow the same format as Table 2.1.

estimates from psychometric functions tended to be consistently greater than with the model-fit and classification boundary methods. However, encoding efficiency remained largely invariant throughout. The changes in internal noise observed using psychometric functions therefore remained robust.

2.4 Experiment II: Experienced listeners

Group mean performance in our naïve listeners (Experiment I) failed to asymptote after four sessions. It may therefore be that sensitivity could be further improved with additional training. It may also be that any such additional learning is limited by factors other than internal noise. To assess these possibilities, two further listeners (not tested in Experiment I) with extensive prior task experience (one of whom was the first author) were tested using the same stimuli.

Furthermore, a potential concern with the methodology of Experiment I is that the external noise (introduced via jittering) may not have been independent of listeners' internal noise. Thus, by randomly varying the stimuli on a trial-by-trial basis, additional variability may have been introduced into listeners' decisions not normally present during traditional (unjittered) frequency discrimination. This could be the case if, for example,

listeners were prone to update their criterion after each trial in a manner contingent upon only the most recent few trials. The two experienced listeners were therefore also tested using unjittered stimuli. Psychometric functions fitted to 'zero noise' data were compared to those derived under jittering. Greater internal noise would be indicated by shallower slopes in the jittered condition.

2.4.1 Methods

The stimuli followed those described in Experiment I, except that all σ_{Hz} and Δ_{Hz} values were halved. This adjustment was necessary since these listeners performed at ceiling when $\Delta_{Hz} > 5$ Hz. Both listeners performed 3 practice sessions, followed by 9 test sessions over two weeks. Each session consisted of 12 blocks, equivalent to the first phase of the session in the main experiment. Listeners then performed 3 additional test sessions in which no external noise was added ($\sigma_{Hz} = 0$).

2.4.2 Results & Discussion

Performance and model estimates

The results of two experienced listeners are summarised in Table 2.3, along with the group-mean data from the final training session of Experiment I for comparison. Given the amount of prior task experience no improvement in sensitivity was expected across test sessions, and none was observed [$F(7) \leq 2.3, p \geq 0.176$]. Because of the different stimulus conditions, d' values were not comparable between experiments. As such, performance was quantified as the mean of listeners' discrimination limens at the 75% and 25% correct levels, FDL_{Hz}.

Both listeners' frequency discrimination limens were significantly lower than in the post-training naïve listeners [$t(14) \geq 4.5, p < 0.001$], indicating that further learning beyond that observed in Experiment I is possible. As per Experiment I, the model fit and psychometric fit techniques were used to estimate internal noise, encoding efficiency, inattention and bias. The pattern of results continued the learning trend observed in Experiment I. Internal noise magnitude was further decreased [$t(14) \geq 4.1, p \leq 0.001$], with no differences in encoding efficiency [$t(14) \leq 0.2, p \geq 0.828$] or bias [$t(14) \leq 0.5, p \geq 0.632$]. This finding corroborates our conclusion that changes in internal noise underlie frequency discrimination learning. Inattentiveness was also lower than the naïve group-mean [$t(14) \geq 4.9, p < 0.001$], suggesting that very highly trained listeners may also benefit from improved sustained attention.

		Model fit		Psychometric fit		
Listener	FDL_{Hz}	σ	η	σ	K	CE
μNaïve	2.8	3.0	0.8	2.9	0.036	-0.330
KM	0.8	0.7	0.8	0.8	0.005	< 0.001
PJ	1.3	1.1	0.8	1.2	0.001	0.016

Table 2.3: Summary of frequency difference limens (FDL) in Hz, and fitted behavioural parameters for group-mean naive listeners (final session) and the experienced listeners KM and PJ. Fitted parameters were internal noise (σ) and encoding efficiency (η), estimated using the model fit; and internal noise (σ), inattentiveness (K) and bias (CE), estimated from psychometric functions

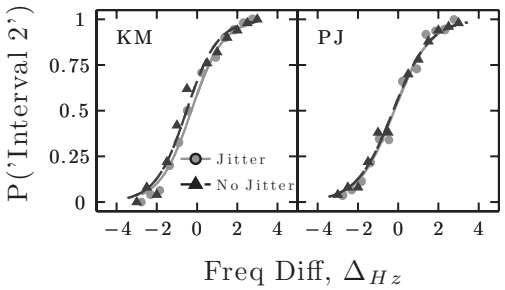


Fig. 2.8: Psychometric functions for Experiment II. Black triangles and dashed-lines indicate raw data and psychometric fits (respectively) given non-jittered stimuli. Gray circles and lines indicate analogous binned raw data and psychometric fits given jittered stimuli. In both cases fits were made to Eq. (2.2).

Internal noise with and without external noise

Figure 2.8 shows psychometric functions with and without external noise. Performance in the two cases was virtually indistinguishable. In one listener (PJ) estimated internal noise was marginally (0.1 Hz) smaller, while in KM estimated internal noise was marginally (0.2 Hz) greater. These results indicate that the use of jittering did not artificially inflate the internal noise estimates, either here or in Experiment I. These results are consistent with Jesteadt et al. (2003), who also observed good agreement between estimates of internal noise derived under jittering, and the slope of a psychometric function fitted to data without external variability.

2.5 Experiment III: Simulations

It has been suggested in the visual literature that perceptual learning represents "re-weighting of stable early sensory representations" (Lu and Dosher, 2009; Mollon and Danilova, 1996). Although we found no evidence of channel re-weighting at the behavioural level (where each stimulus presentation interval was modelled as a channel), our data are consistent with a process of iterative re-weighting of channels at a neural

level of description. Such channel re-weighting has been suggested to occur in visual learning (Law and Gold, 2009; Petrov et al., 2005), and is a plausible explanation for learning on a frequency discrimination task, given that psychophysical thresholds are substantially poorer than would be predicted from the precision of information encoded at the periphery (e.g., Siebert, 1970; Heinz et al., 2001). To investigate whether a process of early sensory re-weighting can produce the observed pattern of learning, a simple neural network model was trained and analysed using the same methods as the human listeners.

2.5.1 Methods

The neural network consisted of a single-layer perceptron (Dayan and Abbott, 2001), with 60 input units innervating a single output unit. The input layer simulated a population of human auditory nerve fibres, with 60 gammatone filters ERB-spaced [Equivalent Rectangular Bandwidth] between 100 and 10,000 Hz (Glasberg and Moore, 1990). This array was constructed using the same model and parameters as described in Heinz et al. (2001). The mean firing rate of each node (i.e., rate-place encoding) was combined in a linear weighted sum by the output node. The decision rule was to select the interval that maximised the output, thus:

$$Out = \begin{cases} \text{'Int 1'}, & if \left(\left[\sum_{i=1}^{n} \omega_i a_i \right] - \left[\sum_{i=1}^{n} \omega_i b_i \right] \right) > 0 \\ \text{'Int 2'}, & otherwise \end{cases}, \tag{2.5}$$

where Out is the system output, a_i and b_i represent the ith input unit's response to the first and second stimulus respectively, and where ω_i represents the strength of the connection between the ith input unit and the output unit (which may be negative). All learning occurred via changes in the connection strengths between the input nodes and output node. The simulations were presented with the same stimuli/protocol as the human listeners. Weight adjustments were made online (i.e., after every trial) via the delta rule (Dayan and Abbott, 2001). The range of learning and starting rates were selected based on a brief period of trial-and-error using a validation dataset, but the precise values were randomly generated at the point of testing.

2.5.2 Results & Discussion

Fifteen independent simulations were run and were analysed in the exact same manner as the human listeners. As expected, connection weights were consistently optimised so as to maximise Fisher information. The resultant pattern of weights formed a roughly sinusoidal pattern, with a

minima offset to the left of the non-target distributions ($< [1000 - \mu\Delta_{Hz}]$) and a maxima offset to the right of the corresponding target distributions ($> [1000 + \mu\Delta_{Hz}]$). The key results regarding performance are summarised in Fig. 2.9. The upper panel expresses how frequency discrimination sensitivity increased as a function of session [$p < .001$]. The lower panel shows the concomitant decrease in internal noise as estimated with the same four methods as described previously [all $p < .001$]. In short, through the re-weighting simulated auditory nerve outputs, the model exhibited at the *functional* level a qualitatively similar pattern of learning to human listeners in terms of increased performance and reduced internal noise. This indicates that the observations of reduced internal noise in human listeners are consistent with the hypothesis of Lu and Dosher (2009) that perceptual learning reflects a re-weighting of early sensory representations. The fact that what at the functional level appears to be a reduction in internal noise magnitude, may at the mechanical level reflect a deterministic process of reweighting, is an oddity, but not a contradiction (in much the same that a response distribution being grossly Gaussian at the behavioural/population level is not inconsistent with subservient elements exhibiting Poisson processes). It does, however, suggest that ascriptions of internal noise at a behavioural level may be of limited explanatory value, and do not necessarily inform or constrain our understanding of how learning proceeds at a physiological level. It is this limited resolution that partly motivated the use, as presented in Chapter 3 and discussed below [$\oint$2.6], of more complex stimuli, where performance is more likely to be limited by functionally quantifiable strategies.

2.6 General Discussion

The purpose of the experiments reported here was to determine the mechanisms underlying auditory perceptual learning. With each of four separate techniques, significant improvements on a frequency discrimination task were best modelled as a decrease in internal noise magnitude. No significant changes in encoding efficiency, bias or inattentiveness were observed. This pattern of results was continued in very highly trained listeners (though these listeners also exhibited less inattentiveness in addition to decreased internal variability and improved frequency discrimination).

The finding that internal noise underlies learning is consistent with recent work in auditory development, where differences in internal noise have also been effective in explaining age-related changes in pure tone discrimination performance. For example, a recent paper by (Buss et al., 2009, see also Buss et al., 2006) concluded, based on the slopes of psychometric fits, that

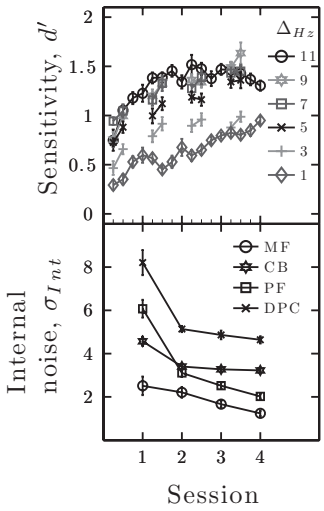


Fig. 2.9: Simulated frequency discrimination learning. The top panel shows changes in d' as a function of block/session for each stimulus condition, in the same format as the human listener data given in Fig. 2.3. The bottom panel shows internal noise estimates as a function of session using each of the following measures: model fit (MF), classification boundary (CB), psychometric function (PF) and double-pass consistency (DPC).

children's poorer intensity discrimination limens were due to elevated levels of internal noise.

However, our finding conflicts with a prominent claim in the visual perceptual learning literature that "Signal [enhancement] but not noise changes with perceptual learning" (Gold et al., 1999, see also Gold et al. 2004). In such papers signal enhancement is conceived as occurring through the appropriate, relative weighting of spatially distributed channels (e.g., by concentrating on those parts of an image that contain the greatest signal-to-external-noise ratios). Such signal enhancement corresponds to our 'encoding efficiency' concept. The claim of 'signal not noise' is therefore diametrically opposed to our finding that internal noise underlies learning. This may indicate qualitative differences between auditory and visual learning. However, the claim by Gold et al. (1999) lacks coherence. In Gold et al. (1999) observers attempted to identify images corrupted by a simultaneous Gaussian masker. Using a model equivalent to the SDT model presented in Eq. (2.1) an increase in signal enhancement was reported. Using a double-pass consistency analysis a constant ratio of internal-to-external noise was reported. However, given the nature of the noise, an optimisation of spatial channel weights implies a reduction in effective external noise. A constant ratio of internal-to-external noise therefore implies a concomitant reduction in internal noise (see Lu and Dosher, 2009 for further discussion)[3].

A more cohesive account of visual perceptual learning is given by Lu and Dosher (e.g., Dosher and Lu, 1999), who argue that learning consists of both internal noise reduction and external noise exclusion. Given that our task precluded external noise exclusion (cf. Lu and Dosher, 2008, for discussion), our finding that internal noise reduction was the primary mechanism of learning is consistent with Lu and Dosher's theory of visual perceptual learning. We predict that our finding would generalise to other pure tone auditory tasks (e.g., see Wright and Fitzgerald, 2005), which, together with frequency discrimination, constitute the substantial majority of the auditory perceptual learning literature. However, it remains an important and open question as to whether external noise reduction also occurs in auditory learning. For example, everyday listening situations often involve a substantial masking noise component. The filtering out of such noise may constitute a distinct and important perceptual learning process. Given the results from visual tasks, we predict that learning in such situations will be subserved by both additive internal noise reduction and an external noise exclusion mechanism.

2.7 Conclusions

(1) Learning on a pure tone frequency discrimination task is subserved by a reduction in internal noise, potentially through re-weighting of early sensory information. Changes in encoding efficiency, bias or attentiveness do not contribute to learning.

(2) Estimates of internal noise derived from four methods (model fit, classification boundary, psychometric function, double pass consistency) yield values in close agreement.

Notes

[1]Jitter was normally distributed on a linear frequency scale. This was intended to introduce Gaussian variance on the underlying decision dimension. For frequency discrimination the decision dimension is likely to correspond most directly to logarithmic frequency (e.g., Wier et al., 1976). Given the very narrow range of frequencies employed in this experiment, we do not believe that this discrepancy has any significant effect on the results. For example, even in the greatest frequency difference condition, the Hellinger distance (Nikulin, 2001) between the linear and logarithmic distributions was slight [$H < .003$; where $0 \leq H \leq 1$].

[2]Classification boundary fits were also made using a support vector machine (Cortes and Vapnik, 1995), but this procedure yielded virtually identical results and as such is not reported.

[3]In contrast, see Appendix 2.A for a description of how multiple information channels *can* 'enhance the signal', independent of external noise level

Acknowledgements

We thank Oliver Zobay for helpful discussions on statistical methods, and Yuxuan Zhang for comments on an earlier draft of this manuscript. We are also indebted to two anonymous reviewers for correcting errors in Equations (2) and (3), and for motivating the analyses reported in Appendix 2.B. This work was supported by the Medical Research Council, UK (Grant: U135097130).

2.A Model derivation

We assume that listeners perform the 2I2AFC task by linearly summing weighted activities (i.e., internal response scalar values) across multiple channels. Here we shall treat each stimulus presentation interval as a channel. We further assume that: (a) a given set of stimuli, $\langle S_1, S_2 \rangle$, generates fixed responses S_1 in channel 1, and S_2 in channel 2; (b) the external noise is a zero-mean Gaussian variable with standard deviation σ_{Hz} [$\phi(0, \sigma^2_{Ext})$], which is independently and identically distributed across both channels; (c) the internal noise is a zero-mean Gaussian variable with standard deviation, σ_{Int} [$\phi(0, \sigma^2_{Int})$], which is independently and identically distributed across both channels; (d) the total activity in each channel is the difference between the signal stimuli and some fixed criterion value $|\lambda - S|$, additively combined with observations from each of the noise distributions; (e) the relative weights given to channels 1 and 2 are denoted by the scalars ω_1 and ω_2 respectively, the squared values of which sum to 1; (f) the observer chooses interval 1 if $([\lambda - S_2 + \phi(0, \sigma^2_{Int}) + \phi(0, \sigma^2_{Ext})] \cdot \omega_1 + [\lambda - S_1 + \phi(0, \sigma^2_{Int}) + \phi(0, \sigma^2_{Ext})] \cdot \omega_2) < 0$ (and interval 2 otherwise); (g) the ideal

weights are given by the values $\langle \alpha_1, \alpha_2 \rangle$, which, when both intervals are equally informative will take the values $[\frac{-\sqrt{2}}{2} \frac{+\sqrt{2}}{2}]$. Given these assumptions, observed sensitivity, d', in the 2AFC case is:

$$d'_{obs} = \frac{\sum |\omega \Delta_{Hz}|}{\sqrt{\sigma^2_{Int} + \sigma^2_{Hz}}}, \tag{2.6}$$

where ω is an array of relative channel weights, and Δ_{Hz} is an array of mean differences between criterion and signal values, $|\lambda - S|$. The performance of an observer limited only by their adopted relative weights is:

$$d'_{weight} = \frac{\sum |\omega \Delta_{Hz}|}{\sigma_{Hz}}. \tag{2.7}$$

While ideal performance is

$$d'_{ideal} = \frac{\sum |\alpha \Delta_{Hz}|}{\sigma_{Hz}} = \frac{\Delta_{Hz}}{\sigma_{Hz}}, \tag{2.8}$$

where Δ_{Hz} is the difference in mean frequency of the two stimulus classes, and α is the ideal weight vector. Following Berg (2004)'s concept of efficiency we can partition overall observed efficiency, η_{total}, into the loss of efficiency due to non-optimal weights, η_{weight}, and due to internal noise, η_{noise}, thus:

$$\begin{aligned} \eta_{total} = & \quad \frac{(d'_{obs})^2}{(d'_{ideal})^2} = \frac{(d'_{obs})^2}{(d'_{weight})^2} \cdot \frac{(d'_{weight})^2}{(d'_{ideal})^2} \\ = & \quad \eta_{noise} \eta_{weight}. \end{aligned} \tag{2.9}$$

where,

$$\eta_{weight} = \left(\frac{d'_{weight}}{d'_{ideal}} \right)^2 = \left(\frac{\sum |\omega \Delta_{Hz}|}{\sum |\alpha \Delta_{Hz}|} \right)^2 \tag{2.10}$$

and

$$\eta_{noise} = \left(\frac{d'_{obs}}{d'_{weight}} \right)^2 = \left(\frac{\sigma_{Hz}}{\sqrt{\sigma^2_{Int} + \sigma^2_{Hz}}} \right)^2. \tag{2.11}$$

Note that by definition $0 \leq \sqrt{\eta_{weight}} \leq 1$. Applying this partitioning of efficiency (2.9 –2.11) to the d' equations (2.6 –2.8):

$$
\begin{aligned}
d'_{obs} &= d'_{ideal}\sqrt{\eta_{total}} && \text{(2.12a)}\\
&= d'_{ideal}\left|\frac{d'_{obs}}{d'_{weight}}\right|\sqrt{\eta_{weight}} && \text{(2.12b)}\\
&= d'_{ideal}\left|\frac{\frac{\sum|\omega\Delta_{Hz}|}{\sqrt{\sigma^2_{Int}+\sigma^2_{Hz}}}}{\frac{\sum|\omega\Delta_{Hz}|}{\sigma_{Hz}}}\right|\sqrt{\eta_{weight}} && \text{(2.12c)}\\
&= d'_{ideal}\frac{\sigma_{Hz}}{\sqrt{\sigma^2_{Int}+\sigma^2_{Hz}}}\sqrt{\eta_{weight}} && \text{(2.12d)}\\
&= \frac{\Delta_{Hz}}{\sigma_{Hz}}\frac{\sigma_{Hz}}{\sqrt{\sigma^2_{Int}+\sigma^2_{Hz}}}\sqrt{\eta_{weight}} && \text{(2.12e)}\\
&= \frac{\sqrt{\eta_{weight}}\Delta_{Hz}}{\sqrt{\sigma^2_{Int}+\sigma^2_{Hz}}} && \text{(2.12f)}
\end{aligned}
$$

For simplicity, d'_{obs} and $\sqrt{\eta_{weight}}$, are henceforth referred to as d' and η, thus:

$$d' = \frac{\eta \cdot \Delta_{Hz}}{\sqrt{\sigma^2_{Int}+\sigma^2_{Hz}}}. \qquad \text{(2.13)}$$

2.B Non-linear slopes in psychometric fits

Several studies concerning 2I2AFC pure tone discrimination tasks (e.g., Dai and Micheyl, 2011; Dai and Richards, 2011) have fitted psychometric functions in which sensitivity is related to signal strength, x, as follows: $d' = \left(\frac{|x|}{\alpha}\right)^{\beta}$. The β term in such models serves to vary the linearity of the psychometric slope (see Fig. 1 of Dai and Richards, 2011). Such non-linearity can be incorporated into the cumulative Gaussian fits described in Eq. (2.2), thus:

$$P(\text{'Int 2'}) = \gamma_{lo} + (\gamma_{up} - \gamma_{lo})\Phi(sign(x)|x|^{\beta}; \mu, \sigma^{\beta}). \qquad \text{(2.14)}$$

The psychometric functions reported in the present study can thus be considered a special case of Eq. (2.14), in which $\beta = 1$. By force-fitting linear ($\beta = 1$) slopes, an alternative explanation of learning may have been occluded. Moreover, since the value of β is liable to affect the other parameter estimates, the values of $\widehat{\mu}$, $\widehat{\sigma}$, $\widehat{\gamma_{lo}}$, and $\widehat{\gamma_{up}}$ may have been biased. To assess these possibilities, Eq. (2.14) was fitted to each listener's

session-by-session data, both when $\beta = 1$, and when β was a free parameter, constrained to be > 0.

Consistent with Dai and Micheyl (2011), estimated values of β did not deviate from unity in any of the four sessions [Hotelling's $T^2; T^2(4, 11) = 12.2, p = 0.11$]. Accordingly, unconstraining β had a minimal effect on the estimates of the other four parameters. In each case, no significant differences were observed when β was allowed to vary [Hotelling's $T^2; T^2(4, 11) = 3.0 - 9.7, p = 0.18 - 0.67$], although, consistent with Dai and Micheyl (2011), there was a general trend towards lower lapse rates (e.g., grand-mean $\widehat{\gamma_{lo}}$ decreased by 0.5%, while $\widehat{\gamma_{up}}$ increased by 0.7%); this difference was not significant, however.

These results suggest that the assumption of linearity is acceptable in the present study, and that the use of a non-linear term, β, would not have substantively effected the reported findings.

CHAPTER 3

Tone detection learning in unpredictable noise

On a pure tone discrimination task, perceptual learning has been shown to represent a decrease in internal noise magnitude (Chapter 2). However, the apparently prominent role of internal noise in learning may reflect the simplicity of such tasks. On a more complex task, in which listeners must appropriately integrate information across multiple channels, the listener's encoding strategy may become the primary limiting factor on performance, and thus the primary driver of learning. In this study we tested this hypothesis by using behavioural models to evaluate the mechanisms of learning in a tone-in-multitone-noise detection task.

Eight listeners practised detecting a 1 kHz pure tone for five sessions (4500 trials total). The target was presented either in quiet or within an unpredictable, notched, 30-tone masker (65 dB SPL; 223–4490 Hz). The amplitudes, phases and frequencies of each masker component were independently randomised prior to every presentation. Performance was indexed by masking level: the difference in 70.7% detection limens between quiet and noise conditions. ***Encoding efficiency****,* η_{enc}*, was indexed by estimating the weight that listeners gave to each spectral region, relative to the ideal. These weights were also used to estimate the decision variable,* DV*, for each trial, under the assumption that listeners responded to the interval containing the greatest sum weighted (dB) level. Psychometric functions were fitted to each listener's responses as a function of this decision variable, from which estimates of* ***internal noise magnitude****,* σ_{int}*, and* ***bias****,* CE*, were derived.*

Group mean masking decreased significantly across training sessions. Concomitant improvements in ***encoding efficiency*** *were also observed, with no changes in* ***internal noise magnitude*** *or* ***bias*** *magnitude. We conclude that practice can substantially improve detection performance in unpredictable noise, and that such learning is subserved by reductions in encoding efficiency (selective attention).*

3.1 Introduction

MEAN detection thresholds for a fixed-frequency (typically 1 kHz) sinusoid have been found to deteriorate by 20–50 dB SPL when a spectrally unpredictable multi-tone complex is presented simultaneously (Neff and Green, 1987; for overviews see Neff and Dethlefs, 1995; Kidd et al., 2007). Such effects cannot be explained by the magnitude of overlapping activity within peripheral auditory filters, since they are occur even when the noise is spectrally distal ('across-channel interference'), and/or energetically weak ('excess-additivity'). Instead, levels of masking appear to be largely driven by the degree of masker uncertainty, and the degree of target-masker similarity. Thus, randomly varying the spectral content of the noise between each presentation increases masking by around 10 dB SPL (Neff and Callaghan, 1988; Neff and Dethlefs, 1995; Tang and Richards, 2003). While introducing spatial, temporal or harmonic dissimilarities between target and masker provides a 10–30 dB SPL masking release (Kidd et al., 1994; Durlach et al., 2003b; Neff, 1995; Oh and Lutfi, 1998, see also Lee and Richards, 2011). The fact that such masking operates across-channels, and is driven by *prima facie* cognitive factors such as similarity and unpredictability, have led many to consider this a form of masking distinct from classical energetic masking (e.g., Tanner, 1958; Durlach et al., 2003a). Accordingly, it is often, though at times contentiously (Durlach et al., 2003a), referred to instead as informational masking (Pollack, 1975).

The first goal of the present work is to establish whether detection thresholds for a tone in unpredictable noise improve with practice. The second goal was to use this task to investigate the mechanisms underlying auditory learning in more complex listening environments.

3.1.1 Evidence of learning effects

Because experimenters often employ extensive practice trials prior to testing (e.g., 4800; Kidd et al., 1994), or otherwise exclude data to minimise learning effects (e.g., Durlach et al., 2003b), evidence of learning on this task is relatively fragmented.

The most unequivocal evidence of learning can be found in studies by Neff and Callaghan (1988) and Neff and Dethlefs (1995), which explicitly examined the effects of practice on performance (see also Oh and Lutfi, 1998). In Neff and Callaghan (1988), four listeners performed 1800 trials of a masked detection task, using interleaved blocks of 2- and 10-component maskers. Only one listener showed an improvement, which was around 30 dB in magnitude, and occurred during the first 600 trials (in

both conditions). Similarly, Neff and Dethlefs (1995) reported masking data from 49 listeners given a 10-component masker. Over 600 trials, eleven of the listeners (22%) exhibited decrements in masking of 10–15 dB. Masking levels in the remaining listeners remaining stable throughout. It therefore appears that learning does occur, but only in a minority of listeners.

However, these studies may have under-represented the degree of learning on this task. In neither study were listeners naïve to the task, having completed 600+ trials prior to testing. This may have occluded the initial phase of learning, where, crucially, learning rates tend to be greatest (see §1.1.4). Moreover, in neither study was the reported test regimen necessarily exhaustive, with listeners completing 600 – 1800 trials. In contrast, it may take several thousand trials to reach peak performance on a simple pure tone discrimination task (Demany, 1985), and potentially many tens of thousands on tasks involving more complex stimuli (Lively et al., 1993). It is therefore possible that given naïve listeners and longer test regimens, learning effects may occur more often and in greater magnitude.

That learning may occur on this task is also supported by the wider literature. For example, practice effects have been found on related 'informational' paradigms, such as intensity discrimination in unpredictable noise (Buss, 2008). More generally, smaller informational masking effects amongst musicians, while causally ambiguous, have often been interpreted as a long-term training effect (Oxenham et al., 2003).

3.1.2 Potential learning mechanisms

A number of changes may underlie improved performance on this task. Listeners may be increasing the effective signal-to-noise ratio, either by decreasing the magnitude of their own internal noise, and/or by developing an encoding strategy that filters out external noise. Changes may also be occurring in non-sensory processes. For example, reductions in decision bias are also liable to manifest as improved detection limens. The present work attempts to evaluate the role of each of these factors in auditory learning.

Encoding efficiency

To quantify a listener's encoding strategy it is necessary to determine what information the listener uses, and how that information is combined to form the decision variable, DV. Given a two-interval two-alternative forced-choice [2I2AFC] detection task, listeners' behaviour appears well characterised by a weighted-linear-sum model, in which each feature

is the level difference (in dB SPL) between the corresponding spectral regions in each interval (Lutfi et al., 2003, see also Tang and Richards, 2003). Thus, each spectral region is considered an independent source of information, and each corresponding weight determines the relative importance of that information within the decision making process. This is shown schematically for a single trial in Fig 3.1, and is formalised as:

$$DV = \sum_{i=1}^{n} \omega_i \Delta L_i, \tag{3.1}$$

where ΔL represents the difference in level at that spectral bin, and ω is the relative weight coefficient.

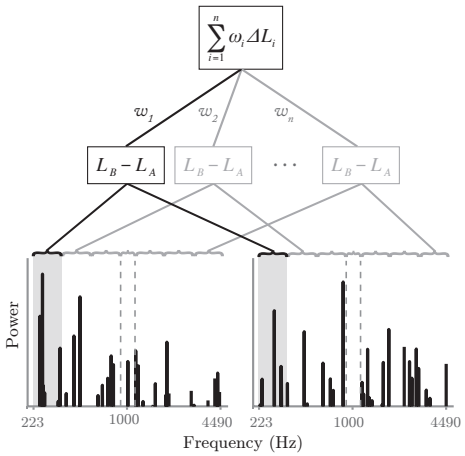


Fig. 3.1: Computing the decision variable. The level (in dB SPL) is calculated independently for each log-distributed bandpass filter, and the difference in level between corresponding bins, ΔL, computed. The decision variable is constructed from the linear weighted sum of ΔL values. The weight values, ω, are empirically estimated using the reverse correlation procedure shown in Fig 3.2.

The optimal strategy given such a model is shown in Fig 3.2b, and consists of giving non-zero weight only to the target spectral region. An alternative strategy would be to uniformly weight the entire spectrum, in which case the listener would be effectively listening for loudest or more intense stimulus. In practice, listeners often exhibit an intermediate strategy, giving greatest weight to the target region, but non-zero weights elsewhere (Alexander and Lutfi, 2004). These non-zero weights often tend to be proximal to the target, resulting in a peaked but gently sloping function. Such a pattern can be intuitively thought of as a broadly tuned 'attentional filter'.

Relative weights can be estimated through a form of molecular analysis often termed reverse correlation (Richards and Zhu, 1994; Lutfi, 1995, for

an overview see Dai and Micheyl, 2010). This consists of determining the degree to which a listener's trial-by-trial responses depend on each stimulus feature. In practice, the stimulus is discretised into N regions, either through binning (e.g., Berg, 2004), or by only employing a discrete set of components (e.g., Alexander and Lutfi, 2004). Multiple regression (or some similar technique) is then used to relate the trial-by-trial difference in level at each region, ΔL, to the listener's response. The resultant regression coefficients can, after normalisation (e.g., such that their magnitudes sum to one), be interpreted as the relative weight, ω, or importance, that the listener attributes to that aspect of the stimulus. A large (relative) weight indicates that the associated spectral region strongly determines the decision variable, DV. Conversely, a small weight indicates that responses are made largely independently of the associated stimulus region. The efficiency of the overall weight vector can be quantified as its (e.g., root-mean-square; RMS) difference from the ideal (Dai and Berg, 1992; Willihnganz et al., 1997; Stellmack et al., 1997; Alexander and Lutfi, 2004).

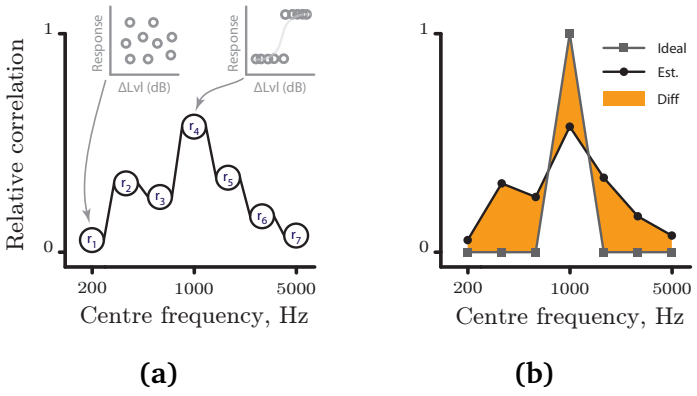


Fig. 3.2: Estimating relative encoding weights/efficiency schema. The left panel shows how a vector of relative weights is derived from the normalised correlation coefficient at each spectral region. The right panel shows how the efficiency of the estimated weights is computed as the (RMS) difference from the ideal. See body text for details.

Internal noise magnitude

Over the last 50 years a number of techniques have been developed to quantify internal noise, such as the multiple-looks approach (Swets, 1959), n-pass consistency (Green, 1964; Burgess and Colborne, 1988), and various external-noise-titrated model fits (see Lu and Dosher, 2008). Probably the simplest approach has been to equate internal noise with the slope of the psychometric function (i.e., with a shallower slope indicating greater internal noise; e.g., Buss et al., 2006). However, the slope is an ambiguous measure in that it is also affected by the efficiency of the encoding strategy. Attributing all inefficiency to internal noise is therefore tantamount to

assuming an optimal encoding strategy, and may lead to internal noise magnitude being overestimated. Berg (2004) presents an elegant two-step solution to this problem. Firstly, encoding weights are calculated for each listener. These values are then used to estimate trial-by-trial DV values. A psychometric curve is then fitted to performance as a function of DV. In this way the relative efficiency of the encoding strategy is partialled out, and the slope parameter (or its equivalent, cf. Gilchrist et al., 2005; Strasburger, 2001) can be interpreted as an unambiguous index of additive internal noise magnitude. Note, however, this 'two-step' approach means that any error in the estimates of the decision weights will be compounded when estimating internal noise.

Bias

While signal strength and internal noise relate to the listener's sensitivity to sensory information, we also considered here a third potential limitation on performance, bias. Bias expresses the fact that listeners are prone to favour some response alternatives, independent of the sensory information. Such bias may occur if the listener mistakenly perceives the relative frequency of a certain event, or the relative utility of a particular response outcome. On a typical psychoacoustical task, where the outcome-likelihoods are uniform and the payoffs symmetrical, any such bias will result in decreased performance relative to the ideal. Levels of (stationary) response bias can be inferred from the degree of lateral shift in the psychometric function. Specifically, by computing constant error, CE: the deviation of the point of subjective equality from the point of physical equality (Gescheider, 1997).

Summary

In short, the first purpose of this study was to establish whether, and to what extent, tone detection thresholds in unpredictable noise are reduced by practice in normal hearing adults. The second purpose of this study was to investigate the mechanisms underlying perceptual learning, by examining to what extent observed improvements in performance can be accounted for by either an increase in signal strength, or by decrements in internal noise or bias. The prevailing assumption is that the efficiency of the encoding strategy is the limiting factor on performance (cf. Oh and Lutfi, 1998). In this case we would predict that encoding weights will become more tightly focused around the target frequency as a function of practice. Alternatively, in simpler, pure tone discrimination tasks, internal noise magnitude has been cited as the driver of learning (Chapter 2) and development (Buss et al., 2006, 2009). If changes in internal noise are responsible for learning on the present tasks then steeper psychometric

slopes would be expected after learning. We also quantified bias and examined its effects on learning and masked threshold performance.

3.2 Methods

3.2.1 Listeners

Eight listeners (five female) participated, aged 19–26. None had any prior experience of psychophysics. All had normal hearing, as assessed by audiometric screening ($\leq$ 20 dB HL bilaterally at 0.25–8 kHz octaves) administered in accordance with BSA standard procedure (British Society of Audiology, 2004). Participants were recruited through advertisements placed around the Nottingham University campus, and received an inconvenience allowance for their time. The study was conducted in accordance with Nottingham University Hospitals Research Ethics Committee approval and informed written consent was obtained from all participants.

3.2.2 Stimuli and Apparatus

The target stimulus in all conditions was a 1 kHz sinusoid, which was randomly assigned to one of two observation intervals (Fig 3.3). In masked trials a 30-component multitone complex was presented simultaneously. All stimuli were 300 ms in duration (including ramps), and were gated on/off by 10 ms cos^2 ramps.

The frequency, phase, and amplitude of each noise component was independently randomised prior to every presentation. Phases and amplitudes were randomly drawn from a rectangular and a Rayleigh distribution, respectively. Frequencies were randomly drawn without replacement from a pool of 715 candidates, log distributed between 223–4490 Hz, excluding a third-octave notch geometrically centred on the target frequency (891–1120 Hz). This notch is similar to, though slightly larger than, the average equivalent rectangular band [ERB] (see also Scharf, 1970 Glasberg and Moore, 1990). The level of target tone varied between 0-80 dB SPL, according to an adaptive track. The masker was always presented at an average power of 60 dB SPL.

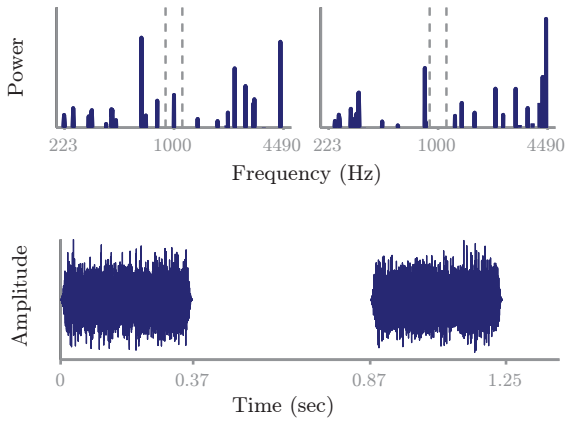


Fig. 3.3: Example stimuli for a single trial, in the frequency (top) and temporal (bottom) domains. See body text for details.

Stimuli were digitally synthesised in Matlab v7.4 (2007a, The MathWorks, Natick, MA) using a sampling rate of 44.1 kHz and 24-bit quantization. Digital-to-analogue conversion was carried out by a PCI sound card (Darla Echo; Echo Digital Audio Corporation, Carpinteria, CA), interfaced via the Psychophysics Toolbox v3 (Brainard, 1997; Pelli, 1997) ASIO wrapper (Steinberg Media Technologies, Hamburg). Stimuli were presented diotically via Sennheiser HD 25-I headphones. Listeners responded via a button box, and were tested individually in a double-walled sound-attenuating booth. Fixation cues and feedback were presented visually on an LCD monitor.

3.2.3 Procedure

The task was 2I2AFC fixed-frequency tone detection, for which participants were asked to "pick the interval containing the target tone".

Each trial consisted of two 300 ms observation intervals separated by a 500 ms interstimulus interval. Listeners were then given an unlimited time to respond, before being presented with 250 ms of visual feedback in the form of a 'happy' or 'sad' smiley face, to which they were instructed to attend.

In each block, a two-down one-up adaptive track (Levitt, 1971) was used to derive an estimate of the listener's detection limen, DL, either in quiet or in the presence of the multitone masker. The level of the target tone was initialised at 60 dB and adapted by an initial step size of 8 dB until the second reversal, and in steps of 2 dB thereafter. Each block consisted of 50 trials. The number of trials was fixed rather than the number of reversals in order to ensure that all listeners received the same amount of practice.

Before each block listeners were reminded of the target signal, which was presented in quiet. At the end of the block listeners were given a score derived from their masked threshold, averaged over the last four reversals (all tracks included > 4 reversals; $\mu = 13.3$; $\sigma = 3.2$ [excluding reversals during lead-in phase]). The next block began when the listener pressed a button.

Each session lasted around 45 minutes, and consisted of 16 noise blocks and two quiet blocks, presented in random order with a rest break after the 10th block. All listeners took part in five sessions over five consecutive days. Before the first session participants also completed one practice trial in quiet and three practice trials in noise. To highlight the task demands the stimuli durations were increased during this practice to 800 ms, and any noise was presented at 50 dB SPL. Any listeners that failed to answer all three noise trials correctly were given an additional two trials, which were answered correctly in all instances.

3.2.4 Analysis

In total, each listener completed 90 adaptive tracks of 50 trials (4500 trials). Ten tracks were performed in quiet (no masker), while in the other 80 tracks the target was simultaneously masked by an unpredictable, 30-tone complex.

Masked detection limens, DL, were independently calculated for each track as the linear mean target level (dB) at the last four reversals. For comparison, DL values were also calculated as the 70.7% point of a logistic psychometric function (fitted to P_C as a function of target level). No substantive differences were observed between either sets of measurements (good agreement $[r = 0.84, p < 0.001]$ with a geometric regression slope indistinguishable from unity $[t(631) = 1.41, p = 0.92, n.s.]$). Accordingly, only the DL values derived using psychometric fits are reported here. Masking level (or: *reception threshold*) was computed for each of the 80 masked tracks as the DL in noise, minus the mean DL in quiet.

The criterion for learning was both: *(i)* a significant negative regression in DL against block, together with *(ii)* a significant difference in mean masking level across sessions. Encoding efficiency, η_{enc}, internal noise magnitude, σ_{int}, and bias, CE, were estimated independently for each listener in each session, as follows.

Relative encoding weights were calculated from the coefficients of a multiple logistic regression (Fig 3.2a). The dependent variable was the listener's binary response ('Interval 1' or 'Interval 2'). The independent variables were the differences in (dB) level between the corresponding

spectral region in each stimulus interval. These regression coefficients were then normalised so that their absolute magnitudes summed to one. Encoding efficiency, η_{enc}, was calculated as one minus the sum root mean square [rms] difference between observed and ideal weights (Dai and Berg, 1992; Stellmack et al., 1997; Willihnganz et al., 1997; Alexander and Lutfi, 2004) – the ideal strategy being to assign a weight of unity to the target bin and zero-weight elsewhere (Fig 3.2b). Thus, η_{enc} ranged from 0 to 1, with efficiencies of 0.0 and 1.0 indicating complete disregard and complete attention to the target region, respectively. Notably, current theories governing the measurement of such weights make a number assumptions (e.g., see Richards and Zhu, 1994), some of which did not hold in the present experiment. However, as discussed in Appendix 3.A, this is unlikely to have qualitatively affected our findings.

Internal noise was assumed to take the form of a zero-mean, normal distribution which combines additively with the listener's internal response on a trial-by-trial basis. The magnitude of the internal noise, σ_{int}, was calculated as the standard deviation of a cumulative normal distribution, fitted to the binned probability of a listener responding 'Interval 2' as a function of estimated DV[1]. Values of DV were in turn computed by multiplying the trial-by-trial stimulus data by the listener's estimated weights, as per Eq 3.1. Psychometric fits were made using PSIGNIFIT version 2.5.6 (see `http://bootstrap-software.org/psignifit/`): a Matlab toolbox which implements the maximum-likelihood method described by Wichmann and Hill (2001).

Psychometric fits were also used to derive the measure of bias, CE (constant error). This was computed as the point of subjective equality minus the point of physical equality (0) on the psychometric function. Thus, positive and negative values indicated a bias in favour of responding 'Interval 1' and 'Interval 2', respectively.

Listeners also completed a short questionnaire regarding family histories of early-onset hearing impairments (non-reported) and musical habits (*N* hours spent playing and listening to music; *N* years music tuition and associated grades).

3.3 Results

3.3.1 Learning

Masked detection thresholds in quiet and noise conditions are plotted for individuals in Fig 3.4. In the quiet condition no learning was observed, with the data in all individuals being well described by a linear regression with a

slope coefficient close to zero ($\mu = 0.01$; $\sigma = 0.01$). In contrast, substantial learning was observed in the masked condition. Linear fits yielded negative slopes in all eight listeners, with improvement rates ranging from -0.02 to -0.18 dB/Block. Data for listeners were better fit by a broken-stick function inflected at the end of session one, suggesting a short initial phase of rapid learning followed by a protracted period of more gradual learning. But with the exception of L5 this improvement was small given the additional degree of freedom [$\Delta_{r^2}/\Delta_{d.f.} < 1$]. Full breakdowns of learning slopes are given in Table 3.1.

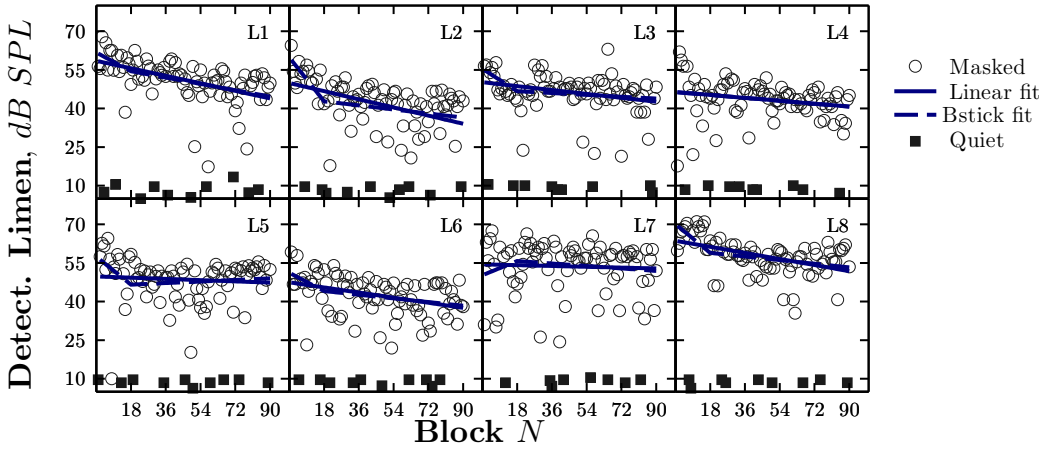
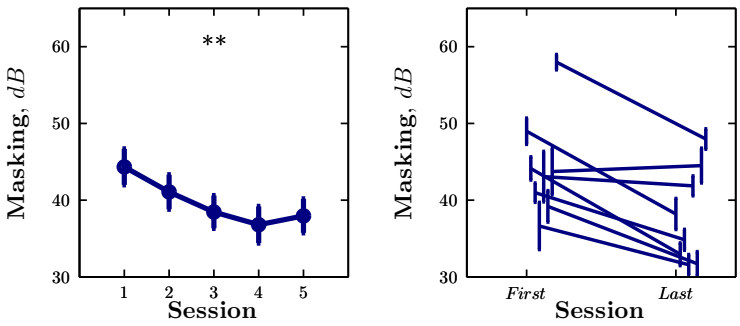


Fig. 3.4: *DLs* for individuals, as a function of block. Detection thresholds in quiet and in noise are shown by filled squares and open circles, respectively. Solid lines represent least square linear fits to the noise data. Dashed lines represent equivalent piecewise linear fits with inflections constrained to lie at the end of session one. Slope coefficients are given in Table 3.1. Tick-marks on the abscissa correspond to the end of each session.

Listener	Regression Slopes			Mean Change, $S1 - S5$		
	$linear$	$bstick_A$	$bstick_B$	ΔdB	$t(14)$	p
O1	-0.16	-0.43	-0.15	-10.78	3.65	< 0.01
O2	-0.18	-1.01	-0.10	-11.18	5.05	< 0.01
O3	-0.08	-0.51	-0.04	-6.15	3.68	< 0.01
O4	-0.06	-0.06	-0.07	-5.06	1.36	0.19
O5	-0.03	-0.64	0.04	-1.20	0.35	0.73
O6	-0.11	-0.42	-0.09	-7.47	3.06	0.01
O7	-0.02	0.32	-0.06	0.78	-0.24	0.82
O8	-0.13	-0.66	-0.09	-10.01	5.67	< 0.01

Table 3.1: Learning parameters for individual listeners. The regression slope data (columns $2 - 4$) show rate of learning per block (see also Fig 3.4). Columns $5 - 8$ show changes in mean masking between the first and last session, and the associated paired-sample t-test statistics (see also Fig 3.5). Values in parentheses for L5 show equivalent statistics when session four was substituted for session five (see body text).

Grand mean masking decreased [$t(7) = 4.09, p = .005$] from 44.3 dB SPL in session one to 38.0 dB SPL in session five (Fig 3.5a). As Fig 3.5b shows, there was substantial individual variability, with changes in masking ranging from +0.8 (L7) to -11.2 dB (L2). Independent t-tests indicated that the reduction in masking was significant in five of eight listeners (see Table 3.1 for breakdown). Notably, the fact that L5 did not display a significant learning effect was due to a marked decline in performance during the final session, most likely caused by a reported loss of concentration. Accordingly, an additional post-hoc t-test was performed for this listener, with the penultimate session's data substituted for session five. This yielded a significant reduction in masking [Bonferroni corrected; $\alpha = 0.05 \rightarrow 0.025$]. For all listeners a significant level of masking remained even after training [$p << .001$].



(a) **(b)**

Fig. 3.5: Mean (± 1 SE) masking as a function of session, averaged between (left panel) and within (right panel) listeners. Individual plots have been jittered along the abscissa for clarity. Mean differences and associated test statistics are shown in Table 3.1.

There was substantial variability in performance between listeners, with masking ranged from 36.6–58.0 dB SPL in session one, and from 31.6–50.0 dB SPL in session five. There was no obvious tendency for interlistener variability to be reduced by practice, with the greatest range of values observed in session four, where group mean masking was lowest (cf. Fig 3.5a). The variability between listeners was not explained by any of the available measures. Thus, there was also no relationship between starting performance or learning, and either sex or amount of musical training [all $p > 0.05$]. Nor was there any relationship between initial masking level and either DL in quiet [$r = .18, p = .671, n.s.$], or change in masking across sessions [$r = -.53, p = .180, n.s.$] (though, as shown in Fig 3.5b, the greatest changes were observed in the two initially poorest listeners). There was some indication that variability within listeners, as indexed by within-session standard deviation in masking, may be decreasing with practice. However, as with inter-listener variability, this decrease was not significant [$t(7) = 2.27; p = .058, n.s.$].

3.3.2 Mechanisms of learning

Encoding strategy

Estimates of encoding weights are shown for individuals in Fig 3.6. All listeners appear to have adopted encoding strategies approximate to the ideal, with the largest weight given consistently to the target bin (1 kHz). However, in session one, every listener negatively weighted the lowest frequency region, and positively weighted the highest frequency region. This may indicate a general strategy in favour of selecting higher (mean) frequency stimulus. In the case of lowest frequencies, all listeners shifted their weights with practice towards the ideal, though in some

cases appeared to over-compensated, resulting in deleteriously positive weightings (e.g., L1, L8).

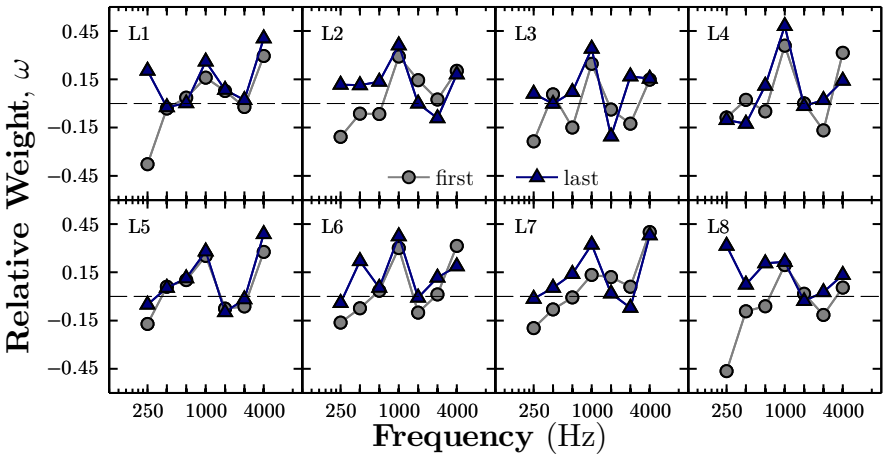


Fig. 3.6: Individual encoding weights, for the first (grey circles) and last (blue triangles) session. Each point represents the geometric centre of the spectral bin. The target signal was always a 1 kHz sinusoid, so the optimal strategy was to give a relative weight of 1.0 to the 1 kHz bin, and zero weight elsewhere. Data from intermediate sessions are omitted for clarity.

A repeated measures ANOVA yielded a significant main effect of session on encoding efficiency [$F(4, 28) = 3.48, p = .020, \eta_p^2 = 0.33$], indicating that listeners' encoding strategies improved (i.e., became closer to the ideal) with practice (Fig 3.7, top).

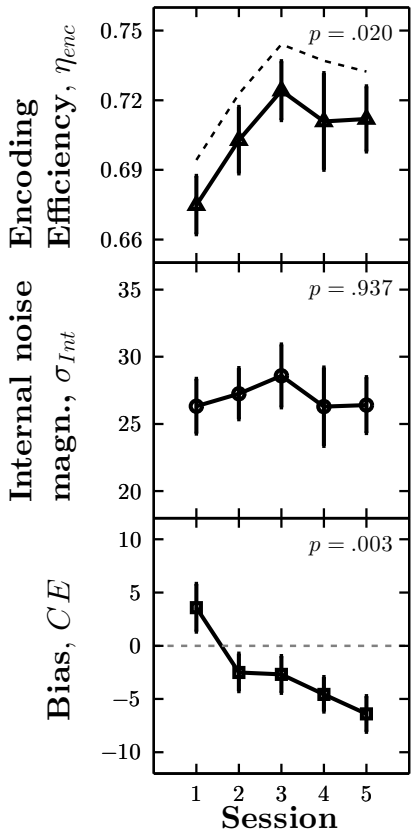


Fig. 3.7: Group mean ($\pm$ 1 SE) learning mechanism parameters as a function of session. From top to bottom: Encoding weight efficiency [$0 \leq x \leq 1$]; Internal noise magnitude [$0 < x$]; and Bias [$-\infty < x < \infty$]. p-values indicate the significance of the associated repeated measures ANOVA (see body text). The dashed line in the top panel indicates η_{enc} after adjusting for underestimation, as per Appendix 3.A.

Internal noise

To evaluate changes in internal noise, cumulative Gaussians were fitted to listeners' response data as a function of the estimated trial-by-trial DV. As shown in Fig 3.7, group-mean estimates of σ_{int} did not systematically vary across session [$F(4, 28) = 0.20, p = .937, \eta_p^2 = 0.03$], indicating that internal noise magnitude was not diminished by practice (Fig 3.7, middle). However, inspection of the individual fits reveals substantial variability between listeners (Fig 3.8). For example, listeners L1 and L4 exhibited a marked decrease in internal noise, as indicated by the steeper psychometric slopes in session five. Conversely, listener L8 shows very little change, despite the substantial learning evident in Fig 3.4. Notably, if ideal weights were assumed (i.e., fits made based on variations on target level only), σ_{int} appeared to change markedly [$F(4, 28) = 8.50, p < .001, \eta_p^2 = 0.55$].

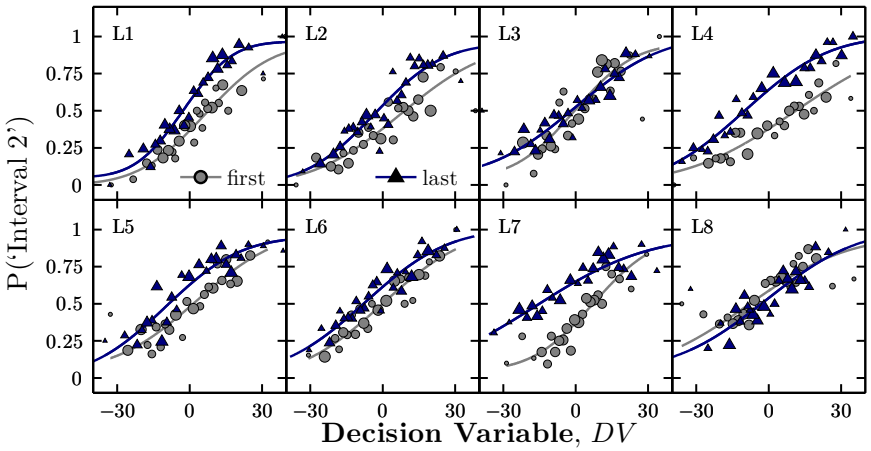


Fig. 3.8: Individual psychometric fits, for the first (grey circles) and last (blue triangles) session. The probability of responding 'Interval 2', plotted as a function of the estimated decision variable and fit with cumulative Gaussian functions. The standard deviation parameter, σ, was taken to represent internal noise magnitude, while CE was computed as an index of bias.

Bias

Changes in group-mean bias were observed [$F(4, 28) = 4.88, p = .004, \eta_p^2 = 0.41$] (Fig 3.7, bottom). However, as in Chapter 2, the session means did not significantly differ from 0 (no bias) [Hotelling's T^2; $T^2(3, 5) = 30.17, p = 0.232, n.s.$]. Moreover, mean bias magnitude did not significantly differ across sessions [$F(4, 28) = 0.40, p = .809, \eta_p^2 = 0.05$]. Changes in bias did not therefore appear to contribute substantively towards learning.

Comparisons between η_{enc} and σ_{int}, and CE

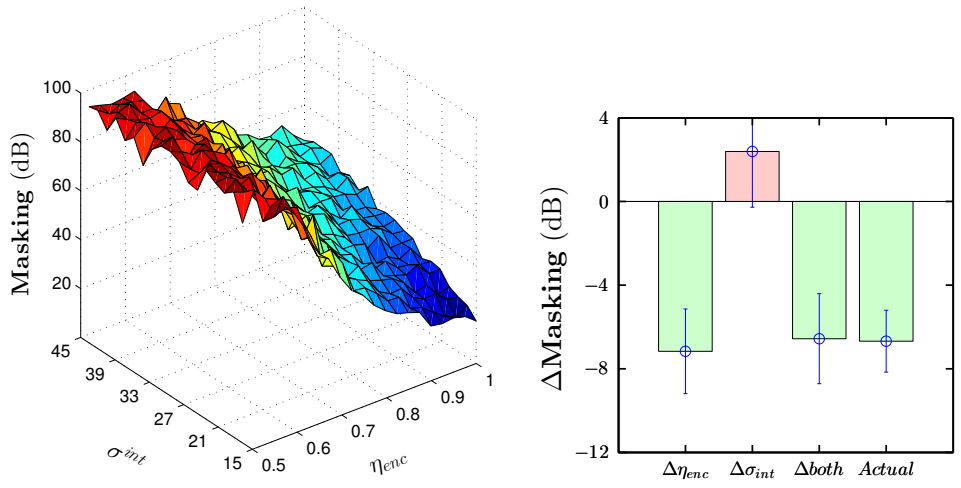
A stepwise multiple linear regression was used to relate session-by-session *changes* in masking to changes in encoding efficiency, internal noise magnitude, and bias. Only encoding efficiency [$p < .001$] and internal noise magnitude [$p = .001$] were significant predictors in this model. The full model, containing these two predictors, explained 18% of the variance in changes in masking. To evaluate the relative importance of each mechanism, partial correlations were performed for each mechanism, controlling for changes in the other two. Encoding efficiency was the strongest predictor of changes in masking [$r^2 = 0.46$], followed by internal noise magnitude [$r^2 = 0.35$]. Changes in bias were a poor predictor of changes in masking [$r^2 = 0.08$]

To further evaluate the relative importance of η_{enc} and σ_{int} in learning, a simulation was used to relate levels of each factor to predicted threshold performance (DL). In this way each improvements in each could be expressed in a common unit – change in masked threshold (dB).

The simulation estimated mean threshold performance for various combinations of σ_{int} and η_{enc}. Mean thresholds were calculated by averaging over 16 thresholds, and this was done for 20 independent values of σ_{int} and η_{enc} (400 conditions total). Each threshold was estimated via the same adaptive procedure used previously with human observers [§3.2.3]. The simulated observer acted as an unbiased signal detector, which responded to the interval producing the greatest internal response. The internal response was calculated as the linear weighted sum of the input values, added to a random valued drawn from a Gaussian internal noise source, as per Eq 3.1. Weight efficiency, η_{enc}, and internal noise magnitude (S.D.), σ_{int}, were determined by the condition. The precise configuration of weights, ω, was determined by linearly interpolating between the ideal and a randomly initialised vector.

The result was the manifold given in Fig 3.9a. Using these values, the first-to-last session variations in σ_{int} and η_{enc} previously observed in human listeners were converted to their predicted change in threshold values. Since the effect of each parameter on threshold is not independent, when evaluating each the other parameter was held constant at that listeners mean level (averaged across sessions).

The results are shown in the first two columns of Fig 3.9a. From these it can be seen that changes in η_{enc} far exceeded changes in σ_{int}, in terms of their impact on masking threshold (DL). Indeed, changes in η_{enc} appeared to fully explain observed changes in threshold (column 4). A secondary function of the simulation was to internally-validate the measures of σ_{int} and η_{enc}. Accordingly, when changes in both were jointly evaluated, the predicted change in threshold closely matched the observed change, suggesting that the parameter estimates reliably captured listeners' abilities.



(a) **(b)**

Fig. 3.9: Simulations relating changes in σ_{int} and η_{enc} to change in threshold (DL). (Left) Simulated manifold showing masking threshold as a bivariate function of σ_{int} and η_{enc} [see body text for details]. (Right) Predicted changes in masking given estimated changes in parameters [columns 1–3], and, for comparison, the empirically observed change in masking [column 4].

3.4 Discussion

3.4.1 Learning

The first aim of the present study was to establish to what extent tone detection limens in unpredictable noise improve with practice. Over five sessions (4500 trials; 4000 in noise), five of eight listeners exhibited significantly lower masking. The magnitude of this change was substantial, with the five learners exhibiting a 9.1 dB reduction in grand mean masking. We therefore conclude that tone detection in unpredictable noise does elicit substantial learning effects.

The proportion of learners is commensurate with other psychoacoustical tasks (for an overview see Zhang and Wright, 2009), such as temporal order discrimination (67-86%; Mossbridge et al., 2006, 2008), interaural level difference (75%; Wright and Fitzgerald, 2001), intensity discrimination in unpredictable noise (75%; Buss, 2008), and frequency discrimination ($81 - 96\%$; Irvine et al., 2000; Demany, 1985)[2]. Since the performance of some listeners (e.g., L1 and L4) had not plateaued by the end of the study, greater reductions in masking may have been possible given further practice. However, any further improvements are likely to be modest given that by the end of the study our cohort was performing at a comparable level to well-trained observers reported previously (Neff and Dethlefs, 1995).

Large individual differences in performance were observed. Previous authors have wondered whether such individual differences can be reduced by training (e.g., Durlach et al., 2003a). Similar to Neff and Callaghan (1988), we find no evidence of that here; between-subject variation in masking was approximately constant across all five sessions. To try to better understand the causes of such variability we examined the effects of sex, musical training (mean hours per week $\times$ N years) and listening habits (mean hours per week) on starting performance and learning, each of which have been reported to affect performance previously (e.g., Oxenham et al., 2003; Spiegel and Watson, 1984). In no case was any relationship evident [Mean $p = .609$], though this may reflect the small and relatively homogeneous sample.

The finding that tone detection in unpredictable noise is improved by training complements an analogous finding by Buss (2008), wherein six of eight listeners also exhibited a partial release from masking on an intensity-discrimination in unpredictable-noise task. However, our conclusion is contrary to that of Neff and Dethlefs (1995), where the authors state that detection limens in unpredictable noise appear 'remarkably stable' as a function of practice. This difference is likely due to the fact that Neff and Dethlefs (1995) excluded the first 600 trials. The most appropriate comparison to the presently reported data would therefore be to the second arm of the piecewise linear fits (which began at 800 trials). When only these latter trials are considered, both incidents and rates of learning are reduced to levels similar to those of Neff and Dethlefs (1995). This is consistent with the fact that auditory learning follows an approximately logarithmic distribution (Molloy et al., 2012), with an initial rapid phase of learning followed by a prolonged period of gradual improvement (similarly Hawkey et al., 2004; Wright and Fitzgerald, 2001).

3.4.2 Learning Mechanisms

The second aim of this study was to investigate the mechanisms underlying perceptual learning. Three potential limiting factors were considered: encoding efficiency, internal noise magnitude, and bias. Of these, learning appeared to be primarily driven by changes in encoding efficiency only (Fig 3.7). A change in bias was also observed, but this did not appear related to learning, resulting as it did from a change in sign only, with the magnitude remaining approximately equal in the first and last session. The stability and consistency of performance in quiet suggest that in the absence of external noise listeners are limited by a factor that

is relatively immutable, such as internal noise from the cardiovascular system[3] (Soderquist and Lindsey, 1971; Shaw and Piercy, 1962).

Increased encoding efficiency can be understood either as an improvement in selective attention to the target frequency, or, equivalently, as a reduction in across-channel interference. This pattern of results is therefore consistent with Lu and Dosher (2008), who argue that (visual) perceptual learning represents a mixture of additive internal noise reduction and external noise exclusion (see also Dosher and Lu, 1998). Similarly, as observed in Dosher and Lu (2005), improvements acquired after training in noise did not transfer to performance in quiet. The reasons for this lack of transfer are not immediate obvious. It may indicate that listeners are in part limited by an internal noise component that interacts multiplicatively with stimulus magnitude. For example, the standard deviation of energy in each spectral region has previously been found to be a good predictor of performance on unpredictable masking tasks (Lutfi, 1993; Oh and Lutfi, 1998); if weights are liable to be affected exogenously, then this variability would cause trial-by-trial jitter in a listener's encoding strategy, which would here alias as increased internal noise. That internal noise results from weight-jitter in this manner could be tested in well-trained listeners by examining whether the spread of residuals in the weighting regression model increases as a function of spectral variability. Alternatively, the fact that learning did not improve performance in quiet may indicate that qualitatively different listening strategies are used in noise and in quiet (i.e., such that the learning only affected the former). Given the blocked design this is certainly possible, and Allard and Cavanagh (2012) have recently argued for just such an interpretation, following related findings in a noisy visual orientation identification experiment. However, their conclusions are predicated on an assumption of early internal noise (prior to integration across channels). Such considerations would therefore require a non-trivial reformulation of the present decision model (Fig 1.3). This lies outside the scope of the present work, but may provide the basis for future research.

Variations in encoding efficiency and internal noise magnitude explained 18% of variance in session-by-session changes in masking. It may be that the remaining variance is accountable in terms of measurement error. Alternatively, other factors may also part-determine learning, such as changes in response contingencies, energy or motivation. Relatedly, Alexander and Lutfi (2004) found that masked performance in hearing impaired listeners was poorly predicted by encoding efficiency alone, implying that other factors are required to explain performance. Alexander and Lutfi (2004) posit that variations in auditory filter width may also affect masking levels by modulating the effective amount of spectral variability.

However, this is unlikely to be a factor during learning in normal-hearing listeners, where auditory filter widths are relatively stable, and are largely uncorrelated with masked thresholds in unpredictable noise (Neff and Dethlefs, 1995).

Although improvements in encoding efficiency alone were shown to underlie learning at the group level, it is unclear from the present data whether this pattern holds for all individuals. Thus, some listeners (e.g., L1, L2) did exhibit decreased internal noise after training. A larger sample is required to determine whether such individual differences reflect distinct approaches to learning, or simply random variation. In favour of the former, previous studies have indicated that encoding weights alone may be neither necessary nor sufficient to explain differences in performance. For example, individuals have been observed who exhibit less efficient weights but better performance than their peers (Berg, 2004), or who perform similarly despite differences in weight efficiencies (Shub, 2012).

As in some previous reports, there was a tendency for listeners to give relatively large weight to the lowest (cf. Buss, 2008; Neff and Odgaard, 2004) and highest (cf. Watson et al., 1976) frequency maskers. This may be due to these spectral regions being perceived louder, either due to decreased (energetic) masking at the masker fringes, and/or in the case of higher components, the relative amplification of acoustic energy around 3–4 kHz by the external auditory meatus (Fletcher and Munson, 1933; Robinson and Dadson, 1956). That the greater distal weights are due to differences in loudness is also consistent with the fact that these weights appear attenuated when maskers are equalised for sensation level (Alexander and Lutfi, 2004), however, even then, distal weights continue to be disproportionately large in some listeners.

In several listeners (L2, L5, and L8) performance declined during the final session (cf. Fig 3.4). We suggest that these decrements are most likely due to changes in non-sensory factors arising from a loss of concentration due to boredom, fatigue, and/or an expectation of completion (i.e., rather than due to a reduction in true sensitivity). Consistent with this Neff and Dethlefs (1995) observed abrupt improvements when listeners began a second, novel paradigm after extensive training on a particular task. Such improvements are difficult to explain purely in terms of changes in perceptual sensitivity, and can be more parsimoniously accounted for in terms of a ‘release from boredom’.

3.5 Conclusions

(1) Masking by unpredictable noise is reduced, but not eliminated, by training. The bulk of this learning occurs rapidly, within the first session (800 noise trials).

(2) Improvements in encoding efficiency alone appeared primarily underlie perceptual learning on this task. This is in contrast to the simpler task presented in Chapter 2, where internal noise magnitude appeared to be the primary mechanism of learning.

Notes

[1]The binning procedure was as follows. The range of stimulus values was initially divided into 50 uniform bins. Any bins containing more than 50 data points was then recursively bisected until it contained $\leq$ 50 points. Each bin was then iteratively evaluated in ascending order, and any bin with fewer than 50 points was merged with the succeeding bin though the removal of its upper boundary.

[2]It is difficult to accurately compare proportions of learners between studies because of the differences in criteria used to determine/report individual learning. At one end of the spectrum, Wright and Fitzgerald (2001) required both significant mean differences (e.g., one-way ANOVA) and negative regression slopes that differed significantly from zero. Conversely, Demany (1985) reports only the sign of regression slopes for individuals.

[3]n.b. such noise would likely be early, but would likely be correlated across channels, making it functionally indistinguishable from late internal noise

Acknowledgements

This work was supported by the Medical Research Council, UK (Grant: U135097130).

3.A Issues concerning weight measurements

The unpredictable masking task considered in the present study was of particular interest due to its real world significance and potential clinical relevance. However, it has a number of properties that complicate the derivation of weights.

Firstly, in the target spectral bin there was always zero energy in one interval. In such instances the dB level was taken to be 0 (rather than $-\infty$), so as to allow a real valued DV to be computed.

Secondly, the methods used for estimating weights assume that variability is normally, identically distributed across channels. This is not the case here. The variability in the noise channels is Rayleigh distributed. While the target variability is dependent on the adaptive tracking procedure, and as a result was both non-normal and distributed differently to that in the noise channels. Small deviations from normality (e.g., as in the noise channels) have been shown not substantively affect weight estimates (Richards and Zhu, 1994). However, the more substantial differences between the target and noise channels, while precedented (Alexander and Lutfi, 2004), may have aversely effected the reliability and/or validity of the weight estimates. The potential level of impact could be assessed in well-trained listeners by

comparing weights derived through adaptive methods, to those where the target was fixed at threshold and jittered as per the noise channels.

Thirdly, the trial-by-trial correlations assume that weights are stationary within each session. Since, *ex hypothesi*, this is not the case during learning, simulations were run to investigate the impact of non-stationary weights on the resultant efficiency measure, η_{enc}. Weights were interpolated between some random initialisation and their ideal values. This interpolation followed either a logarithmic or linear function, with weight values being updated 'online' (after every trial). Every consecutive bin of 800 trials was used to compute an estimate of encoding weights, ω. Figure 3.10 shows the resultant group-mean efficiency estimates for 1000 simulations, given a logarithmic learning model. Both curves followed a two-parameter single-exponential rise to maximum, but estimated weight efficiencies consistently lagged behind the true values. At low levels this disparity was negligible. However, as efficiency increased, empirical values increasingly *underestimated* true efficiency. This is due to the normalisation procedure resulting in small amounts of measurement error being amplified (cf. top-right panel). This disparity followed an approximate power-law, and true efficiency estimates could be recovered by applying a correction of the form $x+1134x^{46.6}$ [least-squares fit]. Analogous results were observed when the learning rate was linear (Fig 3.11).

The principle conclusion to be drawn from these simulations is that weights can be derived during learning, but that changes are likely to be underestimated. More generally, this indicates that the goodness of the encoding strategy is consistently underestimated in more efficient listeners. For example, when $\eta_{enc} = 0.88$, around 50,000 trials were required to derive an weight estimates that accurately represented true efficiency without the need for any explicit correction.

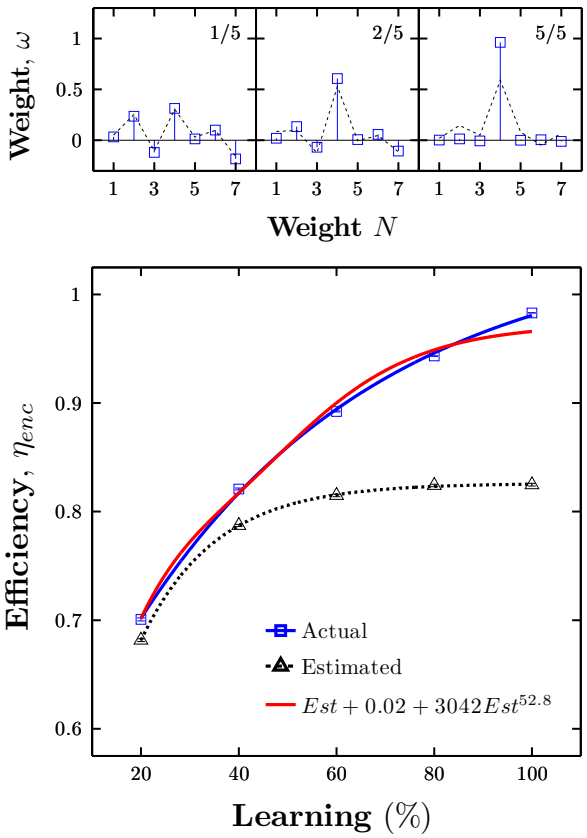


Fig. 3.10: Simulated weight estimation during logarithmic learning. (Top) Selected weight values (blue stems) and estimates (dashed lines) for a single simulation. (Bottom) The resultant group-mean efficiency values, estimated from 1000 simulations.

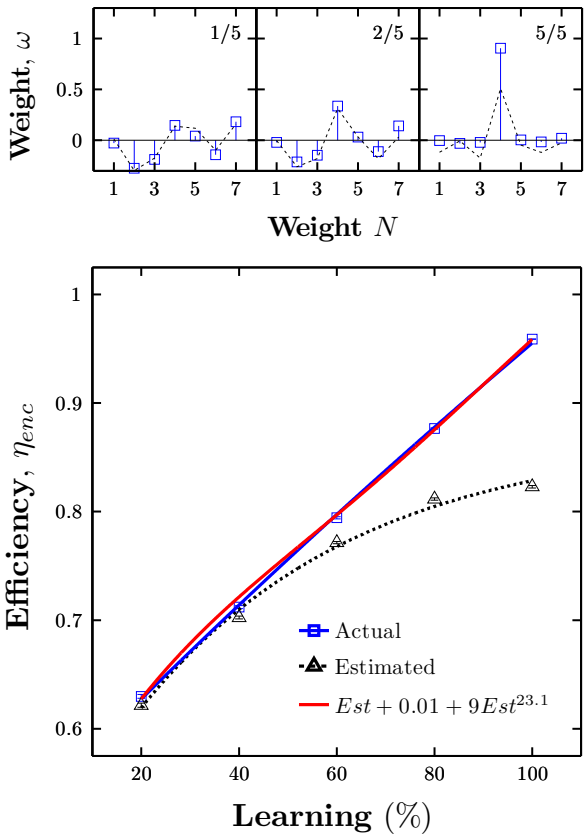


Fig. 3.11: Simulated weight estimation during linear learning. Same format as Fig 3.10.

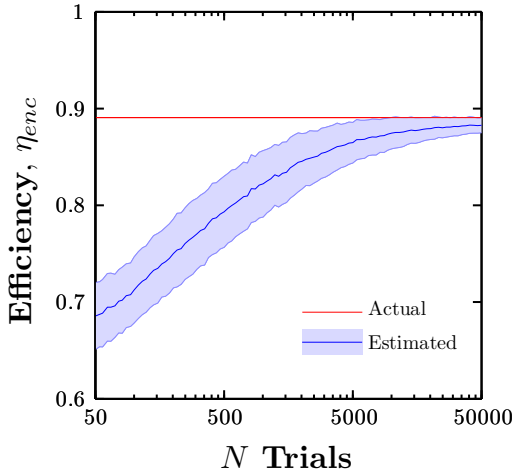


Fig. 3.12: Simulated weight estimation as a function of sample size. Estimated and actual efficiency as a function of N Trials [plotted on a semilog(x) axis]. Efficiency was derived from weight estimates, computed independently using 100 N Trials values uniformly log spaced between 50–50000. The solid blue line is the estimated mean ($\pm S.D.$) efficiency from 1000 simulations. The dashed red line shows the true encoding efficiency, which was held constant at $\eta_{enc} = 0.88$.

CHAPTER 4

Development: Evidence that learning recapitulates ontogeny

This study examined whether the mechanisms subserving developmental improvements in psychoacoustic performance are the same as those responsible for perceptual learning in adults. Over three experiments, 143 school-aged children (4 – 11 years) and 15 adults performed a two-interval two-alternative forced choice [2AFC] tone detection task, in which a 1 kHz tone was simultaneously masked by a spectrally unpredictable multitone complex. Measures of childrens' 'attentional filters' were constructed from relative-weight coefficients derived through reverse-correlation. Indices of further constructs (internal noise magnitude, bias, inattentiveness) were derived from psychometric fits. Levels of masking were found to decrease with age, reaching adult-like performance by 12 years. These changes were explained by improvements in selective attention, with older listeners exhibiting more narrowly tuned attentional filters. Consistent with this, age-related differences in masking were abolished when the masker components were constrained to be more spectrally distal from the target. These changes appear to recapitulate the learning effects observed previously in normal hearing adults (Chapter 3).

4.1 Introduction

THERE has recently been considerable interest in perceptual learning as a tool for enhancing or remediating the auditory abilities of children (Halliday et al., 2008; Merzenich et al., 1996; Tallal et al., 1996). The potential utility of such an approach depends on the degree of commonality between the mechanisms of learning and development. If what changes during learning is the same as what limits performance in children, then perceptual learning may afford a powerful, non-invasive technique for

improving children's hearing (cf. Moore, in press). Conversely, if the mechanisms of learning and development differ, then perceptual learning may be of little utility in children. Here we examine the extent to which children's auditory judgements, given an ecologically relevant task, are limited by the same factors that underlie perceptual learning (Chapters 2 & 3).

An essential everyday task for the auditory system is to extract relevant information from a complex and unpredictable input, such as a teacher's voice in a noisy classroom. This task is made particularly difficult by the *similarity* of signal and noise (Kidd et al., 1994; Durlach et al., 2003b; Lee and Richards, 2011), and also by the *unpredictability* of the noise (Neff and Callaghan, 1988; Tang and Richards, 2003). Adult detection limens deteriorate markedly under such conditions, and in children this effect is even more pronounced (Wightman et al., 2010). Here we modelled this task using a tone in spectrally-unpredictable multitone noise detection task, similar to that used previously in pre-school children (Wightman et al., 2003; Oh et al., 2001), older children (Lutfi et al., 2003), normal-hearing adults (Neff and Green, 1987; Neff and Dethlefs, 1995; Oh and Lutfi, 1998), and hearing impaired listeners (Alexander and Lutfi, 2004). This psychoacoustical task captures the essential elements of similarity and unpredictability, but is easier to parametrically manipulate than speech stimuli.

In adults, performance on such a task has been previously demonstrated to improve with practice (Chapter 3). These improvements were shown to represent increased encoding efficiency, manifesting as effective external-noise reduction. After training, listeners gave more weight to the target spectral region, and relatively less weight to the noise regions. In effect, listeners were learning to selectively attend (or: 'tune their attentional filter') to the signal component, and ignore the task-irrelevant noise components.

There is some evidence that children may be similarly limited by their ability to selectively attend to task-relevant, spectrally distributed information. For example, Oh et al. (2001) found that children's threshold performance was well described by a model in which the free parameters were the number and (spectral) range of auditory filters over which information was integrated (cf. Lutfi, 1993; Oh and Lutfi, 1998). The range of integration was determined by the width of a rectangular 'window of attention', which was observed to be substantially greater in children [$\mu \approx 7$ kHz] than in adults [$\mu \approx 1.5$ kHz][1]. That a common factor limits decision-making in both children and adults is also consistent with Lutfi et al. (2003). Therein, the authors modelled unpredictable-masking performance using

data from both children and adults, and observed that a single principle component was able to explain the majority of the variance. However, other authors have conjectured that children may be limited by mechanism distinct from selective-attention. For example, Buss et al. (2006) suggest that children are impaired by greater internal noise magnitude in the auditory system (a reduction of which was also observed amongst some listeners in Chapter 3), while Viemeister and Schlauch (1992) argue for the importance of non-sensory factors, such as levels of bias or inattentiveness.

The present study aimed to establish whether children are limited in their ability to selectively attend to spectrally distributed information. Experiment I consolidated previously observed age differences in unpredictably-masked detection performance. Masked detection limens were measured in adults and school-aged children (4–11 y.o.). Detection limens were found to be substantially increased at 4 y.o., but converged on adult-like performance by 11. Experiment II examined the factors underlying these age-related differences. Reverse correlation was used to estimate the relative weight listeners gave to binned spectral regions, and the profiles of the resultant attentional filters were compared across age groups. Psychometric fits were also used to derive indices of additional decision-efficiency-limiting constructs, such as internal noise magnitude and bias. The results indicated that differences in selective attention were (solely) responsible for the developmental differences in masking. Finally, Experiment III tested a key prediction from Experiment II by measuring performance as the spectral similarity between target and masker was manipulated. Differences in masking were abolished when a wide spectral notch was employed, consistent with the notion that younger children are limited by a broader attentional filter.

4.2 General Methods

Here we describe those methods that were common across all three experiments; aspects of the listeners and stimuli that differ between experiments are discussed in the context of the relevant experiment. The methods are similar to those used in Chapter 3, with the notable addition that the signal was cued (in quiet) prior to every trial. The use of a cue follows Wightman et al. (2003), and was intended to encourage listeners to use a consistent listening strategy (and specifically to militate against listeners forgetting what the signal was as the adaptive track approached the listener's threshold).

4.2.1 Listeners

Listeners were school aged children (4–11 y.o.) and normal hearing adults. Each experiment used an independent cohort of listeners. The children were recruited through the Nottingham University 'Summer Scientist' event, in which children are invited to attend the University to participate in a number of scientific studies[2]. This event was advertised through schools and newspapers in the local area, and resulted in children from a range of socioeconomic backgrounds [Deprivation Index; $Q_{.25} = 53\%, Q_{.75} = 94\%$]. Children were not screened in advance, but listeners with 1 kHz pure tone thresholds $>$ 20 dB HL were excluded post hoc[3]. Adult listeners were recruited through advertisements placed around the Nottingham University campus, and received an inconvenience allowance for their time. All adult listeners had normal hearing, as assessed by audiometric screening ($\leq$ 20 dB HL bilaterally at 0.25 kHz to 8 kHz octaves), administered in accordance with BSA standard procedure (British Society of Audiology, 2004). Written consent was obtained from all participants (adults) or the responsible caregiver (children), and the study was conducted in accordance with Nottingham School of Psychology Research Ethics Committee approval.

4.2.2 Stimuli & Apparatus

The target signal in all conditions was a 1 kHz sinusoid, 370 ms duration (including ramps), gated on/off by 20 ms cos^2 ramps. In signal+noise conditions, an N-component multitone noise was presented simultaneously with the target tone, and both signal and noise were gated together. The frequency, phase, and amplitude of each component of the multitone complex were independently randomised in each interval (i.e., within-trials). Phases and amplitudes followed rectangular and Rayleigh random distributions, respectively. To minimise energetic masking, distractor components were not permitted to fall within a rectangular band centred at the signal frequency. The level of target tone varied between 0–80 dB SPL, according to an adaptive track. The masker was presented at an average total level of 60 dB SPL.

Stimuli were digitally synthesised in Matlab v7.4 (2007a, The MathWorks, Natick, MA) using a sampling rate of 44.1 kHz and 24-bit quantisation. Digital-to-analogue conversion was carried out by an external USB sound card (Experiment I & II: Custom built in-house hardware. Experiment III: M-Audio Fast Track Pro), interfaced via the Psychophysics Toolbox v3 (Brainard, 1997; Pelli, 1997) ASIO wrapper (Steinberg Media Technologies, Hamburg). Stimuli were presented monaurally to the left ear only, via Sennheiser HD 25-I headphones.

Adults were tested individually in a double-walled sound-attenuating booth. Children were tested in a single-walled sound-attenuating booth. In a minority of occasions the child was accompanied by a caregiver, who sat outside the listener's field of vision, and who, like the experimenter, was blind to the stimuli. With both children and adults, the experimenter was present throughout to provide instruction and encouragement.

Listeners responded by pressing one of two buttons on a button box, while graphical elements (see below) were presented on an LCD monitor.

4.2.3 Procedure

The task was cued, two-interval, two-alternative forced choice [2I2AFC], fixed-frequency tone detection, couched as a game in which the player must "listen for where the special [1 kHz] alien sound is hiding" , and "ignore the clones" (Fig 4.1). During the task explanation, listeners were required to successfully complete one practice trial in quiet and three practice trials in noise [signal + distractor complex]. To highlight the task demands, the target was here fixed at ceiling (66 dB SPL), stimulus durations were increased to 800 ms, and total noise level was reduced to 45 dB SPL.

Each trial consisted of a 400 ms presentation of the tone in quiet (the cue), followed by a 700 ms pause, and two 370 ms stimulus observations separated by a 500 ms interstimulus interval. Listeners were then given an unlimited time to respond via a button box, before being presented with 1000 ms of visual feedback in the form of 'happy' or a 'sad' smiley face, together with a corresponding auditory stimulus. The order of noise and signal+noise intervals was independently randomised on each trial.

In each block, a two-down one-up adaptive track (Levitt, 1971) was used to measure detection threshold, either in quiet or in the presence of the multitone masker. The level of the target tone was initialised at 66 dB SPL and adapted by an initial step size of 6 dB, reduced to 3 dB after the second reversal[4]. Each block consisted of four reversals, or 35 trials (whichever occurred first).

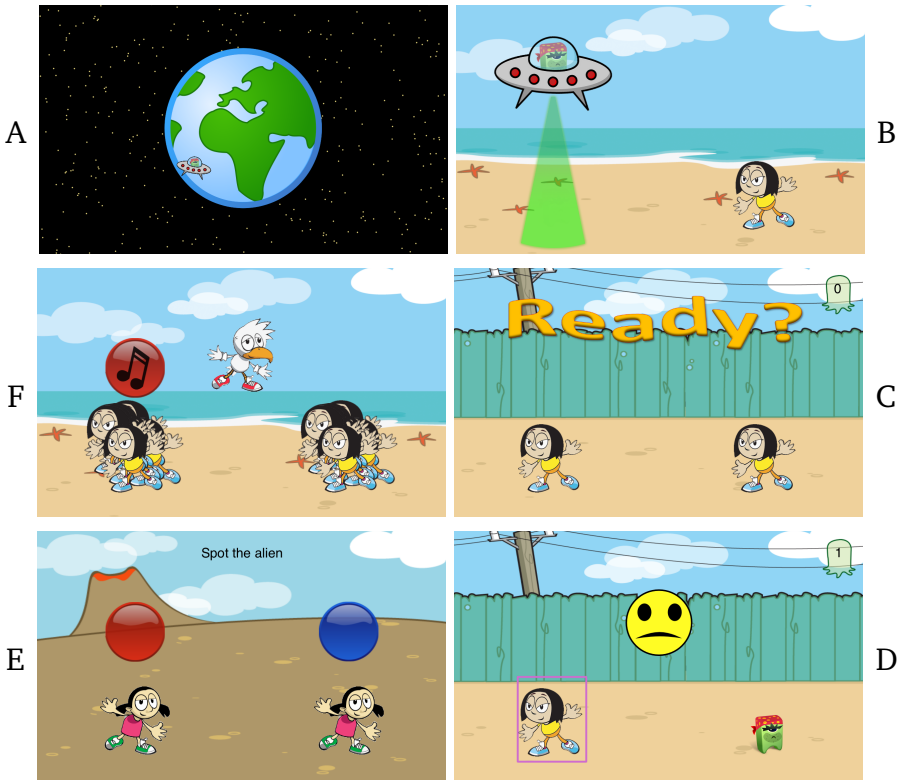


Fig. 4.1: Screenshots of the listening game. Clockwise from top left: [A,B] Children were introduced to the game using a brief animation in which they are told they will be 'catching aliens'. During this time they were introduced to the various characters, and practiced using the buttons and listening for the 1 kHz target tone in quiet and in noise. [C] Each 'level' consisted of a single adaptive track, and featured a new location and new characters. [D] Veridical auditory and visual feedback was given after each response. [E] Responses were made by pressing coloured buttons corresponding to each interval. [F] Before each trial listeners were reminded of the target sound by a bird character, who they were encouraged to name and pay attention to.

4.2.4 Measures

Detection limens, DLs, were determined by averaging the signal level at the last two reversals. Separate adaptive tracks were used to calculate absolute detection limens in quiet, DL_{quiet}, and in noise, DL_{noise}. Masking level (or: *reception threshold*) was calculated as $DL_{noise} - DL_{quiet}$.

A number of additional, more general measures were also taken. Attention deficit hyperactivity [ADHD] was measured using the SWAN Rating Scale (Polderman et al., 2007; Swanson et al., 2005), and socioeconomic status (McLennan et al., 2011), and BPVS vocabulary scores (Dunn et al., 1997) were also assessed.

4.3 Experiment I

The purpose of this experiment was to establish whether or not significant age-related differences in masking could be observed amongst school-age children and adults, given a spectrally unpredictable masker. Such effects have previously been demonstrated for preschool children versus adults (Oh et al., 2001; Wightman and Allen, 1992), and have also been explored in school-aged children by Lutfi et al. (2003) and Leibold and Neff (2007). This experiment used a larger combination of listeners and masker conditions than in previous studies in order to examine in greater depth the changes occurring during childhood. This was necessary to inform the design of Experiments II and III.

4.3.1 Methods

Listeners

Forty-nine children (4.08–11.46 y.o.) and fifteen adults (21.94–32.35 y.o.) participated. Two children (4%) were excluded on the basis of their hearing thresholds in quiet. Six further children (12%) were excluded for not completing all blocks. The remaining children were binned into three age groups: 5–6 $[5.0 < x \leq 6.7; \mu_{age} = 6.14; n = 11]$, 7–8 $[7.2 < x \leq 8.9; \mu_{age} = 8.01; n = 17]$, and 9–11 y.o. $[9.2 < x \leq 11.4; \mu_{age} = 10.15; n = 13]$. The adult listeners constituted a fourth group $[21.9 < x \leq 22.3; \mu_{age} = 25.99; n = 15]$.

Stimuli & Design

Maskers consisted of N distractor tones, selected from the components of a randomly selected bandpassed Gaussian noise (0.1–10 kHz), as per Neff and Callaghan (1988). On each presentation 1 of 50 noise samples was randomly selected and a fast Fourier transform [FFT] was used to decompose the noise into 2.7 Hz-spaced spectral components. Any components falling within a 160 Hz notch arithmetically centred on the target frequency (the 'protected region') were removed. N components were randomly selected, and their amplitude, frequency and phase information was used to synthesise the multitone complex. The candidate tones were thus equally spaced between 100 and 10000 Hz on a linear scale. Selecting all components ($\sim$3500) was equivalent to synthesising a notched white noise.

The number of distractor tones, N, increased in ascending order: $0, 2, 10, 30, 300, 906, \sim 3500$, with each condition being presented in a separate block. The order of the blocks was intended to ensure that task difficulty increased gradually.

4.3.2 Results

The results of this experiment are summarised in Fig 4.2. In quiet, detection limens did not significantly differ across age [Unbalanced 1-way ANOVA; $F_{52} = 1.09, p = 0.360, n.s.$]. To assess differences in masking, a mixed-effects ANOVA was conducted with AGE as a between-subject factor, and N DISTRACTORS as a within-subject factor. To ensure linearity only the first three masker conditions [2 10 30] were included as levels[5]. Significant differences in masking were observed as a function of both N DISTRACTORS [$F_{(2,104)} = 10.98, p < 0.001, \eta_p^2 = 0.91$] and AGE [$F_{(3,52)} = 3.30, p = 0.027, \eta_p^2 = 0.16$]. From inspection of Fig 4.2 it can be seen that the youngest children consistently exhibited the greatest masking, while the adults exhibited least masking.

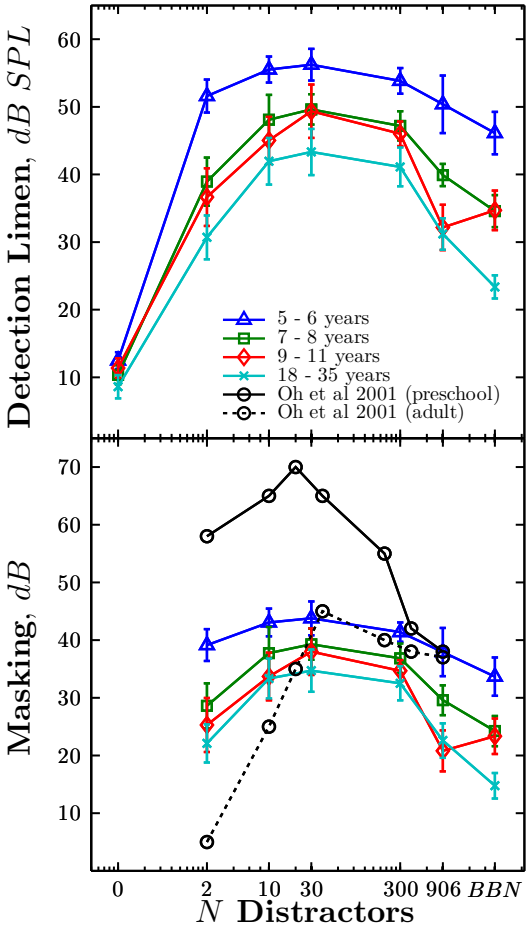


Fig. 4.2: Experiment I: Absolute detection (top) and masking (bottom) limens as a function of N distractors. The broadband (BBN) condition is equivalent to ~3500 distractors. The quiet condition ($N = 0$) is omitted from the bottom panel, since by definition masking in quiet equals zero. The different curves represent different age groups (see key). Error bars represent ± 1 SE (here and in all subsequent figures). Masking data from Oh et al. (2001) are plotted for comparison.

The overall performance of the oldest (9–11) children was indistinguishable from that of adults [$F_{(1,162)} = 1.82, p = .179$], though a post-hoc comparison did indicate a specific difference in the broadband noise condition, where the children were ∼8 dB poorer [$t_{26} = 2.31, p = .029$]. The 7–8 year olds formed an intermediate group in terms of masking levels, but were generally more similar to the older children than the younger children. All age groups exhibited markedly less masking than the preschool children of Oh et al. (2001), the results of which are plotted in Fig 4.2 for comparison. Differences were also observed between the adult group and those in Oh et al. (2001), with the latter exhibiting less masking with sparse multitone complexes ($N = 2, 10$).

As shown in Fig 4.3, there was substantial individual variability in masking levels between adult listeners (similar variability was also observed in younger listeners, cf. Appendix 4.A). Moreover, there was also substantial variability in masking across conditions, with the maxima varying in both height and location. This resulted in a flattening of the mean masking curve relative to those for individuals. Thus, the standard deviation of the mean curve across N distractors was significantly less than for individual listeners [one sample t-test; $t_{14} = 4.30, p < 0.001$].

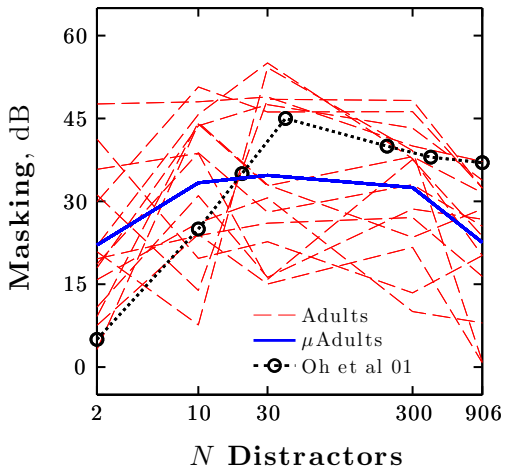


Fig. 4.3: Experiment I: Individual adult masking profiles as a function of N distractors. Red dashed lines show individual listeners. The flatter, solid blue line indicates the group mean masking function. This figure largely follows the format of Fig 4.2.

4.3.3 Discussion

Pure tone detection was significantly impaired by the simultaneous presence of spectrally-uncertain maskers, and this effect was greater in younger children ($\mu_{age} = 6.1$) than in either of the older age groups ($\mu_{age} = 8.01, 10.2$). In unpredictable noise, mean performance in the

9–11 y.o. age group was indistinguishable from that of adults, suggesting that the ability to filter-out unpredictable distractors is largely mature by adolescence. Even in the youngest group, masking levels (i.e., reception thresholds) were considerably smaller than in the preschool children of Oh et al. (2001), indicating that listeners' decision-making faculties develop substantially during early childhood.

The data support a number of previous findings. Group-mean masking curves varied non-monotonically as a function of N Distractors, with maxima at 20–40 components. This pattern is consistent with with the results from a number of previous studies (e.g., Oh and Lutfi, 2000; Oh et al., 2001; Lutfi et al., 2003). Adults exhibited approximately 12 dB *less* masking than in Chapter 3, where the target was not cued, indicating that a pretrial target cue confers a 12 dB release from masking. This is consistent with Richards and Neff (2004), who observed that pre-trial cuing improved detection limens by up to 18 dB ($\mu = 5.4$ dB), given a similar task (2–10 component random-frequency masking). With respect to the developmental trajectory, 5–6 years old exhibited 9.7 dB more masking than adults, while 7–8 years old performed at an intermediate level ($+$ 4.4 dB). These developmental differences are similar to, albeit slightly smaller than, those observed by Leibold and Neff (2007), who used a 10-tone random-frequency masker, and found increased masking of around 12 dB and 6 dB in 5–7 and 8–10 years old, respectively.

Conversely, some aspects of the data deviate from those reported previously. Within each age group, masking varied less with N Distractors (i.e., were 'flatter') than in Oh et al. (2001) (Fig 4.2). As reported in the results, masking curves tend to become flattened when averaging over multiple profiles with large inter-individual variability. The relative flatness of the present mean masking curves may therefore be an artefact of averaging over larger sample sizes (around twice those of Oh et al., 2001). Alternatively, order effects may also have had a homogenising effect on observed masking. Thus, the sparsest distractors always occurred earlier in the session, when listeners were least practiced at the task. Since learning is known to occur on this task (cf. Chapter 3), masking may have been inflated at low N Distractor levels, where masking is typically low, and deflated at higher N Distractor levels, where masking is typically high. If this was the case then one would expect masking to be greater than Oh et al. (2001) at low N, and lesser at high N. The observed data were consistent with this pattern (cf. Fig 4.3).

The substantial age-related differences in broadband noise masking may also represent an order effect, with younger children becoming disproportionately fatigued towards the end of the session. However, we

cannot rule out a genuine age-related differences in broadband masking, and it is certainly not uncommon for individual children to exhibit elevated masking even in spectrally predictable noise (e.g., Leibold and Bonino, 2009). We return to this question in Experiment III, the results of which favour the former interpretation.

Absolute detection thresholds in noise were smaller than some that have been observed previously. These differences are of less import than the relative masking effects across age, and in many cases may be explained by vagaries in the design (e.g., more spectrally proximal distractors; Leibold and Neff, 2007).

4.4 Experiment II

This experiment investigated *why* younger children exhibit greater masking. Estimates of encoding efficiency, internal noise magnitude, bias, and inattentiveness were compared across age groups. Since Experiment I indicated relatively little difference in masking between 7–8 and 9–11 years old, in Experiment II we compared only 'younger' (4–7 y.o.) and 'older' (8–11 y.o.) school-aged children.

Listeners

Fifty-nine children participated. Nine children (15%) were excluded on the basis of poor hearing thresholds in quiet. Two further children (3%) were excluded due to not completing any blocks in noise. The remaining children were binned into two groups: 4–7 $[4.3 < x \leq 7.8; \mu_{age} = 6.52; n = 28]$, 8–11 $[8.0 < x \leq 11.1; \mu_{age} = 9.15; n = 20]$. The first, younger group roughly corresponded to the youngest age in Experiment I. The mean age of this group was marginally greater than in Experiment I (0.38), but this difference was not significant $[t_{37} = -1.16, p = 0.253, n.s.]$[6].

Stimuli & Design

The stimuli were the same as those in Experiment I, with the following exceptions. The range of distractors was restricted to 223–4490 Hz, and the frequencies were drawn from 715 values uniformly distributed on a log scale. The use of a log scale was to ensure that the signal was geometrically centred in the complex, and so as to discourage listeners from using the overall pitch of the complex as a cue. The notch was a $\frac{1}{3}$ octave in width, and was geometrically centred on the target tone (891–1120 Hz). Since fewer distractor tones were used, the distractor complex was constructed simply by summing together 30 pure tones of random frequency, phase and amplitude.

Only the quiet condition and the $N = 30$ condition were used. The first block was always in quiet. Listeners then completed as many masked tracks as they felt comfortably able to, up to a maximum of eight ($\mu = 5.0 \pm 1.6$).

Analysis

The methods were similar to those described in Chapter 3. Relative weights were derived for each individual by correlating listeners' responses with the trial-by-trial (inter-interval) difference in intensity at each $\frac{1}{3}$ octave bin [MATLAB'S GLMFIT]. Encoding efficiency was calculated by computing the RMS difference between the observed weight vectors and the ideal: $[0\ 0\ 0\ 1\ 0\ 0\ 0]$. Mean weight-vectors were computed for each age group by averaging over each individual's signed weight coefficients. This average was weighted relative to the number of trials that each listener completed. This was done in order to maximise overall accuracy, under the assumption that listeners who completed more trials yielded more reliable behavioural estimates.

Further parameters were derived from psychometric fits to the probability of responding 'Interval 2', as a function of the decision variable, DV:

$$DV = \sum_{i=1}^{n} \omega_i \Delta L_i, \tag{4.1}$$

where ΔL represents the difference in level at that spectral bin, and ω is the relative weight coefficient from the associated group-mean weight functions. The psychometric fits were cumulative-Gaussian, and constrained to fall within $0.05 - 0.95$. Internal noise magnitude, inattentiveness and bias were estimated from the standard deviation, lapse-rate and constant error of the fits, respectively, in the manner described previously in ∮2.2.4 and ∮3.2.4.

4.4.1 Results

As shown in Fig 4.4, the pattern of performance followed that of Experiment I. No significant difference in detection limens was observed between age-groups in quiet [$t_{46} = 0.89, p = 0.379, n.s.$], but younger children exhibited significantly greater masking [$t_{46} = 2.86, p = 0.006$].

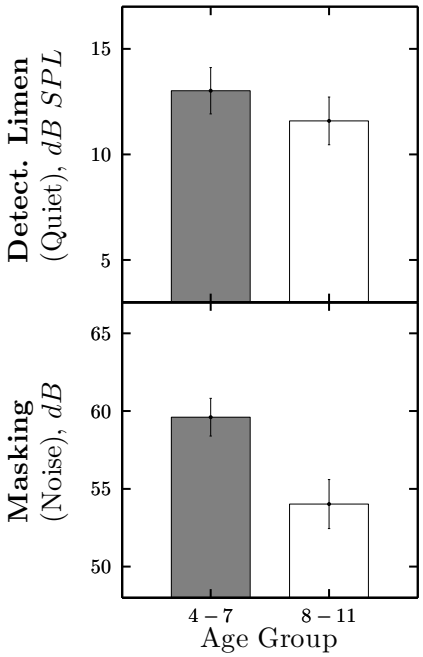


Fig. 4.4: Experiment II: Detection and masking thresholds for younger and older children.

To explore why younger children were more adversely affected by noise, spectral-weights were calculated for each individual (Appendix 4.B). The (weighted) group-means of these weight functions are shown for younger and older children in Fig 4.5[7]. Younger children exhibited a flatter profile, giving greater relative weight to the spectral regions flanking the target (1 kHz). Conversely, the older children produced a mean weight function that was closer to the ideal, and in good agreement with the naïve adult data presented in Chapter 3, the results of which are reproduced in Fig 4.5 for comparison. Encoding efficiency, η_{enc}, was significantly [$t_{46} = 3.9, p < 0.001$] higher in the older [$\mu = 0.64$] children than the younger children [$\mu = 0.60$].

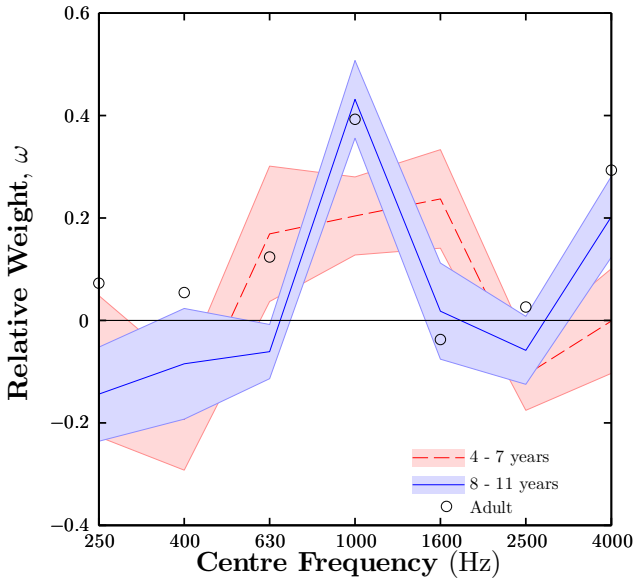


Fig. 4.5: Experiment II: Weight vectors for younger (red, dashed) and older (blue, solid) children. Bold, coloured lines indicate mean weight coefficients. Shaded regions indicate $\pm$ 1 SE, and are linearly interpolated between each point. Circles are the mean, normalised weight coefficients from the naïve adults in Chapter 3.

Psychometric functions were fitted to the trial-by-trial DV values, computed as per Eq 4.1 using the estimated weights (cf. Fig 4.5). No significant differences in internal noise magnitude, inattentiveness or bias were observed between age groups [all $p \geq 0.265$]. This result continued to hold after those psychometric fits of questionable validity (i.e., those best fitted by negative, $\sigma < 0$, or very large, $\sigma > 500$, slopes) were excluded [all $p \geq 0.178, n.s.$]. Fits for the remainder of the (valid) data are shown for both age groups in Fig 4.6 (top panels), together with the more traditional fits made to the unweighted signal level (bottom panels). The effects of using the weighted stimulus data, rather than simply the signal level, when fitting psychometric functions were two-fold. Firstly, in both sets of listeners the psychometric functions became substantially less variable between individuals [both $p < 0.001$][8]. This indicates that some of the (considerable) individual variability in unpredictable-masker detection tasks can be accounted for by the trial-by-trial fluctuations in the (weighted) noise. Secondly, when the weights were not used, younger children appeared to have greater internal noise [$t_{32} = 2.69, p = 0.011$], as manifest in their shallower psychometric slopes.

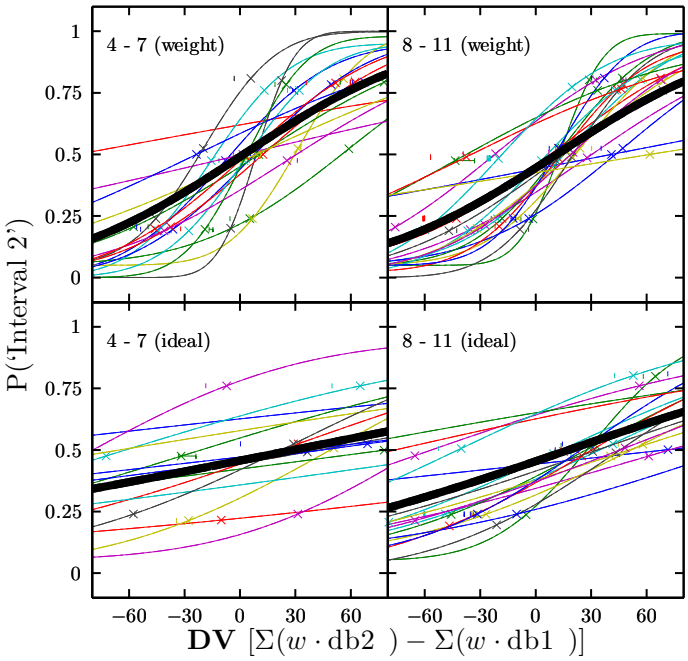


Fig. 4.6: Experiment II: Psychometric functions for younger (left) and older (bottom) children. The top panels are fits to the trial-by-trial data, weighted by the group-mean weights given in Fig 4.5. The bottom panels are fits to the trial-by-trial data given the ideal weight vector $[0\ 0\ 0\ 1\ 0\ 0\ 0]$ (i.e., fits to signal level). Fits with negative slopes or extremely high σ parameters were deemed invalid and are not shown (see body).

4.4.2 Discussion

The results showed that younger children (< 8 y.o.) differed from older children in terms of the efficiency with which they weighted spectral information. No differences in internal noise magnitude, bias or general inattentiveness were observed. Weights can be intuitively thought of as indexing the amount of attention paid to each information channel (Berg and Green, 1990). Accordingly, these data indicate that younger are primarily limited by their ability to selectively attend to information distributed within a narrow spectral range. In particular, they appear to be unable to 'filter out' information similar in frequency to the target sound, with younger children exhibiting elevated weights in the regions flanking the target. The fact that the developmental differences in encoding weights were localised proximal to the target is consistent with the notion from Lutfi (1993) of an 'attentional band' (see also Green, 1958) that narrows during development.

The concept of an attentional band predicts that masking will rise as the noise becomes increasingly proximal (e.g., in frequency) to the signal. This appears to be the case. Thus, substantially more masking was observed

here than in Experiment I, which used a wider range of maskers (~ 4.5 versus ~ 5.6 octaves). While substantially less masking was observed than in Leibold and Neff (2007), who used an even narrower noise range (3.3 octaves), and listeners of very similar age [$\mu_{age} = 6.6, 9.0$] to those reported here [$\mu_{age} = 6.5, 9.2$]. Experiment I and Leibold and Neff (2007) both used maskers that were spectrally centred on a linear scale. This may have also affected masking levels relative to Experiment II, where the maskers with sampled from a logarithmic scale. However, it cannot explain why masking was reduced in Experiment I, but increased in Leibold and Neff (2007).

The weight vectors in Fig 4.5 predict that the developmental differences in masking would be attenuated if the protected region around the target were increased. We tested this prediction in Experiment III.

4.5 Experiment III

Masking levels were measured in younger and older children using four, progressively wider protected regions. This experiment may be considered an attentional analog of the 'notched noise' method used to estimate auditory filter shapes (Glasberg and Moore, 1990). It was predicted that with narrow protected regions younger children would be disadvantaged relative to older children (replication of Experiment II), but that this disadvantage would diminish as the region increased (i.e., as the noise was progressively limited to regions that younger and older children weight equally).

A potential confound of this approach arises from the fact that, if the range and number of maskers are held constant, wider protected regions will result in less masker variability. Masker variability has been found to be a critical determinant of the amount of masking (Lutfi, 1993; Oh and Lutfi, 1998). Moreover, changes in variability appear to interact with age, such that under low variability conditions, even very young children perform distinguishably from adults (Oh et al., 2001). Increasing the protected region may also abolish the developmental differences observed in Experiment II, via a reduction in variability, independent of the listener's underlying encoding strategy *per se*. Accordingly, in this experiment the number of maskers was covaried with protected-region width, so as to maintain an approximately constant standard deviation of dB–energy in each spectral bin.

Potential differences in peripheral auditory filter widths pose a second confound. Thus, if younger children have broader peripheral filters, then this would also manifest as a progressive convergence in masking levels as the width of the protected region increases, irrespective of any

developmental differences in attention. In fact, substantial developmental differences in auditory filter widths are unlikely given that the human peripheral auditory system appears anatomically (Pujol et al., 1991) and functionally (Schneider et al., 1990) well-developed by childhood. Nonetheless, to exclude such an explanation, masking was also assessed using a broadband noise masker. If younger listeners have wider auditory filters then they should exhibit greater masking in broadband noise. This condition also provided a potential indicator of whether younger listeners were paying attention in the noise condition.

4.5.1 Methods

Listeners

Seventy-nine children participated. Twelve listeners (15%) were excluded on the basis of their hearing thresholds in quiet. Thirteen further listeners (16%) were excluded for completing fewer than three blocks. This exclusion rate was higher than in previous experiments, due principally to a reduction in the amount of testing time permitted per child. As a result, children were asked if they wished to continue after the second block, and only those willing to proceed continued. This minimised the number of breaks required between blocks. The children were binned into two groups: 4–7 y.o. $[4.4 < x \leq 8.0; \mu_{age} = 6.6; n = 38]$ and 8–11 y.o. $[8.0 < x \leq 11.3; \mu_{age} = 9.7; n = 16]$.

Stimuli & Design

The stimuli were the same as those in Experiment II, except for the size of the protected region, which was manipulated between blocks. A notched broadband noise (BBN) condition was also added, analogous to that in Experiment I (a white Gaussian noise with a brick-wall notch at 891–1120 Hz).

The block conditions are shown in Table 4.1. The independent variable of interest was the width of the protected region. However, the number of masker components, N, was covaried so as to maintain a constant level of energetic variability in each $\frac{1}{3}$ octave bin. In every listener the first block was always in quiet. The order of the remaining blocks was then randomised. Most, but not every, child completed every block (41 of 54).

	Noise Range (Hz)		Protected Region (Hz)	
N Components	lower	upper	lower	upper
0 (*Quiet*)		$n.a.$		
15	223	4490	891	1122
20	··		707	1414
25	··		561	1782
30	··		445	2245
715	··		891	1122

Table 4.1: Experiment III stimulus conditions. Tones drawn from a log-uniform distribution within noise range, except for protected region.

4.5.2 Results

In quiet

No difference in detection limens in quiet was observed between age groups [$t_{52} = 0.41, p = 0.683, n.s.$].

In unpredictable (multitone) noise

From inspection of Fig 4.7a, the pattern of results was similar in the two widest and two narrowest protected regions. Therefore, to simplify comparisons, masking levels were mean aggregated across these conditions. The results are shown in Fig 4.7b. Masking levels were similar when using wide notches [$t_{50} = -0.09, p = 0.927, n.s.$], but younger children exhibited significantly greater masking when the notch was narrow [$t_{48} = 2.19, p = 0.033$]. A t-test was used to evaluate whether the *difference* in masking across condition significantly differed between the two age groups. This difference was not significant [$t(46) = -1.40, p = .168, n.s.$]. Accordingly, we cannot conclude that masking was significantly attenuated by the increase in notch width, though (as predicted) the younger listeners were only significantly poorer in the narrow notch condition.

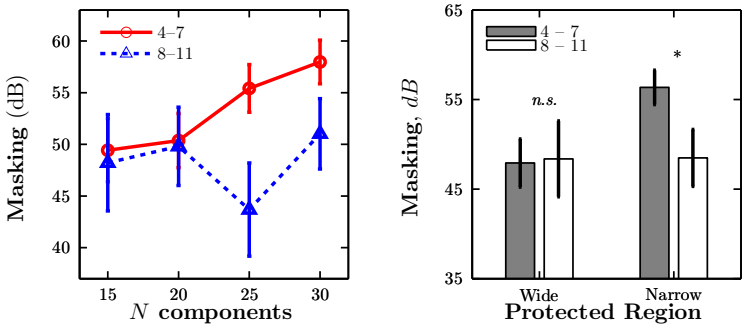


Fig. 4.7: Experiment III: Masking levels across notch conditions in younger and older children. (Left) Grand-mean ($\pm$ 1 SE) as a function of N components (N components $\propto$ notch width; cf. Table 4.1). (Right) Same data, aggregated over the two widest and two narrowest notches.

In predictable (broadband) noise

No difference in detection limens in broadband, notched-noise was observed between age groups [$t_{45} = 0.27, p = 0.789, n.s.$], indicating that energetic masking is not elevated in younger children.

4.5.3 Discussion

These data corroborate the patterns of encoding weights derived in Experiment II. When the protected region was as narrow as in Experiment II, younger listeners again exhibited greater masking by unpredictable-noise. In contrast, when the protected region was increased (i.e., so as to contain the spectral regions in which younger listeners exhibited less optimal weights), no differences were observed between older and younger listeners. This is consistent with our interpretation of Experiment II, that younger listeners are primarily impaired by their ability to filter out spectrally proximal, unpredictable noise.

The lack of an age-related difference in masking when the masker was predictable (broadband) suggests that differences in masking by unpredictable noises cannot be accounted for using a traditional 'sum energy detector' model. It also indicates that the developmental differences in broadband masking observed in Experiment I were an artefact of the testing procedure.

4.6 Individual Differences

Unpredictable masking is characterised by large variability between individuals (e.g., as much as 60 dB; Neff and Dethlefs, 1995). To quantify how much variance is explained by differences in age, and to

explore other potential predictors of performance, a number of regressions were performed using the data from all three experiment. Masking was computed from the $N = 30$ condition in each experiment. The results are summarised in Table 4.2. As expected, age was a significant predictor [$p = 0.003$], with masking thresholds decreasing by around 2.5 dB per year. However, contrary to Oxenham et al. (2003), no effects of sex were observed [$p = 0.978, n.s.$]. Socioeconomic status [SES], vocabulary [BPVS], and level of attention deficit hyperactivity [ADHD] were similarly poor predictors [$p \geq 0.268, n.s.$]. That even 1% of variance on a basic psychoacoustic task is explained by socioeconomic status may be of note, but this result did not reach significance. Overall, the vast majority of variance (94%) remained unaccounted for.

Predictor	p	β	R^2	F	range
*Age	0.003	-2.39	0.06	9.24	$4.3 - 11.5$
Sex	0.978	0.07	0.00	0.00	$0.0 - 1.0$
SES	0.268	4.70	0.01	1.23	$0.0 - 1.0$
BPVS	0.783	-0.02	0.00	0.08	$0.0 - 99.0$
ADHD	0.814	0.38	0.00	0.06	$-2.9 - 1.8$
Full Model	0.104	-2.29	0.07	1.87	

Table 4.2: Regressions statistics for predictors of masking in listeners aged $4.33 - 11.46$ y.o.. All listeners from the three experiments reported here were included in these analyses. For each of the five independent variables the regression was performed independently. A multiple-regression model was also run using all five independent variables.

4.7 General Discussion

This study demonstrates that tone detection in unpredictable noise improves between 4–11 y.o., by which point performance is broadly adult-like. This difference was explained by superior selective attention (encoding efficiency) in older children (8–11). Younger children (4–7) were less able to 'filter out' noise that lay spectrally proximal to the target tone. As Leibold and Neff (2007) note, this developmental trend is consistent with a wide literature indicating that attentional faculties development progressively within the first seven years of life (e.g., Ruff and Rothbart, 1996).

The mechanism underlying development appears the same as that driving perceptual learning in adults. Thus, in Chapter 3 naïve adults exhibited similar encoding weights to the older children here, which they were able

to optimise through practice. After practice adults exhibited significantly more efficient weights, and a 3.1 dB mean decrease in masking.

It remains an open question as to whether children would exhibit similar reductions in masking with practice. Halliday et al. (2008) demonstrated within-session frequency-discrimination learning in 6–11 y.o. listeners, but this was largely restricted to those listeners with initially poor performance. Conversely, Huyck and Wright (2011) had 11 y.o. children practice a temporal-interval discrimination task, and observed a *deterioration* in performance over 10 sessions. It may therefore be that while basic sensory faculties are largely mature by puberty, additional factors continue to limit childrens' ability to exploit supervised reinforcement signals to optimise their decision making.

Although there was substantial individual variability, levels of internal noise magnitude, inattentiveness or bias were not observed to differ consistently across age groups. Notably, this was not the case when differences in encoding weights were not taken into account, whereupon internal noise magnitude was significantly greater in younger listeners (and was generally inflated in all age groups). This suggests that, within the developmental literature, internal noise may have been overestimated, both in terms of its magnitude, and its importance during the maturation of hearing. This conclusion is *prima facie* inconsistent with Buss et al. (2006), who argued that internal noise accounts for developmental differences in pure tone intensity discrimination. However, this disparity may reflect differences in task demands. In the present study, performance was constrained by the amount of effective external noise, which listeners were able to regulate through the optimisation of attentional weights. In contrast, when external noise is negligible, as in Buss et al. (2006), listeners may be limited by internal sources of noise. This would parallel analogous results in the perceptual learning literature (e.g., compare chapters 2 and 3).

4.8 Conclusions

(1) Tone detection in unpredictable, multi-tone noise improves between 4–11 y.o., by which point performance is adult-like.

(2) Older children's decreased masking is due to improved selective attention, with younger listeners disproportionately failing to filter out noise falling in the spectral proximity of the target. This mechanisms is the same as that previously observed to underlie learning in adults (cf. Chapter 3).

(3) No developmental changes in internal noise magnitude, inattentiveness or bias were observed.

Notes

[1]As discussed in Oh et al., 2001, these estimates of window-widths are rough estimates only.

[2]The salient point is that the children were not in attendance specifically to participate in this study. As such, some children, especially those tested in the latter portions of the day, were slightly fatigued and/or did not have sufficient time to complete the task. A number of children also participated who had diagnosed hearing impairments. Both of these facts are reflected in the relatively high exclusion rate.

[3]Single-tone screening has been demonstrated to provide a relatively robust method of screening for hearing impairments; (Maxwell and Davidson, 1961).

[4]A failure to accurately correct for non-linearity in the level calibration meant that the step sizes were inflated at low levels ($<$ 20 dB SPL) in Experiment I. This did not appear to have a substantive impact on the results, and was corrected in Experiments II and III.

[5]No effects were substantively altered by including additional masker levels [p values decreased in all cases].

[6]The mean age of the older group fell between the means of the two senior school-aged groups of Experiment I.

[7]Compound weight functions were also estimated by performing a single multiple-regression using all the raw data from each age group. The resultant functions did not differ substantively from those shown in Fig 4.5.

[8]Variability was quantified as the standard deviation of the difference limen [DL] at chance ($P(\text{'Interval 2'}) = 0.5$). Paired-bootstrapping ($N = 2000$) was used to derived 95% Confidence Intervals. The equality of the two standard deviations was tested using the CI and mean of the paired-differences to compute a z-score, which was converted to a p value using the normal distribution (Altman and Bland, 2011).

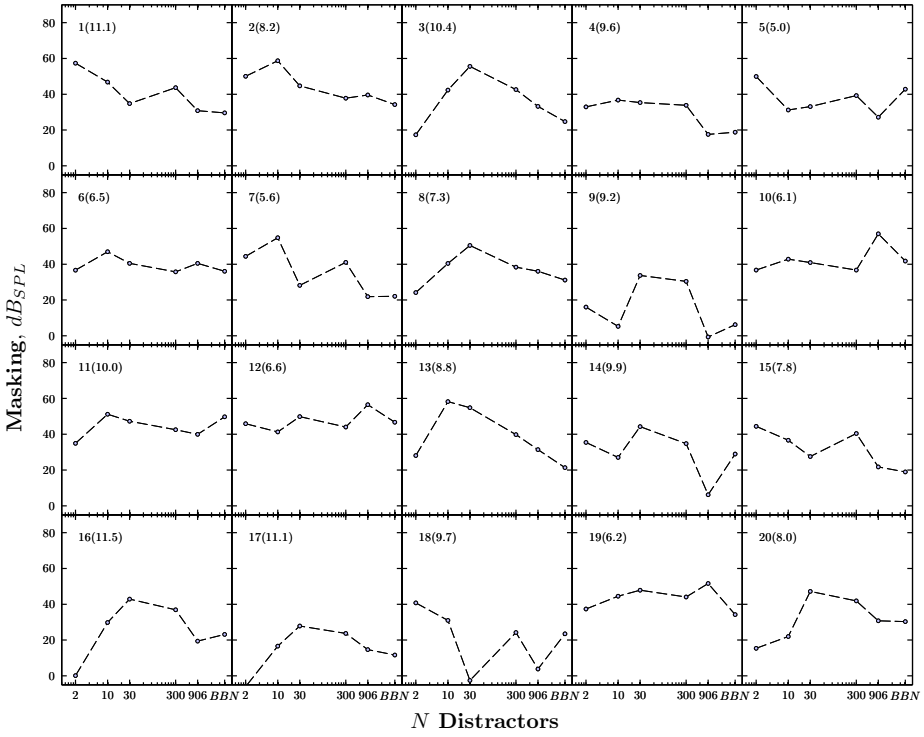
Acknowledgements

We are indebted to the organisers and volunteers of the Nottingham Summer Scientist Week (http://www.summerscientist.org/), who recruited all the children reported in this study. We would also like to thank Natasha Ratcliffe, who assisted with data collection in Experiment III. This work was supported by the Medical Research Council, UK (Grant: U135097130).

4.A Individual masking functions

Individual masking functions (Experiment I). Most individuals exhibited a distinctly non-monotonic profile. In most listeners peak masking occured at

$N = 30$. But in some listeners it occurred at lower (e.g., 37, 55) or greater (e.g., 10, 33) values.



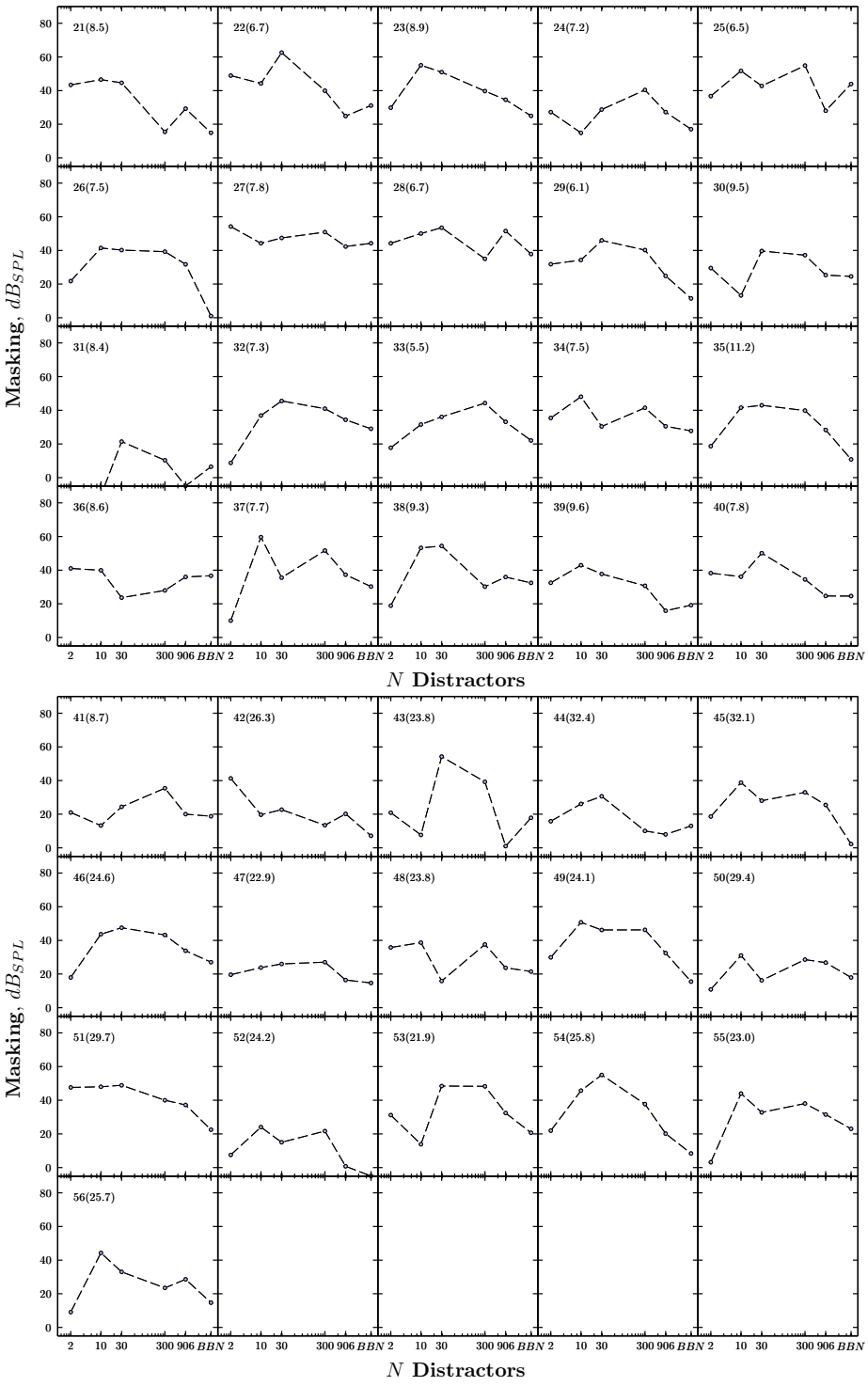


Fig. 4.8: Experiment I: Masking limens for individual listeners as a function of N distractors, for comparison with the group-mean fits given in Fig 4.2. Text annotations give the id(age) of the listener shown in that panel.

4.B Individual weight functions

Fig 4.9 shows individual weights for a selection of listeners (Experiment II). In many cases listeners approximated the ideal strategy (blue circles). Some listeners exhibited a 'near miss' to this approximation, either by systematically weighting a bin other than the target (green squares), or by exhibiting a more broadly tuned function that also gave weight to bins proximal to the target (red crosses). In contrast, some listeners appeared to use distinct strategies. For example, the listener given by black triangles appeared to be more generally responding to the interval with the greater energy at high frequencies.

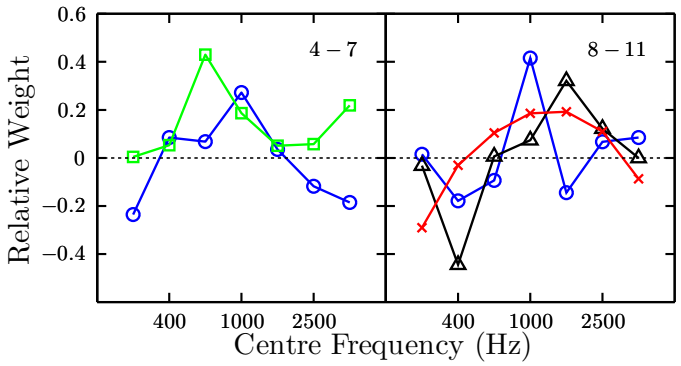


Fig. 4.9: Experiment II: Weight vectors for individual listeners, for comparison with the group-mean fits given in Fig 4.5. See body text for more.

CHAPTER 5

Bias in yes/no (detection) task learning

Auditory perceptual learning is often assumed to reflect an increase in perceptual sensitivity. However, the performance measures typically used to quantify learning do not disambiguate sensitivity from bias. In this study we investigate to what extent learning on an auditory yes/no task represents a decrease in bias. Thirteen normal hearing adult listeners practised a sinusoidal amplitude modulation [SAM] yes/no detection task for seven sessions. Naïve listeners tended to be biased in favour of responding 'yes' (liberal). With practice detection limens consistently decreased, and this was accompanied by a concomitant reduction in bias. We conclude that reductions in bias partially account for the improvements observed in perceptual learning. Simulations suggest around one-third of learning on this task was explained by reductions in bias. Our conclusions differ from recent reports in the visual learning literature, where practice has been suggested to increase bias (Wenger and Rasche, 2006). We demonstrate that these differences arise from a flawed analyses in previous papers.

5.1 Introduction

LEARNED improvements in auditory perceptual ability are typically evidenced by a reduction in the listener's difference limen, DL. However, the DL constitutes a relatively gross measure of the underlying decision process, in that it confounds a number of performance-determining factors (cf. Swets, 1973). One class of factors relates to 'sensitivity', and concerns the efficiency with which information is extracted (i.e., when mapping from the sensory input to the decision variable). Sensitivity is affected by considerations such as internal noise magnitude: the amount of intrinsic variability in the decision process, and encoding efficiency: the ability of the listener to appropriately attend to

the task-relevant information. Changes in sensitivity manifest as changes in the slope of the psychometric function. A second class of factors is primarily non-sensory, and relates to the listener's propensity to respond in a manner independent of the sensory input. Such propensities may result in a systematic preference for a certain response, in which case we term them bias. Alternatively, they may be random, in which case we term them inattentiveness. Changes in bias and inattentiveness manifest as changes in the lateral shift and asymptotes of the psychometric function, respectively. In this study we examined changes in bias, and their role in perceptual learning.

Differences in DL cannot distinguish between changes in sensitivity and bias. This fact is illustrated in Fig 5.1 (see also Appendix 5.C). Listener *A* is more sensitive than listener *B*, but due to *B*'s bias they both exhibit the same DL values at the 74.7% level. Contrawise, listeners *A* and *C* exhibit the same sensitivity, but produce different DL values at the 74.7% level. Improvements in DL alone therefore do not uniquely specify differences in sensitivity or bias, and changes in either may be responsible for learning, in part or in full.

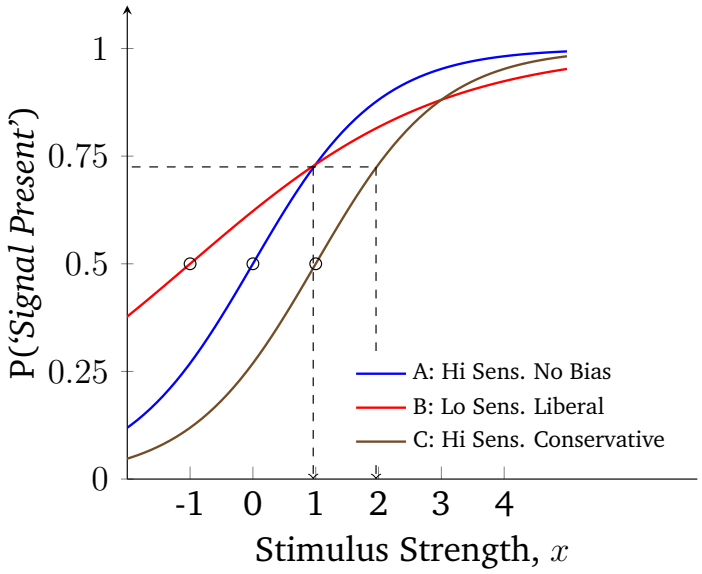


Fig. 5.1: Schema showing the ambiguity of psychometric limens. Each curve is a cumulative normal distribution, $\Phi(\mu, \sigma_{int})$, evaluated at x. Changing μ and σ_{int} will vary the bias and internal noise magnitude, respective. See body text for details.

Despite this inherent ambiguity, it is common practice to assume (often implicitly) that bias has a negligible impact on learning, and to accordingly attribute all learning to improvements in perceptual sensitivity (e.g., van Wassenhove and Nagarajan, 2007; Hawkey et al., 2004). If incorrect, this

assumption may have led to systematic errors in models of learning (e.g., Gold et al., 2004; Lu and Dosher, 2008), and may have encouraged a misleading view of learning as a purely low-level, and/or feedforward, phenomenon. In the present work we tested whether the assumption of zero bias is correct. We predicted that learning represents more than a change in perceptual sensitivity, and that naïve listeners will exhibit a response bias that is subsequently minimised through practice.

To our knowledge, the only studies to systematically investigate the role of bias in perceptual learning are those by Wenger and Rasche (2006) and Wenger et al. (2008). In Wenger and Rasche (2006), nine observers performed a visual yes/no contrast detection task over ten, 600-trial sessions. Across the six observers that exhibited reliable improvements in performance, levels of bias were found to significantly differ across sessions. Surprisingly, this shift resulted in bias being *greater* after training (cf. Fig 3 of Wenger and Rasche, 2006), with observers becoming increasingly liberal (predisposed to say 'yes') as a function of practice. This suggests that bias does change with practice, but that the nature of this change is negative in terms of overall decision efficiency. However, as we shall discuss below (§5.2), such a finding may be predicated on a flawed method of analysis. We tested this empirically in the present study, by comparing the previous method of analysis with an approach which we argue is more theoretically justified. The first aim of the present work was therefore to replicate the observations of Wenger and Rasche (2006) using an analogous auditory task. The second aim was to study changes in bias using a novel method of analysis that more appropriately reflects the efficiency of the listener's decision strategy.

The task used in the present study was sinusoidal amplitude modulation [SAM] detection, applied to a bandpassed white noise carrier. This task was selected in order to emulate visual contrast detection, in which observers must distinguish a sinusoidal, Gaussian-windowed, luminance pattern (a Gabor patch) from a uniform luminance field. Thus, SAM noise and Gabor patches are conceptually similar, though in the former the modulations are distributed in time rather than in space. The precise parameters of the stimulus were modelled after Fitzgerald and Wright (2011), who recently demonstrated learning effects given a SAM detection task (see also Kong et al., 2004).

A central contention of the present work is that Wenger and Rasche (2006) employed a flawed measure of bias. We therefore begin by presenting a brief theoretical exposition of how bias may be operationalised. Some readers may wish to proceed directly to the empirical work in §5.3.

5.2 Measuring bias

5.2.1 Bias definition

As illustrated in Fig 5.2, bias is defined as the distance between the listener's decision criterion location, λ_{obs}, and the ideal, λ_{ideal}.

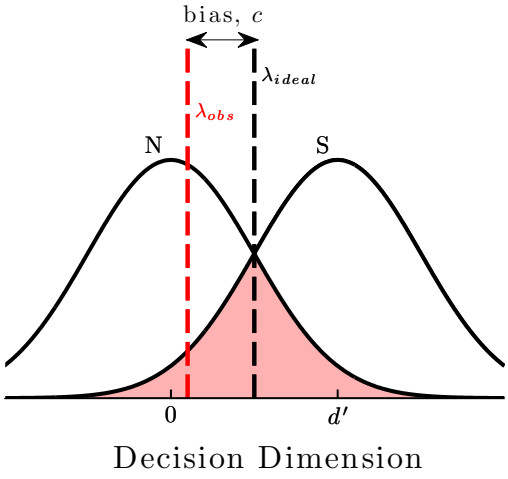


Fig. 5.2: Basic signal detection theory bias schema. Bias, c, is the distance between the listener's decision criterion location, λ_{obs} (red dashed), and the ideal decision criterion location, λ_{ideal} (black dashed). When the noise (N) and signal (S) distributions are equal in variance and frequency, λ_{ideal} is located at the midpoint between the two distribution means, as shown here. The decision dimension is unspecified, but on a psychophysical task is typically assumed to be a straightforward function of the physical stimulus variable. Given two equal-Gaussian internal response distributions, λ_{ideal} will intersect the noise (N) and signal (S) distributions. In the example shown, the listener is liberal (biased to indicate a signal was present). Performance will also be limited by the listener's sensitivity, which is inversely-proportional to the common area under the two distributions, highlighted here in red.

Formally, bias is given by:

$$bias = \lambda_{obs} - \lambda_{ideal}, \tag{5.1}$$

where λ_{obs} is estimated from the listener's false-alarm rates (Wickens, 2002). Thus:

$$\widehat{\lambda}_{obs} = Z(1-f) = -Z(f). \tag{5.2}$$

With two conditions $\lambda_{ideal} = \frac{1}{2}d'$, in which case the amount of bias may be indexed by the term c:

$$c = \lambda_{obs} - \frac{1}{2}d' = -\frac{1}{2}[Z(f) + Z(h)]. \tag{5.3}$$

More generally the ideal criterion is that which maximises the probability of a correct response, P_C. In turn, P_C, is the sum of the probability of a hit

and the probability of a correct rejection,

$$P_C = P(hit) + P(correct\ rejection). \tag{5.4a}$$

In turn, the probability of a hit is the joint probability of a signal trial occurring, $\mathrm{P}(S)$, and listener responding 'yes', $\mathrm{P}(\text{'yes'})$. Likewise, the probability of a correct rejection is the joint probability of a noise trial occurring, $\mathrm{P}(N)$, and the listener responding 'no', $\mathrm{P}(\text{'no'})$:

$$= P(\mathrm{S}, \text{'yes'}) + P(\mathrm{N}, \text{'no'}). \tag{5.4b}$$

Using the chain rule, this probability can be calculated from the conditional probability of a correct response given that trial type, together with the probably of that trial type occurring:

$$= P(\mathrm{S})P(\text{'yes'} \mid \mathrm{S}) + P(\mathrm{N})P(\text{'no'} \mid \mathrm{N}). \tag{5.4c}$$

Given a Gaussian detection model, the conditional probability of a correct response can be derived from the cumulative Gaussian distribution, thresholded at a particular criterion value, λ:

$$= P(\mathrm{S})\Big[1 - \Phi(\lambda; \mu_{signal}, \sigma_{signal})\Big] + P(\mathrm{N})\Big[\Phi(\lambda; \mu_{noise}, \sigma_{noise})\Big]. \tag{5.4d}$$

For the equal, unit variance model this becomes:

$$= P(\mathrm{S})\Big[1 - \Phi(\lambda; d', 1)\Big] + P(\mathrm{N})\Big[\Phi(\lambda; 0, 1)\Big]. \tag{5.4e}$$

Finally, when using m signal conditions, this generalises to:

$$= \sum_{i=1}^{m} \Big(P(\mathrm{S})_i\Big[1 - \Phi(\lambda; {d'}_i, 1)\Big]\Big) + P(\mathrm{N})\Big[\Phi(\lambda; 0, 1)\Big] \tag{5.4f}$$

Combining Eq 5.4f with the basic bias formula given in Eq 5.1 yields:

$$c_T = \lambda_{obs} - \arg\max_{\lambda} \left(\sum_{i=1}^{m} \Big(P(\mathrm{S})_i\Big[\Phi(\lambda; {d'}_i, 1)\Big]\Big) + P(\mathrm{N})\Big[\Phi(\lambda; 0, 1)\Big] \right) \tag{5.5}$$

The subscript in c_T serves to highlight the fact that bias is here computed using a total criterion applied to multiple signals, and to differentiate this, more general measure of bias from the more common metric, c, given in Eq 5.3 (i.e., which implicitly assumes a two-distribution situation, and which was used by Wenger and Rasche, 2006).

5.2.2 Bias in Wenger and Rasche (2006)

In Wenger and Rasche (2006) bias was measured for each individual listener as follows. Immediately after session one, a psychometric function was fitted to the listener's *hit rate* data (the proportion of correct detections at each of the 10 signal levels), as derived using the Method of Constant Stimuli. From this hit rate function [HRF], the minimum physical difference necessary to achieve a hit rate of 50% , $DL_{(H=50|sess1)}$, was calculated. This value served as the physical reference point for this individual, in all sessions. In each session a new HRF was calculated, and from this fit the hit rate at $DL_{(H=50|sess1)}$ was computed. This value, together with the false alarm rate from the noise-only trials, was then used to calculate bias as per Eq 5.3.

Our contention is that this method of calculating bias is conceptually flawed and potentially misleading. It is flawed because it uses an inappropriate value of λ_{ideal} as a reference point. To see that this is the case consider that the index of bias, c, which is illustrated in the left column of Fig 5.3, assumes that the ideal criterion, λ_{ideal}, always lies halfway between the noise and signal distributions ($\frac{1}{2}d'$). Such a strategy will maximise performance at a particular signal level. However, Wenger and Rasche (2006) employed multiple signal levels, and the goal was to maximise performance over *all* conditions. In such circumstances the ideal criterion will only correspond to $\frac{1}{2}d'$ for any particular signal level, if the listener maintained a separate criterion for *every* signal level (or by negligible chance). However, the use of independent criteria is implausible when using the Method of Constant Stimuli. Without any *a priori* means of identifying the trial-by-trial signal distribution, tailoring one's criterion to the current signal distribution would beg the question that listeners are attempting to answer. Moreover, that listeners do not employ independent criteria is implied by the shape of the observed psychometric function, which would otherwise extend below chance (50%) only if listeners were highly suboptimal in their criteria placements (see Appendix 5.A). Instead, it is most parsimonious to assume that the listener maintains only a single, total criterion for all conditions, as shown in the right of column of Fig 5.3, and as formalised in Eq 5.5.

By comparing the left and right columns of Fig 5.3, it can be seen that the inappropriate assumption of independent criteria has a number of undesirable corollaries. Firstly, it may indicate the presence of bias even when the total criterion was ideally placed (i.e., so as to maximise total percent correct). Secondly, it may indicate a change in bias even if the listener's criterion remained invariant. Thirdly, bias magnitude in session one will be determined by the arbitrary choice of performance threshold.

For example, by measuring bias at the 50% hit rate in session one, listeners will *by definition* be conservatively biased. This is necessarily the case since the criterion leading to a 50% hit rate is located at the mean of the signal distribution, which will always lie to the right of the midpoint between the noise and signal distributions[1]. Higher hit rate values will lead to even more conservative starting values, while lower values will cause c to decrease, eventually becoming negative ('liberal'). Fourth and finally, the sign of any change in bias is liable to be reversed, relative to when a total criterion is used.

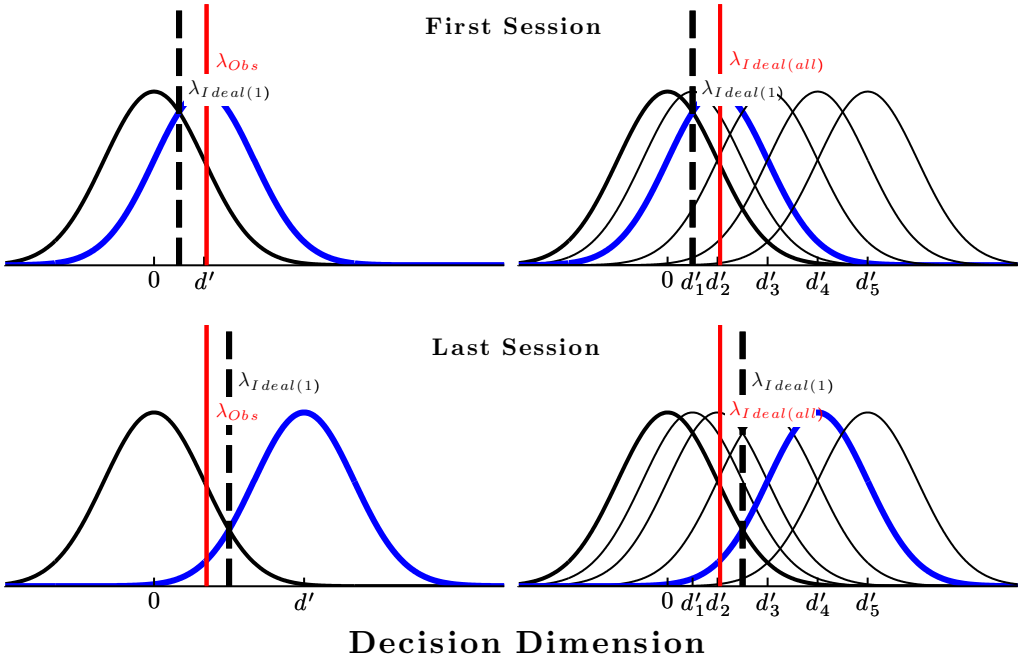


Fig. 5.3: Schematic description of the results of Wenger and Rasche (2006), as reported (left panels) and as argued here (right panels). λ_{obs} is the subject's observed criterion. $\lambda_{ideal(1)}$ is the ideal criterion according to Eq 5.3. $\lambda_{ideal(all)}$ is the ideal criterion according to Eq 5.5. By comparing the top-left and bottom-left panel, the observer appears to shift from a conservative bias to a liberal bias. However, by comparing the solid red lines in the left column (λ_{obs}) with those in the right column ($\lambda_{ideal(all)}$), it can be seen that the observed criteria are actually identical with the ideal, once all signal distributions are taken into account.

In short, the method of calculating bias implemented in Wenger and Rasche (2006) is flawed in that it uses an inappropriate value of λ_{ideal} as a reference point. As a result, it remains unclear to what extent naïve listeners are biased on a yes/no detection tasks, whether levels of bias change with practice, and what the sign and magnitude of any such changes are. In this present study we attempted to answer these questions by measuring bias relative to the (single) criterion that maximises percent correct given the listener's observed d' values (Eq 5.5).

5.3 Methods

The experiment reported here was designed to replicate that of Wenger and Rasche (2006), using an analogous auditory task: SAM detection (Fitzgerald and Wright, 2011).

5.3.1 Listeners

Thirteen normal hearing listeners participated, none of whom had any prior experience of auditory psychophysics. Four were female ($\mu_{age} = 20.8$ y.o.), nine were male ($\mu_{age} = 22.4$ y.o.). Normal hearing was assessed by audiometric screening, administered in accordance with the BSA standard procedure ($\leq$ 20 dB HL bilaterally, at 0.5 kHz to 4 kHz octaves; British Society of Audiology, 2004). Listeners were recruited through advertisements placed around Nottingham University campus, were not screened (except for normal hearing), and received an inconvenience allowance for their time. The study was conducted in accordance with Nottingham University Hospitals Research Ethics Committee approval and informed written consent was obtained from all participants.

One of the 13 listeners was excluded from all analyses due to an apparent loss of concentration over the seven sessions. Despite initially good performance (session one: $DL_{79\%} = 0.26$), over the course of the regimen the listener appeared to resort to a strategy of guessing. By session seven performance failed to reach the 79% threshold even at full modulation, and there was no significant point-biserial correlation between target and response [$r(598) = 0.06, p = 0.175, n.s.$].

5.3.2 Stimuli & Procedure

The task was one-interval, yes/no, sinusoidal amplitude modulation [SAM] detection, in which participants were asked "was the sound modulated?"

The stimulus was a sinusoidally amplitude modulated bandpassed Gaussian noise, approximate to those used in Fitzgerald and Wright (2011). The carrier was a 3–4 kHz bandpassed white Gaussian noise. The amplitude modulator was an 80 Hz sinusoid. As shown in Fig 5.4, the modulation depth (or: *index*) was varied between 0 [no modulation] and 1 [full modulation], with the trial-by-trial value determined by the stimulus condition. The stimuli were 400 ms in duration (including ramps), gated on/off by 10 ms cos^2 ramps, and presented at 70 dB SPL in all conditions (irrespective of modulation depth).

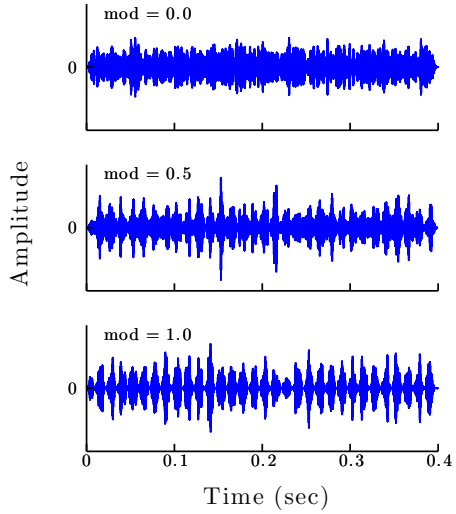


Fig. 5.4: Example stimuli, given zero, intermediate and full modulation depths. Samples of the zero modulation stimulus were played on the noise trials. The signal trials were evenly divided between 10 modulation depths, uniformly log spaced between α and β. Where $\alpha > 0$, $\beta \leq 1$, and their precise values were determined by the performance on the previous session. See body text for details.

Each trial commenced with a 300 ms warning interval, during which a visual fixation cross was displayed. This was followed by a single 400 ms stimulus observation. The modulation depth of the signal trials was drawn randomly without replacement from a discrete set of stimulus conditions, the elements of which were adaptively varied between sessions (see below). Participants were then given an unlimited time to respond, after which visual feedback was presented for 300 ms prior to the next trial onset. Feedback was included since it has been shown previously to promote learning (Ball and Sekuler, 1987). Feedback was not present in Wenger and Rasche (2006), but its inclusion was subsequently shown not to substantively affect the results (Wenger et al., 2008).

Each session consisted of 600 trials, with short breaks after the 200^{th} and 400^{th} trial. Half (300) of the trials were noise trials [depth = 0] and half were signal trials [$0 <$ depth ≤ 1]. The order of trials was randomised as per the Method of Constant Stimuli. The 300 signal trials consisted of 30 trials at each of 10 modulation depths, which were uniformly spaced between α and β on a logarithmic scale. Initially, in session one, α and β were preset at 0.1 and 1, respectively. In the subsequent sessions α and β were adapted in a manner contingent on the preceding session's performance, as follows. In each session, α and β were set to the modulation depths required to attain 5% and 95% correct detection performance, respectively, as estimated from a psychometric fit to the preceding session's hit rate data (see ∮5.3.4). This session-by-session adaptive procedure replicated that of Wenger and Rasche (2006), where it was used to minimise floor/ceiling effects.

Each test regimen consisted of seven sessions, which were completed over two weeks with no more than one session per day. Before the first session listeners were given three examples of an unmodulated noise, and three examples of a fully modulated noise [depth = 1].

5.3.3 Apparatus

Stimuli were digitally synthesized in Matlab v7.4 (2007a, The MathWorks, Natick, MA) using a sampling rate of 22.05 kHz and 24-bit quantization. Digital-to-analog conversion was carried out by a PCI sound card (Darla Echo; Echo Digital Audio Corporation, Carpinteria, CA), interfaced via the Psychophysics Toolbox v3 (Brainard, 1997) ASIO wrapper (Steinberg Media Technologies, Hamburg).

Stimuli were presented monaurally (left ear only) via Sennheiser HD 25-I headphones. Listeners were tested individually in a double-walled sound-attenuating booth, and made responses by pressing one of two buttons on a button box. Visual pre-trial fixation cues and response feedback were presented on an LCD monitor.

5.3.4 Measures & Analysis

Listeners' binary responses were used to compute false alarm and hit rates. These values were used to derive measures of each listener's detection limen, DL, sensitivity, d', and bias, c_T. One measure of each was computed per listener, per session.

DLs were estimated by fitting cumulative Gaussian psychometric functions to the hit rate data, and computing the smallest log modulation index necessary for 79% correct performance.

Sensitivity, d', was estimated for each signal level in the typical manner, via the difference in observed hit and false-alarm rates (e.g., see Wickens, 2002).

The bias metric c_T was estimated as per Eq 5.5, using the observed hit rate, false alarm rate, and d' values, together with the known *a priori* probability of each condition. For comparison, the bias metric c, was also calculated as per Wenger and Rasche (2006), in the manner detailed previously in ∮5.2.2.

A measure of non-verbal IQ was also taken, using the Matrix Reasoning subtest of the Weschler Abbreviated Scale of Intelligence (Wechsler, 1999).

5.4 Results

Learning

As shown in Fig 5.5, mean DL improved across sessions [$F(6, 66) = 5.80, p < .001, \eta_p^2 = 0.35$]. This improvement in performance was observed in 10 of 12 listeners. There was no consistent relationship between starting performance and magnitude of change [Spearman's rho; $r(10) = -0.04, p = .921, n.s.$]. As Fitzgerald and Wright (2011) observed, the majority of learning occurred during the first session.

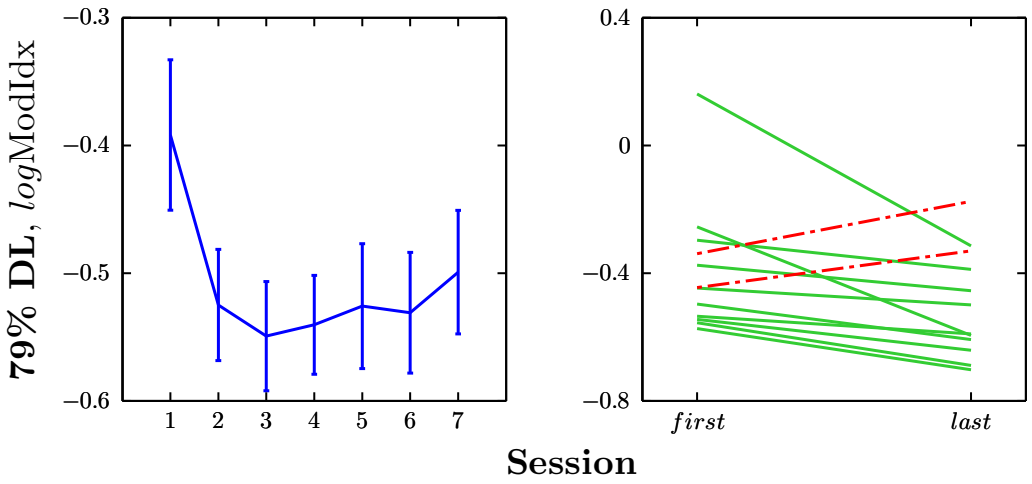


Fig. 5.5: Learning. Group mean (± SE) $DL_{79\%}$ values as a function of session (left), and for individuals in sessions 1 and 7 (right). Learners and non-learners are shown by solid-green and dashed-red lines, respective.

Bias

Bias, computed as per Eq 5.5, is shown in Fig 5.6. In session one, listeners tended to be negatively biased (liberal; predisposed to say 'yes') [$t(11) = -6.16, p < .001$]. In the final session no significant bias was observed [$t(11) = -0.14, p = .888, n.s.$]. This reduction in group-mean bias was significant [$F(6, 66) = 5.11, p < .001, \eta_p^2 = 0.32$]. As can be seen in the individual data, most (nine) listeners exhibited this pattern of reduced liberalism, though two listeners displayed the opposite trend (initially conservative, with a liberal shift). One listener was initially conservative and became more so after practice. Since we were primarily interested in bias magnitude rather than direction, the group analysis was also repeated with the signs of *all* the session-by-session bias values reversed for the three initially conservative listeners. The results are overlaid in Fig 5.6 for comparison, and did not differ substantively from the initial analysis in terms of the amount of bias reduction observed. These normalised results

do, however, provide a closer approximation to the shape of the learning curve given previously in Fig 5.5 [$r^2_{stand} = -0.65$; $r^2_{norm} = -0.85$].

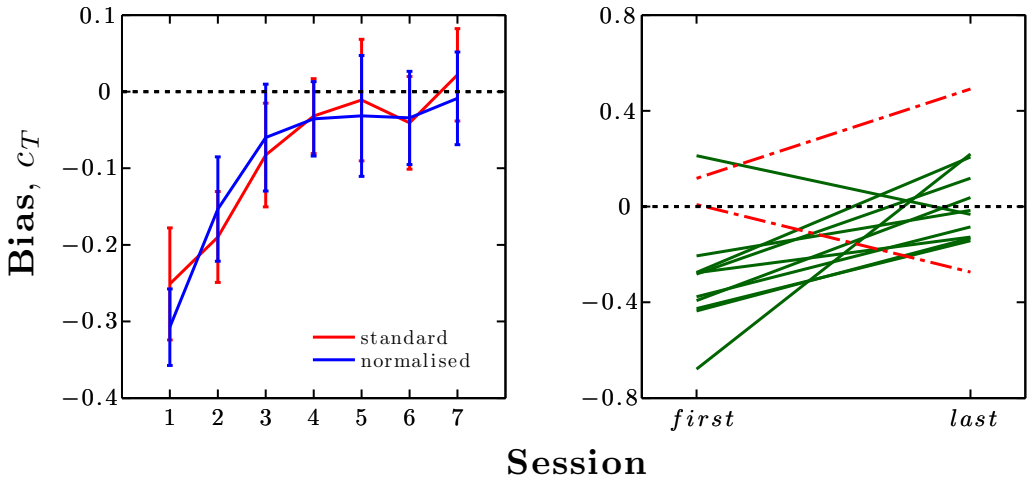


Fig. 5.6: Bias. Group mean (± 1 SE) bias, c_T, values as a function of session (left), and for individuals in sessions one and seven (right). Decreases and increases in bias magnitude are shown by solid-green and dashed-red lines, respective. The horizontal dashed line indicates zero-bias (ideal).

To determine what proportion of learning was due to changes in bias, $DL_{79\%}$ values were recalculated without bias for each listener, given observers' estimated sensitivities, d'_{obs}. A reduction in bias across sessions would manifest as a convergence between biased (i.e., empirical) and unbiased thresholds. Conversely, an improvement in sensitivity would manifest as a parallel reduction in thresholds. Inspection of Fig 5.7 shows that there was a convergence in thresholds, indicating a reduction in bias. This was supported by a significant interaction between Session and curve-type [RMANOVA: $F(6, 66) = 3.10, p = .010, \eta_p^2 = 0.19$]. The change in sensitivity [B] was approximately three times greater than the change in bias [A]. This indicates that approximately one quarter of the change in $DL_{79\%}$ was due to reductions in bias. More formally, sensitivity and bias can be converted to z-score units for direct comparison. Group-mean bias, as indexed by c_T, decreased by 0.30 z-units. Conversely, group-mean sensitivity, as indexed by d' values computed at a fixed modulation depth (a depth corresponding to a 79% hit rate in session one), increased by 0.65 z-units. This indicates that approximately 31% of learning consisted of a reduction in bias.

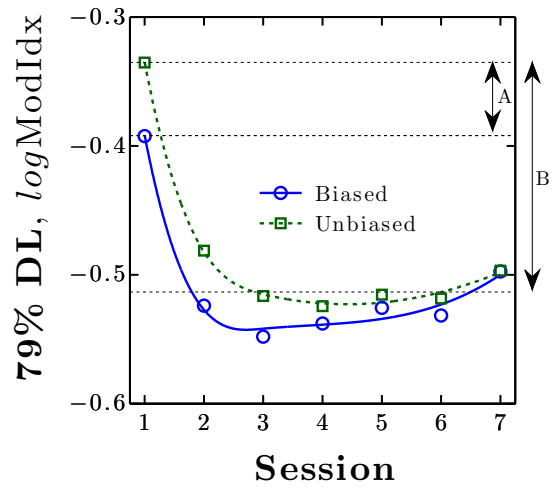


Fig. 5.7: Comparison of performance ($DL_{79\%}$) with and without bias. The 'biased' data show the listener's observed $DL_{79\%}$ values, shown previously in Fig 5.6. The 'unbiased' data were calculated from the observed d' values, thresholded by the ideal decision criterion, λ_{ideal}. Note that because $DL_{79\%}$ was calculated from the hit rate function (i.e., signal trials only), a liberal bias will actually result in thresholds appearing elevated ('improved').

There was no indication of a relationship between non-verbal IQ and initial bias magnitude [Pearson's r; $r = -0.0, p = 0.979, n.s.$], or between non-verbal IQ and change in bias magnitude across sessions [$r = -0.2, p = 0.625, n.s.$].

Comparison with Wenger and Rasche (2006)

For comparison, bias was also analysed in the manner described by Wenger and Rasche (2006), both at their reported hit rate level (50%), and also at a higher (70%) and lower (30%) level. As shown in Fig 5.8, using this method listeners initially appeared conservative, and gradually became more liberal with practice. This 'liberalising' shift was identical in sign to that reported by Wenger and Rasche (2006), though was not as great in magnitude. Accordingly, the change in bias failed to reach significance when analysed at the 50% hit rate level [$F(6, 66) = 1.10, p = .369, n.s.$], but was significant at the 70% level [$F(6, 66) = 3.42, p = .005, \eta_p^2 = 0.24$]. As the reference hit rate was increased from 30% to 70%, starting performance appeared increasingly conservative.

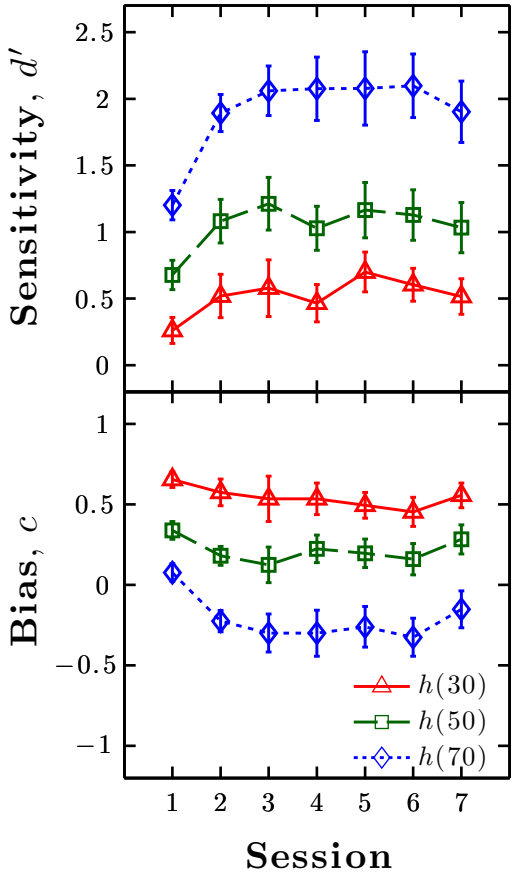


Fig. 5.8: Group mean sensitivity (top), d', and bias (bottom), c, scores as a function of session, as derived using the analysis technique of Wenger and Rasche (2006). Each line pertains to a given hit rate reference level ($30\%, 50\%, 70\%$). See §5.2.2 for details.

5.5 Discussion

Listeners are often assumed to be unbiased agents, making responses conditional on sensory evidence only, with no *a priori* constraints. The present results indicate that, in a yes/no SAM detection task, this is an acceptable assumption only for listeners who have completed around 1800 practice trials. In contrast, naïve listeners tended to exhibit significant bias, generally in favour of responding 'yes'. Practice-induced (conservative) criterion shifts led to this bias being largely eradicated by session three. Reduced bias therefore constitutes one of the mechanisms of learning on this task, and appeared to account for approximately one third of the improvements in SAM detection limens.

This finding is contrary to Wenger and Rasche (2006), who, antithetically to the present study, found observers to be initially conservative, to shift their criterion liberally, and to exhibit *greater* bias post-practice. However, these differences were entirely due to differences in the method of analysis.

Thus, when the analysis of Wenger and Rasche (2006) was applied to the present data, participants appeared to behave as in Wenger and Rasche (2006), to the extent that they became more liberal, and in some cases *more* biased, with practice. We have argued, however, in ∮5.2 that this method of analysis is logically flawed, and potentially misleading. It is therefore likely that the observers in Wenger and Rasche (2006) underwent a reduction in bias similar in nature to that reported here.

The cause(s) of the bias observed in naïve listeners remain uncertain. One possibility is that the asymmetry in responding stems from a corresponding asymmetry in how listeners perceive the statistics of the task. Thus, listeners may perceive the relative utility (or: *payoff*) of each response outcome to be unequal (e.g., if they believe the penalty of missing a *signal* outweighs the benefit of spotting a *noise*; cf. Maddox and Bohil, 1998), or they may perceive the relative probability of each trial-type occurring to be unequal. We speculate that the latter may have been influenced in the present study by the fact that the number of signal conditions (i.e., *types* not *tokens*) greatly outnumbered noise-alone types (which, by definition, is always one). This may have led listeners to assume that signal observations were more likely to occur than noise observations. If this were the case then bias may be attenuated if the likelihood of drawing from the noise distribution was equal to that of each signal distribution. Alternatively, naïve listeners' bias may result from systematic errors in their perception of their underlying distributions. For example, a liberal bias may result if listeners underestimate their sensitivity to one or more stimulus contrast, or if they underestimate the magnitude of additive internal noise (see Appendix 5.B).

The use of yes/no detection task was used in the present study for consistency with Wenger and Rasche (2006), and because of its real world relevance. However, some aspects of the paradigm make it suboptimal as a model of auditory learning more generally. Its simplicity may have caused levels of bias, and by implication the importance of bias in auditory learning, to be underestimated. This is the case since in most real world situations the payoffs and probabilities associated with each alternative are often much more complex, and in many cases nonstationary. Learning, and maintaining, the correct criterion placement is therefore a far more complex task than in the present study, where payoffs and probabilities were constant and balanced. Contrawise, in other respects the present task may have led to the role of bias being overestimated in the type of learning typically evidenced by psychoacoustical studies. This is so since many perceptual studies (including previous examinations of SAM detection; Fitzgerald and Wright, 2011) specifically militate against response bias,

either through the use of forced-choice paradigms, and/or via metrics such as d' that explicitly attempt to partial out bias effects (Macmillan and Creelman, 2005). This latter point is addressed in Chapter 6, where bias effects are shown to occur in forced-choice tasks also, and are similarly shown to decrease with practice

5.6 Conclusions

(1) Listeners exhibit learning on a yes/no SAM detection task, as evidenced by a significant decrease in modulation depth detection limens.

(2) A substantial minority of this learning ($\sim 31\%$) is due to a reduction in bias. Naïve listeners tend to be predisposed to respond 'yes' (liberal). This bias is eradicated after 7×600 trials of practice.

(3) The analysis method of an analogous visual perceptual learning study (Wenger and Rasche, 2006) is flawed, in that it uses an inappropriate reference point when calculating bias. The key results of that study, in terms of the direction of the starting bias and of the change in bias with practice, are consistent with those presented here.

Notes

[1]Except in the limiting case where the two distributions are identical and the task is impossible

Acknowledgements

We would like to thank Natasha Ratcliffe for assistance in recruitment and testing. This work was supported by the Medical Research Council, UK (Grant: U135097130).

5.A Empirical evidence of a single criterion

The present data, and those reported previously by Wenger and Rasche (2006), imply that listeners maintain a single decision criterion across all trials/stimulus-conditions. To wit, consider Fig 5.9. The bottom panel shows the distributions of internal responses associated with a noise stimulus, N_0, and each of five signal stimuli, $S_1 \ldots S_5$. The top panel shows the corresponding psychometric hit rate function [HRF] when a single decision criterion of $\lambda = d'_2$ is maintained. The result is a monotonic cumulative Gaussian curve, ranging from 0 and 1.

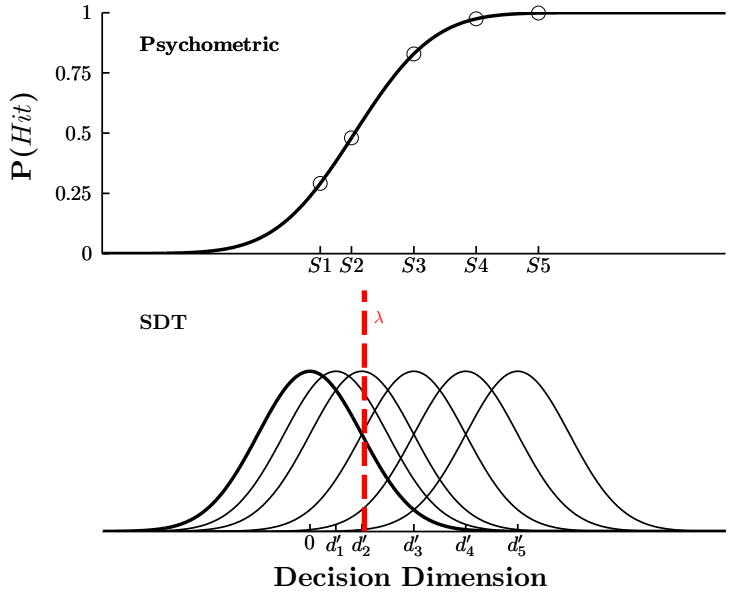


Fig. 5.9: Schematic representation of a multi-signal signal detection scenario, and the corresponding psychometric function when observations are thresholded by a single decision criterion value. See body text for details.

Conversely, Fig 5.10 depicts the analogous situation when the listener maintains a single criterion for each noise-signal, N_0S_i, combination. Note,

this would require the listener to know the signal distribution from which a given observation, x, was drawn. The result is a positive, unipolar, convex HRF. Since the individual criteria are locally ideal, the hit rate never falls below the $P = 0.5$ level.

Since the empirical data is consistent with Fig 5.9, and inconsistent with Fig 5.10, we conclude that listeners maintain a single, total decision criterion.

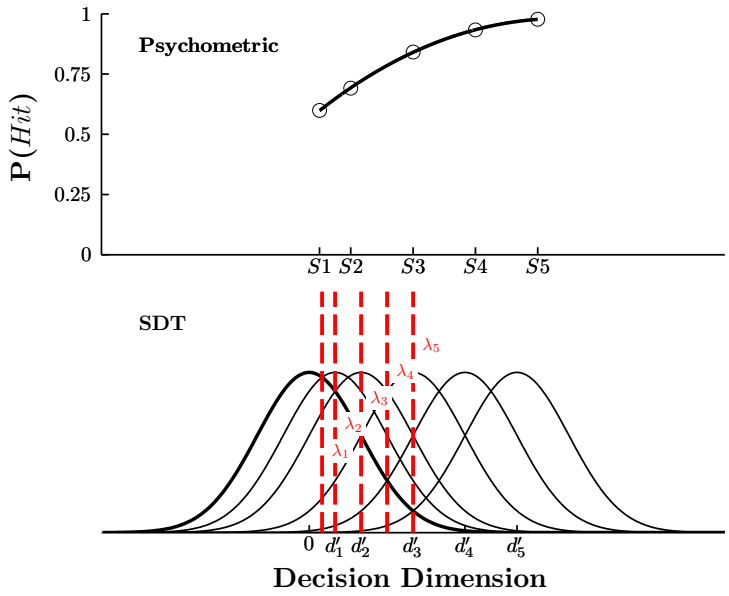


Fig. 5.10: Same as Fig 5.9, for situations when an independent criterion is maintain for each signal distribution.

5.B Relating perceived sensitivity to bias

Bias, in the sense of suboptimal criterion placement, may result from listeners misestimating one or more parameters in the underlying distributions upon which their responses are predicated. To see how this is the case, we simulated idealised signal detectors with varying signal magnitude, S, and internal noise, σ_{int}, parameters. As shown in Fig 5.11, reductions in either result in a liberal shift in decision criterion. The observed behaviour of naïve listeners may therefore reflect listeners underestimating either parameter. Notably, small samples sizes typically lead to the standard deviation of a normally distributed variable being underestimated. It may therefore be that listeners initially have insufficient information to accurately estimate their own internal noise levels.

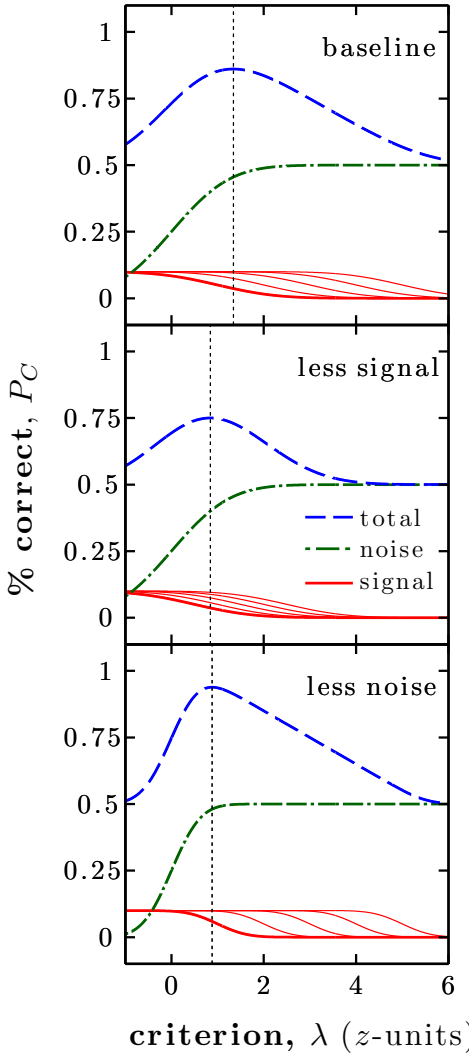
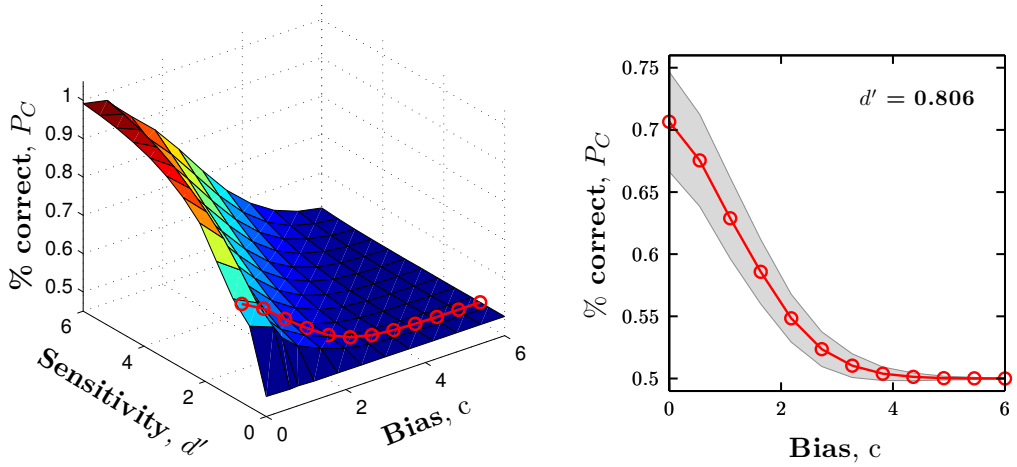


Fig. 5.11: Effect of changes in signal magnitude (middle), S, and internal noise (bottom), σ_{int}, to the ideal criterion location, λ_{ideal}. The top panel represents an arbitrary baseline condition. The solid red and dot-dashed green lines describe percent correct, P_C, rates for individual signal and noise distributions, as a function of threshold value, λ. These values are scaled relative to the number of trials from each distribution. The dashed blue line indicates total P_C, and is the sum of the other curves. The dashed vertical line indicates the ideal decision criterion, as per Eq 5.5.

5.C Effects of bias and sensitivity on P_C

Shown in Fig 5.12 is an alternative depiction of how bias can negatively affect performance. Contrast this with Fig 5.1.



(a) **(b)**

Fig. 5.12: Effects of bias, c, and sensitivity, d', on P_C, for a yes/no detection task. **(a)** Manifold relating each combination of c, $[0 \ldots 6]$, and d', $[0 \ldots 6]$. Values are the mean of 10000 simulations. This result is symmetric for negative c values. **(b)** A single cross section of *(a)*, for a single fixed d' value. Error bars indicate $\pm$ 1 SD.

CHAPTER 6

Bias in Forced-Choice learning

Bias is the tendency to favour one response alternative, independent of the sensory evidence. The widely used forced-choice [mAFC] paradigm is often assumed to preclude such bias. However, if this assumption is incorrect then differences in performance may actually reflect differences in bias. In this study, we assessed bias as a function of session for 57 naïve listeners, given a two interval, two alternative, forced choice (2I2AFC), pure tone discrimination task (frequency or intensity). Nonstationary bias was quantified by analysing responses conditional on previous trials, using only trials where the discrimination was very difficult or impossible. Naïve listeners were found to exhibit significant levels of bias, tending to perseverate after correct responses and alternate after incorrect responses. These biases were reduced by practice. Consistent with this, more experienced listeners also exhibited response sequences that better approximated a time-series of independent events. Simulations showed that the reductions in bias were sufficient to explain some but not all of observed improvements in discrimination thresholds. These results indicate that hearing sensitivity may be underestimated in listeners who lack prsior task experience, and that perceptual learning in part represents a reduction in response bias.

6.1 Introduction

Listeners' performance on sensory tasks often improves with practice (Wright and Fitzgerald, 2001; Fine and Jacobs, 2002; Goldstone, 1998). Such learning is commonly assumed to reflect increased sensitivity to the task-relevant sensory information (e.g., van Wassenhove and Nagarajan, 2007; Hawkey et al., 2004). Accordingly, models of learning have tended to attribute improvements to sensitivity-limiting constructs, such as the magnitude of internal noise or the efficiency with which perceptual features are combined (e.g., Gold et al., 2004; Lu and Dosher, 2008). However, task

performance is constrained not just by the strength of the sensory evidence, but also by the efficiency of the wider decision process that the sensory evidence informs. One crucial step in this process is to compare the sensory evidence to a criterion, λ, in order to determine the appropriate response. Ideally λ should be placed so as to maximise some payoff metric, such as percent correct. However, a listener's criterion may deviate from the ideal (cf. §5.2). This may occur if the listener misjudges the frequency with which a particular stimulus-type occurs, or misestimates the value of a particular response outcome. For example, a listener may require more evidence to respond 'A' if they believe that it is more important to spot B's, or that B's are more likely to occur. Here we term any deviation from the ideal criterion: *bias*. Since bias will diminish expected performance, some or all perceptual learning may represent a *reduction* in bias. In this study we evaluated this possibility by quantifying the extent to which bias is present in naïve listeners, and is reduced through practice.

Within the perceptual learning literature, bias has not been extensively investigated (though see Wenger and Rasche, 2006; Wenger et al., 2008). This neglect is partly practical. Learning studies often employ adaptive tracks and more than two response options, both of which make bias metrics difficult to compute. More importantly though, learning studies tend to utilise m-alternative forced-choice [mAFC] designs, which are thought to preclude bias. Indeed, when bias has been reported for such tasks, levels have tended to be near zero and invariant across sessions (e.g., Schoups et al., 1995; Ben-David et al., 2011; Campbell and Small, 1963, though for exceptions see Koyama et al., 2004; Halliday et al., 2011). Crucially though, the bias measures used in such studies assume that the listener's criterion remains stationary throughout the session. Thus, both the constant error [CE] term used in psychophysics (Gescheider, 1997), and the Signal Detection Theory metrics c and $log\beta$ (Macmillan and Creelman, 1990; Dusoir, 1975), only index a pervasive bias to always favour one response alternative. However, bias may also be nonstationary (or: 'dynamic'; Atkinson et al., 1962). It may fluctuate randomly, for example if the observer is unable to maintain a stable criterion (Kubovy and Healy, 1977). Or it may vary systematically, for example in a manner determined by previous trials. Such biases are not necessarily discouraged in mAFC designs – an 'alternating' listener may be just as inclined to respond 'A' after 'B' as they are to respond 'Yes' after 'No'. Moreover, such biases may have an internal symmetry, with 'B' following 'A' as often as 'A' follows 'B' (cf. Table 6.3). Such processes are therefore liable to cancel out at the molar level, manifesting instead as decreased sensitivity.

That observers' responses may be systematically influenced by preceding trials has long been noted[1]. In the 19^{th} century Fechner observed sequential effects which he attributed to "[interference] in the memory of the observer" (Fechner, 1966), while Green (1964) reports "a tendency among all observers to choose the interval opposite the one on which they had just been correct." Such response dependencies have been quantified using a variety of techniques. For example, Verplanck et al. (1952) used a serial-correlation procedure (cf. Wald and Wolfowitz, 1943) to assess the statistical independence of luminance detection responses made at chance threshold. Runs of identical responses were observed to be greater in length (and thus fewer in number) than would be expected if each response was made independently, indicating that observers were biased towards repeating their previous response (hereafter: 'presponse'). Garner (1953) found an analogous result in the auditory domain (see also Garner and Hake, 1951). Using an information analytic approach, he found that the presponse on a loudness identification task explained some of the variance in trial-by-trial responses, and that the nature of this effect was again to bias responses in favour of repetition. Similar dependencies have also been demonstrated using multiple regression (Jesteadt et al., 1977). Such sequential effects are nuanced and not fully understood. They vary in strength according to the use of feedback (Ward and Lockhead, 1970; Tanner et al., 1967; Atkinson et al., 1962), the number of response options (Garner, 1953) and the stimulus magnitude (McGill, 1957). Moreover, there is not always consensus as to the direction (Parducci et al., 1966), time-course (Petzold and Haubensak, 2001; Staddon et al., 1980), or underlying causes of these effects. Nonetheless, it is clear that sequential response dependencies do occur, and that listeners are liable to be influenced in their responding by the events of previous trials.

That nonstationary bias may explain practice effects is not a new concept to the psychophysical literature (Kubovy and Healy, 1977). But perhaps the best evidence that sequential response dependencies are reduced by practice comes from studies of the gambler's fallacy (Laplace, 1995; Ayton and Fischer, 2004; Anderson, 1960; Jarvik, 1951; Clotfelter and Cook, 1993; Croson and Sundali, 2005), and other related recency effects. For example, Lindman and Edwards (1961) constructed shuffled decks of cards, composed in equal part of Red and Green cards. Observers were presented with each card in turn and asked to predict the colour of the next one. Alternation (or: 'negative recency') was observed, whereby observers tended to avoid guessing the most recently occurring outcome. In a split-half analysis, this negative recency was reduced, with observers becoming less inclined to predict 'Green' after a run of 'Red' cards (see also Edwards, 1961). This suggests that response dependencies can be

modified through practice. However, it is not clear how well these results – obtained using tasks where outcomes are predicted *a priori* – generalise to psychophysical tasks, where judgements are made *a posteriori*, and where the use of information from previous trials is discouraged (often explicitly).

Here we examined the extent to which a systematic, nonstationary bias is present in naïve listeners (Experiment I), is reduced by practice (Experiments II), and can explain auditory perceptual learning (Experiment III). We begin by describing how nonstationary bias was computed.

6.2 General analysis methods

6.2.1 Sequential response dependency

Multiple regression was used to test whether observer responses comprise a time-series of independent events. The identity, I (here equal to the interval selected in a 2AFC trial: 1 or 2), and correctness, C (0 or 1), of the previous N responses were used to linearly predict the current response identity. The predicted identity of the response on trial t, I_t, was:

$$I_t = \left(\sum_{i=0}^{N} \alpha_i I_{t-i} + \beta_i C_{t-i} \right) + \gamma + \epsilon, \tag{6.1}$$

where α, β, and γ are the estimated regression coefficients, and ϵ is a Gaussian error term. When only considering the immediate presponse[2], this simplifies to:

$$I_t = \alpha I_{t-i} + \beta C_{t-i} + \gamma + \epsilon \tag{6.2}$$

If a significant proportion of response variability is explained by the presponse then this would indicate that successive judgments were not made independently. This approach is identical to that of Jesteadt et al. (1977), with the following exceptions. Firstly, we did not enter signal magnitude into the model, since we only applied it to cases where all stimuli were identical (impossible discrimination). Secondly, we did enter the correctness of the preceding responses as well as their identity. This inclusion was motivated by the results of the χ^2 analysis (Appendix 6.B).

6.2.2 Bias

Interval bias, c, was calculated as per traditional Signal Detection Theory (cf. Wickens, 2002, , p.100), via the (relative) probability of correct 'Interval 1' and 'Interval 2' responses:

$$c = \widehat{\lambda} - \widehat{\lambda}_{ideal} = \frac{1}{2}\Big[Z(P_{C\langle NS\rangle}) - Z(P_{C\langle SN\rangle}) \Big], \tag{6.3}$$

where Z is the inverse of the cumulative Gaussian distribution (Φ^{-1}), $P_{C\langle NS\rangle}$ is the proportion of correct 'Interval 2' (noise-signal) responses, and $P_{C\langle SN\rangle}$ is the corresponding proportion of correct 'Interval 1' (signal-noise) responses.

6.2.3 Nonstationary Bias

Nonstationary bias was assessed by classifying trials by the pattern of preceding responses, and applying Eq 6.3 independently to each subset of data.

When partitioning the data set, only the identity and the correctness of the presponses were considered (see below). However, even with only two variables, patterns of presponses are often too sparse to analyse. To reduce this sparseness we made the Markov assumption that the location of the listener's criterion was conditional upon only the previous N responses. Thus, when N = 0 bias was calculated in the standard manner (§6.2.2), with no regard for the preceding trials. When $N = 1$, bias was calculated for only those trials where the preceding trial was of a given identity and correctness. As N is backed off to higher values, progressively more previous trials were taken into account.

A second challenge is that response-dependencies are capricious. Presponses are likely to affect λ differently across listeners, or even within listeners across trials. We therefore concentrated on a subset of patterns where the presponse effects are likely to interact constructively and in a consistent manner. Accordingly, we only measured bias for instances where all the presponses were identical, both in terms of the interval selected [IDENTITY: 1 or 2] and whether or not that selection was correct [CORRECTNESS: 0 or 1]. Thus, $N = 2$ yields the following four bias estimates: $\{c|1_{\checkmark}1_{\checkmark}\}$, $\{c|1_{\times}1_{\times}\}$, $\{c|2_{\checkmark}2_{\checkmark}\}$, $\{c|2_{\times}2_{\times}\}$, where $\{c|1_{\checkmark}1_{\checkmark}\}$ is bias given correct 'Interval 1' responses on the two preceding trials.

Notably, the terminal items at any given value of $N+1$ form a subset of the terminal items at N. This may lead to bias being overestimated at lower values of N. Therefore, as Table 6.1 illustrates, when calculating bias each individual response was only evaluated once, at the highest possible value of N.

Response	2	2	1	1	2	2	1	2	2	2	1	1	2	1	1	1	2	1
Correct	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	0	1	1
$N = 0$	†	†		†	†	†		†				†	†	†	†	†	†	
$N = 1$			†				†		†									†
$N = 2$										†								
$N = 3$											†							

Table 6.1: Schema for selecting trials conditional on correct 'Interval 2' presponses. The first two rows show the target and response intervals for 18 hypothetical trials. Trials that would be considered when calculating bias at N-back levels 0–3 are marked with an obelisk (†). Analogous subsets of trials (not shown here) were also constructed for those trials preceded by incorrect and/or 'Interval 1' responses.

Trials where the signal magnitude was greater than chance threshold were not used to calculate bias, but were used when constructing presponse chains. Thus, bias was assessed by examining all those trials where the signal magnitude was below a predetermined cut-off, x, and where the preceding N responses were identical, both in terms of identity and correctness (irrespective of the signal magnitude on these trials).

It was assumed that listeners would not be influenced by presponses in previous test blocks. Thus, in instances were data blocks were aggregated for analysis, presponse chains were not permitted to cross test blocks.

As N is backed-off to higher values, the number of available observations generally declines. To ensure sufficient observations for a valid estimate of c, N was limited to ≤ 3. The tendency for observations to decrease with N is a potential confound for analyses in which N is an independent variable. However, this would most likely have led, if anything, to the effects reported here being *underestimated* (see Appendix 6.A).

6.3 Experiment I: Bias in naïve listeners

The purpose of this experiment was to establish the extent to which levels of systematic, nonstationary bias are present in listeners naïve to psychoacoustical testing. As a preliminary step, levels of sequential dependencies were also quantified.

6.3.1 Methods

Thirty listeners each completed 500 trials of a two-interval, two-alternative, forced choice [2I2AFC], pure tone discrimination task, in which both tones were identical on every trial (impossible discrimination). Half (15) of

the listeners were instructed to 'pick the higher tone', while the other 15 listeners were instructed to 'pick the louder tone' (as shown in the Results, this difference in instruction did not appear to affect the present results). On every trial, both tones were 1 kHz sinusoids, 100 ms in duration (including ramps), and gated on/off by 10 ms cos^2 ramps. The two tones were separated by a 500 ms inter-stimulus interval, and were presented diotically at 80 dB SPL over Sennheiser HE 60 headphones. Trial-by-trial feedback was presented visually for 500 ms during the inter-trial interval.

In this and all subsequent experiments, the 'target' tone (for the purposes of scoring and feedback) was randomly assigned to one of the two intervals with uniform probability on every trial. In cases where both tones were identical, the target tone was arbitrarily, randomly determined. Listeners were naïve to the task and had normal audiometric profiles ($\leq$ 20 dB HL at octave frequencies 1–4 kHz). The study was conducted in accordance with Nottingham University Hospitals Research Ethics Committee approval, and informed written consent was obtained from all participants.

6.3.2 Results

Sequential dependencies

To test for sequential response dependencies, the regression model of Eq 6.2 was applied to individuals' time-series response data. The identity and correctness of the immediately presponse were found to significantly predict listeners' responses in 19 of 30 cases.

To examine whether this sequential dependency effect extended beyond the immediately preceding trial, further regression models were constructed in which progressively greater numbers of presponses were included as independent variables (cf. Eq 6.1). As Table 6.2 shows, the immediate presponse was shown to explain 3.3% of the variability in responses. For some individuals, including a greater number of presponses increased the amount of variance explained by as much as 10%. However, the improvements beyond $N = 1$ were not significant at the group level [all; $F \leq 0.16, p \leq 0.86, n.s.$].

ΔR^2	N presponses			
	1	2	3	4
mean	.033	.012	.007	.007
std	.041	.019	.006	.008
min	.000	-.000	-.001	-.001
max	.184	.101	.018	.037
median	.015	.006	.006	.005

Table 6.2: Increments in the coefficient of determination, R^2, produced by including additional presponse trials in the multiple regression model, Eq 6.1.

Bias

Bias was analysed contingent on the number of identical presponses. Table 6.3 shows the results for $N = 1$. Listeners tended to alternate their responses after incorrect presponses, and perseverate after correct presponses.

Presponse		Target Interval		**bias, c**
Interval	Correct	1	2	
all	*all*	48.8	50.9	**0.03**
1	0	45.0	56.2	**0.14**
	1	58.6	43.5	**-0.19**
2	0	54.1	45.8	**-0.11**
	1	37.9	58.0	**0.25**

Table 6.3: Percent correct responses to each interval, and the resultant bias index, c, for $N = 0$ and $N = 1$. In the first row all the data is aggregated together ($N = 0$). The near-zero value of c indicates no bias. In rows $2 - 5$, the same data is partitioned contingent upon the immediately preceding response ($N = 1$). Positive and negative c values indicate 'Interval 2' and 'Interval 1' preferences, respectively. The data is a subset of that shown in Fig 6.1.

This result is extended to longer presponse runs in Fig 6.1. As shown in the top-left panel, as the number of correct presponses progressively increased, listeners became increasingly biased in favour of repeating the same response. Thus, repeated 'Interval 1' responses were likely to be followed by a further 'Interval 1' response, while repeated 'Interval 2' responses were likely to be followed by a further 'Interval 2' response. To compare 'Interval 1' (bottom curve) and 'Interval 2' (top curve) presponses, the values from

the top curve were compared to the additive inverse of the values from the bottom curve. A repeated-measures analysis of variance [rmANOVA] found no significant difference [$F(1, 24) < 0.01, p = 0.966, n.s.$], indicating that presponse identity affected the direction but not the magnitude of the bias.

The bottom-left panel shows mean absolute bias magnitudes, averaged across presponse identity. These values consistently increased for $N = 0$ to 2 [rmANOVA: $F(2, 54) = 9.86, p < 0.001$]. $N = 3$ was not included as a level in the ANOVA analysis as some listener provided few (< 12) responses at this level. A significant difference between listeners was also observed [$F(27, 30) = 2.21, p = .007$]. Notably, absolute levels of (stationary) bias calculated in the traditional manner ($N = 0$, Eq 6.3) were also significantly greater than zero [$t(54) = 2.98, p = .004$].

Responses made after repeated incorrect presponses are shown top-right. For incorrect presponses, the relationship between N and bias was non-monotonic. After only one incorrect presponse ($N = 1$), responses were biased in favour of the alternate interval. However, after three identical, incorrect responses ($N = 3$), listeners were inclined to perseverate. Again, mean bias magnitude (bottom-right) was found to increase as a function of N [rmANOVA: $F(2, 52) = 10.85, p < 0.001$]. Note that in this format, unlike with the signed values (top-right), substantial bias was observed in the $N = 2$ condition. This is because some listeners exhibited biases that were approximately equal in magnitude but opposite in sign (i.e., some alternated, while some perseverated), and thus signed bias values largely canceled at the group level.

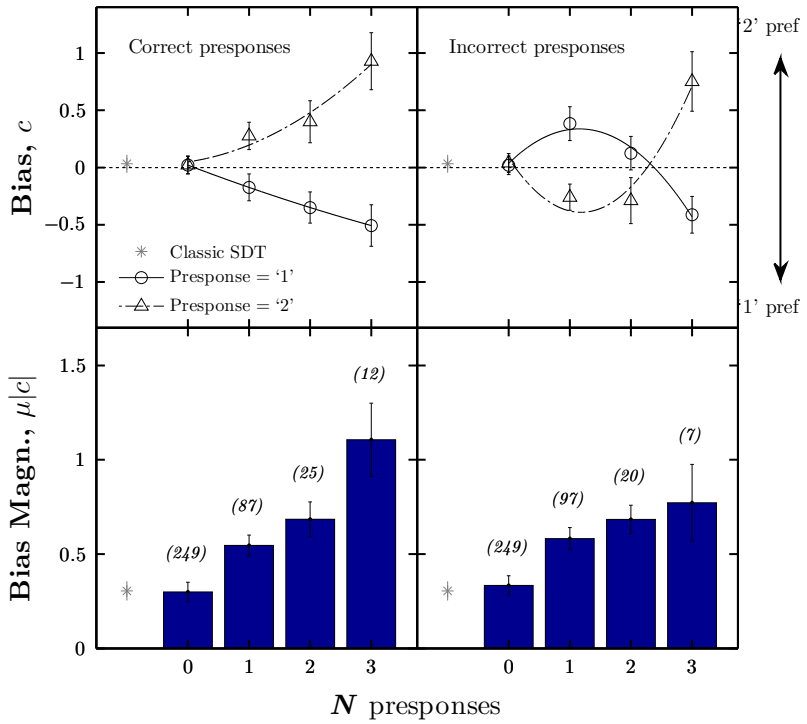


Fig. 6.1: Group mean ($\pm$ 1 SE) bias as a function of N identical presponses [Experiment I]. The left column shows data for identical, correct presponses. The right column shows data for identical, incorrect presponses. The upper row shows signed c values (Eq 6.3) for 'Interval 1' (solid, circles) and 'Interval 2' (dashed, triangles) presponses. The lower row shows absolute bias magnitude, $|c|$, averaged across response identities. The numbers in parentheses give the mean number of observations (averaged over intervals and listeners). The grey marker (far left) shows bias as estimated using all trials, as per classic SDT. Curves represent least-square 2nd-degree polynomial fits.

Frequency- versus intensity-discrimination instructions

Since the task was impossible (identical tones), the stimuli were invariant of whether listeners were instructed to perform frequency discrimination or intensity discrimination. However, to investigate whether bias differed depending on the initial task instructions, mean bias magnitude was analysed in a mixed-effects ANOVA, with N PRESPONSES as a within-subjects factor, and INSTRUCTION TYPE as a between-subjects factor (two levels: frequency discrimination; intensity discrimination). No significant difference was observed between the two groups [$F(1, 28) = 2.10, p = 0.160, n.s.$], indicating that the task instructions did not affect bias.

6.3.3 Discussion

The results showed that the trial-by-trial responses of naïve listeners are dependent on the events of the preceding trials. Repeated identical

responses were liable to be continued given positive feedback, or discontinued given negative feedback. Moreover, this effect increased in a roughly additive cumulative fashion as the number of repeated identical responses increased. This result was symmetric with respect to the two response intervals. Thus, repeated correct 'Interval 1' responses promoted a further 'Interval 1' response, just as repeated correct 'Interval 2' responses promoted a further 'Interval 2' response. With molar measures that assume trial-by-trial independence, such equal and opposite biases will cancel out, aliasing as lower sensitivity rather than higher bias.

The immediate presponse was shown to explain 3.3% of the variability in responses. This figure is similar to the 2.9% value reported by Jesteadt et al. (1977). Also similar to Jesteadt et al. (1977) was the fact that including longer runs of presponses did not significantly improve the power of the model. This suggests that response dependencies only extend over a single trial. However, the N-back analysis demonstrated that, in fact, for certain subsets of responses, sequential dependencies can be long lived and cumulative. Bias was greatest after three identical presponses, and showed no sign of asymptoting. These long-range dependencies are obscured in the gross regression analysis due to noise from other trials.

6.4 Experiment II: Bias & learning

Experiment I evidenced the presence of nonstationary bias in naïve listeners. In this experiment we examined whether such bias decreases with learning. Analogous analyses were performed on two multi-session datasets, one using adaptive tracks [AT] and one using fixed frequencies [FF], the values of which were jittered prior to presentation. In each case an independent cohort of naïve listeners practiced a pure tone frequency discrimination task. In order to evidence learning these datasets did not use identical tones. This resulted in proportionally fewer runs of incorrect responses. Accordingly, unlike Experiment I, bias was only measured following runs of correct presponses. As listeners do not necessarily maintain a fixed criterion across stimulus levels, only the most difficult trials ($<$ 2 Hz) were used when calculating bias. The frequency discrimination limen [FDL] was calculated as the smallest frequency required for 70.7% correct performance.

6.4.1 Methods

Adaptive frequency discrimination

This dataset is a subset of that detailed previously in Amitay et al. (2005). Twelve listeners performed seven test blocks across four sessions. Each

block consisted of five interleaved adaptive tracks consisting of 100 trials, interleaved with 50 catch trials in which stimulus differences were much greater, yielding a total of 550 trials per block.

On each trial, listeners were presented with two sinusoids separated by a 500 ms inter-stimulus interval. Each tone was 100 ms in duration (including ramps), gated on/off by 20 ms cos^2 ramps and presented diotically at 70 dB SL via Sennheiser HD480II headphones. The test tone frequency was always greater than or equal to the standard tone frequency, which was fixed at 1 kHz. On adaptive trials the frequency difference was determined by a two-down one-up staircase (Levitt, 1971). The initial frequency difference, ΔF, was 20% of the 1 kHz standard (0.2 kHz). This difference decreased in steps of 4% until the seventh reversal, in steps of 1% for a further four reversals, and 0.2% thereafter. Steps sizes were decreased where necessary to prevent $\Delta F < 0$. Trial-by-trial feedback was presented visually for 500 ms after each response.

During analysis the first and last three blocks of data were aggregated together for comparison. Two participants were excluded from all analyses since they provided very little data below the $<$ 2 Hz cutoff ($n < 75$).

Jittered frequency discrimination

This dataset is a subset of that detailed in Chapter 2. Fifteen listeners performed four sessions of 1600 trials. On each trial listeners were presented with two sinusoids separated by 400 ms. Each tone was 300 ms in duration (including ramps), gated on/off by 10 ms cos^2 ramps and presented diotically at 70 dB SPL via Sennheiser HD 25-I headphones. The two tones were arranged symmetrically about 1 kHz, with an initial frequency difference of either 1, 3, 5, 7, 9, or 11 Hz (depending on block). However, the frequency of each tone was independently jittered prior to presentation by a random value drawn from a zero-mean Gaussian with a standard deviation equal to half the initial (mean) frequency difference (i.e., $0.5, 1.5, \ldots, 5.5$ Hz). Trial-by-trial feedback was presented visually for 500 ms after each response.

6.4.2 Results

The results for the AT and FF data did not differ substantively, and so are presented together.

Significant learning was observed (Fig 6.2). Group mean FDL improved from 10.5 Hz to 4.6 Hz [$t(9) = 3.76, p = 0.005$] in the AT listeners, and from 4.5 Hz to 3.6 Hz [$F(3, 42) = 10.6, p < 0.001, \eta_p^2 = 0.43$] in the FF listeners.

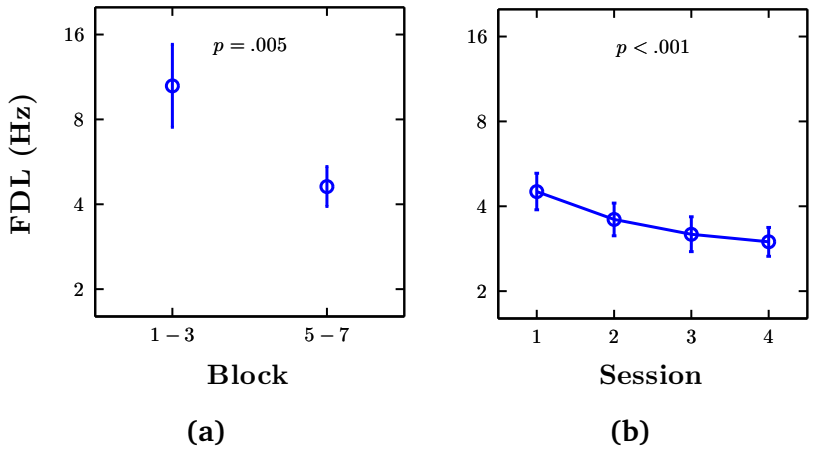


Fig. 6.2: Group mean (± 1 SE) learning effects for the AT (left) and FF (right) frequency discrimination tasks (Experiment II). Frequency discrimination limens [FDL] were evaluated at the 70.7% correct level.

The improved FDLs were accompanied by a reduction in bias in both the AT [$F(1,9) = 12.6, p = 0.006, \eta_p^2 = 0.58$] and FF [$F(3,42) = 4.00, p = 0.014, \eta_p^2 = 0.22$] listeners (Fig 6.3). As in Experiment I, bias also increased with increasing N [AT: $F(3,27) = 5.0, p = 0.007, \eta_p^2 = 0.29$; FF: $F(3,42) = 12.34, p < 0.001, \eta_p^2 = 0.73$]. An interaction between this effect and the learning was apparent. Thus, it can be seen in Fig 6.3 that in the experienced listeners bias increased more gradually as a function of N. However, this was interaction was marginal in the AT listeners [$F(3,27) = 3.0, p = 0.048, \eta_p^2 = 0.15$], and non-significant in the FF listeners [$F(9,126) = 1.7, p = 0.090, n.s.$].

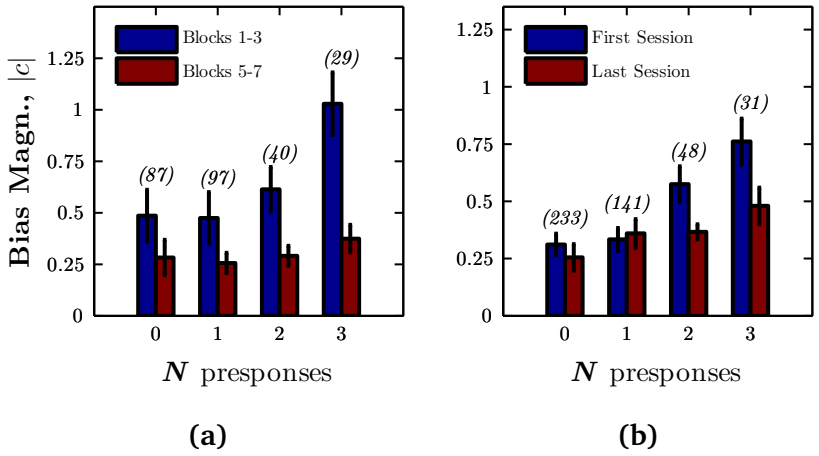


Fig. 6.3: Group mean (± 1 SE) bias magnitudes before and after learning, as a function of N correct presponses (Experiment II). The mean number of trials (per listener) are shown above each bar in parentheses. The left and right bars show bias before and after training, respectively. The left and right panels show the AT and FF data, respectively.

The reduction in bias was consistently observed within individuals (Fig 6.4). Thus, at $N = 3$ reductions in bias were observed in 9 of 10 AT listeners (90%), and 11 of 15 FF listeners (73%).

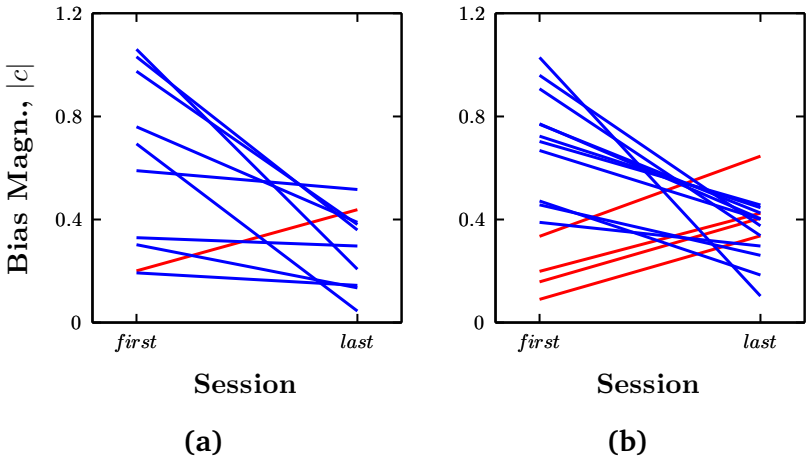


Fig. 6.4: Individual changes in bias magnitude before and after practice, evaluated at $N = 3$ (Experiment II). Reductions and increases in bias are indicated in blue and red, respectively. This represents a subset of the grouped data given in Fig 6.3.

6.4.3 Discussion

Both datasets demonstrated that listeners learn to reduce nonstationary bias. Thus, the response dependencies observed previously in naïve listeners (Experiment I) were replicated. But these effects decreased with practice, along with improvements in frequency discrimination performance. The smaller decrease in bias observed in the FF listeners is consistent with the smaller learning effect.

These results indicate that some of the perceptual learning effect is due to a reduction in nonstationary bias. In Experiment III, we attempted to quantify *how much* learning is accounted for by changes in bias.

6.5 Experiment III: Simulations of learning

To relate the levels of bias observed in human listeners to changes in psychophysical performance, we simulated two paradigmatic auditory learning tasks: temporal-interval discrimination (Wright et al., 1997; see also Karmarkar and Buonomano, 2003), and frequency discrimination (the 200 Hz condition of Demany, 1985). Discrimination limens, DL, were estimated from the trial-by-trial responses of simple signal-discriminators, which were either ideal in their decision making, or inhibited by a systematic nonstationary bias. In one condition ('Estimated Bias'), bias magnitude was fixed at levels approximate to those observed in naïve

human listeners (Experiments I & II). In a second condition ('High Bias'), bias magnitude was a free parameter which was fitted to provide thresholds equivalent to those observed in the naïve human listeners of Wright et al. (1997) and Demany (1985). The difference between DLs, with and without bias, was taken as a measure of how much learning could potentially be attributed to changes in nonstationary bias. To the extent that DLs are greater in the High Bias condition than the Estimated Bias condition, learning cannot be explained by a reduction in the form of nonstationary bias currently under consideration.

A systematically changing criterion can be formalised in a number of ways. The magnitude of criterion-shifts may be random (e.g., Dorfman et al., 1975) or deterministic (e.g., Dorfman and Biderman, 1971). Similarly, the occurrences of these shifts may be deterministic (e.g., Kac, 1969, 1962) or probabilistic (e.g., Thomas et al., 1982), and may be contingent on various factors, such as the occurrences of incorrect responses (e.g., Kac, 1969), the occurrences of either incorrect or correct responses (Dorfman and Biderman, 1971), or in accordance with some stochastic optimisation procedure (e.g., Erev, 1998). Here we assume a relatively simple, additive, deterministic process. We assume that each observer maintains a single fixed baseline criterion (which may be more or less ideal), but that this criterion is additively shifted depending on previous responses and their outcomes (cf. Dorfman and Biderman, 1971; Kac, 1962). It is important to note that we are not suggesting that listeners are unable to maintain a stable criterion, but that listeners systematically shift their criterion in a manner dependent on previous responses.

6.5.1 Method

Listeners were simple signal detectors, ideal except for a zero-mean Gaussian internal noise that was independently, additively combined with each stimulus observation. The standard deviation of the internal noise was set so as to produce thresholds that qualitatively corresponded to the post-training thresholds observed in human listeners. Performance thresholds were estimated from adaptive tracks as per the parameters and procedures reported in the associated paper (e.g., starting value, step sizes, n trials, n reversals, n listeners, etc.).

To assess the impact of bias, thresholds were estimated both when the criterion was invariantly ideal, and when the criterion was conditional on presponses. In the presponse-conditional case, the criterion location was determined as follows. On the first trial of a block, and after any response that differed from the immediately preceding response, the criterion was set to the ideal value ($\frac{d'}{2}$). After every correct response the criterion was

cumulatively shifted x z-score units towards the responded interval (i.e., making repetition more likely). After every incorrect response, the criterion was cumulatively shifted x z-score units away from the responded interval (i.e., making alternation more likely). In the Estimated Bias simulations, the value of x was set to roughly approximate the cumulative bias values observed in Experiments I and II [$0.25 \pm 0.2SE$; cf. Fig 6.1]. In the High Bias simulations, x was optimised to reproduce the starting performance observed in Wright et al. (1997) and Demany (1985).

The simulations of Wright et al. (1997) and Demany (1985) were executed independently, and each simulation was repeated 256 times.

6.5.2 Results and Discussion

The results are shown in Fig 6.5. The Estimated Bias condition assessed how much learning would result if the bias observed in naïve listeners was removed. In both simulations, DL*s* improved after the 0.25 cumulative bias term was removed [both $p < 0.001$]. However, the rates of improvement were not as great as were observed in human listeners. In both cases, the bias-unbiased change in DL was 38% as great as the observed improvements.

The High Bias condition assessed how much bias would be required for *all* learning to be explained by changes in bias. The Wright et al. (1997) data were well fitted by a cumulative shift in bias of ~ 0.68 after every identical response [$t(501) = -217.64, p < 0.001$], while the Demany (1985) data were fit by a cumulative bias of ~ 0.75 [$t(501) = -107.28, p < 0.001$].

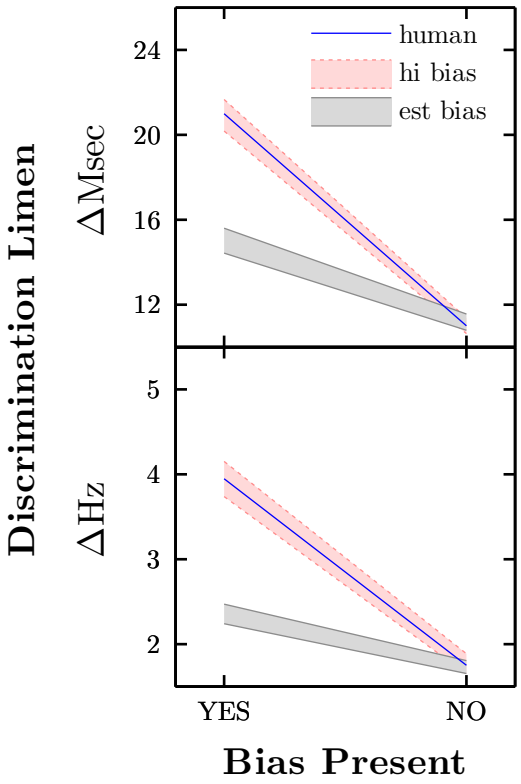


Fig. 6.5: Effect of bias on simulated discrimination limens, as measured by adaptive tracks. Independent simulations were run with parameters from Wright et al. (1997) (top panel) and Demany (1985) (bottom panel). The solid blue line shows the observed group-mean performance of human listeners. Shaded regions show mean ($\pm$ 1 SD) simulated DLs with and without bias. In the Estimated Bias condition (black solid) the level of cumulative bias (c per N) was similar to that observed in naïve listeners (cf. Fig 6.1). In the High Bias condition (red dashed), bias level was a free parameter, fitted to the group-mean human data (see body text).

These results indicate that changes in bias can explain typical learning effects, but that, given the levels in bias estimated in Experiment I, the removal of bias can explain only 38% of the observed learning effects . The remaining learning effect may be due to changes in perceptual sensitivity, or due to reductions in other forms of bias. The full extent to which bias underlies learning may be difficult to codify due to the potential for response-dependencies that are too sparse to measure within listeners, and too esoteric to measure across listeners.

Here we have taken the relatively crude approach of modelling criterion shifts in a purely deterministic manner. In reality, human listeners probably act more like weighted finite-state automatons (cf. Speeth and Mathews, 1961). Herein , a criterion shift is modelled as a probabilistic event, and it

becomes necessary to model both the magnitude and the relative likelihood of a shift given preceding events. Such an approach would likely yield a picture in which criterion shifts occur less frequently, but with greater effect. Unfortunately, such a model would likely be under-constrained by a typical perceptual learning dataset.

6.6 General Discussion

This study investigated the extent to which reductions in bias underlie auditory learning. By analysing trials conditional on their presponses, significant levels of bias were evidenced in naïve listeners. Such bias was found to decrease with practice. Simulations suggested that the observed decreases in bias explained some but not all of the improvements seen in typical learning studies. The remaining (majority) of learning may be due either to changes in sensitivity (cf. Chapter 2), or via forms of bias other than those measured here.

An obvious concern with the present work is whether the levels of bias evidenced were to some extent a reaction to the difficulty of the task (which in some cases was impossible), and to what extent the effects generalise to situations where there are substantial stimulus differences between intervals. Here it is important to distinguish between two disparate considerations. Firstly, it may be argued that the observed shifts in criterion are negligibly small given the decision variable magnitudes expected under more realistic conditions. While this is trivially true for very large signal magnitudes, most psychophysical studies are concerned principally with threshold performance, where signal magnitudes are by definition small. Accordingly, it was demonstrated in Experiment III that the levels of bias observed in Experiment I were sufficient to cause difference limens to be significantly underestimated by an adaptive staircase. Secondly, it may be argued that the observed shifts only occurred as a reaction to the difficult (at times impossible) conditions, and that such shifts simply do not occur when stimulus differences are apparent. In the extreme this hypothesis is unlikely given the data in Experiment II. Thus, the use of interleaved adaptive tracks in the adaptive data, and jittering in the fixed frequency data, meant that listeners rarely experienced long runs of sub-threshold stimuli. Nonetheless, levels of bias were smaller than in Experiment I (where discriminations were impossible). This may indicate that shifts in criterion are to some degree modulated by the perceived difficulty of the task. Alternatively, this reduction may be a statistical artefact, resulting from the fact that bias tends to be progressively underestimated as listener sensitivity increases (especially when sample sizes are small; see ∮6.A).

Robustly disambiguating these effects is a non-trivial task, and would likely require datasets orders of magnitude greater than those reported here.

The causes of presponse-conditional bias in naïve listeners are potentially multifarious. One set of drivers can be termed 'statistical'. Listeners may assume that trials are autocorrelated, such that the outcome of one trial affects the likelihood of its subsequent recurrence. This may occur if, for example, listeners believe (in some cases correctly; e.g., Lindman and Edwards, 1961) that stimuli are drawn without replacement from a balanced set. Relatedly, it may be that biased behaviour arises from a poor understanding of randomness, with listeners misestimating the probability of runs occurring. Demand characteristics may also promote biased behaviours. For example, some listeners may believe that a particular pattern of responses may give the impression that they are being inattentive, uncooperative or are otherwise malingering, and may adjust their responding accordingly. Given that these considerations are largely non-specific to the present task, we predict that the reported results will also hold for mAFC tasks in other domains and modalities. It is also likely that this effect generalises to other closed-set response paradigms (cf. Garner, 1953), though data sparsity may make bias prohibitively difficult to evidence as the number of response alternatives increases.

If listeners are indeed learning to generate independent responses, then one might expect learning to transfer between sequential response tasks. In contrast, learning is often considered to be stimulus specific (e.g., see Fahle, 2005; Green and Bavelier, 2008). For example, improvements may be specific to a particular temporal interval (Karmarkar and Buonomano, 2003; Wright et al., 1997) or visual orientation (Fahle and Edelman, 1993; Fiorentini and Berardi, 1980; Karni and Sagi, 1991). It is not clear why bias should reset between stimulus configurations. However, the problem of reconciling these facts may be obviated by a growing body of evidence, which suggests that a portion of learning does in fact transfer between tasks (Wright and Zhang, 2009). For example, Irvine et al. (2000) found that training to discriminate frequencies around 5 kHz induced similar, though smaller, improvements at 8 kHz, and vice versa (see also Demany, 1985). Similarly, Jeter et al. (2010) reported analogous results for visual orientation discriminations that differed either in terms of orientation or retinal location. Interestingly, the proportion of transfer was greatest in observers who had trained least (1248 trials). This timescale is consistent with the timescales for bias reduction reported both here and in the gambler's fallacy literature (e.g., Ayton and Fischer, 2004; Anderson, 1960; Jarvik, 1951). This timescale is also – as Jeter et al. (2010) notes – consistent with the early, rapid stage of perceptual learning (Hawkey et al.,

2004; Poggio et al., 1992). It may therefore be that not only is perceptual learning constituted by a number of different mechanisms, but that these mechanisms exhibit different temporal dynamics. Thus, we conjecture that there may exist an early phase in which learning is fast, generalisable, and may involve non-sensory elements, as well as a second, more protracted phase in which the learning is more gradual, more specific to a particular sensory configuration, and which may principally involve physiological changes in primary-sensory networks.

Perceptual learning is often considered a potential tool for enhancing or remedying everyday sensory abilities. If changes in bias underlie observed learning effects, then its usefulness may be contingent on the degree to which such biases arise spontaneously in everyday life. Alternatively, sequential effects may be an artefact of the psychophysical testing procedure, where a small and closed set of outcomes are repeatedly sampled over a timeseries. There are tentative reasons to favour the latter. For example, Barron and Leider (2010) found that observers were biased in their predictions of future events when a binary sequence was presented sequentially over time. This bias was attenuated when the entire preceding sequence was presented simultaneously. This suggests that the type of response dependencies considered here may only be relevant in the quintessentially psychophysical situation where the listener is presented with a Bernoulli process. This fact may also have a bearing on why listeners with greater working memory commonly exhibited enhanced sensory thresholds (e.g., Ahissar and Hochstein, 1997). That is, it may be that since they are able to integrate over a greater number of trials, and so are less prone to be mislead by localised vagaries such as the occurrence of runs of identical outcomes.

Differences in response-dependent bias may also be extended to explain some of the developmental differences observed in basic sensory tasks. This proposition is particularly germane for two reasons. Firstly, because children anecdotally appear to be highly influenced by the events of preceding trials when making responses. And secondly, since even static interval biases in single interval (yes/no) designs are often (e.g., Werner et al., 1992; Trehub et al., 1991), though not always (Werner and Marean, 1991), inflated in younger listeners relative to adults. It would therefore be instructive to examine whether response-dependencies decrease as a function of age as well as a function of learning.

6.7 Conclusions

(1) In a 2AFC design, listeners' responses are conditional on the events of previous trials. Successive identical, correct responses increasingly encourage perseverance. Short runs (N = 1) of identical incorrect responses encourage alternation, though this reverts to perseverance after longer chains (N = 3).

(2) Sequential response dependencies represent a systematic inefficiency that limits performance in naïve observers.

(3) Sequential bias is significantly reduced through practice, though it is not eradicated in all listeners even after several thousand trials.

(4) A reduction in bias explains some, but not all of the perceptual learning effect.

Notes

[1]Moreover, a large number of studies have demonstrated that listeners are certainly *capable* of integrating over sequential observations (e.g., Swets, 1959; Berg, 1990)

[2]If only the identity of the immediate presponse were used then this analysis is virtually equivalent to performing a serial-correlation (e.g., Hoel, 1947).

Acknowledgements

We thank Yuxuan Zhang for comments on an earlier draft of this manuscript. This work was supported by the Medical Research Council, UK (Grant: U135097130).

6.A Statistical bias in bias measures

Small sample sizes have been shown to statistically bias estimates of d' given a fixed (ideal) criterion, λ. For example, Miller (1996) showed that with small numbers of observation, low values of d' tend to be overestimated, while high values tend to be underestimated. An analogous statistical bias for estimates of c may be a confound for the present experiment, since the number of samples tended to vary with N presponses. To examine how sample size affects the sampling distribution of c, numerical simulations analogous to those of Miller (1996) were run. Monte Carlo estimates of c were made as both d', c, and the number of trials were independently varied.

6.A.1 Results and Discussion

The matrix of results is shown in Fig 6.6, with the associated mean values given in Table 6.4. The first column of values essentially recapitulates the finding of Miller (1996) that as the number of observations decreases, higher levels of bias become progressively underestimated. As with Miller (1996), there is also a tendency for low levels of bias to be overestimated when using small numbers of observations ($\sim$32). However, this effect is marginal, with c being overestimated by $< 10\%$ in the greatest case here examined. Moving from left to right across the panels, it is also possible to see how this effect varies as sensitivity increases. Namely, as d' increases, expected values of bias are increasingly underestimated. Variance in c estimates tend to decrease as the number of observations increase, and as values of c and/or d' increase.

These findings suggest that statistical bias is unlikely to have greatly affected our conclusion that bias increases as a function of N. In some cases bias may have been overestimated. However, at higher levels of N, where observations were fewer and true values of c probably greater, bias may actually have been substantially underestimated.

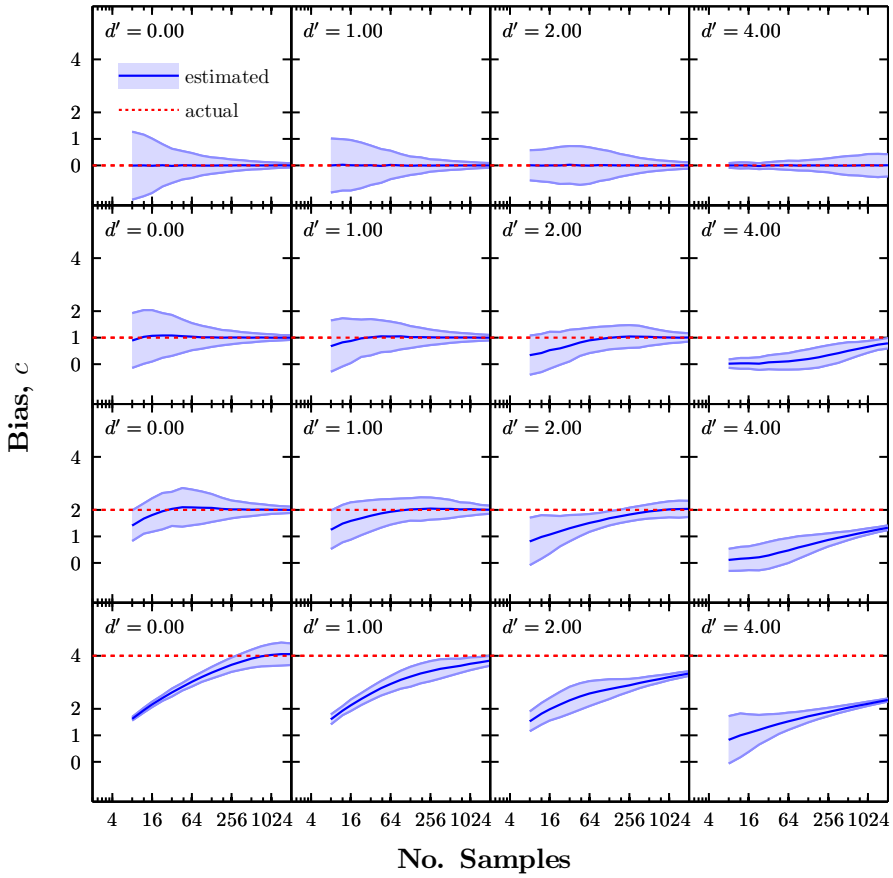


Fig. 6.6: Estimates of bias, c, in a forced choice task, as a function of d' and the number of samples. Mean estimates of bias ($\pm$ 1 SD) are shown in blue. The horizontal dashed lines show the true bias value. No. Samples is the total number of samples, of which half contained the signal in the first interval (SN). Bias was estimated at each trial level from 10,000 Monte Carlo samples. Associated mean values are given in Table 6.4.

		N Samples								
c	d'	8	16	32	64	128	256	512	1024	2048
0.0	1.0	-0.01	-0.01	-0.02	-0.00	-0.01	0.00	0.00	0.00	-0.00
	2.0	-0.00	0.01	0.01	0.02	-0.01	-0.01	-0.00	0.00	-0.00
	3.0	0.01	0.00	0.02	-0.01	0.00	0.00	-0.00	-0.00	-0.00
	4.0	-0.00	-0.00	-0.00	0.00	0.00	0.00	-0.01	0.00	0.00
1.0	1.0	0.89	*1.07*	*1.08*	*1.04*	*1.01*	*1.01*	*1.00*	*1.00*	*1.00*
	2.0	0.68	0.88	*1.02*	*1.04*	*1.02*	*1.02*	*1.01*	*1.01*	*1.00*
	3.0	0.34	0.53	0.70	0.90	0.99	*1.05*	1.03	1.01	1.01
	4.0	0.02	0.03	0.07	0.11	0.20	0.34	0.50	0.66	0.79
2.0	1.0	1.41	1.81	*2.04*	*2.10*	*2.08*	*2.02*	*2.01*	*2.01*	*2.00*
	2.0	1.25	1.58	1.78	1.93	*2.02*	*2.05*	*2.04*	*2.02*	*2.01*
	3.0	0.81	1.07	1.31	1.51	1.68	1.83	1.95	*2.02*	*2.04*
	4.0	0.11	0.18	0.29	0.47	0.68	0.87	1.03	1.19	1.33
4.0	1.0	1.62	2.16	2.61	3.01	3.36	3.66	3.89	*4.03*	*4.06*
	2.0	1.60	2.13	2.57	2.93	3.21	3.42	3.57	3.70	3.81
	3.0	1.53	1.98	2.32	2.57	2.74	2.89	3.05	3.19	3.33
	4.0	0.83	1.08	1.32	1.53	1.71	1.88	2.04	2.19	2.33

Table 6.4: Mean estimates of bias, c, in a forced choice task, as a function of d' and the number of samples. Overestimates of non-zero bias values are shown in italics. These values are shown graphically in Fig 6.6.

6.B χ^2 analyses of sequential dependencies

In addition to multiple regression analyses, chi-square contingency tables were also used to assess statistical independence between responses. Responses were categorised according to the selected interval [1 or 2], the presponse interval [1 or 2], and whether the presponse was incorrect or correct [0 or 1]. The chi-square test was used to test whether the $4x2$ contingency table of observed values significantly differed from the table of uniformly distributed values that would be expected given independent trial-by-trial responses.

Naïve listeners

The group-aggregate contingency table for Experiment I is given in Table 6.5. It indicates that the responses of naïve listeners were conditional on the immediately preceding trial. Specifically, listeners tended to alternate their responses after incorrect presponses, and perseverate after correct presponses. These deviations from a uniform response pattern were significant [$\chi^2(3, 14883) = 303.3, p < .001, V = 0.14$].

Presponse		Response Identity							
		Group		Best		Median		Worst	
Identity	Correct	1	2	1	2	1	2	1	2
1	0	1633	2048	10	82	42	69	160	36
	1	2074	1527	57	37	37	61	152	33
2	0	2044	1728	93	56	71	65	40	17
	1	1531	2298	26	139	58	96	29	32

Table 6.5: Number of responses, contingent on presponse identity and correctness (Experiment I). The group data is aggregated over all listeners. The Best, Median, and Worst data show individual data, fitted to the idealised group-aggregate response-pattern (see body text).

At the individual level significant contingencies were also found in 23 of 30 listeners [$p < .01$]. These deviations generally followed the same pattern as the group aggregate responses, though to a varying degree. To quantify the similarity between individual listeners and the group-aggregate profile, the observed responses of each listener were compared, via the chi-square statistic, with those predicted by a listener who always alternated when incorrect and perserverated when correct. The values for the best, median and worst fitting individuals are given in Table 6.5. As per the group-aggregate, the best and median fitting individuals alternated after

incorrect presponses, and perseverated after correct presponses. The worst fitting individual exhibited a more general 'Interval 1' preference.

Learning

The group-aggregate response counts for Experiment II are shown in Table 6.6. The distribution of responses was initially similar to that found in Experiment I (Table 6.5), though in the AT cohort a stationary 'Interval 1' bias was also apparent. In both cohorts responses were dependent on the previous trial before [AT: $\chi^2(3, 4413) = 61.83, p < 0.001$; FF: $\chi^2(3, 10075) = 157.73, p < 0.001$] and after [$\chi^2(3, 10196) = 53.00, p < 0.001$] practice. However, the size of this effect decreased with practice [Cramer's V; AT: $V_{1st} = 0.12$; $V_{2nd} = 0.07$; FF: $V_{first} = 0.13$; $V_{last} = 0.07$], indicating a reduction in response dependencies. Deviations from chance were also observed in fewer listeners after training [$p < .01$; AT: $6 \rightarrow 4$; FF: $9 \rightarrow 6$]. The regression p values were strongly correlated with those from the chi-square test [$r = .74, p < .001$], and are consistent with the presponse-dependent bias analyses.

Presponse		Response Identity							
		AT dataset				FF dataset			
		First Half		Second Half		First Session		Last Session	
Identity	Correct	1	2	1	2	1	2	1	2
1	0	313	236	207	267	683	1051	581	884
	1	1120	664	1058	941	1713	1376	1645	1752
1	0	210	162	265	332	973	912	905	848
	1	847	861	976	1127	1464	1903	1768	1813

Table 6.6: Group-aggregate number of responses, contingent on presponse identity and correctness, before and after practice (Experiment II).

CHAPTER 7

General Discussion

In this chapter the experimental results of Chapters 2 – 6 are reviewed, and their implications discussed. It is concluded that auditory perceptual learning is subserved by multiple mechanisms, which: Operate in parallel; vary in importance depending on the task demands; and include both sensory and non-sensory processes. Limitations of the present work are discussed, and future research suggested.

7.1 Mechanisms of auditory perceptual learning

THE aim of this thesis was to establish what changes underlie auditory perceptual learning. A model of decision making was proposed, containing four potential limiting constructs (internal noise magnitude, encoding efficiency, bias, and inattentiveness). These constructs were modelled using a variety of behavioural techniques, and changes evaluated as a function of practice. The results pertaining to each construct are summarised in Fig 7.1, and will be discussed in turn.

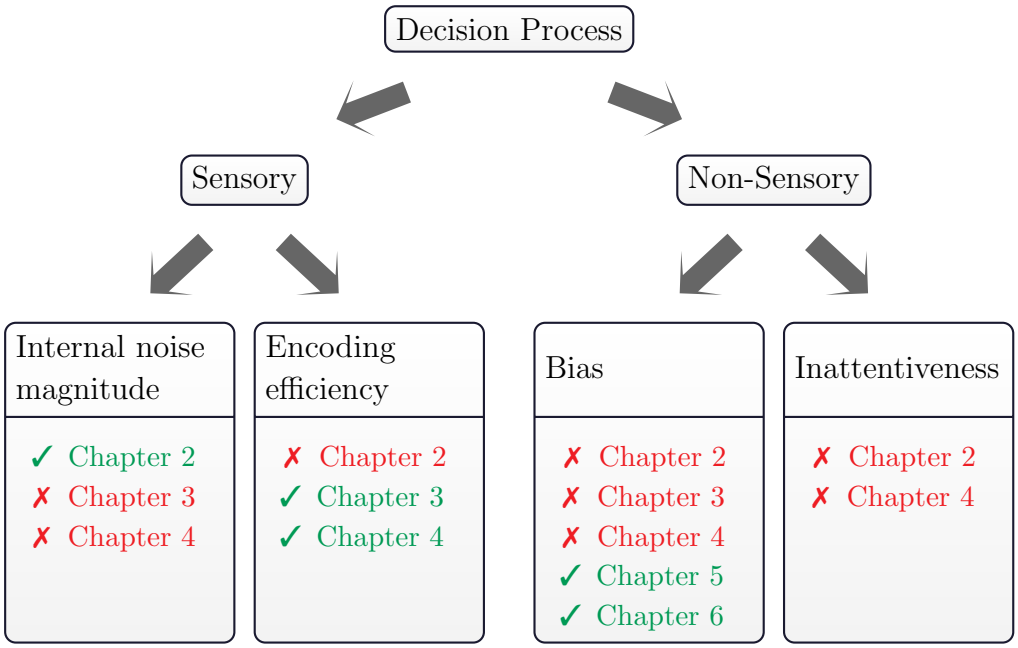


Fig. 7.1: Summary of findings from all studies, with regards to the four principle mechanisms of internal noise magnitude, encoding efficiency, bias, and inattentiveness. Improvements are indicated by green ticks, while red crosses denote no significant change.

7.1.1 Internal noise magnitude

Quantities of internal noise were indexed using four methods in Chapter 2. The four methods produced estimates that were in remarkably good agreement. Accordingly, estimates of internal noise magnitude were derived solely from psychometric fits in Chapter 3. In a simple pure tone discrimination task a reduction in additive internal noise reduction appeared to be the primary (Chapter 2), though not sole (Chapter 6), mechanism of learning. However, a more complex task, *prima facie* differences in internal noise magnitude disappeared once changes in encoding strategy were accounted for (Chapters 3 and 4). This highlights how unhelpful, and potentially misleading, it can be to rely on internal noise as the sole explanation of performance differences.

7.1.2 Encoding efficiency

Encoding efficiency was conceptualised as the appropriateness with which listeners weighted various information channels. When, as in Chapter 3, these channels were spectrally distributed, encoding efficiency was interpreted as an index of selective attention. Efficiency was estimated either by comparing performance to the ideal (Chapter 2), or by attempting to directly estimate listeners' listening strategies (Chapter 3, Chapter 4). In a very simple task no improvements in stimulus encoding were observed

(Chapter 2), potentially because even naïve listeners were already close to optimal. On a more complex task in which the stimulus was masked by unpredictable noise, changes in encoding efficiency were the primary mechanism of learning (Chapter 3, Chapter 4). It may therefore be that as the complexity of the stimuli increases, changes in encoding efficiency progressively swamp any changes in internal noise magnitude.

7.1.3 Bias

No substantive bias effects were observed in the first three studies. Thus, although in Chapters 2 and 3 a slight change in bias direction was observed, actual bias magnitudes differed relatively little, and were close to zero throughout. Similarly, in 4, naïve listeners exhibited no consistent predilections towards any particular response alternative. However, significant changes in bias were evidenced in Chapter 5 and Chapter 6. These changes were capable of explaining a substantial proportion of the observed learning.

That substantive effects were only observed in Chapters 5 and 6 was due to differences in the experimental paradigm and the form of bias considered, respectively. In Chapter 5, bias was assessed in a yes/no paradigm. In accordance with the wider literature (e.g., see Macmillan and Creelman, 2005), naïve listeners were more prone to biases here than with the mAFC paradigm used in the earlier studies. Accordingly, listeners tended to be liberal initially, favouring 'yes' over 'no' responses. This bias effect was eradicated through practice. Chapter 6 moved beyond the traditional measure of bias (a stationary response preference), and examined nonstationary bias, to which end response preferences were measured conditional on events of previous trials. Such effects are sometimes termed 'sequential dependencies' rather than bias, but the upshot was nonetheless that listeners became inclined to favour a particular response alternative, independent of the sensory evidence. This form of bias was shown not to be apparent given common measures of decision making (e.g., c; d'), was shown to be present even in forced-choice (mAFC) designs, and was shown to be attenuated (though not eliminated) by practice. The size of the observed effects were especially noteworthy given that only one subset of dependencies were considered; the reported estimates therefore represented only a lower limit on stationary bias.

7.1.4 Inattentiveness

The potential role of inattentiveness was considered primarily in Chapter 2, and also to a lesser extent in Chapter 4. No significant changes were observed.

Consistent deterioration was observed towards the end of most studies. These changes were not significant within any given study, but taken together these observations may be considered noteworthy. Such deterioration can most parsimoniously be attributed to inattentiveness due to boredom or fatigue. However, the size of this effect was generally small relative to the learning[1]. The role of inattentiveness in auditory learning therefore appears slight at most, and more related to fatigue than the learning *per se*. However, the apparent unimportance of inattention may to some degree reflect the nature of the listeners tested (see ∮7.3.2).

7.1.5 Overview

One corollary of these results is that there is no unitary mechanism of learning. Multiple factors, both sensory and non-sensory in nature, are liable to contribute within any given task. Moreover, it appears from Chapters 2 and 3 that the mechanisms of learning are also liable to differ across tasks, implying that it makes limited sense to inquire into 'the mechanisms of learning' *per se*, independent of any specific task.

That the listener's encoding strategy became the limiting factor when complexity increased may reflect differences in task demands, with lower order factors progressively dominating as complexity decreases (cf. the Reverse Hierarchy Theory of Ahissar and Hochstein, 2004). Alternatively, such differences in complexity may be more germane to the researcher than the listener. To wit, on very simple tasks it becomes difficult to formulate or measure, in behavioural terms, what the listeners encoding strategy may be, and in such cases internal noise becomes the *de facto* explanation.

To see how this may be the case, consider how the information channels were conceived in Chapters 2 and 3. In Chapter 3 each channel was a frequency-specific filter, and the relative weights given to these filters determined how much (task-relevant) information was extracted from each incoming sound. Conversely, in Chapter 2 each channel was an observation interval, and their relative weights determined how effectively information was integrated across observations. The former is a deeper level of explanation than the latter, since if no information is extracted from either sound then the efficiency with which each observation is integrated becomes moot. For an equitable comparison, the weights in Chapter 2 could be recast so as to similarly pertain to frequency-specific channels. When such a model was simulated, improvements in spectral weights *were* shown to be capable of explaining the observed learning (but given the nature of the stimuli no such changes could be evidenced in human observers).

[1]The one possible exception to this was one listener in Chapter 5, who underwent a marked deterioration in performance over seven practice sessions.

One implication of this is that just as one cannot talk about which mechanism underlie learning independent of a specific task, it may be similarly senseless to talk about the mechanisms themselves, independent of a reference to a specific level of description. The level of description is more gross in Chapter 2 than Chapter 3, and had they been equivalent then similar changes may have been observed in both. Given the most basic level of description available, learning in Chapter 2 was legitimately seen as resulting from decreased internal noise. But physiological techniques, or more elaborate behavioural paradigms, may yet be able to re-describe this learning in terms of modified channel weights. This point is further expounded in §7.3.3, where it is argued that such lower-level explanations are generally preferable, and should be sought where possible.

7.2 Implications

The work of this thesis has a number of implications, both for the auditory learning literature and more generally.

7.2.1 The design of learning tasks

There is a growing interest in using perceptual learning as a tool for enhancement, remediation or acclimatisation (e.g., Levi and Li, 2009; Merzenich et al., 1996; Tallal et al., 1996). The fact that the mechanisms of learning vary depending on the training task makes it imperative that the desired learning outcomes are carefully considered when designing such materials. For example, the mechanisms of learning appeared to fundamentally differ depending on whether the signal was presented in quiet, or obscured by external noise. Since understanding speech in noisy situations is particularly difficult for hearing-impaired listeners (Plomp, 1978), it is recommended that training materials for speech comprehension, such as those administered to recipients of hearing aids, should similarly focus on masked stimuli (n.b. see also Dosher and Lu, 2005).

This teleological approach to design also brings into question the fundamental usefulness of training paradigms in which listeners are asked to repeatedly judge a sequentially unfolding Bernoulli process. In such situations, a substantial proportion of the learning appears to concern the statistical independence of each trial (§7.1.3). This learning is unlikely to be of real-world benefit, and so may be redundant, or even distracting, in therapeutic contexts. It may therefore be more efficient to design training materials in which the independence of each trial is intuitive, and where bias effects do not arise in the first place. This may be achievable through careful task instructions, the use of more realistic tasks (see §7.3.1), and/or

by varying the set of response options on each trial, so that listeners are less likely to intuit a connection between the outcome of one trial and the answer of the next.

7.2.2 Obtaining pure measures of hearing sensitivity

The results from Chapters 5 and 6, concerning bias, highlight the importance of training for investigators wishing to obtain a 'pure' measure of sensory aptitude. In this respect, the auguries are mixed. In both studies, significant bias was evident in naïve listeners. In Chapter 5, the stationary bias effects observed in yes/no detection were effectively eradicated after 3×600 trials practice. This suggests that a relatively small amount of practice is sufficient to obviate any confound of stationary bias on performance[2]. However, the nonstationary biases observed in Chapter 6 persisted, albeit at a diminished level, even after several thousand trials of practice, suggesting that the inclination to perceive contingencies between trials is deep rooted in listeners. It remains to be seen whether such bias can ever be truly eliminated, and thus whether a truly pure measure of sensory ability derived. However, for most purposes, the behaviour of listeners given one week's practice may provide a sufficient approximation to a purely empirical decision process.

7.2.3 Explaining population differences

More generally, the work concerning nonstationary bias clearly disproved the widely asserted assumption that forced-choice paradigms are 'bias free' (Kingdom and Prins, 2009; Gescheider, 1997). This knowledge may aid our understanding of the performance-differences that exist between certain populations. For example, children, clinical populations, and older listeners, all often exhibit lower performance on sensory judgement tasks than normal hearing adults. Amongst such listeners, it is not usual to 'train-out' bias effects prior to testing. The differences in performance may therefore represent non-sensory differences in bias. Hitherto, such explanations have been largely discounted on the basis that forced-choice tests are typically employed, and that "[since such tests are bias free] differences in response criteria are not... plausible" (Halliday, 2005, p. 181). The results of Chapter 6 demonstrate that the premise of this argument is false. Bias does occur even amongst normal-hearing adult listeners, and such bias may be even greater amongst younger listeners, who may struggle with the concept of independent, randomly sampled trials, or in hearing-impaired listeners, who may have learned to place less reliance on sensory evidence in favour

[2]It may be further possible to consolidate the requisite practice into a single session (though perhaps not, Molloy et al., 2012)

of *a priori* assumptions. It would therefore be instructive to examine the relative extent of nonstationary bias amongst psychoacoustically naïve cohorts of children, hearing impaired, and other such populations, where asymptotic practice is impractical and not de rigueur.

7.2.4 Similarities between audition and vision

Where commonalities have existed, the findings of this thesis have tended to be in good agreement with the those of visual learning studies. As in vision, learning in quiet was subserved by internal noise reduction, while learning in noise was subserved by a combination of internal noise reduction and external noise exclusion (Dosher and Lu, 1998). Moreover, both forms of learning were shown to be explainable by a neuronal process of incremental reweighting, as has been suggested to occur in vision (Petrov et al., 2005). When our conclusions have differed with those in the vision literature, those differences have been explained entirely by the interpretation of the data (cf. Chapter 2, Chapter 5).

The investigations of non-sensory factors are without comparison in the visual literature. However, it seems plausible that similar results would be found using visual stimuli. I conclude therefore, in line with other researchers (Nahum et al., 2010), that the principles of learning are common across vision and audition.

7.3 Limitations and future work

7.3.1 Other tasks

The present work has been almost exclusively preoccupied with frequency discrimination (the ability to detect changes over time) and selectivity (the ability to resolve the frequency components of a complex sound). In the first instance, it would therefore be instructive to verify that the same findings apply given other psychoacoustical tasks, such as temporal or spatial discrimination (cf. §1.1.1). However, beyond simply constructing a compendium of tasks, there are also a number of specific task-differences, for which it may be of particular interest to apply the present techniques.

More realistic tasks

In the present work, auditory perception was evaluated using basic psychoacoustical tasks. Such methods afford accurate, parametric control over the task parameters. However, learning in more realistic situations may be governed by different principles. In particular, it is unclear whether learning in more realistic situations would be more or less driven by non-sensory mechanisms. Take, for example, the nonstationary bias effects

observed in Chapter 6. The response dependencies observed essentially reflect a logical fallacy concerning the autocorrelation of sequential events. Instances of similar logical fallacies have been shown to be greatly attenuated when problems are presented in a more realistic context (Cosmides and Tooby, 1992; Carraher et al., 2011). For example, in the classic Wason selection task, participants are presented a rule of the form 'if x, then y'. They are then presented with four pairs of values. In two pairs the antecedent is hidden, and in two the consequent is hidden. The participants must indicate which of the four pairs of values potentially falsify the rule. When the relations are relatively abstract, people overwhelmingly answer incorrectly[3]. However, Carraher et al. (2011) showed that participants find the task trivial when it is presented in a familiar social context[4]. Similarly, it may be that in more realistic listening tasks, the need to reduce bias is obviated and the mechanisms of learning are chiefly sensory. Conversely, other authors have indicated that as the complexity of the stimuli increase, the demands on memory and attention are greater (Amitay et al., 2005), in which case non-sensory effects may actually be increased in more realistic listening situations.

Tasks with a temporal dimension

In the tasks reported here, listeners were given an unlimited time to respond, and memory constraints were assumed to be negligible. These tasks represent one extreme of a spectrum, where speed is irrelevant and the only concern is to maximise accuracy. Listeners in such situations can be reasonably assumed to integrate over all available information when making a decision. However, in many situations the speed of the response is a competing, sometimes primary, concern. In such cases the decision variable takes on a temporal dimension, and it becomes meaningful to ask how efficiently listeners integrate information over time (Selen et al., 2012), how efficiently listeners trade-off the competing interests of speed and accuracy (Juni et al., 2012), and how both of these processes vary with practice. These aspects of learning may be considered a generalisation of encoding efficiency, and may be investigated using some the same methods as described in this thesis. Thus, the efficiency with which sequentially presented information is combined may be estimated via detection theoretic models (Swets, 1959), or computed directly from estimated weights (Berg,

[3] e.g., each card has a number on one side, and a colour on the other. Given the cards $\langle 3, 8, red, blue \rangle$, which card(s) must you turn over to test the proposition, 'all even numbers must be red on the other side'. Common answer: '8, red'. Correct answer: '8, blue'

[4] e.g., there are four people in a bar. A is 17 y.o., B is 22 y.o., C is drinking beer, D is drinking water. Which people must you ID to ensure that, 'all people drinking alcohol are over 18'.

1989), in much the same way as spectrally distributed information was studied in the present work.

Structural learning tasks

Studies of auditory perceptual learning, both here and more generally, have tended to concentrate on tasks where the decision strategy is relatively simple, and the task-relevant features clearly prescribed ('this is the signal, ☞, listen for it'). Learning on such tasks can be readily conceived in terms of the optimisation of parameters, such as encoding weights and response preferences. However, in the real world it is not always obvious what the stimulus features are, and how they relate to the task objectives. Thus, listeners must learn the relevant inputs, and the functional form of the equations relating the inputs to the decision variable. Such learning is termed 'structural learning' (Wolpert et al., 2011), and presents a number of challenges distinct from those hitherto considered. For example, there is the problem of feature extraction: how to cluster acoustic features in order to derive appropriate cues. There is also the problem of credit assignment: how to back-propagate the gross feedback in order to prune out the irrelevant features and bolster the useful ones. In both of these cases, the requisite additions to the current decision model (Fig 1.3) are not trivial, and a new lexicon may be required when discussing the mechanisms of such learning. In some cases, the demands of such structural learning may actually conflict with the typical process of outcome-maximisation through parameter optimisation. For example, Gureckis and Love (2009) present a task in which rewards are contingent on previous responses, such that the best overall strategy is always to choose the response that yields the smallest immediate reward. Such tradeoffs between short- and long-term gains has been a fruitful line of inquiry in the developmental literature (Kidd et al., in press; Rodriguez et al., 1989), and could similarly provide fresh insights into the relative priorities of listeners during learning. Tasks involving latent contingencies may also help us to more fully understand the apparent inefficiencies in human decision making. Thus, whereas in the present thesis internal noise was considered a nuisance to be eliminated, in Gureckis and Love (2009) the presence of internal noise effectively forced the listener to explore the decision space by guaranteeing (short-term) suboptimal responding. In some instances internal noise may therefore have beneficial properties, such as helping to prevent listeners becoming trapped in local minima of the decision space.

7.3.2 Other listeners

The adult listeners reported in this thesis were all 18–35 years old, normal-hearing, with no diagnosed behavioural problems. They also

tended to be well motivated, and of above-average IQ. The homogeneity of the samples may have led to the amount of individual variability being suppressed. It also raises questions as to whether the mechanisms of learning remain invariant across other populations. For example, as discussed in §7.2.3, the role of bias may be greater in hearing-impaired listeners who have grown to distrust their sensory input, while phenomena suggestive of inattentiveness have previously been reported in younger listeners (Huyck and Wright, 2011). Such hypotheses could be straightforwardly tested by applying the same measures to more diverse cohorts. However, the commonalities observed between learning and development suggest that learning mechanisms may actually be relatively well conserved, even across populations that vary widely in performance.

7.3.3 Other, deeper mechanisms

In essence, this thesis has been concerned with partitioning observed changes in decision making amongst various limiting constructs. This process is shown schematically in Fig 7.2. In mapping LEVEL I to LEVEL II, sensory factors are delineated from non-sensory factors. At the next level (LEVEL III), each of these factors are subdivided into their random (internal noise, inattention) and deterministic (encoding efficiency, bias) components. For the type of decisions considered here, such a scheme appears reasonably exhaustive in scope, but can be further specified in detail. This is illustrated in LEVEL IV. Here, encoding efficiency is reduced to the goodness of each individual channel weight, bias is reduced to the perceived payoffs and likelihoods associated with each response alternative, and internal noise is reduced to various potential subcomponents, such as those relating to criterion jitter, the mapping of decisions to responses, and the neural representations of sound. During this thesis only a few tentative steps have been made towards the degree of understanding depicted in LEVEL IV. Thus, in Chapter 3 and Chapter 4 actual encoding weights were estimated, while in Chapter 5 it was speculated that bias may have arisen through a misperception of relative frequencies. Boring down into constructs in this manner implies a deeper level of understanding, and allows for much more powerful predictions to be made. For example, from the weight estimates in Chapter 4 it was possible to characterise not just differences in efficiency, but how these differences were manifested across the spectral domain, and thus how performance was likely to differ as a function of stimulus frequency. An obvious next step would therefore be to further subdivide each of the principle constructs into their respective constituents, and to examine to what extent changes in each of these underlies learning. For example, response errors could be assessed by asking listeners to report accidental button presses given an impossible task.

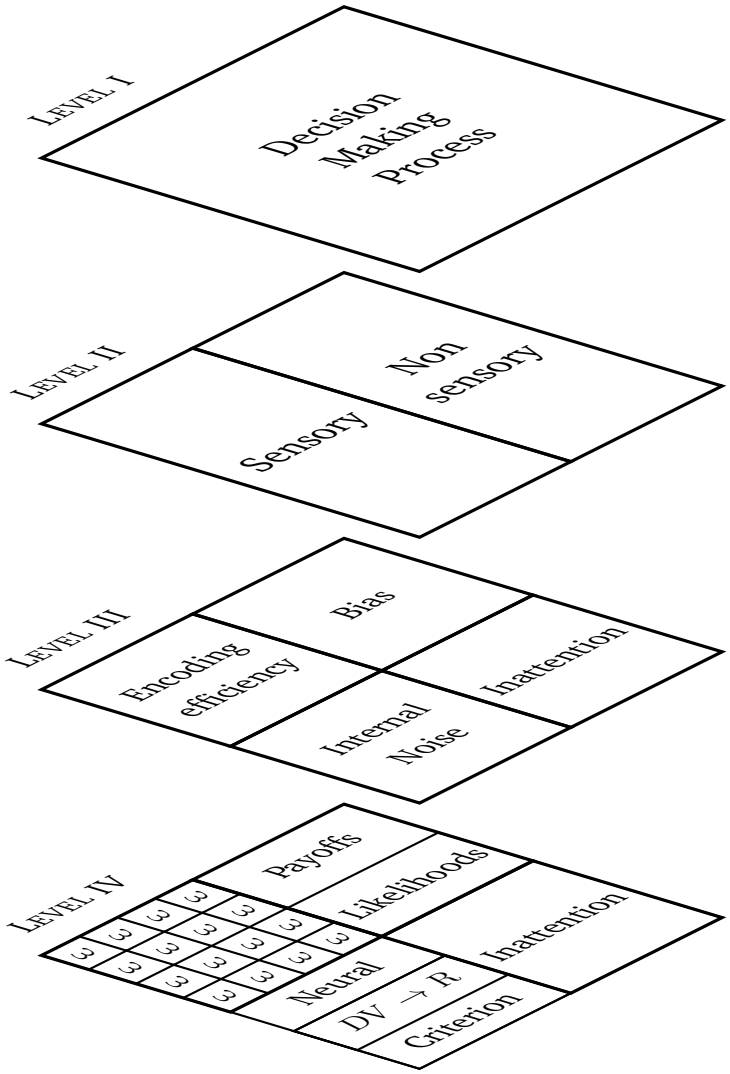


Fig. 7.2: Schema demonstrating how the decision process may be partitioned. Level I to Level III correspond to levels depicted previously in Fig 7.1.

Note, however, that Fig 7.2 follows popular convention in presenting random and deterministic factors as distinct, competing mechanisms, either of which may be hypothesised to underlie learning. Conversely, a possible outcome of this thesis is that random and deterministic factors may also be viewed, in some instances at least, to be different levels of description for the same phenomenon, with deterministic factors potential being the more elemental of the two.

This can be seen most directly Chapter 2, where changes in internal noise evidenced at the behavioural level were modelled as changes in frequency-tuned channel-weights at a pseudo-neuronal level. It was also demonstrated in Chapters 3 and 4, where age- and practice-related differences in internal noise magnitude, apparent when ideal encoding strategies were assumed, disappeared when estimated patterns of

channel-weights were factored into the model. Finally, that changes in random processes can be re-described in deterministic terms was also implied, albeit more implicitly, in the study of mAFC bias reported in Chapter 6. To see how this is so, consider that another (precedented; Kubovy and Healy, 1977) approach to modelling those data would be to estimate the a global (mean) criterion for each listener, and observe when and how often this cutoff rule was breached (i.e., how often listeners responded contrary to what their mean criterion would predict). This would have led to a broadly Gaussian distribution of errors, centred on the average criterion location; giving the impression of a 'noisy' criterion that was randomly jittered on a trial-by-trial basis. In contrast, it was shown that the pattern of results could be explained by a fundamentally deterministic system, in which the criterion was shifted in favour of one or other response, in a manner contingent on the events of previous trials.

It therefore appears that many changes in behaviour can be described either in terms of random or deterministic factors (cf. §7.1.5). Of the two, deterministic account appear preferable, since these afford better predictions of trial-by-trial behaviour. Conversely, measures of internal noise have little explanatory power, and often serve more as descriptions than as explanations of behaviour. In this light, research into perceptual decision-making may be best served by striving to dispense with internal noise as an explanatory variable altogether.

7.3.4 Other dependant variables

Neurometric measures

Many of the constructs studied here have been operationalised using measures derived from psychometric functions. Such functions are fitted to the final outputs of the decision making process; listeners' responses. However, analogous fits can also be made using levels of neural activity at the various stages of the auditory system. Such fits have been derived in humans using MEG imaging (Witton et al., 2002), and in animals using multi-unit physiological recordings (Alves-Pinto et al., 2010). In this way it may be possible to relate the behavioural changes reported here, to their underlying neuronal implementations.

Listening effort

In typical perceptual learning studies listeners are required to maximise a response outcome, such as percent correct. Accordingly, the dependant measures of learning studies overwhelming relate to accuracy or sensitivity. However, this need not necessarily be the case, and other measures may also be of interest. For example, as discussed above, an alternative goal

could be to minimise response times, in which case the measure of learning would be the amount of time the listener requires to complete the task. A related concept is the amount of *effort* listeners must exert in order to perform the task. Listening effort may also be improved through practice, and such improvements may be of particular benefit to those individuals who find listening strenuous, such as some elderly and/or hearing impaired listeners (Gatehouse and Gordon, 1990; Hicks and Tharpe, 2002; Larsby et al., 2008). Understanding these changes may require a new theoretical framework, including a new set of potential learning mechanisms. For example, it may be necessary to consider concepts such as perceptual load (Lavie, 2005), and the degree to which its reduction may produce non-specific benefits by freeing up resources for other tasks. A principle difficulty with this work would be how to index listening effort. This would require the development of measures beyond those discussed hitherto, such as those relating to reaction times, galvanic skin response, pupillomotry, and memory recall (Sarampalis et al., 2009).

7.3.5 Individual differences

A limitation of the present work is that the key analyses have tended to aggregate across many listeners. Thus, while data from individuals have been presented where feasible, all the principle conclusions are predicated on group-mean differences in performance (i.e., across sessions or age). This was motivated largely by necessity. Parametric comparisons require estimates of both the parameter in question, and the sampling distribution from which it was drawn. Learning effects within an individual simply do not, in most instances, afford enough data to reliably constrain such estimates[5] (cf. Appendix A). As a result, I have been unable to address certain questions, and some nuances may have been missed altogether. For example, Chapter 3 demonstrated that changes in both σ_{int} and η_{enc} underlie learning. However, it remains unclear the extent to which the two factors covaried within an individual, or, conversely, whether different listeners concentrated on improving one or the other. Similarly, while inattentiveness appears in the round to play no role in learning, it is possible that it may nonetheless be a significant factor in a minority of listeners (in regards to which, the deterioration observed in one listener in Chapter 5 may be apposite). These problems could be addressed either through the use of tasks in which the rate of learning is diminished by an order of magnitude (Lively et al., 1993), allowing for detailed within-subject analyses, or by examining patterns amongst greatly enlarged cohorts of

[5]Though bootstrapping techniques may be used in some instances to attenuate the problem

listeners, using clustering techniques such as principal component analysis. However, both of these approaches constitute substantial undertakings.

7.3.6 The potential costs of learning

At this point, it may be instructive to consider *why* perceptual learning occurs. By definition improved performance on auditory judgement tasks is 'a good thing'. But so much so one wonders why improved performance doesn't come fitted as standard.

Possibly it is a matter of economy. It may be biologically more efficient to allow expertise in perceptual decision making to emerge 'epigenetically', through interaction with the environment, rather than to fully specify the operation of the system in advance (for further discussion on this point, see Elman, 1997). Under this interpretation, perceptual learning is simply about providing the requisite experience in order for listeners to fulfil their potential.

Alternatively, in a system of finite resources learning may represent a trade-off. Thus, previous papers have demonstrated that learned improvements in the trained stimulus region may also be offset by declinations in performance elsewhere (e.g., see Fig 4a of McGovern et al., 2012). The results of Chapter 3 and of Chapter 4, in particular, invite speculation as to what the latent costs of learning may be. In these studies, learning was subserved by the tightening of 'attentional tuning curves' around the spectral region of interest. These curves are reminiscent of those derived using probe-signal analysis (Greenberg and Larkin, 1968; Arbogast and Kidd Jr, 2000), in which attention is quantified as the relative *insensitivity* of the listener to sounds spectrally distal to a cued central region. It may therefore be that inexperienced listeners place a premium on monitoring the whole auditory scene, and that expertise in target-detection gradually emerges at the cost of decreased sensitivity outside the region of interest. This possibility could be investigated experimentally by using techniques such as probe-signal analysis to correlate increments in target detection limens with decrements in detection limens for spectrally remote stimuli.

Similarly, learning may not be viewed as an explicit trade-off between finite alternatives, but may be seen nonetheless as a *commitment* to perceiving the world in a certain manner. To see how this may be the case, consider that a listener's encoding strategy and bias can both be thought of as hypotheses concerning the structure of the world. For example, the listeners in Chapter 6 can be seen as rejecting one belief (trials are mutually conditional), in favour of another (trials are independent). The question

then becomes to what extent listeners are 'learning to learn', versus learning a specific, potentially arbitrary, solution to a particular problem. This could be empirically tested by manipulating the statistics of the stimuli once performance has reached asymptote, and observing the timecourse of any learning that followed. It may be that listeners who are less committed to the first hypothesis are able to adapt more easily to the second. Alternatively, it may be that listeners who have learnt one pattern are quicker or better able to learn a second. Such analyses may also be illuminating in regards to individual differences in learning *rates*. In the context of the present work, lower learning rates were suboptimal, but when the task statistics are nonstationary a degree of conservatism may represent the ideal learning strategy.

7.4 Final Conclusions

Here the main conclusions of the thesis are summarised. These findings draw on the information summarised in this chapter, as well as on the relevant experimental discussions.

1. Learning is the product of multiple mechanisms. Namely, additive internal noise reduction, bias reduction, and increased encoding efficiency.

2. The importance of any given mechanism is contingent on the task. In very simple tasks internal noise and bias are the predominant learning mechanisms. In external noise situations changes in encoding efficiency (here interpretable as selective attention) are also important.

3. Perceptual learning is not purely perceptual. Deterministic non-sensory considerations, such as stationary interval biases and sequential dependencies are also minimised by practice, and contribute towards reported measures of learning.

4. Learning and development share common mechanisms. The differences between older and younger children paralleled the changes between naïve and practiced adults. In both instances the improved performance of more experienced listeners was due to improved integration the input variables (encoding efficiency).

5. Audition and vision share common mechanisms. The same mechanisms were found to operate in auditory learning as have been reported in vision. Apparent inconsistencies were explained purely in terms of analysis/interpretation. Some of the changes in bias have not been examined in vision, but similar results are predicted.

APPENDIX A

Learning Effect Size

Learning effect sizes are contingent on the mean difference in performance before and after training, relative to the estimated measurement error. There is therefore a balance to be struck in terms of the number of trials used to estimate performance. Fewer trials will maximise both the mean difference, but also the measurement error. Conversely, more trials will cause both factors to decrease. Simple simulations were run to establish how effect size varies as a function of sample size given various learning rates. The observer was an ideal signal detector that performed a simple yes/no task based on normally distributed observations of noise, $\mathcal{N}(0,1)$, or signal, $\mathcal{N}(d',1)$, stimuli. As shown in Fig A.1, the learning dynamic was either linear (dashed line) or logarithmic (solid line), and proceeded at a rate corresponding to a typical auditory learning paradigm [see ∮1.1.4]. A more rapid logarithmic learning curve was also tested (dotted line). Each simulation was run 10,000 times. Mean ($\pm$ 1 SE) sensitivity was calculated using each of 40 log-spaced sample (bin) sizes. Effect sizes were calculated as per Cohen's d, by computing the first/last session differences in performance, and dividing the mean difference by the standard deviation[1], thus:

$$Effect\ Size = \frac{Mean[d'_{last} - d'_{first}]}{Std[d'_{last} - d'_{first}]} \tag{A.1}$$

[1] Equivalent results were obtained using a pooled variance term

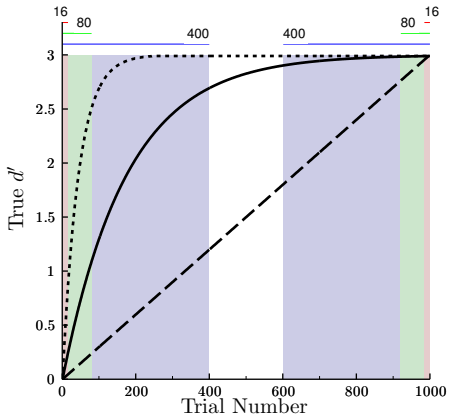


Fig. A.1: Schema showing example learning rates and bin sizes. The bin colours correspond to the example estimates shown in Fig A.2, Fig A.3, and Fig A.4.

When the learning progressed linearly (Fig A.2), the effect size was maximised by using the greatest number of samples available (split-half analysis).

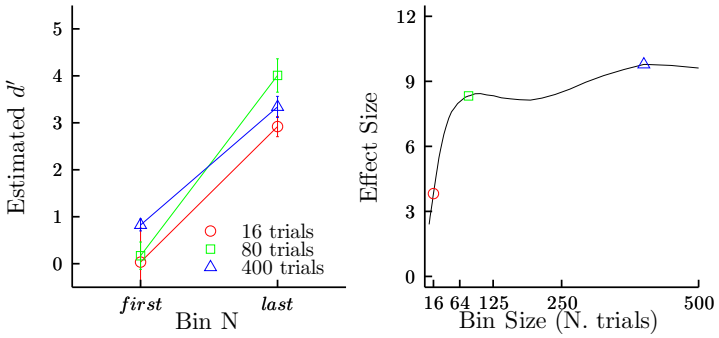


Fig. A.2: Learning effect size as a function of sample size, given a linear learning rate (the dashed line in Fig A.1). The left panel shows mean $\pm$ 1 SE estimates of performance using the example small ($N = 16$), medium ($N = 80$), and large ($N = 400$) sample sizes shown previously in Fig A.1. The right panel shows the resultant effect sizes as a function of each sample size. Markers highlight the three effect sizes corresponding to the data in the left panel.

Conversely, when sensitivity increased in the logarithmic fashion more typical of auditory learning (Fig A.3), an intermediate number of samples ($64 - 125$) provided the most sensitive measure of learning. The difference between the right tail of the linear and log effect size graphs can be explained by noting that the denominator of Eq A.1 is not strongly affected by the manipulation; in both cases measurement error decreased the number of observations increases. However, in the linear case the mean differences are greater in general, and also tend to peak at a greater number of observation. Thus, in the linear case there continues to be a substantial mean difference in performance even at the largest bin sizes

(cf. blue triangles in Fig A.2 versus Fig A.3). Since in this region the denominator error term is relatively small (< 0.5), such differences are strongly emphasized when computing effect sizes. One corollary of this is that while it is important to have sufficient data points to robustly measure d' (e.g., $\gtrsim 50$), 'split-halves' type analyses are likely to be insensitive to learning in cases where learning occurs logarithmically (i.e., such that the majority of learning is completed within the first half).

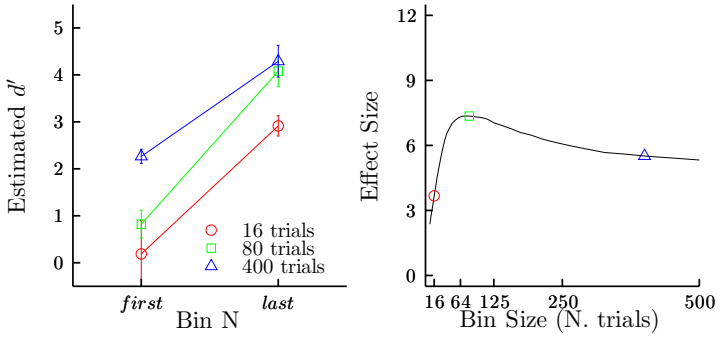


Fig. A.3: Same as Fig A.2, but given a logarithmic learning rate (the solid line in Fig A.1).

As shown in Fig A.4, the ideal number of observations continues to diminish as the learning rate increases further. Thus, given a very rapid learning rate, the learning effect curve is shifted downwards, indicating diminished effect sizes for all numbers of observations, and is also compressed leftwards, indicating that the maximal effect size is found at a smaller number of observations. The overall pattern of change can be summarised by the rough, but intuitive, heuristic that the ideal number of trials per estimate is approximately equal to the 50^{th} percentile on the learning curve (cf. Fig A.1).

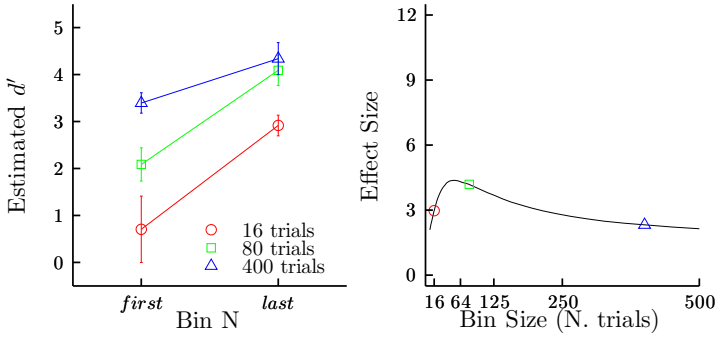


Fig. A.4: Same as Fig A.2, but given a very rapid logarithmic learning rate (the dotted line in Fig A.1).

References

Ahissar, M. and Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, **387**(6631), 401–406. 1, 146

Ahissar, M. and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, **8**(10), 457–464. 158

Ahissar, M., Nahum, M., Nelken, I., and Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **364**(1515), 285–299. 4, 21

Alexander, J. and Lutfi, R. (2004). Informational masking in hearing-impaired and normal-hearing listeners: Sensation level and decision weights. *J. Acoust. Soc. Am.*, **116**(4), 2234–2247. 31, 58, 59, 64, 74, 75, 77, 82

Allard, R. and Cavanagh, P. (2012). Different processing strategies underlie voluntary averaging in low and high noise. *J. Vis.*, **12**(11), 1–12. 74

Altman, D. and Bland, J. (2011). How to obtain the p value from a confidence interval. *Br. Med. J. (Clin. Res. Ed.)*, **343**, d2304. 103

Alves-Pinto, A., Baudoux, S., Palmer, A., and Sumner, C. (2010). Forward masking estimated by signal detection theory analysis of neuronal responses in primary auditory cortex. *J. Assoc. Res. Otolaryngol.*, **11**(3), 477–494. 166

Amitay, S., Halliday, L., Taylor, J., Sohoglu, E., and Moore, D. (2010). Motivation and intelligence drive auditory perceptual learning. *PLoS One*, **5**(3), e9816. 6

Amitay, S., Hawkey, D., and Moore, D. (2005). Auditory frequency discrimination learning is affected by stimulus variability. *Atten. Percept. Psychophys.*, **67**(4), 691–698. 2, 4, 5, 6, 33, 137, 162

Amitay, S., Irwin, A., Hawkey, D., Cowan, J., and Moore, D. (2006). A comparison of adaptive procedures for rapid and reliable threshold assessment and training in naive listeners. *J. Acoust. Soc. Am.*, **119**(3), 1616–1625. 2, 4

Anderson, N. (1960). Effect of first-order conditional probability in a two-choice learning situation. *J. Exp. Psychol.*, **59**(2), 73–93. 129, 145

Arbogast, T. and Kidd Jr, G. (2000). Evidence for spatial tuning in informational masking using the probe-signal method. *J. Acoust. Soc. Am.*, **108**(4), 1803–1810. 168

Atkinson, R., Carterette, E., and Kinchla, R. (1962). Sequential phenomena in psychophysical judgments: A theoretical analysis. *IRE Trans. Inf. Theory*, **8**(5), 155–162. 128, 129

Ayton, P. and Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Mem. Cognit.*, **32**(8), 1369–1378. 129, 145

Ball, K. and Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vis. Res.*, **27**(6), 953–965. 4, 7, 115

Barron, G. and Leider, S. (2010). The role of experience in the gambler's fallacy. *J. Behav. Dec. Making*, **23**(1), 117–129. 146

Barsz, K. (1996). Accuracy of same/different judgments of sequences of complex tones differing in tonal order under various levels of fundamental frequency range, listener training, and type of standard sequence. *J. Acoust. Soc. Am.*, **99**(3), 1660–1669. 3

Ben-David, B., Campeanu, S., Tremblay, K., and Alain, C. (2011). Auditory evoked potentials dissociate rapid perceptual learning from task repetition without learning. *Psychophysiology*, **48**(6), 797–807. 128

Berg, B. (1989). Analysis of weights in multiple observation tasks. *J. Acoust. Soc. Am.*, **86**(5), 1743–1746. 162

Berg, B. (1990). Observer efficiency and weights in a multiple observation task. *J. Acoust. Soc. Am.*, **88**(1), 149–158. 148

Berg, B. (2004). A molecular description of profile analysis: Decision weights and internal noise. *J. Acoust. Soc. Am.*, **115**(2), 822–829. 30, 35, 37, 52, 59, 60, 75

Berg, B. and Green, D. (1990). Spectral weights in profile listening. *J. Acoust. Soc. Am.*, **88**(2), 758–766. 30, 95

Brady, M. and Kersten, D. (2003). Bootstrapped learning of novel objects. *J. Vis.*, **3**(6), 413–422. 21

Brainard, D. (1997). The psychophysics toolbox. *Spat. Vis.*, **10**(4), 433–436. 32, 62, 84, 116

Braun, D., Aertsen, A., Wolpert, D., and Mehring, C. (2009). Motor task variation induces structural learning. *Curr. Biol.*, **19**(4), 352–357. 7

British Society of Audiology (2004). Recommended procedure for pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels. Technical report, British Society of Audiology. 39, 61, 84, 114

Brown, M., Irvine, D., and Park, V. (2004). Perceptual learning on an auditory frequency discrimination task by cats: association with changes in primary auditory cortex. *Cereb. Cortex*, **14**(9), 952–965. 26

Burgess, A. and Colborne, B. (1988). Visual signal detection. iv. observer inconsistency. *J. Opt. Soc. Am. A*, **5**(4), 617–627. 38, 59

Buss, E. (2008). Across-channel interference in intensity discrimination: The role of practice and listening strategy. *J. Acoust. Soc. Am.*, **123**(1), 265–272. 3, 4, 5, 57, 72, 73, 75

Buss, E., Hall III, J., and Grose, J. (2006). Development and the role of internal noise in detection and discrimination thresholds with narrow band stimuli. *J. Acoust. Soc. Am.*, **120**(5), 2777–2788. 48, 59, 60, 83, 101

Buss, E., Hall III, J., and Grose, J. (2009). Psychometric functions for pure tone intensity discrimination: Slope differences in school-aged children and adults. *J. Acoust. Soc. Am.*, **125**(2), 1050–1058. 16, 20, 30, 48, 60

Campbell, R. and Small, Jr., A. (1963). Effect of practice and feedback on frequency discrimination. *J. Acoust. Soc. Am.*, **35**(10), 1511–1514. 3, 4, 5, 128

Carcagno, S. and Plack, C. J. (2011). Pitch discrimination learning: specificity for pitch and harmonic resolvability, and electrophysiological correlates. *Journal of the Association for Research in Otolaryngology*, **12**(4), 503–517. 3

Cardozo, B. and Ritsma, R. (1968). On the perception of imperfect periodicity. *IEEE Trans. Aud. Electroacoust.*, **16**(2), 159–164. 17

Carraher, T., Carraher, D., and Schliemann, A. (2011). Mathematics in the streets and in schools. *Br. J. Dev. Psy.*, **3**(1), 21–29. 162

Chung, S., Levi, D., and Tjan, B. (2005). Learning letter identification in peripheral vision. *Vis. Res.*, **45**(11), 1399–1412. 32

Clotfelter, C. and Cook, P. (1993). Notes: The "gambler's fallacy" in lottery play. *Manag. Sci.*, **39**(12), 1521–1525. 129

Cortes, C. and Vapnik, V. N. (1995). Support vector machines. *Mach. Learn.*, **20**(3), 273–297. 51

Cosmides, L. and Tooby, J. (1992). Cognitive adaptations for social exchange. In Barkow, J., Cosmides, L., and Tooby, J., editors, *The adapted mind: Evolutionary psychology and the generation of culture*, pages 163–228. Oxford University Press, USA. 162

Craig, A. (1976). Signal recognition and the probability-matching decision rule. *Atten. Percept. Psychophys.*, **20**(3), 157–162. 13

Creelman, C. (1965). Discriminability and scaling of linear extent. *J. Exp. Psychol.*, **70**(2), 192–200. 26

Croson, R. and Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *J. Risk. Uncert.*, **30**(3), 195–209. 129

Cuddy, L. (1970). Training the absolute identification of pitch. *Atten. Percept. Psychophys.*, **8**(5), 265–269. 3

Dai, H. and Berg, B. (1992). Spectral and temporal weights in spectral-shape discrimination. *J. Acoust. Soc. Am.*, **92**(3), 1346–1355. 30, 59, 64

Dai, H. and Micheyl, C. (2010). Psychophysical reverse correlation with multiple response alternatives. *J. Exp. Psychol. Hum. Percept. Perform.*, **36**(4), 976–993. 59

Dai, H. and Micheyl, C. (2011). Psychometric functions for pure-tone frequency discrimination. *J. Acoust. Soc. Am.*, **130**(1), 263–272. 37, 53, 54

Dai, H. and Richards, V. (2011). On the theoretical error bound for estimating psychometric functions. *Atten. Percept. Psychophys.*, **73**(3), 919–926. 37, 53

Dayan, P. and Abbott, L. (2001). *Theoretical neuroscience: Computational and mathematical modeling of Neural Systems*, pages 313–322. MIT Press, Cambridge, Massachusetts. 47

Demany, L. (1985). Perceptual learning in frequency discrimination. *J. Acoust. Soc. Am.*, **78**(3), 1118–1120. 3, 4, 5, 6, 26, 31, 57, 72, 77, 140, 141, 142, 143, 145

Denk, W., Webb, W., and Hudspeth, A. (1989). Mechanical properties of sensory hair bundles are reflected in their brownian motion measured with a laser differential interferometer. *Proc. Natl. Acad. Sci. U. S. A.*, **86**(14), 5371–5375. 13, 30

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Mem. Cognit.*, **24**(4), 523–533. 26

Donovan, J. J. and Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, **84**(5), 795–805. 6

Dorfman, D. and Biderman, M. (1971). A learning model for a continuum of sensory states. *J. Math. Psychol.*, **8**(2), 264–284. 141

Dorfman, D., Saslow, C., and Simpson, J. (1975). Learning models for a continuum of sensory states reexamined. *J. Math. Psychol.*, **12**(2), 178–211. 141

Dosher, B. and Lu, Z. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc. Natl. Acad. Sci. U. S. A.*, **95**(23), 13988–13993. 23, 29, 74, 161

Dosher, B. and Lu, Z. (1999). Mechanisms of perceptual learning. *Vis. Res.*, **39**(19), 3197–3221. 50

Dosher, B. and Lu, Z. (2005). Perceptual learning in clear displays optimizes perceptual expertise: Learning the limiting process. *Proc. Natl. Acad. Sci. U. S. A.*, **102**(14), 5286–5290. 24, 74, 159

Dosher, B. and Lu, Z. (2006). Level and mechanisms of perceptual learning: Learning first-order luminance and second-order texture objects. *Vis. Res.*, **46**(12), 1996–2007. 24

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.*, **2**(11), 820–829. 23

Dunn, L., Dunn, L., Whetton, C., and Burley, J. (1997). The british picture vocabulary scale, 2nd edn (windsor: Nfer-nelson). 86

Durlach, N. and Braida, L. (1969). Intensity perception. i. preliminary theory of intensity resolution. *J. Acoust. Soc. Am.*, **46**(2), 372–383. 13, 32

Durlach, N., Mason, C., Kidd Jr, G., Arbogast, T., Colburn, H., and Shinn-Cunningham, B. (2003a). Note on informational masking (l). *J. Acoust. Soc. Am.*, **113**(6), 2984–2987. 56, 73

Durlach, N., Mason, C., Shinn-Cunningham, B., Arbogast, T., Colburn, H., and Kidd Jr, G. (2003b). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *J. Acoust. Soc. Am.*, **114**(1), 368–379. 56, 82

Dusoir, A. (1975). Treatments of bias in detection and recognition models: A review. *Atten. Percept. Psychophys.*, **17**(2), 167–178. 128

Edwards, W. (1961). Probability learning in 1000 trials. *J. Exp. Psychol.*, **62**(4), 385–394. 26, 129

Elman, J. (1997). *Rethinking innateness: A connectionist perspective on development*, volume 10. The MIT press, Cambridge, Massachusetts. 7, 168

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychol. Rev.*, **105**(2), 280–298. 141

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Curr. Opin. Neurobiol.*, **15**(2), 154–160. 1, 7, 145

Fahle, M. and Edelman, S. (1993). Long-term learning in vernier acuity: Effects of stimulus orientation, range and of feedback. *Vis. Res.*, **33**(3), 397–412. 4, 145

Fechner, G. (1966). *Elements of Psychophysics: Translated by Helmut E. Adler. Edited by Davis H. Howes and Edwin G. Boring, with an introduction by Edwin G. Boring. (Original work published 1860)*. Holt, Rinehart and Winston. 129

Fine, I. and Jacobs, R. (2002). Comparing perceptual learning across tasks: A review. *J. Vis.*, **2**(2), 190–203. 1, 5, 127

Finomore, V., Matthews, G., Shaw, T., and Warm, J. (2009). Predicting vigilance: A fresh look at an old problem. *Ergonomics*, **52**(7), 791–808. 27

Fiorentini, A. and Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature*, **287**(5777), 43–44. 25, 145

Fitzgerald, M. and Wright, B. (2005). A perceptual learning investigation of the pitch elicited by amplitude-modulated noise. *J. Acoust. Soc. Am.*, **118**(6), 3794–3803. 3

Fitzgerald, M. and Wright, B. (2011). Perceptual learning and generalization resulting from training on an auditory amplitude-modulation detection task. *J. Acoust. Soc. Am.*, **129**(2), 898–906. 3, 109, 114, 117, 121

Fletcher, H. and Munson, W. (1933). Loudness, its definition, measurement and calculation. *J. Acoust. Soc. Am.*, **5**(1), 82–108. 75

Garner, W. (1953). An informational analysis of absolute judgments of loudness. *J. Exp. Psychol.*, **46**(5), 373–380. 129, 145

Garner, W. and Hake, H. (1951). The amount of information in absolute judgments. *Psychol. Rev.*, **58**(6), 446–459. 129

Gatehouse, S. and Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *Br. J. Audiol.*, **24**(1), 63–68. 167

Gescheider, G. (1997). *Psychophysics: The fundamentals*, pages 73–124. Lawrence Erlbaum Associates, Mahwah, New Jersey. 30, 31, 60, 128, 160

Gifford, R. and Bacon, S. (2000). Contributions of suppression and excitation to simultaneous masking: Effects of signal frequency and masker-signal frequency relation. *J. Acoust. Soc. Am.*, **107**(4), 2188–2200. 17

Gilbert, C., Sigman, M., and Crist, R. (2001). The neural basis of perceptual learning. *Neuron*, **31**(5), 681–697. 26

Gilchrist, J., Jerwood, D., and Ismaiel, H. (2005). Comparing and unifying slope estimates across psychometric function models. *Percept. Psychophys.*, **67**(7), 1289–1303. 60

Glasberg, B. and Moore, B. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, **47**(1-2), 103–138. 47, 61, 96

Glasberg, B., Moore, B., and Peters, R. (2001). The influence of external and internal noise on the detection of increments and decrements in the level of sinusoids. *Hear. Res.*, **155**(1-2), 41–53. 37

Gold, J., Bennett, P., and Sekuler, A. (1999). Signal but not noise changes with perceptual learning. *Nature*, **402**(6758), 176–178. xv, 21, 22, 23, 24, 29, 30, 32, 49

Gold, J., Sekuler, A., and Bennett, P. (2004). Characterizing perceptual learning with external noise. *Cog. Sci.*, **28**(2), 167–207. 11, 22, 23, 32, 49, 109, 127

Goldstone, R. (1998). Perceptual learning. *Annu. Rev. Psychol.*, **49**(1), 585–612. 1, 127

Green, C. and Bavelier, D. (2008). Exercising your brain: A review of human brain plasticity and training-induced learning. *Psychol. Aging*, **23**(4), 692–701. 145

Green, C., Bavelier, D., et al. (2003). Action video game modifies visual selective attention. *Nature*, **423**(6939), 534–537. 27

Green, D. (1958). Detection of multiple component signals in noise. *J. Acoust. Soc. Am.*, **30**(10), 904–911. 95

Green, D. (1964). Consistency of auditory detection judgments. *Psychol. Rev.*, **71**(5), 392–407. 16, 26, 30, 37, 38, 59, 129

Green, D. (1992). The number of components in profile analysis tasks. *J. Acoust. Soc. Am.*, **91**(3), 1616–1623. 3

Green, D. (1995). Maximum-likelihood procedures and the inattentive observer. *J. Acoust. Soc. Am.*, **97**(6), 3749–3760. 19, 20, 31, 37

Green, D. and Swets, J. (1974). *Signal detection theory and psychophysics*, pages 1–455. Krieger, Melbourne, Florida. 8, 18, 30

Greenberg, G. and Larkin, W. (1968). Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method. *J. Acoust. Soc. Am.*, **44**(6), 1513–1523. 168

Grimault, N., Micheyl, C., Carlyon, R., Bacon, S., and Collet, L. (2003). Learning in discrimination of frequency or modulation rate: generalization to fundamental frequency discrimination. *Hear. Res.*, **184**(1-2), 41–50. 4

Grimault, N., Micheyl, C., Carlyon, R., and Collet, L. (2002). Evidence for two pitch encoding mechanisms using a selective auditory training paradigm. *Atten. Percept. Psychophys.*, **64**(2), 189–197. 3

Gundy, R. (1961). Auditory detection of an unspecified signal. *J. Acoust. Soc. Am.*, **33**(8), 1008–1012. 3

Gureckis, T. and Love, B. (2009). Learning in noise: Dynamic decision-making in a variable environment. *J. Math. Psychol.*, **53**(3), 180–193. 163

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scand. Audiol.*, **11**(2), 79–87. 3

Halliday, L. (2005). *Are auditory processing deficits linked to literacy problems: a comparison of specific reading disability and mild to moderate sensorineural hearing loss*. PhD thesis, University of Oxford. 160

Halliday, L., Moore, D., Taylor, J., and Amitay, S. (2011). Dimension-specific attention directs learning and listening on auditory training tasks. *Atten. Percept. Psychophys.*, **73**(5), 1329–1335. 128

Halliday, L., Taylor, J., Edmondson-Jones, A., and Moore, D. (2008). Frequency discrimination learning in children. *J. Acoust. Soc. Am.*, **123**(6), 4393–4402. 81, 101

Hartman, E. (1954). The influence of practice and pitch-distance between tones on the absolute identification of pitch. *Am. J. Psychol.*, **67**(1), 1–14. 3

Hawkey, D., Amitay, S., and Moore, D. (2004). Early and rapid perceptual learning. *Nat. Neurosci.*, **7**(10), 1055–1056. 6, 7, 31, 73, 108, 127, 145

Heinz, M., Colburn, H., and Carney, L. (2001). Evaluating auditory performance limits: I. one-parameter discrimination using a computational model for the auditory nerve. *Neural Comput.*, **13**(10), 2273–2316. 47

Herzog, M. and Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vis. Res.*, **37**(15), 2133–2141. 4, 21

Hicks, C. and Tharpe, A. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *J. Speech Lang. Hear. Res.*, **45**(3), 573–584. 167

Hoel, P. (1947). *Introduction to Mathematical Statistics*. John Wiley & Sons, New York, New York. 148

Houtsma, A. J. M. (1995). Pitch perception. In Moore, B., editor, *Hearing*, pages 267–296. Academic Press, San Diego, California. 31

Humes, L. and Jesteadt, W. (1989). Models of the additivity of masking. *J. Acoust. Soc. Am.*, **85**(3), 1285–1294. 17

Huyck, J. and Wright, B. (2011). Late maturation of auditory perceptual learning. *Dev. Sci.*, **14**(3), 614–621. 7, 101, 164

Irvine, D. (2007). Auditory cortical plasticity: Does it provide evidence for cognitive processing in the auditory cortex? *Hear. Res.*, **229**(1-2), 158–170. 26

Irvine, D., Martin, R., Klimkeit, E., and Smith, R. (2000). Specificity of perceptual learning in a frequency discrimination task. *J. Acoust. Soc. Am.*, **108**(6), 2964–2968. 1, 72, 145

Jarvik, M. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. *J. Exp. Psychol.*, **41**(4), 291–297. 129, 145

Javel, E. and Viemeister, N. (2000). Stochastic properties of cat auditory nerve responses to electric and acoustic stimuli and application to intensity discrimination. *J. Acoust. Soc. Am.*, **107**(2), 908–921. 13, 30

Jesteadt, W., Luce, R., and Green, D. (1977). Sequential effects in judgments of loudness. *J. Exp. Psychol. Hum. Percept. Perform.*, **3**(1), 92–104. 129, 130, 137

Jesteadt, W., Nizami, L., and Schairer, K. (2003). A measure of internal noise based on sample discrimination. *J. Acoust. Soc. Am.*, **114**(4), 2147–2157. 16, 17, 30, 33, 35, 46

Jeter, P., Dosher, B., Liu, S., and Lu, Z. (2010). Specificity of perceptual learning increases with increased training. *Vis. Res.*, **50**(19), 1928–1940. 26, 145

Juni, M., Gureckis, T., and Maloney, L. (2012). Effective integration of serially presented stochastic cues. *J. Vis.*, **12**(8), 1–16. 162

Kac, M. (1962). A note on learning signal detection. *IRE Trans. Inf. Theory*, **8**(2), 126–128. 141

Kac, M. (1969). Some mathematical models in science. *Science*, **166**(3906), 695–699. 141

Karmarkar, U. and Buonomano, D. (2003). Temporal specificity of perceptual learning in an auditory discrimination task. *Learn. Mem.*, **10**(2), 141–147. 3, 140, 145

Karni, A. and Sagi, D. (1991). Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. *Proc. Natl. Acad. Sci. U. S. A.*, **88**(11), 4966–4970. 145

Karni, A. and Sagi, D. (1993). The time course of learning a visual skill. *Nature*, **365**(6443), 250–252. 1

Kidd, C., Palmeri, H., and Aslin, R. (in press). http://www.bcs.rochester.edu/people/ckidd/papers/KiddPalmeriAslin2012_Cognition.pdf. 163

Kidd, Jr., G., Mason, C., Deliwala, P., Woods, W., and Colburn, H. (1994). Reducing informational masking by sound segregation. *J. Acoust. Soc. Am.*, **95**(6), 3475–3480. 56, 82

Kidd, Jr., G., Mason, C., Richards, V., Gallun, F., and Durlach, N. (2007). Informational masking. In Christophe Micheyl, S. A. S. and Oxenham, A. J., editors, *Auditory perception of sound sources*, pages 143–189. Springer Science+Business Media, New York, New York. 56

Kingdom, A. and Prins, N. (2009). *Psychophysics: a practical introduction*. Academic Press, London, England. 160

Klein, S. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Atten. Percept. Psychophys.*, **63**(8), 1421–1455. 30

Kong, Y.-Y., Lu, Z.-L., Dosher, B. A., and Zeng, F.-G. (2004). Mechanisms of perceptual learning in amplitude modulation detection. In *ARO Midwinter Meeting*, New York, New York. Association for Research in Otolaryngology. 109

Koyama, S., Harner, A., and Watanabe, T. (2004). Task-dependent changes of the psychophysical motion-tuning functions in the course of perceptual learning. *Perception*, **33**(9), 1139–1147. 128

Krakauer, J., Ghilardi, M., and Ghez, C. (1999). Independent learning of internal models for kinematic and dynamic control of reaching. *Nat. Neurosci.*, **2**(11), 1026–1031. 7

Kubovy, M. and Healy, A. (1977). The decision rule in probabilistic categorization: What it is and how it is learned. *J. Exp. Psychol. Gen.*, **106**(4), 427–446. 26, 128, 129, 166

Kurt, S. and Ehret, G. (2010). Auditory discrimination learning and knowledge transfer in mice depends on task difficulty. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(18), 8481–8485. 21

Laplace, P. (1995). *Philosophical Essays on Probabilities*. transl. AI Dale, Springer-Verlag, originally published 1814. 129

Larsby, B., Hällgren, M., and Lyxell, B. (2008). The interference of different background noises on speech processing in elderly hearing impaired subjects. *Int. J. Audiol.*, **47**(S2), 83–90. 167

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends Cog. Sci.*, **9**(2), 75–82. 167

Law, C.-T. and Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature neuroscience*, **12**(5), 655–663. 47

Lee, T. and Richards, V. (2011). Evaluation of similarity effects in informational masking. *J. Acoust. Soc. Am.*, **129**(6), EL280–EL285. 56, 82

Leek, M. and Watson, C. (1984). Learning to detect auditory pattern components. *J. Acoust. Soc. Am.*, **76**(4), 1037–1044. 3

Leek, M. and Watson, C. (1988). Auditory perceptual learning of tonal patterns. *Atten. Percept. Psychophys.*, **43**(4), 389–394. 3

Leibold, L. and Bonino, A. (2009). Release from informational masking in children: Effect of multiple signal bursts. *J. Acoust. Soc. Am.*, **125**(4), 2200–2208. 91

Leibold, L. and Neff, D. (2007). Effects of masker-spectral variability and masker fringes in children and adults. *J. Acoust. Soc. Am.*, **121**(6), 3666–3676. 87, 90, 91, 96, 100

Levi, D., Klein, S., and Chen, I. (2005). What is the signal in noise? *Vis. Res.*, **45**(14), 1835–1846. 34

Levi, D. and Li, R. (2009). Perceptual learning as a potential treatment for amblyopia: A mini-review. *Vis. Res.*, **49**(21), 2535–2549. 159

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J. Acoust. Soc. Am.*, **49**(2), 467–477. 5, 62, 85, 138

Lindman, H. and Edwards, W. (1961). Supplementary report: Unlearning the gambler's fallacy. *J. Exp. Psychol.*, **62**(6), 630–630. 26, 129, 145

Linkenhoker, B. and Knudsen, E. (2002). Incremental training increases the plasticity of the auditory space map in adult barn owls. *Nature*, **419**(6904), 293–296. 4

Lively, S., Logan, J., and Pisoni, D. (1993). Training japanese listeners to identify english /r/ and /l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.*, **94**(3), 1242–1255. 6, 57, 167

Lu, Z., , and Dosher, B. (2000). Spatial attention: Different mechanisms for central and peripheral temporal precues? *Journal of Experimental Psychology Human Perception and Performance*, **26**(5), 1534–1548. 23

Lu, Z., Chu, W., and Dosher, B. (2006). Perceptual learning of motion direction discrimination in fovea: Separable mechanisms. *Vis. Res.*, **46**(15), 2315–2327. 23

Lu, Z. and Dosher, B. (1998). External noise distinguishes attention mechanisms. *Vis. Res.*, **38**(9), 1183–1198. 23

Lu, Z. and Dosher, B. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *J. Opt. Soc. Am.*, **16**(3), 764–778. xv, 22, 24

Lu, Z. and Dosher, B. (2004). Perceptual learning retunes the perceptual template in foveal orientation identification. *J. Vis.*, **4**(1), 44–56. 7, 23, 24

Lu, Z. and Dosher, B. (2008). Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychol. Rev.*, **115**(1), 44–82. 16, 17, 21, 30, 38, 50, 59, 74, 109, 127

Lu, Z. and Dosher, B. (2009). Mechanisms of perceptual learning. *Learn. Per.*, **1**(1), 19–36. 23, 32, 46, 48, 49

Lutfi, R. (1993). A model of auditory pattern analysis based on component-relative-entropy. *J. Acoust. Soc. Am.*, **94**(2), 748–758. 74, 82, 95, 96

Lutfi, R. (1995). Correlation coefficients and correlation ratios as estimates of observer weights in multiple-observation tasks. *J. Acoust. Soc. Am.*, **97**(2), 1333–1334. 58

Lutfi, R., Kistler, D., Oh, E., Wightman, F., and Callahan, M. (2003). One factor underlies individual differences in auditory informational masking within and across age groups. *Atten. Percept. Psychophys.*, **65**(3), 396–406. 58, 82, 87, 90

Macmillan, N. and Creelman, C. (1990). Response bias: Characteristics of detection theory, threshold theory, and "nonparametric" indexes. *Psychol. Bull.*, **107**(3), 401–413. 128

Macmillan, N. and Creelman, C. (2005). *Detection theory: A user's guide*, pages 1–495. Lawrence Erlbaum Associates, Mahwah, New Jersey. 8, 30, 31, 34, 122, 157

Maddox, W. and Bohil, C. (1998). Base-rate and payoff effects in multidimensional perceptual categorization. *J. Exp. Psychol. Learn. Mem. Cogn.*, **24**(6), 1459–1482. 13, 26, 121

Maxwell, W. and Davidson, G. (1961). Limited-frequency screening and ear pathology. *J. Speech Hear. Disord.*, **26**(2), 122–125. 103

McArthur, G. and Hogben, J. (2012). Poor auditory task scores in children with specific reading and language difficulties: Some poor scores are more equal than others. *Sci. Stud. Read.*, **16**(1), 63–89. 27

McGill, W. (1957). Serial effects in auditory threshold judgments. *J. Exp. Psychol.*, **53**(5), 297–303. 129

McGill, W. and Goldberg, J. (1968). A study of the near-miss involving weber's law and pure-tone intensity discrimination. *Atten. Percept. Psychophys.*, **4**(2), 105–109. 10

McGovern, D., Roach, N., and Webb, B. (2012). Perceptual learning reconfigures the effects of visual adaptation. *The Journal of Neuroscience*, **32**(39), 13621–13629. 168

McLennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E., and Dibben, C. (2011). The english indices of deprivation 2010. Technical report, London, Department for Communities and Local Government. 86

Mednick, S., Drummond, S., Arman, A., and Boynton, G. (2008). Perceptual deterioration is reflected in the neural response: fmri study between nappers and non-nappers. *Perception*, **37**(7), 1086–1097. 7

Merzenich, M., Jenkins, W., Johnston, P., Schreiner, C., Miller, S., and Tallal, P. (1996). Temporal processing deficits of language-learning impaired children ameliorated by training. *Science*, **271**(5245), 77–81. 81, 159

Meyer, M. (1899). Is the memory of absolute pitch capable of development by training? *Psychol. Rev.*, **6**(5), 514–516. 3

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hear. Res.*, **219**(1-2), 36–47. 3

Miller, J. (1996). The sampling distribution of d'. *Atten. Percept. Psychophys.*, **58**(1), 65–72. 149

Millward, K., Hall, R., Ferguson, M., and Moore, D. (2011). Training speech-in-noise perception in mainstream school children. *Int. J. Pediatr. Otorhinolaryngol.*, **75**(11), 1408–1417. 3

Mollon, J. and Danilova, M. (1996). Three remarks on perceptual learning. *Spat. Vis.*, **10**(1), 51–58. 26, 46

Molloy, K., Moore, D., Sohoglu, E., and Amitay, S. (2012). Less is more: Latent learning is maximized by shorter training sessions in auditory perceptual learning. *PLoS One*, **7**(5), e36929. 1, 6, 21, 33, 73, 160

Moore, B. and Vickers, D. (1997). The role of spread excitation and suppression in simultaneous masking. *J. Acoust. Soc. Am.*, **102**(4), 2284–2290. 17

Moore, D. (in press). Listening difficulties in children: Bottom-up and top-down contributions. http://www.sciencedirect.com/science/article/pii/S0021992412000706. 82

Moore, D., Ferguson, M., Halliday, L., and Riley, A. (2008). Frequency discrimination in children: Perception, learning and attention. *Hear. Res.*, **238**(1-2), 147–154. 27, 31

Mossbridge, J., Fitzgerald, M., O'Connor, E., and Wright, B. (2006). Perceptual-learning evidence for separate processing of asynchrony and order tasks. *J. Neurosci.*, **26**(49), 12708–12716. 3, 72

Mossbridge, J., Scissors, B., and Wright, B. (2008). Learning and generalization on asynchrony and order tasks at sound offset: implications for underlying neural circuitry. *Learn. Mem.*, **15**(1), 13–20. 3, 72

Nahum, M., Nelken, I., and Ahissar, M. (2010). Stimulus uncertainty and perceptual learning: Similar principles govern auditory and visual learning. *Vis. Res.*, **50**(4), 391–401. 161

Nakamoto, K., Jones, S., and Palmer, A. (2008). Descending projections from auditory cortex modulate sensitivity in the midbrain to cues for spatial position. *J. Neurophysiol.*, **99**(5), 2347–2356. 26

Neff, D. (1995). Signal properties that reduce masking by simultaneous, random-frequency maskers. *J. Acoust. Soc. Am.*, **98**(4), 1909–1920. 56

Neff, D. and Callaghan, B. (1988). Effective properties of multicomponent simultaneous maskers under conditions of uncertainty. *J. Acoust. Soc. Am.*, **83**(5), 1833–1838. 2, 56, 73, 82, 87

Neff, D. and Dethlefs, T. (1995). Individual differences in simultaneous masking with random-frequency, multicomponent maskers. *J. Acoust. Soc. Am.*, **98**(1), 125–134. 7, 56, 57, 72, 73, 75, 82, 99

Neff, D. and Green, D. (1987). Masking produced by spectral uncertainty with multicomponent maskers. *Atten. Percept. Psychophys.*, **41**(5), 409–415. 56, 82

Neff, D. and Odgaard, E. (2004). Sample discrimination of frequency differences with distracters. *J. Acoust. Soc. Am.*, **116**(5), 3051–3061. 75

Neri, P. (2013). The statistical distribution of noisy transmission in human sensors. *Journal of Neural Engineering*, **10**(1), 16014–16026. 10

Nikulin, M. (2001). Hellinger distance. In Hazewinkel, M., editor, *Encyclopaedia of Mathematics*. Kluwer Academic Publishers, Norwell, Massachusetts. 51

Oh, E. and Lutfi, R. (1998). Nonmonotonicity of informational masking. *J. Acoust. Soc. Am.*, **104**(6), 3489–3499. 56, 60, 74, 82, 96

Oh, E. and Lutfi, R. (2000). Effect of masker harmonicity on informational masking. *J. Acoust. Soc. Am.*, **108**(2), 706–709. 90

Oh, E., Wightman, F., and Lutfi, R. (2001). Childrens detection of pure-tone signals with random multitone maskers. *J. Acoust. Soc. Am.*, **109**(6), 2888–2895. 82, 87, 88, 89, 90, 96, 103

Oxenham, A. and Buus, S. (2000). Level discrimination of sinusoids as a function of duration and level for fixed-level, roving-level, and across-frequency conditions. *J. Acoust. Soc. Am.*, **107**(3), 1605–1614. 13

Oxenham, A., Fligor, B., Mason, C., and Kidd Jr, G. (2003). Informational masking and musical training. *J. Acoust. Soc. Am.*, **114**(3), 1543–1549. 57, 73, 100

Parducci, A., Marshall, L., and Degner, M. (1966). Interference with memory for lifted weight. *Atten. Percept. Psychophys.*, **1**(2), 83–86. 129

Parducci, A. and Sandusky, A. (1965). Distribution and sequence effects in judgment. *J. Exp. Psychol.*, **69**(5), 450–459. 13

Parks, T. and Kellicutt, M. (1968). The probability-matching decision rule in the visual discrimination of order. *Atten. Percept. Psychophys.*, **3**(5), 356–360. 13

Patterson, R., Handel, S., Yost, W., and Datta, A. (1996). The relative strength of the tone and noise components in iterated rippled noise. *J. Acoust. Soc. Am.*, **100**(5), 3286–3294. 17

Pelli, D. (1991). Noise in the visual system may be early. In Movshon, M. L. . J. A., editor, *Computational Models of Visual Processing*, pages 147–152. MIT Press, Cambridge, Massachusetts. 11

Pelli, D. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spat. Vis.*, **10**(4), 437–442. 62, 84

Peterson, W., Birdsall, T., and Fox, W. (1954). The theory of signal detectability. *IRE Trans. Inf. Theory*, **4**(4), 171–212. 8

Petrov, A., Dosher, B., and Lu, Z. (2005). The dynamics of perceptual learning: an incremental reweighting model. *Psychol. Rev.*, **112**(4), 715–743. 26, 47, 161

Petzold, P. and Haubensak, G. (2001). Higher order sequential effects in psychophysical judgments. *Atten. Percept. Psychophys.*, **63**(6), 969–978. 129

Plack, C., Oxenham, A., and Fay, R. (2005). *Pitch: neural coding and perception*, volume 24. Springer Science+Business Media, New York, New York. 17

Plomp, R. (1978). Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J. Acoust. Soc. Am.*, **63**(2), 533–549. 159

Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, **256**(5059), 1018–1021. 7, 146

Polderman, T. J., Derks, E. M., Hudziak, J. J., Verhulst, F. C., Posthuma, D., and Boomsma, D. I. (2007). Across the continuum of attention skills: a twin study of the swan adhd rating scale. *J. Child Psychol. Psychiatry*, **48**(11), 1080–1087. doi:10.1111/j.1469-7610.2007.01783.x. 86

Pollack, I. (1975). Auditory informational masking. *J. Acoust. Soc. Am.*, **57**(S1), S5–S5. 56

Pujol, R., Lavigne-Rebillard, M., and Uziel, A. (1991). Development of the human cochlea. *Acta Otolaryngol. (Stockh.)*, **111**(S482), 7–13. 97

Recanzone, G., Schreiner, C., and Merzenich, M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *J. Neurosci.*, **13**(1), 87–103. 26

Richards, V., Heller, L., and Green, D. (1991). The detection of a tone added to a narrow band of noise: The energy model revisited. *Q. J. Exp. Psychol.*, **43**(3), 481–501. 17

Richards, V. and Neff, D. (2004). Cuing effects for informational masking. *J. Acoust. Soc. Am.*, **115**(1), 289–300. 90

Richards, V. and Zhu, S. (1994). Relative estimates of combination weights, decision criteria, and internal noise based on correlation coefficients. *J. Acoust. Soc. Am.*, **95**(1), 423–434. 11, 58, 64, 77

Robinson, D. and Dadson, R. (1956). A re-determination of the equal-loudness relations for pure tones. *Brit. J. Appl. Phys.*, **7**(5), 166–181. 75

Robinson, K. and Summerfield, Q. (1996). Adult auditory learning and training. *Ear Hear.*, **17**(3), 51S–66S. 6

Rodriguez, M., Mischel, W., and Shoda, Y. (1989). Cognitive person variables in the delay of gratification of older children at risk. *J. Pers. Soc. Psychol.*, **57**(2), 358–367. 163

Rosenberg, A. (1966). Pitch discrimination of jittered pulse trains. *J. Acoust. Soc. Am.*, **39**(5), 920–928. 17

Rowan, D. and Lutman, M. (2007). Learning to discriminate interaural time differences at low and high frequencies. *Int. J. Audiol.*, **46**(10), 585–594. 3

Ruff, H. and Rothbart, M. (1996). *Attention in early development: Themes and variations*. Oxford University Press (New York). 100

Sarampalis, A., Kalluri, S., Edwards, B., and Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *J. Speech Lang. Hear. Res.*, **52**(5), 1230–1240. 167

Scharf, B. (1970). Critical bands. In Tobias, J., editor, *Foundations of Modern Auditory Theory*, volume 1, chapter 5, pages 159–202. Academic Press, New York. 61

Schneider, B., Morrongiello, B., and Trehub, S. (1990). Size of critical band in infants, children, and adults. *J. Exp. Psychol. Hum. Percept. Perform.*, **16**(3), 642–652. 97

Schoups, A., Vogels, R., and Orban, G. (1995). Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. *J. Physiol.*, **483**(3), 797–810. 25, 128

Schulman, A. and Greenberg, G. (1970). Operating characteristics and a priori probability of the signal. *Atten. Percept. Psychophys.*, **8**(5), 317–320. 13

Schweizer, K. and Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence*, **32**(4), 329–347. 27

Seitz, A. and Dinse, H. (2007). A common framework for perceptual learning. *Curr. Opin. Neurobiol.*, **17**(2), 148–153. 21

Selen, L., Shadlen, M., and Wolpert, D. (2012). Deliberation in the motor system: reflex gains track evolving evidence leading to a decision. *J. Neurosci.*, **32**(7), 2276–2286. 162

Shaw, E. and Piercy, J. (1962). Physiological noise in relation to audiometry. *J. Acoust. Soc. Am.*, **34**(5), 745–745. 13, 74

Shinn-Cunningham, B. (2000). Adapting to remapped auditory localization cues: A decision-theory model. *Atten. Percept. Psychophys.*, **62**(1), 33–47. 13

Shonle, J. and Horan, K. (1980). The pitch of vibrato tones. *J. Acoust. Soc. Am.*, **67**(1), 246–252. 17

Shub, D. E. (2012). personal communication. 75

Siebert, W. (1970). Frequency discrimination in the auditory system: Place or periodicity mechanisms? *Proc. IEEE*, **58**(5), 723–730. 10, 47

Soderquist, D. and Lindsey, J. (1971). Physiological noise as a low-frequency masker: The cardiac cycle. *J. Acoust. Soc. Am.*, **50**(1), 143–143. 13, 30, 74

Speeth, S. and Mathews, M. (1961). Sequential effects in the signal-detection situation. *J. Acoust. Soc. Am.*, **33**(8), 1046–1054. 143

Spiegel, M. and Green, D. (1981). Two procedures for estimating internal noise. *J. Acoust. Soc. Am.*, **70**(1), 69–73. 34, 38

Spiegel, M. and Watson, C. (1984). Performance on frequency-discrimination tasks by musicians and nonmusicians. *J. Acoust. Soc. Am.*, **76**(6), 1690–1695. 73

Staddon, J., King, M., and Lockhead, G. (1980). On sequential effects in absolute judgment experiments. *J. Exp. Psychol. Hum. Percept. Perform.*, **6**(2), 290–301. 129

Stellmack, M., Willihnganz, M., Wightman, F., and Lutfi, R. (1997). Spectral weights in level discrimination by preschool children: Analytic listening conditions. *J. Acoust. Soc. Am.*, **101**(5), 2811–2821. 59, 64

Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Atten. Percept. Psychophys.*, **63**(8), 1348–1355. 60

Stratton, G. (1897). Vision without inversion of the retinal image. *Psychol. Rev.*, **4**(4), 341–360. 1

Swanson, J., Schuck, S., Mann, M., Carlson, C., Hartman, K., and Sergeant, J. (2005). Categorical and dimensional definitions and evaluations of symptoms of adhd: The snap and the swan ratings scales [draft]. available at: `http://www.adhd.net/SNAP_SWAN.pdf`. 86

Swets, J. (1959). Multiple observations of signals in noise. *J. Acoust. Soc. Am.*, **31**(4), 514–521. 10, 16, 30, 59, 148, 162

Swets, J. (1973). The relative operating characteristic in psychology. *Science*, **182**(4116), 990–1000. 107

Swets, J., Tanner Jr, W., and Birdsall, T. (1961). Decision processes in perception. *Psychol. Rev.*, **68**(5), 301–340. 8

Tallal, P., Miller, S., Bedi, G., Byma, G., Wang, X., Nagarajan, S., Schreiner, C., Jenkins, W., and Merzenich, M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, **271**(5245), 81–84. 81, 159

Talwar, S. and Gerstein, G. (1999). A signal detection analysis of auditory-frequency discrimination in the rat. *J. Acoust. Soc. Am.*, **105**(3), 1784–1800. 10

Tang, Z. and Richards, V. (2003). Examination of a linear model in an informational masking study. *J. Acoust. Soc. Am.*, **114**(1), 361–367. 56, 58, 82

Tanner, Jr., T., Haller, R., and Atkinson, R. (1967). Signal recognition as influenced by presentation schedules. *Atten. Percept. Psychophys.*, **2**(8), 349–358. 13, 129

Tanner, Jr., W. (1958). What is masking? *J. Acoust. Soc. Am.*, **30**(10), 919–921. 37, 56

Tanner, Jr., W. and Birdsall, T. (1958). Definitions of d' and n as psychophysical measures. *J. Acoust. Soc. Am.*, **30**(10), 922–928. 30

Tanner, Jr., W. and Rivette, C. (1963). Learning in psychophysical experiments. *J. Acoust. Soc. Am.*, **35**(11), 1896–1896. 3

Teich, M. and Khanna, S. (1985). Pulse-number distribution for the neural spike train in the cat's auditory nerve. *J. Acoust. Soc. Am.*, **77**(3), 1110–1128. 10

Thomas, J., Gille, J., and Barker, R. (1982). Simultaneous visual detection and identification: theory and data. *J. Opt. Soc. Am.*, **72**(12), 1642–1651. 141

Trehub, S., Schneider, B., Thorpe, L., and Judge, P. (1991). Observational measures of auditory sensitivity in early infancy. *Dev. Psychol.*, **27**(1), 40–49. 146

Tremblay, K., Kraus, N., and McGee, T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, **9**(16), 3557–3560. 3, 5

Tucker, A., Williams, P., and Jeffress, L. (1968). Effect of signal duration on detection for gated and for continuous noise. *J. Acoust. Soc. Am.*, **44**(3), 813–816. 3

Turner, C., Zeng, F., Relkin, E., and Horwitz, A. (1992). Frequency discrimination in forward and backward masking. *J. Acoust. Soc. Am.*, **92**(6), 3102–3108. 17

Van Wanrooij, M. and Van Opstal, A. (2005). Relearning sound localization with a new ear. *J. Neurosci.*, **25**(22), 5413–5424. 3

van Wassenhove, V. and Nagarajan, S. (2007). Auditory cortical plasticity in learning to discriminate modulation rate. *J. Neurosci.*, **27**(10), 2663–2672. 108, 127

Verplanck, W., Collier, G., and Cotton, J. (1952). Nonindependence of successive responses in measurements of the visual threshold. *J. Exp. Psychol.*, **44**(4), 273–282. 129

Viemeister, N. (1970). Intensity discrimination: Performance in three paradigms. *Atten. Percept. Psychophys.*, **8**(6), 417–419. 10

Viemeister, N. F. and Schlauch, R. S. (1992). Issues in infant psychoacoustics. In Werner, L. and Rubel, E., editors, *Developmental Psychoacoustics*, pages 191–210. American Psychological Association, Washington, D.C. 19, 31, 83

Virsu, V., Oksanen-Hennah, H., Vedenpaa, A., Jaatinen, P., and Lahti-Nuuttila, P. (2008). Simultaneity learning in vision, audition, tactile sense and their cross-modal combinations. *Exp. Brain Res.*, **186**(4), 525–537. 3

Vogels, R., Spileers, W., and Orban, G. (1989). The response variability of striate cortical neurons in the behaving monkey. *Exp. Brain Res.*, **77**(2), 432–436. 13, 30

Wald, A. and Wolfowitz, J. (1943). An exact test for randomness in the non-parametric case based on serial correlation. *Ann. Math. Stat.*, **14**(4), 378–388. 129

Ward, L. and Lockhead, G. (1970). Sequential effects and memory in category judgments. *J. Exp. Psychol.*, **84**(1), 27–34. 129

Warren, R. (1974). Auditory temporal discrimination by trained listeners. *Cognit. Psychol.*, **6**(2), 237–256. 3

Watson, C., Kelly, W., and Wroton, H. (1976). Factors in the discrimination of tonal patterns. ii. selective attention and learning under various levels of stimulus uncertainty. *J. Acoust. Soc. Am.*, **60**(5), 1176–1186. 75

Watson, C., Miller, J., Kewley-Port, D., Humes, L., and Wightman, F. (2008). Training listeners to identify the sounds of speech: I. a review of past studies. *Hear. J.*, **61**(9), 26–31. 2

Watson, C., Rilling, M., and Bourbon, W. (1964). Receiver-operating characteristics determined by a mechanical analog to the rating scale. *J. Acoust. Soc. Am.*, **36**(2), 283–288. 10

Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence*. Psychological Corporation. 116

Wenger, M., Copeland, A., Bittner, J., and Thomas, R. (2008). Evidence for criterion shifts in visual perceptual learning: Data and implications. *Atten. Percept. Psychophys.*, **70**(7), 1248–1273. 27, 109, 115, 128

Wenger, M. and Rasche, C. (2006). Perceptual learning in contrast detection: Presence and cost of shifts in response criteria. *Psychonomic Bull. Rev.*, **13**(4), 656–661. xiii, xvi, 2, 7, 27, 107, 109, 111, 112, 113, 114, 115, 116, 119, 120, 121, 122, 123, 128

Werner, L. and Marean, G. (1991). Methods for estimating infant thresholds. *J. Acoust. Soc. Am.*, **90**(4), 1867–1875. 146

Werner, L., Marean, G., Halpin, C., Spetner, N., and Gillenwater, J. (1992). Infant auditory temporal acuity: Gap detection. *Child Dev.*, **63**(2), 260–272. 146

Wichmann, F. and Hill, N. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Atten. Percept. Psychophys.*, **63**(8), 1293–1313. 37, 64

Wickens, T. (2002). *Elementary signal detection theory*, pages 114–118. Oxford University Press (USA), New York, New York. ix, 13, 34, 110, 116, 130

Wier, C., Jesteadt, W., and Green, D. (1976). A comparison of method-of-adjustment and forced-choice procedures in frequency discrimination. *Atten. Percept. Psychophys.*, **19**(1), 75–79. 51

Wightman, F. and Allen, P. (1992). Individual differences in auditory capability among preschool children. In Werner, L. and Rubel, E., editors, *Developmental Psychoacoustics*, pages 113–134. American Psychological Association, Washington, D.C. 19, 27, 37, 87

Wightman, F., Callahan, M., Lutfi, R., Kistler, D., and Oh, E. (2003). Children's detection of pure-tone signals: Informational masking with contralateral maskers. *J. Acoust. Soc. Am.*, **113**(6), 3297–3305. 82, 83

Wightman, F., Kistler, D., and O'Bryan, A. (2010). Individual differences and age effects in a dichotic informational masking paradigm. *J. Acoust. Soc. Am.*, **128**(1), 270–279. 82

Willihnganz, M., Stellmack, M., Lutfi, R., and Wightman, F. (1997). Spectral weights in level discrimination by preschool children: Synthetic listening conditions. *J. Acoust. Soc. Am.*, **101**(5), 2803–2810. 59, 64

Winter, I. and Palmer, A. (1991). Intensity coding in low-frequency auditory-nerve fibers of the guinea pig. *J. Acoust. Soc. Am.*, **90**(4), 1958–1967. 10

Witton, C., Stein, J., Stoodley, C., Rosner, B., and Talcott, J. (2002). Separate influences of acoustic am and fm sensitivity on the phonological decoding skills of impaired and normal readers. *J. Cogn. Neurosci.*, **14**(6), 866–874. 27, 166

Wolpert, D., Diedrichsen, J., and Flanagan, J. (2011). Principles of sensorimotor learning. *Nat. Rev. Neurosci.*, **12**(12), 739–751. 163

Wright, B., Buonomano, D., Mahncke, H., and Merzenich, M. (1997). Learning and generalization of auditory temporal–interval discrimination in humans. *J. Neurosci.*, **17**(10), 3956–3963. 2, 3, 4, 5, 25, 140, 141, 142, 143, 145

Wright, B. and Fitzgerald, M. (2001). Different patterns of human discrimination learning for two interaural cues to sound-source location. *Proc. Natl. Acad. Sci. U. S. A.*, **98**(21), 12307–12312. 3, 5, 72, 73, 77, 127

Wright, B. and Fitzgerald, M. (2003). Sound-discrimination learning and auditory displays. In *9th International Conference on Auditory Display (ICAD03)*. 5

Wright, B. and Fitzgerald, M. (2005). Learning and generalization of five auditory discrimination tasks as assessed by threshold changes. In Pressnitzer, D., de Cheveigne, A., McAdams, S., and Collet, L., editors, *Auditory signal processing: Physiology, psychoacoustics, and models*, pages 509–515. Springer Science+Business Media, New York, New York. 3, 50

Wright, B. and Zhang, Y. (2005). Insights into human auditory processing gained from perceptual learning. In Gazzaniga, M. S., editor, *The cognitive neurosciences IV*, pages 353–366. MIT Press, Cambridge, Massachusetts. 5

Wright, B. and Zhang, Y. (2006). A review of learning with normal and altered sound-localization cues in human adults. *Int. J. Audiol.*, **45**(S1), 92–98. 3

Wright, B. and Zhang, Y. (2009). A review of the generalization of auditory learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **364**(1515), 301–311. 1, 26, 31, 145

Yost, W., Berg, K., and Thomas, G. (1976). Frequency recognition in temporal interference tasks: A comparison among four psychophysical procedures. *Atten. Percept. Psychophys.*, **20**(5), 353–359. 17

Young, E. and Barta, P. (1986). Rate responses of auditory nerve fibers to tones in noise near masked threshold. *J. Acoust. Soc. Am.*, **79**(2), 426–442. 10

Zhang, Y. and Wright, B. (2009). An influence of amplitude modulation on interaural level difference processing suggested by learning patterns of human adults. *J. Acoust. Soc. Am.*, **126**(3), 1349–1358. 5, 72

Zwislocki, J., Maire, F., Feldman, A., and Rubin, H. (1958). On the effect of practice and motivation on the threshold of audibility. *J. Acoust. Soc. Am.*, **30**(4), 254–262. 3