INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI®

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Absolute Identification of Loudness: Theory and Experiment

by

Elad Sagi

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Physiology and the Institute of Biomedical Engineering
University of Toronto

Absolute Identification of Loudness: Theory and Experiment

Master of Science 1998

Elad Sagi

Department of Physiology and the Institute of Biomedical Engineering

University of Toronto

Abstract

An absolute identification experiment on loudness requires that the subject classify as well as he/she can the loudness of a stimulus tone. The intensity level is randomly selected from **m** stimulus categories within a fixed stimulus range, **R**. The matching of a stimulus category with a response category comprises one trial. **N** such trials are compiled into an **m x m** "Confusion" matrix. Subject performance is quantified from the matrix using an information measure **I(N, m, R)**. The dependence of **I(N, m, R)** on **N** includes a small sample bias. The dependence of **I(N, m, R)** on **m** is a mathematical property, while the dependence on **R** is a property of the individual. Utilizing the observation that a subject's responses are governed by a normal distribution of constant variance, a model has been developed that accounts for these three dependencies.

Acknowledgments

I would like to thank my supervisor, Dr. Ken Norwich, for his unforgettable inspiration, patience and support.

I would also like to thank my supervisory committee, Dr. Peter Hallett, Dr. Hans Kunov and Dr. Ken Norwich for taking the time to oversee my development throughout the program.

I am greatly indebted to the late Professor Poul Madsen for all of his technical support. His ingenuity and unique knowledge made an invaluable contribution to the Institute of Biomedical Engineering (IBME) at the University of Toronto and he will be greatly missed. I am also indebted to the Bioacoustics Group for all their assistance and especially to Dr. Hans Kunov for allowing full access to the IBME bioacoustics laboratory.

This work was supported by the Natural Sciences and Engineering Research Council of Canada. I am grateful to the University of Toronto for providing me with a fellowship, without which this research would not be possible.

Finally, I would like to thank my family and friends for their enduring support from beginning to end.

Contents

List of Figures

Chapter 1

Introduction

1.1 Information and Hearing

How well do we perceive? Perception is a very difficult concept to define; however, it is used in our everyday lives. As explained by Norwich (1993, pg. 12),

> "Lest's start with the ancients—at least with the ancient Romans. I was miffed, a few years ago, when I learned from the Department of Classics at the University of Toronto that the Roman-in-the-street would not likely have said "percipio" if he or she meant "I perceive" (I had dutifully ferreted this out of a Latin dictionary), but rather "intellego." Wherefore intellego? Well, etymologically the word is made up of two simpler Latin words: *inter* meaning *between* and *lego* meaning *I choose* or *gather*. To *perceive* was, then *to choose between* alternatives."

I would like to adopt this more exoteric approach, that of *intellego perception*, in addressing the question at hand. That is, How well do we perceive, or, how well do we choose between alternatives?

Let us first consider the amount of uncertainty we have regarding the selection of an alternative. The more uncertain one is about the selection of an alternative, the

more information one gains when the uncertainty associate with that selection is reduced completely. In our imperfection, however, we are unable to reduce uncertainty completely and must be satisfied with the information gained in the reduction of uncertainty up to our physiological limits. This type of information gain is called the *transmitted information*. Hence, we can describe our ability to make a selection among a set of possible alternatives in terms of the amount of information transmitted to us in the process of making the selection.

In the area of hearing, these ideas were applied to an experimental setting in the form of the absolute identification paradigm (Garner and Hake, 1951). Absolute identification of loudness describes a subject's ability in the task of identifying the intensity of a stimulus tone (in decibels [dB]) selected from a set of possible stimulus categories. A subject's performance is assessed by how well he/she is able to classify the stimulus category in terms of a corresponding response category. For simplicity, we will consider an equal number of stimulus and response categories.

The subject's task in the absolute identification paradigm is simply the assignment of an appropriate response category to a stimulus category. If each assignment is considered as one trial, N such trials can be compiled into a stimulus/response matrix or *confusion matrix*. The overall information transmitted to the subject can be estimated from the confusion matrix in the form of an information measure.

In an absolute identification experiment on loudness, the information measure is sensitive to three variables: the number of trials, N; the number of stimulus response categories, m; and the fixed range, R, (in dB) over which stimulus categories are constructed.

Only after a sufficiently large number of trials does the information measure, or $I(N, m, R)$, approach the information transmitted to the subject, $I_t(m, R)$, in that absolute identification experiment. For small N, a small sample bias is incurred and $I(N, m, R)$ overestimates $I_t(m, R)$. Hence, there is a requirement for a method of overcoming small sample bias.

Both $I(N, m, R)$ and $I_t(m, R)$ are affected by the number of stimulus/response categories used for experimentation in that both tend to increase for an increasing number of categories. The increase, however, saturates for large enough m. In particular, the upper bound of $I_t(m, R)$ corresponds to an infinitely large number of categories and is completely determined by the stimulus range. This upper bound can be considered as the "channel capacity" of the transmitted information associated with the absolute identification experiment, $I_t(R)$.

Finally, each of $I(N, m, R)$, $I_t(m, R)$ and $I_t(R)$ are affected by the fixed stimulus range selected for the experiment in that all three tend to increase for a larger stimulus range. As with the increase in $I_t(m, R)$ due to m, the increase in $I_t(R)$ due to R tends to saturate towards an upper limit. This upper limit can be described as the channel capacity of the transmitted information due to the sensory limits of the subject, I_∞.

This thesis focuses on improving information estimates. A subject's error in the selection of response categories to a given stimulus category is modelled after an underlying normal distribution of constant variance, σ^2 [CV-model]. By confining this distribution to the stimulus range selected for experimentation, the distributions of responses as found in actual confusion matrices are accounted for. This includes the apparent "edge effects" or anchoring caused by stimulus intensities more to the extremes of the stimulus range. Through computer simulation, the model is able to predict and track the small sample bias incurred in $I(N, m, R)$. This is achieved by three essential steps. First, an estimate of the average variance is obtained from all the rows of an actual confusion matrix (including those influenced by edge effects). Second, the CV-model is used to estimate σ^2 associated with the distribution thought to underlie a subject's performance. Finally, σ can be used to generate pseudo-stimulus/response pairs and create simulated confusion matrices.

Essentially, σ only depends on the stimulus range such that $\sigma(R) = aR + b$ (a, b positive constants). The significance of σ is that it can be used to estimate $I_t(m, R)$, $I_t(R)$ and I_∞. Also, σ can be used to account for the affects of m and R described

earlier. We have found that stimulus range and number of stimulus categories both affect the information transmitted to the subject; however, the effect of stimulus categories is purely mathematical while the effect of range is determined by the sensory properties of the individual.

Chapter 2

Background

2.1 Information Theory

2.1.1 A Measure of Uncertainty

Information can be defined using a concept of uncertainty. The more uncertain we are about the occurrence of some event, the more information we obtain when that event is revealed. For example, imagine I am holding a playing card in my hand and that I always tell the truth. If I express that the card in my hand is none other than the Queen of Hearts, more uncertainty has been removed than if I only revealed the suit, i.e. Hearts. Also, more information has been gained in the former over the latter. We can therefore see a gain in information as a reduction in uncertainty. To quantify information, we must obtain a measure for the amount of uncertainty associated with some event.

Lets consider an event X with m discrete, independent outcomes, defined as $X = \{X_1, ...X_m\}$. Each X_j can be associated with a probability of occurrence p_j such that $\sum_j p_j = 1$. We would like to develop a function $\psi(p_j)$ that measures the amount of uncertainty associated with the occurrence of X_j.

The independence of any two outcomes, X_j and X_k, implies that the occurrence of one in no way affects the probability of occurrence of the other. Consider now the uncertainty

associated with the joint occurrence of X_j and X_k, i.e. $\psi(p_j \cdot p_k)$. If we remove from this the uncertainty of X_j, the condition of independence would require us to be left with the uncertainty of X_k. We can formally express this property of the uncertainty measure as

$$\psi(p_j \cdot p_k) = \psi(p_j) + \psi(p_k).$$

Without getting into too much rigor, this property leads the uncertainty measure to take on the form

$$\psi(p_j) = C \log p_j$$

where C is some arbitrary constant. If we further impose the condition $\psi \geq 0$, then we can be satisfied with the form

$$\psi(p_j) = -\log p_j.$$

Notice that when the occurrence of X_j is certain, i.e. $p_j = 1$, then no uncertainty is to be found in its outcome, i.e. $\psi(p_j) = 0$.

Now that a measure of the uncertainty in X_j has been established, the *total uncertainty* or "entropy" in the set of events X can be defined as the average uncertainty for all the outcomes in X, i.e. $H(X)$. Hence,

$$H(X) = \sum_{j=1}^{m} \psi(p_j) p_j$$

or in expanded notation,

$$H(X) = -\sum_{j=1}^{m} p(X_j) \log p(X_j).$$

As a special case, consider the situation where the X_j's are equally likely. The corresponding probabilities would therefore be uniform such that $p(X_j) = \frac{1}{m}$ and the entropy would become

$$H(X) = -\sum_{j=1}^{m} \frac{1}{m} \log \frac{1}{m} = \log m.$$

This tells us that the entropy associated with equally probable events can be described as the logarithm of the number of events.

Finally, the units used to describe the entropy of an event depends on the base of the logarithm. For base 2, we use "bits" and for base e, we use "natural units". In our special case, the bit becomes useful in describing, theoretically, the number of two-choice discriminations required to reduce the uncertainty. For example, if I hold in my hand 4 Queens from a fair deck, the uncertainty can be calculated as $H(X) = \log_2 4 = 2$ bits. If you are asked to describe a process that would reveal the Queen of Hearts, we are assured that the process requires at most two steps notwithstanding correct guesses in any step.

2.1.2 Transmission of Information

This development follows the work of Shannon (1948) and Wiener (1948), the originators of modern Information Theory.

Imagine that we are sending a discrete message $X = \{X_1, ..., X_m\}$ that is received in the form of $Y = \{Y_1, ..., Y_{m'}\}$. To our misfortune, the message is not noise free and more than one possibility in Y registers upon receiving the element X_j. The amount of noise incurred can be described in terms of the probability of response Y_k conditioned upon the receival of X_j, or $p(Y_k|X_j)$. Using the previous section, we can measure the total amount of uncertainty in Y inflicted by noisy transmission of X_j as

$$H(Y|X_j) = -\sum_{k=1}^{m'} p(Y_k|X_j) \log p(Y_k|X_j).$$

Furthermore, this uncertainty can be averaged over all of X, i.e.

$$H(Y|X) = \sum_{j=1}^{m} H(Y|X_j)p(X_j)$$

or equivalently,

$$H(Y|X) = -\sum_{j=1}^{m}\sum_{k=1}^{m'} p(X_j)p(Y_k|X_j)\log p(Y_k|X_j).$$

Albeit cumbersome, this expression is the amount of uncertainty that is unresolvable due to noisy transmission. Typically, $H(Y|X)$ is called the "response equivocation".

Previously, information was described as a reduction in uncertainty. The amount of information transmitted in the process of sending the message X can be measured by the degree to which the uncertainty in Y has been reduced by the receiver. Let us represent the amount of information received as $I(Y|X)$. We state, without proof, that the amount of information received equals the amount of information transmitted, i.e. $I(Y|X) = I(X|Y)$. In any case, perfect transmission of information occurs when the receiver is able to reduce all of the uncertainty in Y, i.e. the "response entropy" $H(Y)$. We can represent perfect transmission as

$$I(Y|X) = H(Y).$$

Typically, information transmission is not noise free and the receiver is only able to reduce the response entropy up to the response equivocation. We therefore present the information transmitted as

$$I(Y|X) = H(Y) - H(Y|X) \tag{2.1}$$

or in expanded notation,

$$I(Y|X) = -\sum_{k=1}^{m'} p(Y_k)\log p(Y_k) + \sum_{j=1}^{m}\sum_{k=1}^{m'} p(X_j)p(Y_k|X_j)\log p(Y_k|X_j). \tag{2.2}$$

Qualitatively, the transmitted information expresses how well the Y_k reflect the accurate transmission of the corresponding X_j.

2.2 Absolute Identification and the Confusion Matrix

The absolute identification paradigm measures how well a subject is able to classify a stimulus as belonging to a specific category. The stimulus must originate from a finite set of possibilities called "stimulus categories". If the stimulus lies along a continuum, then the continuum must be artificially discretized. The observer identifies the stimulus by making a selection from a finite set of response alternatives or "response categories". A subject's performance is measured by how well he/she matches a stimulus category with its corresponding response category. Each stimulus/response category differs along one or more dimensions. For example, let us conduct an experiment where the subject is required to classify the strength of a sugar solution as "unsweet", "slightly sweet", "moderately sweet", "sweet" or "very sweet". These categories differ along the single dimension of intensity. As stimulus categories, these alternatives take the form of *physical* intensity. As response categories, these alternatives take the form of *perceived* intensity. The experiment can be extended to two dimensions by allowing the solution to be one of "sweet", "salty", "bitter" or "sour", while varying the intensity as before. In doing this, we have provided the subject with 20 alternatives where previously, only 5 were available. For simplicity, this thesis will focus on absolute identification experiments that vary along one dimension.

Stimulus and response categories can be described set theoretically. The set of possible stimuli, X, is finite and contains the category X_j; namely, $X = \{X_1, ..., X_m | 1 \leq j \leq m\}$. Similarly, the set of possible responses, Y, is finite and contains the category Y_k; namely $Y = \{Y_1, ..., Y_{m'} | 1 \leq k \leq m'\}$. In the above example, we could have written $X = Y = \{$unsweet,slightly sweet,moderately sweet,sweet,very sweet$\}$, where $X_1 = Y_1 =$ "unsweet", $X_2 = Y_2 =$ "slightly sweet", $X_3 = Y_3 =$ "moderately sweet", etc. One should note that the number of stimulus categories does not have to equal the number of response categories.

In any case, as long as we are able to define a discrete set of stimuli and "identify" them using a discrete set of responses, we can set up the absolute identification paradigm. Let us now define one trial as the pairing of a given stimulus category X_j with a response category Y_k. We can compile N such trials in a "stimulus/response" or "confusion" matrix where the element n_{jk} represents the number of times X_j was identified as Y_k. A constructed example is shown below.

	Y_1	Y_2	Y_3	Y_4	Y_5	X_j^{total}
X_1	8	2				10
X_2	2	7	1			10
X_3		2	6	2		10
X_4			1	7	2	10
X_5				2	8	10
Y_k^{total}	10	11	8	11	10	50

In this example, the number of times a moderately sweet solution (X_3) was identified as a slightly sweet solution (Y_2) was twice ($n_{32} = 2$). The matrix, thus, provides a qualitative description of the subject's performance. For every row, the correct responses lie along the main diagonal of the matrix while incorrect responses fall to the left and/or right.

Subject performance can also be described quantitatively by applying an information measure to the confusion matrix (Garner and Hake, 1951). First, let's consider the generalized confusion matrix outlined below.

	Y_1	...	Y_k	...	$Y_{m'}$	X_j^{total}
X_1	n_{11}	...	n_{1k}	...	$n_{1m'}$	$n_{1.}$
:	:		:		:	:
X_j	n_{j1}	...	n_{jk}	...	$n_{jm'}$	$n_{j.}$
:	:		:		:	:
X_m	n_{m1}	...	n_{mk}	...	$n_{mm'}$	$n_{m.}$
Y_k^{total}	$n_{.1}$	...	$n_{.k}$	...	$n_{.m'}$	N

From Section 2.1, we recall that

$$I(Y|X) = -\sum_{k=1}^{m'} p(Y_k) \log p(Y_k) + \sum_{j=1}^{m} \sum_{k=1}^{m'} p(X_j) p(Y_k|X_j) \log p(Y_k|X_j).$$

Each probability can be measured from the matrix using the following maximum likelihood estimates:

$$\begin{aligned} p(Y_k) &= \frac{n_{.k}}{N} \text{ where } n_{.k} = \sum_{j=1}^{m} n_{jk} \\ p(X_j) &= \frac{n_{j.}}{N} \text{ where } n_{j.} = \sum_{k=1}^{m'} n_{jk} \\ p(Y_k|X_j) &= \frac{n_{jk}}{n_{j.}}. \end{aligned}$$

With a little algebra, substituting these into $I(Y|X)$ gives the maximum likelihood estimate $\hat{I}(Y|X)$;

$$\hat{I}(Y|X) = \log N + \frac{1}{N} \left(\sum_{j=1}^{m} \sum_{k=1}^{m'} n_{jk} \log n_{jk} - \sum_{j=1}^{m} n_{j.} \log n_{j.} - \sum_{k=1}^{m} n_{.k} \log n_{.k} \right). \qquad (2.3)$$

In our constructed example above, $\hat{I}(Y|X) = 0.9$ n.u. or about 1.3 bits. A qualitative interpretation of these values will be postponed to Section 2.4. For now, the absolute

identification paradigm has been outlined and the application of an information measure has been described.

We must, however, introduce a caveat: the amount of information transmitted to the subject in performing the absolute identification experiment is measured from the *a posteriori* values in the confusion matrix. The confusion matrix is a statistically sensitive entity that develops progressively with increasing experimental trials. Hence, the amount of information calculated from the confusion matrix is merely a maximum likelihood estimate of $I(Y|X)$. From this point onward, the information as measured from the confusion matrix will be represented symbolically as $I(N, m, R)$. This estimate is sensitive to the number of experimental trials, N, and the number of categories, m. We mostly consider experiments conducted where the number of stimulus categories equals the number of response categories. Similarly, all of the experiments considered involve the discretization of a range R that lies on a continuum. As a final note, instead of $I(Y|X)$, the transmitted information will be designated by a small subscript "t"; for example, I_t.

2.3 Overcoming Small Sample Bias

The purpose of the absolute identification experiment is to measure the information transmitted to the subject, I_t, given a range R that has been discretized into m categories. The information as measured from the resulting m x m confusion matrix, $I(N, m, R)$, is not an unbiased estimate of I_t. $I(N, m, R)$ is sensitive to the number of experimental trials, N. Previous investigators have determined that insufficiently large values of N tend to yield values of $I(N, m, R)$ that overestimate I_t and have attempted methods to overcome this "small sample bias".

Miller (1955) developed an approximation to correct the small sample bias, but the approximation requires that N be in excess of five times the square of the number of categories used for the confusion matrix. This would mean that a 50 x 50 matrix would

require at least 12,500 trials to obtain an adequate estimate of I_t from $I(N, m, R)$. In the absolute identification experiments conducted in our laboratory (please see section 3.1 below), we found that subjects fatigue after approximately 200 trials of experimentation in one sitting. More than 60 sittings would still be required before an adequate estimate of I_t is to be obtained using Miller's correction. In any case, estimating I_t from $I(N, m, R)$ requires far too many trials and is difficult to obtain from one subject. For this reason, previous investigators pooled data from several subjects in order to overcome the small sample bias.

Carlton (1969) developed an approximation that defines the bias in information estimates and is far more accurate than that proposed by Miller. Carlton's approximation, however, requires an *a priori* knowledge of the probability density functions that govern the information measure. This is difficult in the case of $p(Y_k|X_j)$, the probability of response Y_k, conditioned upon the stimulus X_j, since some assumption is required about the response properties of the individual. If $p(Y_k|X_j)$ were known, not only could we accurately track the overestimation of I_t, but we could also calculate I_t directly from Equation 2.2, free from any bias.

Houtsma (1983) proposed a method for overcoming small sample bias using a computer algorithm to simulate an absolute identification experiment on a trial by trial basis. The stimulus was represented as a uniformly distributed pseudorandom integer, X, generated such that $1 \leq X \leq m$. The response was represented as $Y = X + T$ whereby T was also a uniformly distributed integer such that $-S \leq T \leq S$ and $1 \leq Y \leq m$. The resulting stimulus/response pair, (X, Y), would be compiled into the corresponding m x m confusion matrix. Performance would be reflected by small or large values of S. That is, large values for S would result in more non-zero elements off the main diagonal of the matrix. This would yield lower values of information reflecting poor performance.

Houtsma conducted his own absolute identification experiment and plotted the estimated information $I(N, m, R)$ (hereafter, $I(N)$) as a function of the number of stimulus trials (in the thousands). $I(N)$ decreased monotonically towards an asymptote with in-

creasing N. The monotonic decrease was also found to occur in plots of $I(N)$ measured from simulation. As a method for overcoming small sample bias, Houtsma proposed to determine the best S-value such that the experimental and simulated plots of $I(N)$ with increasing N would overlap. Subsequently, $I(N)$ would be extended through simulation to yield unbiased estimates of I_t.

Extending the work of Houtsma, Wong and Norwich (1997) proposed a computer simulation that models the responses as having a normal distribution about the stimulus. This simulation generated confusion matrices that resemble those found from experiment more closely than Houtsma's approach. They made the further observation that the responses along the more central rows of the confusion matrix tend to a common variance value, σ, which could be used to simulate the data and overcome small sample bias. The σ-value was estimated directly from the confusion matrix by taking a geometric mean of the *a posteriori* variances along all the rows. One should note that σ was not found through a curve-fitting procedure.

The main drawback to the approach of Wong and Norwich was that apparent anchor effects were accounted for in a preliminary fashion. Although the central rows tend to a normal distribution of common variance, the distribution of responses along the extreme rows of the matrix display heavy skewing or "anchoring". Since the variances of these rows were incorporated into the estimate for σ, the accuracy of the resulting information estimates was affected. In Chapter 3 of this thesis, the approach of Wong and Norwich is extended to incorporate edge effects yielding even more accurate information estimates.

2.4 Transmitted Information and Number of Stimulus Categories

An absolute identification experiment is typically conducted over some range R that has been discretized into stimulus and response categories. For now, let's consider the case where the number of stimulus categories, m, equals the number of response categories.

$I(N, m, R)$ is calculated from the resulting m x m confusion matrix after N trials of experimentation. For N sufficiently large, the small sample bias incurred in the information measure is overcome and $I_t(m, R)$ represents the information transmitted to the subject for that absolute identification experiment. We can represent this process as

$$N \to \infty \Longrightarrow I(N, m, R) \to I_t(m, R).$$

Effectively, $I_t(m, R)$ provides a theoretical measure for the number of categories a subject can identify without error. For example, consider a range of $1-90$ dB discretized into three categories represented by three stimuli spread equally across the range. In our case, category 1 would be represented by 1 dB, category 2 by 45 dB and category 3 by 90 dB. One could imagine that a subject would likely not err in identifying absolutely the category to which each stimulus belongs. If no errors are made, information transmission is noise free and Equation 2.1 above becomes

$$I(Y|X) = H(Y).$$

Also, if the presentation of stimuli are equally likely, the distribution of responses, $p(Y_k)$, is uniform and noise free transmission becomes

$$I(Y|X) = -\sum_{j=1}^{m} p(Y_k) \log p(Y_k) = \log m.$$

In our example above, we would therefore expect $I_t(3, 90) = \log 3$. In natural units, this would be expressed as $I_t(3, 90) = \ln 3 = 1.099$ n.u.

We can extend this example to the situation where the range $1-90$ dB is discretized, in a similar fashion, into 5 categories. The stimulus corresponding to its category in increasing order would be 1 dB, 23 dB, 45 dB, 67 dB and 90 dB. A subject who is able to absolutely identify these categories would yield an information value of $I_t(5, 90) = \ln 5 = 1.609$ n.u. We can imagine that as the number of categories used to discretize

the range increases, subjects will make more mistakes in the absolute identification task. In these situations, $I_t(m, R)$ describes, theoretically, how many categories the subject would be able to identify absolutely. For example, say one finds experimentally that $I_t(20, 90) = 1.65$ n.u. This value corresponds to $\exp(1.65) \cong 5.2$ categories of absolute identification. Although 20 categories were used to discretize a range of 90 dB, $I_t(20, 90)$ tells us that the subject would be able to absolutely identify only around 5.2 categories.

Many investigators have conducted experiments whereby $I_t(m, R)$ was estimated after the number of categories used to discretize a fixed range was progressively increased. Miller (1956) outlined several of these experiments in his celebrated paper, "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". Typically, $I_t(m, R)$ initially increased linearly with $\ln m$ indicating perfect or near perfect transmission. After about 4 or 5 categories, errors in absolute identification became more prominent such that the increase in $I_t(m, R)$ slowed markedly and saturated after about 20 categories. This effect is outlined diagrammatically in Figure 2.1. The saturation in $I_t(m, R)$ has been commonly referred to as the human "Channel Capacity" for the information processed per stimulus and reaches a value of around 1.8 n.u. or $\exp(1.8) \cong 6$ categories of absolute identification. The quoted values of channel capacity reported by Miller slightly fluctuated around this number, hence his title of 7 ± 2 (7 was probably used for dramatic appeal).

Since its discovery nearly 50 years ago, the form of the increase outlined in Figure 2.1 has been quoted in many psychology textbooks (for example, Coren & Ward (1989) pg. 32) as the standard description of the informational processing limits of categorical judgements. That is, $I_t(m, R)$ will increase monotonically and saturate at channel capacity as m increases. Upon review of some experimental data from previous investigators, one finds that estimates of $I_t(m, R)$ do not necessarily follow the monotonic increase. In fact, some investigators found that $I_t(m, R)$ estimates decrease for large m (Miller (1956), Erikson and Hake (1955)). In those cases, the peak information estimate was considered as the channel capacity.

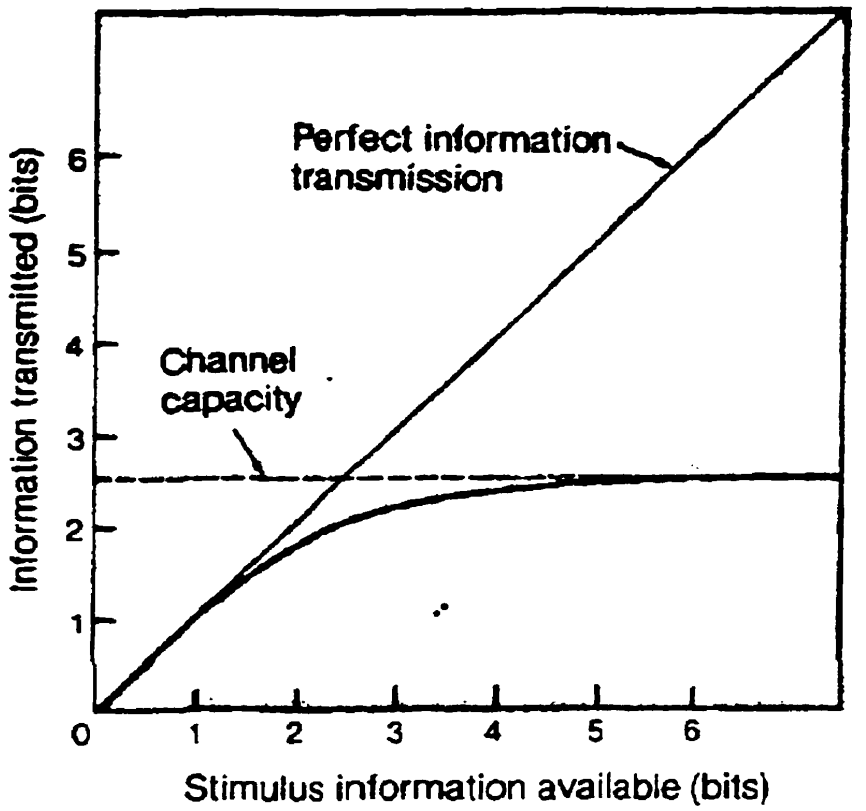

Figure 2-1: A typical description for the increase in transmitted information, $I_t(m, R)$ (measured in bits), with the number of stimulus categories m over a fixed range R. The straight diagonal represents perfect transmission. The curve represents typical performance. The dotted horizontal line is channel capacity (this is what I define as $I_t(R)$). "Stimulus information" is equal to $\log_2(m)$. From Coren and Ward (1989, pg. 32).

Different views have been taken towards this phenomenon. First, estimates of $I_t(m, R)$ contain not only experimental error, but may be biased depending on how many trials were used relative to the number of categories used. For example, if not enough trials were used for experiments utilizing fewer categories, the resulting information estimates will be too high. On the other hand, MacRae (1970) attempted to correct for small sample bias and maintained that the fall in estimates of $I_t(m, R)$ for large m is not an experimental artifact and should be considered. However, MacRae's bias corrections were made on experiments of previous investigators that likely pooled subject data in their original attempts in obtaining unbiased information estimates. This implies that "poor" performance could be averaged with "good" performance. Since most subjects perform very well for few categories, intersubject differences will be amplified mostly for larger values of m. Hence, a "poor" performance would cause more of a decrease in averaged estimates of $I_t(m, R)$ in experiments that use larger numbers of categories. In any case, the standard Miller-type curve in Figure 2.1 is considered to represent the informational properties of the subject.

We contend that the generally held view is in error and that the increase of $I_t(m, R)$ is a mathematical artifact of the information measure itself. In most experiments of the type, "$I_t(m, R)$ vs. m for fixed R", a point stimulus was chosen to represent the stimulus category. Typically, point stimuli were selected in order to optimize information transfer, but one must ask if this adequately reflects how the subject actually discretizes the range. If we adopt the following notation,

$$m \to \infty \Longrightarrow I_t(m, R) \to I_t(R),$$

then $I_t(R)$ represents the channel capacity that is achieved in the Miller-type experiments and reflects the theoretical case in which the subject uses an infinite number of categories. We contend that $I_t(R)$ reflects the informational properties of the subject in

response to the stimulus range; however, the approach of $I_t(m, R)$ towards $I_t(R)$ reflects the improvement in the resolving power of the information measure. This issue will be addressed in more detail in Section 4.2.

2.5 Transmitted Information and Stimulus Range

Using a larger number of stimulus and response categories in an absolute identification experiment tends to increase the information transmitted to the subject. The increase is, however, bounded from above. The information transmitted to the subject will stop increasing beyond a certain number of categories. How is this "channel capacity" affected by the size of the stimulus range used for experimentation?

Erikson and Hake (1955) conducted absolute identification experiments involving judgements on visual size. Subjects were told that the purpose of the experiment was to determine how well people could distinguish between different sized squares. Two sizes for stimulus range were studied: "large" and "small". In each case, the range was discretized into 5, 11, or 21 stimulus and response categories. The number of stimulus categories did not necessarily equal the number of response categories and all 9 possibilities for each range were tested. The large range consisted of different sized squares between $2 - 82$ mm^2 and the small range was between $2 - 42$ mm^2. They found that information estimates for the larger range were consistently greater than those for the smaller range with the exception of the 21 x 21 experiment. In this case the information estimates were equal for both ranges.

For absolute identification experiments on loudness, three groups of investigators have independently tested the effect of stimulus range on $I_t(m, R)$. All three investigators used stimulus tones fixed at 1000 Hz and varying in intensity over a fixed range. Braida and Durlach [BD] (1972) used several ranges of stimuli. Each range was discretized into 10 categories with a maximum intensity value of 86 dB SPL. Similarly, Luce, Green and Weber [LGW] (1976) tested ranges discretized into 10 categories; however, their ranges

were centered at 60 dB SPL spanning a width of up to 45 dB. Norwich, Wong and Sagi [NWS] (1998) tested five ranges; $1-10$ dB HL, $1-30$ dB HL, $1-50$ dB HL, $1-70$ dB HL and $1-90$ dB HL. Each range had a minimum intensity value at the average population threshold (about 8 dB SPL at 1000 Hz). Also, each range was discretized into single decibel categories. For example, the range $1-10$ dB HL was broken down into 10 single decibel categories. Each group of investigators (BD, LGW and NWS) had found that increasing the stimulus range increases $I_t(m, R)$ and that the increase saturates for large ranges.

Braida and Durlach [BD] addressed the range effect using a model based in signal detection theory to analyze their results. A subject's response to a stimulus of intensity I_j was thought to be governed by a normal probability density function with a mean and variance of $\mu(I_j)$ and σ^2 respectively. The number of stimulus categories used would determine the number of overlapping Gaussian curves that would lie along the total "decision axis". A sensitivity index, d', was developed and measured the separation between any two density functions as $d'(I_i; I_j) = \frac{\mu(I_i)-\mu(I_j)}{\sigma}$. The total sensitivity, $\Delta' = d'(I_{\max}; I_{\min})$, would measure the subject's overall resolution or ability to identify categories absolutely. The subject's transmitted information relates to the logarithm of the total sensitivity whereby Δ' was found to increase monotonically and saturate as the stimulus range increased. Essentially, BD proposed that the increase and subsequent saturation in $I_t(10, R)$ with R was governed by the behaviour of σ. They proposed an "internal-noise" model that predicted two components for σ. One component of 'memory'-noise that was assumed proportional to the stimulus range and another component of 'sensation'-noise that was assumed constant. These two noise components combine to give $\sigma^2 = \alpha^2 R^2 + \beta^2$. Notice for a fixed range, the variance is assumed constant for the set of overlapping gaussians.

In the experiments of Norwich, Wong and Sagi (1998) [NWS], estimates of $I_t(m = R$ dB$, R)$ were obtained using the matrix simulation described in Wong and Norwich (1997). As with BD, the probability distribution governing a subject's response Y conditioned

upon the stimulus X_j, i.e. $p(Y|X_j)$, was assumed to be Gaussian with constant variance for all values of X_j. The constant value for variance, σ, was estimated directly from the matrix, but in the form of an arithmetic mean of the row variances as opposed to a geometric mean. The assumption of normality underlying the $p(Y|X_j)$'s along with the assumption of constant variance allows one to express the transmitted information as follows:

$$\begin{aligned} I_t(R) &= \ln R - \frac{1}{2}\ln\left(2\pi e\sigma^2\right) \\ &= \ln\left(\frac{R}{\sigma}\right) - \frac{1}{2}\ln\left(2\pi e\right). \end{aligned}$$

Notice that $I_t(R)$ would grow without bound if σ did not increase with increasing R. In the language of BD, this expression would become:

$$I_t(R) = \ln\left(\Delta'\right) - \frac{1}{2}\ln(2\pi e).$$

Hence, NWS are in agreement with BD in that the increase and subsequent saturation in I_t with R can be understood in terms of the growth of σ with R, or $\sigma(R)$; however, no assumption of noise infrastructure was required or made.

2.6 Edge Effects in the Stimulus/Response Matrix

The common assumption underlying the models of both BD and NWS is that the error in a subject's response to a given stimulus is governed by a normal distribution with a variance, σ^2, that is assumed fixed in accordance with the stimulus range used for experimentation. Upon closer inspection, one finds that performance in absolute identification is significantly better for stimulus intensities located more to the extremes of the stimulus range (Garner and Hake [1951]). That is, the loudest and softest categories were easiest to identify. This phenomenon has been described as an "anchoring" or "edge" effect.

Braida et. al. (1984) have proposed a revision of their model to incorporate perceptual

anchors. They postulated that a subject identifies the intensity of a stimulus by mapping it onto some theoretical context that represented the range of stimulus intensity. Coding the sensation relative to the context was described as a psychological process in which subjects would measure the distance from the sensation to the internal references or perceptual anchors. "Distance" was thought to be measured in discrete noisy steps, whereby the noise directly contributed to the subject's response variance, σ.

The existence of intrinsic anchors imposed onto the subject by the edges of the stimulus range is substantiated by the work of Berliner et al. (1978). In large range identification experiments, tones of standard intensity were introduced to see if resolution would improve. Resolution did not improve when the standards were located near the extreme intensities of the stimulus range, but did improve for standards corresponding to the mid-range intensities. If we are to assume that standards tend to improve resolution, then the lack of improvement for standards located at the extreme intensities suggests that the standards already existed intrinsically.

Two fundamental drawbacks were found with the perceptual anchor model of Braida et. al.. First, sensitivity measurements, d', were too difficult to calculate directly and had to be estimated. Second, a closed-form expression for the distribution that predicted a subject's response and that incorporated edge effects was not obtainable (Braida et. al. [1984]).

This thesis extends the work of Wong and Norwich (1997) to incorporate edge effects and introduces a model that provides a closed-form expression for a subject's response Y_k conditioned upon a stimulus X_j, i.e. $p(Y_k|X_j)$. This model has been named the "Constant-Variance (CV) model" and is outlined explicitly in Section 3.2 below.

Chapter 3

Methods

3.1 Absolute Identification Experiments Conducted in Our Laboratory

Some of the experiments to be discussed have already been published (Norwich, Wong and Sagi (1998)). Absolute identification experiments were conducted on loudness. Stimuli were in the form of pure tones of varying intensity given at 1000 Hz for a duration of 1.5 s each. All intensity measurements were made in hearing level (HL), which are decibels (dB) above threshold. HL is taken with respect to a standard population threshold and not with respect to the threshold of each participant. The threshold of most participants was actually below 1 dB HL, while that of one subject was about 10 dB HL.

Several ranges were tested on each participant. Primarily, these were $1-10$ dB, $1-30$ dB, $1-50$ dB, $1-70$ dB and $1-90$ dB. There were six participants, "B", "C", "E", "J", "R" and "W", whose ages at the time of experimentation were 52, 19, 22, 18, 19 and 25 years respectively. Not all participants were tested over all the ranges mentioned. Specifically, since the threshold of subject "B" was around 10 dB, the range used for testing began at 11 dB. Each participant completed a given range before progressing to the next range.

In each experimental session, stimulus tones of varying intensity were selected at random from R. Upon the presentation of a stimulus, a subject was required to estimate the loudness to the nearest dB. Each range, R, was discretized into m categories of integral decibels such that $m = R$ dB. For example, category 76 corresponded to a stimulus tone of 76 dB. The random presentation of stimuli was governed by a data set generated by computer, prior to experimentation. The data set consisted of positive, pseudorandom integers uniformly distributed over a fixed range.

For all but one subject, 500 trials were given over a period of three separate days for each range (approximately 480 or 160 for participant "B"). A trial constitutes the pairing of a stimulus tone with its response category. An inter-trial interval was set at 20 s, rather than the 7 s used by Garner (1953), in order to minimize adaptation effects. In each experimental session, testing did not extend beyond 200 trials. Stimulus-response pairs were compiled into the corresponding m x m matrix and $I(N, m, R)$ calculated using Equation 2.3.

A Subject was allowed to study and practise within a designated range for as long as he/she desired. During practice sessions, feedback was given after each stimulus tone. The participant would learn to match tones in the specified range into corresponding categories of integral decibel value, as well as he/she was able. Typically, participants felt that learning was non-contributory after around twenty minutes. During the course of the experiment, no feedback was permitted following each unknown stimulus.

Testing took place within a sound attenuated room in the bioacoustics laboratory in the Institute of Biomedical Engineering at the University of Toronto. In most cases, the experiment was conducted by a trained experimenter who was present in the room with the subject. Stimulus tones of 1000 Hz were generated by an OB70, two-channel audiometer, produced by Madsen Electronics, Toronto, Canada. Tones were delivered binaurally using TDH-39 headphones. Each tone was 1.5 s in duration with 50 millisecond ON/OFF ramps.

Attempts were made to automate the experiment, obviating the need for the exper-

imenter. Automation was developed using a LABVIEW software protocol. Each trial consisted of a 20 second loop whereby a uniformly distributed pseudorandom number in R was generated followed by a stimulus tone of corresponding intensity. Stimulus tones of 1000 Hz, 1.5 second duration with 50 millisecond ON/OFF ramps were generated using a TAHITI sound card. Within a trial loop, subjects were given 20 seconds to input a response corresponding to the stimulus intensity using the computer keyboard. Timing and response input was displayed graphically on the monitor. Stimulus-response pairs were stored in a file that was inaccessible to the subject. The subject could terminate the session at will. A similar protocol was used for practice sessions that included feedback. The automated procedure was used on the following subjects and ranges: subjects "C" and "R" were both tested at $1 - 10$ dB and $1 - 30$ dB; subject "E" was tested at $1 - 10$ dB, $1 - 30$ dB and $1 - 50$ dB. For these subjects, the experiment over the range $1 - 90$ dB was conducted with an experimenter present in the manner described above.

3.2 A Constant Variance (CV) Model for the Confusion Matrix

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	X_j^{total}
X_1	20	7	12	9	4	1	0	0	0	0	53
X_2	10	10	15	10	10	1	1	0	0	0	57
X_3	14	11	7	10	6	1	2	1	0	0	52
X_4	1	5	14	16	9	2	4	0	0	0	51
X_5	4	1	4	14	11	8	2	1	0	0	45
X_6	0	0	5	8	14	12	10	6	0	1	56
X_7	0	0	0	2	8	7	8	14	7	1	47
X_8	0	0	0	1	2	7	17	9	9	4	49
X_9	0	0	0	0	3	4	11	18	11	4	51
X_{10}	0	0	0	1	1	0	2	4	18	13	39
Y_k^{total}	49	34	57	71	68	43	57	53	45	23	500

Shown above is one of the 10 x 10 confusion matrices measured from absolute identification experiments conducted in our laboratory over the fixed range $1 - 10$ dB HL. Notice how the middle rows tend towards a Normal distribution while the extreme rows display skewed distributions with data that accumulates toward the edges. Suppose that responses along row vectors to the middle of the Confusion matrix tend to a Normal distribution of **constant variance**. This concept dates back to Thurston as a special case of his "Law of Categorical Judgements" (Thurston (1927), Durlach and Braida (1969)).

As a form of checking, consider only rows 4 through 7 in the matrix above. Furthermore, let us combine rows 4 through 7 as if they all represent the same distribution with means shifted. Before combining the rows, the means need to be aligned. If rows 5 and 6 are considered to already be in alignment, it should be sufficient to shift row 4 with row 5 and row 7 with row 6. Shown in Figure 3.1 is a frequency plot of rows 4 through

7 superimposed into one row of 10 columns. Each point represents the frequency value of its corresponding column. The smooth curve is the Normal frequency distribution that corresponds to a variance value measured from the data. This demonstration is mainly for illustrative purposes, the goal being to express to first order how response data tend to follow a normal distribution of constant variance; however, one is still lead to inquire about the extreme rows. *More generally, how is it possible to account for all row distributions of the confusion matrix using a constant value for a subject's response error?*

Please recall that the simulator described in Wong and Norwich (1997) also uses a constant value for σ. This value is used as an input and enables simulation of a confusion matrix in its entirety including the skewed distributions toward the edges. The simulator calls up two random numbers: a uniform integer x, $x \in \{1, 2, ..., R\}$, and a normally distributed integer y, $y \in \{1, 2, ..., R\}$, centered about x with constant variance σ^2. The resulting simulated confusion matrix appears quite similar to the one above. That is, even though a constant value for variance is used in simulation, the resulting confusion matrix will also have middle rows that tend to a common normal distribution with skewed distributions appearing in the extreme rows.

With this in mind, let σ^2 represent the constant variance used as an input for simulation. Let σ^2_{eff} represent the arithmetic average of the variances across all rows including the extreme distributions. If σ is sufficient to account for all row distributions of the confusion matrix, it should be possible to relate σ^2_{eff} to σ^2. The only sample statistic of subject error available to the experimenter is found along the rows of the confusion matrix. This estimate is strongest when all trials are incorporated, so an arithmetic mean of all row variances could be used as an estimator for subject error.

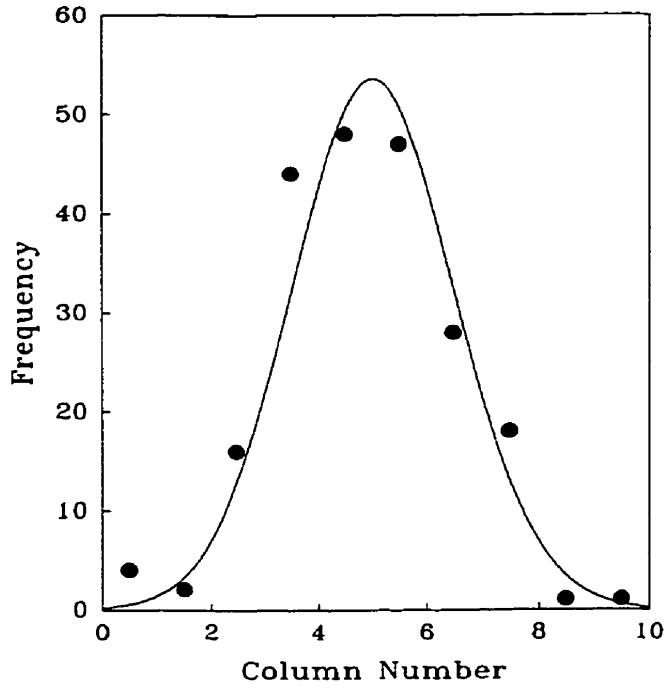

Figure 3-1: Frequency plot of rows 4 through 7 superimposed into one row of 10 columns. Data taken from 10 x 10 confusion matrix measured for Subject W. The frequency of superimposed data tends to a Normal distribution.

3.2.1 Modeling the row distributions; p(y|x)

Finding a relationship between σ^2_{eff} and σ requires a model of the row distributions. For ease in calculation, the set of discrete responses, Y, will be represented by the continuous random variable, y, and the set of discrete stimuli, X, will be represented by the continuous random variable, x. One should note the necessary correction of the form $x = X_j - \frac{1}{2}$ when transforming from the continuous domain to the discrete domain.

We start by representing distributions along any row of the confusion matrix as a conditional probability of response y given a mean value of x. This distribution, $p^*(y|x)$ is a continuous normal distribution with variance σ^2.

$$p^*(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] \tag{3.1}$$

$p^*(y|x)$ is, however, unbounded in y such that $y \in (-\infty, \infty)$. To conform to the boundaries of the matrix more realistically, one can confine $p^*(y|x)$ to the width of the matrix; namely, define y such that $y \in [0, R]$ where y is a continuous random variable. If we now integrate over the space of y,

$$\begin{aligned}
\int_0^R p^*(y|x)dy &= \int_0^R \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] dy \\
&= \frac{1}{2}\left[\operatorname{erf}\left(\frac{R-x}{\sigma\sqrt{2}}\right) + \operatorname{erf}\left(\frac{x}{\sigma\sqrt{2}}\right)\right] \\
&= I(x)
\end{aligned}$$

dividing $p^*(y|x)$ by $I(x)$ renormalizes 3.1 over the range, giving an expression, $p(y|x)$, for the row distributions as follows:

$$\begin{aligned}
p(y|x) &= \frac{1}{I(x)\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] \\
x, y &\in [0, R]
\end{aligned} \tag{3.2}$$

$$I(x) = \frac{1}{2}\left[\text{erf}\left(\frac{R-x}{\sigma\sqrt{2}}\right) + \text{erf}\left(\frac{x}{\sigma\sqrt{2}}\right)\right]$$

This procedure is outlined graphically in Figure 3.2. A continuous normal distribution whose mean lies at the centre of the second column is renormalized over the Range. The renormalized distribution, $p(y|x)$, is now sufficient to describe the distribution of responses along the rows of a Confusion matrix as seen in Figure 3.3. Notice how the distributions skew as the mean approaches the edges.

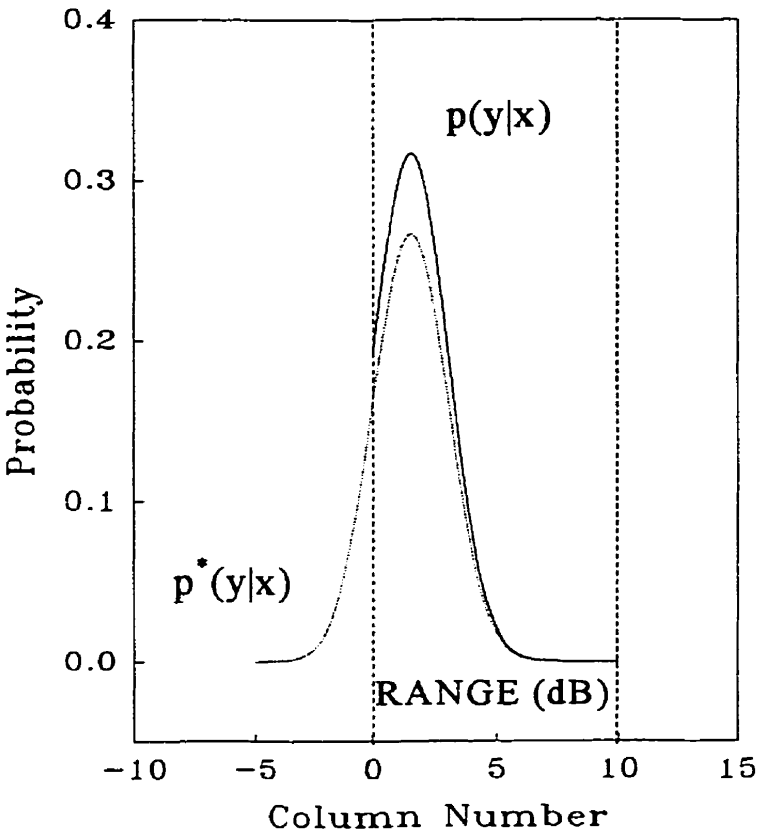

Figure 3-2: A model for the probability distribution, $p(y|x)$, of the second row of a 10 x 10 Confusion matrix over $R = 10$ dB. $p(y|x)$ is obtained by placing the mean of the underlying normal distribution, $p^*(y|x)$, at the centre of the second column and renormalizing over all column numbers y in $[0, R]$.

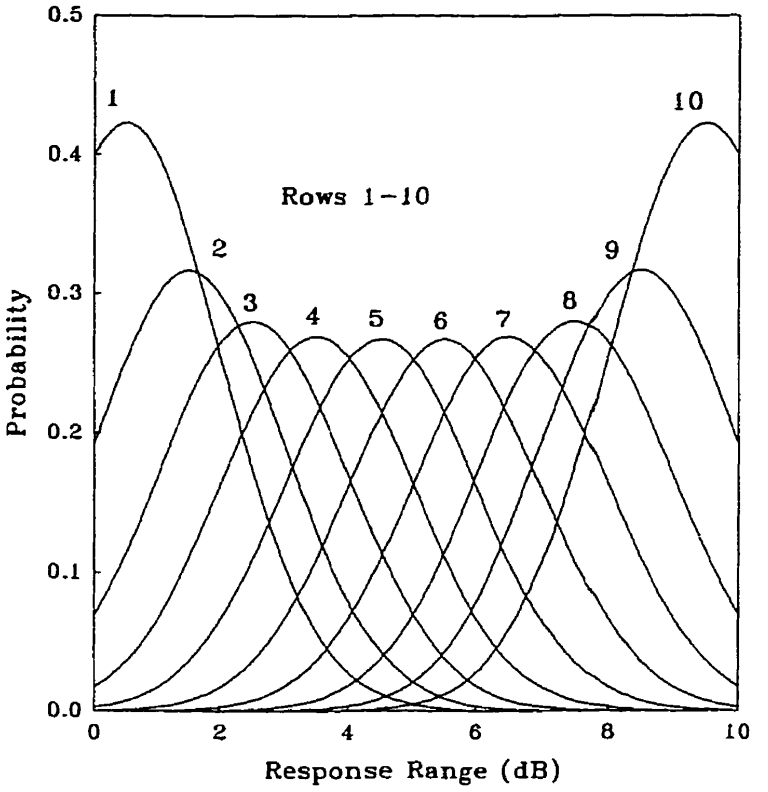

Figure 3-3: The probability distributions, $p(y|x)$'s, found along rows 1 through 10 in a typical 10 x 10 confusion matrix over $R = 10$ dB. $p(y|x)$ is obtained by placing the mean of the underlying distribution, $p^*(y|x)$, at the centre of the appropriate column and then renormalizing over $y \in [0, 10]$. $p^*(y|x)$ is the normal distribution of constant variance σ^2 thought to underlie a subject's responses.

3.2.2 σ^2 as a Function of σ^2_{eff}

As demonstrated in Figure 3.2 and Figure 3.3, bounding $p^*(y|x)$ over the range of stimuli has two effects on the appearance of the row distributions. First, all distributions have been "squeezed" as a result of having to be normalized over a smaller region. Second, skewing occurs in distributions whose mean approaches the edges. These distributions can't cross the boundary and will therefore become asymmetrical. These two effects directly affect the variance measured in each row. That is, a "squeezed" distribution will exhibit a smaller variance than an "unsqueezed" distribution. Also, a "squeezed" and skewed distribution will exhibit a smaller variance than a "squeezed" and unskewed distribution. Since σ^2_{eff} is measured as an arithmetic mean of the row variances, one would expect that it would be smaller than σ^2. The power of modeling the row distributions using $p(y|x)$ is that one can also model how σ^2 relates to σ^2_{eff}. This is very useful if one wants to essentially unwrap the matrix and yield a value for σ that can be introduced into the simulator to "recreate" the matrix.

To model how σ^2 varies with σ^2_{eff}, we need to calculate the first two moments of $p(y|x)$; namely $\langle y \rangle$ and $\langle y^2 \rangle$ (please refer to appendix I).

$$\begin{aligned} \langle y \rangle &= \int_0^R y p(y|x) dy \\ &= \int_0^R \frac{y}{I(x)\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] dy \end{aligned}$$

at this point it is necessary to implement the substitution $t = \frac{y-x}{\sqrt{2}\sigma}$ or $y = \sqrt{2}\sigma t + x$. This will greatly simplify calculations as follows:

$$\begin{aligned} var[p(y|x)] &= \left\langle y^2 \right\rangle - \langle y \rangle^2 \\ &= 2\sigma^2(\left\langle t^2 \right\rangle - \langle t \rangle^2) \end{aligned}$$

making the necessary substitutions:

$$\begin{aligned}\langle t\rangle &= \int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t\exp(-t^2)}{I(x)\sqrt{\pi}}dt \\ &= \frac{\sigma}{I(x)\sqrt{2}}\frac{\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right)-\exp\left(-\frac{1}{2}\left(\frac{R-x}{\sigma}\right)^2\right)}{\sqrt{2\pi\sigma^2}}\end{aligned}$$

this can be conveniently expressed in a shorthand using the normalizing factor $I(x)$:

$$\begin{aligned}\langle t\rangle &= \frac{\sigma}{I(x)\sqrt{2}}\frac{dI(x)}{dx} \\ I(x) &= \frac{1}{2}\left(\mathrm{erf}(\frac{R-x}{\sqrt{2}\sigma})+\mathrm{erf}(\frac{x}{\sqrt{2}\sigma})\right)\end{aligned}$$

one can take a similar approach for the second moment:

$$\begin{aligned}\left\langle t^2\right\rangle &= \int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t^2\exp(-t^2)}{I(x)\sqrt{\pi}}dt \\ &= \frac{1}{I(x)2\sqrt{\pi}}\left(-t\exp(-t^2)|_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}}\right)+\frac{1}{2}\end{aligned}$$

using the same shorthand above:

$$\left\langle t^2\right\rangle = \frac{\sigma^2}{I(x)2}\frac{d^2I(x)}{dx^2}+\frac{1}{2}$$

we now can express the variance of any row distribution as a function of its mean; namely:

$$\begin{aligned}var[p(y|x)] &= 2\sigma^2\left(\left\langle t^2\right\rangle-\langle t\rangle^2\right) \\ &= \frac{\sigma^4}{I(x)}\frac{d^2I(x)}{dx^2}+\sigma^2-\frac{\sigma^4}{(I(x))^2}\left(\frac{dI(x)}{dx}\right)^2\end{aligned}$$

One should note that

$$\left(\frac{I'(x)}{I(x)}\right)' = \frac{I''(x)}{I(x)}-\frac{1}{(I(x))^2}(I'(x))^2$$

so the variance of any row distribution becomes:

$$var[p(y|x)] = \sigma^4 \left(\frac{I'(x)}{I(x)} \right)' + \sigma^2$$

This is still a complicated expression and difficult to work with. However, if we measure the average row variance over the mean, we can obtain an expression for σ^2_{eff}. In the discrete case, taking an arithmetic mean for the row variance simply involves adding the variances of rows 1 to R and dividing by the range R. Deriving a closed expression for this using $var[p(y|x)]$ above would be quite difficult. It is possible, however, to consider jumping from the discrete case to the continuous case over the mean x. Specifically, σ^2_{eff} in the continuous case takes the form:

$$\begin{aligned} \sigma^2_{eff} &= \frac{1}{R} \int_0^R var[p(y|x)] dx \\ &= \frac{1}{R} \int_0^R \left[\sigma^4 \left(\frac{I'(x)}{I(x)} \right)' + \sigma^2 \right] dx \end{aligned}$$

It is easily verified that

$$\begin{aligned} \left(\frac{I'(x)}{I(x)} \right)_0^R &= \frac{\sqrt{2}}{\sigma} \left(\frac{\frac{2}{\sqrt{\pi}} \exp(-\frac{R}{\sqrt{2}\sigma})^2 - \frac{2}{\sqrt{\pi}}}{\text{erf}(\frac{R}{\sqrt{2}\sigma})} \right) \\ &\simeq -\frac{4}{\sqrt{2\pi\sigma^2}} \text{ for } R \gg \sigma \end{aligned}$$

so to a strong approximation,

$$\sigma^2_{eff} = \sigma^2 - \frac{4}{R\sqrt{2\pi}} \sigma^3 \tag{3.3}$$

3.2.3 The CV Model Underlies Matrix Simulation

The simulator produces a Confusion Matrix using a single input value, σ. The CV Model attempts to account for how the distributions resulting along the rows of this Confusion matrix arise from a common underlying Normal distribution of constant variance σ.

By design, the simulator generates normally distributed responses that are confined to the stimulus range. The CV Model describes what happens to the underlying Normal distribution when it is confined to the stimulus range. As data are arranged into the simulated matrix on a trial by trial basis, the probability distributions found along the rows of the matrix, should conform to the distributions predicted by the CV model.

Figure 3.4 demonstrates how distributions from a simulated matrix approach those distributions predicted by the CV model. In particular, probabilities calculated from rows 4 and 9 are compared with the model after 100 trials and after 10,000 trials of simulation. All simulations use the same value for σ. At 100 trials, the simulator deviates from the CV model, but by 10,000 trials it converges very closely. It seems clear that the CV model accurately describes the probability distributions along the rows of the confusion matrix resulting from simulation with large N.

The correspondence between the model and the simulator comes as no surprise since both are constructed using a single normal distribution of constant variance confined to the stimulus range. How well the CV model, and therefore simulation, corresponds to experimentally measured confusion matrices is another matter. This depends on the degree to which the constant variance assumption describes the participant's responses. The CV model does, however, provide a methodology for obtaining an estimate of the hypothesized constant variance from confusion matrices that are obtained experimentally. Subsequently, this estimate can be used to simulate the matrix and observe how closely the resulting measure of information compares with that of the experiment.

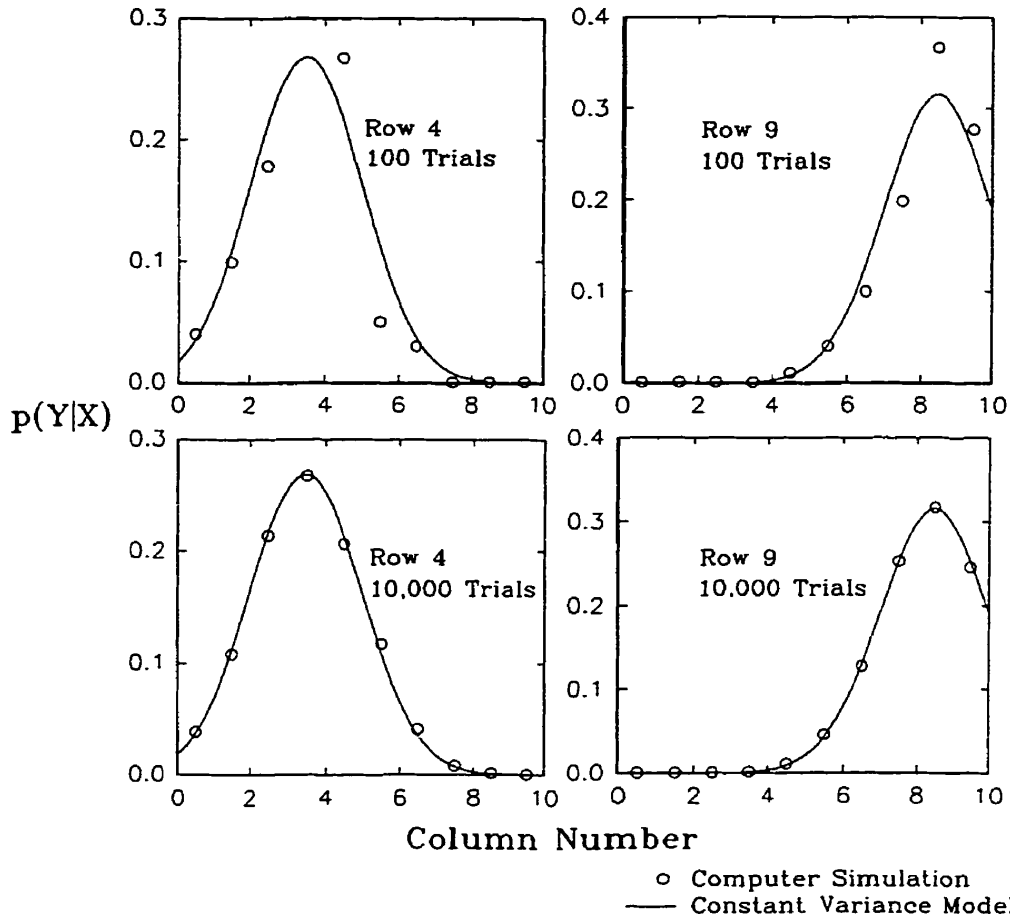

Figure 3-4: Comparison between the row distributions predicted by the CV-model and the row distributions calculated from the computer simulation described in Wong and Norwich (1997). Distributions are plotted for rows 4 and 9 after 100 and 10,000 trials of simulation.

3.3 Application of CV Model to Experiment

How Equation 3.3 can be used to simulate the matrix presented at the beginning of this chapter will now be explained. Measuring the average row variance from the matrix gives $s^2_{eff} = 2.21$. The notation s^2_{eff} is used to express our measure of subject error as the corresponding sample estimate of σ^2_{eff}. Similarly, s^2 will be used to estimate σ^2. Substituting $\left(s^2_{eff}, s^2\right)$ for $\left(\sigma^2_{eff}, \sigma^2\right)$ respectively into Equation 3.3 and solving for s gives $s^2 = 3.07$. This value is now used as an input into the simulator. In Figure 3.5 are graphs that compare simulated Information with experimentally measured Information as a function of the number of trials. Twenty averages were used for each simulation. Also, to demonstrate the practicality of the CV model, one simulation uses an input value of $s^2_{eff} = 2.21 = s^2$ while the other uses $s^2 = 3.07$. The experimental curve uses the trial by trial data corresponding to the matrix above. Notice how the simulator that implements Equation 3.3 to obtain $s^2 = 3.07$ as an input conforms more closely to the experimental curve, thus improving upon the method of overcoming small sample bias developed by Wong and Norwich (1997).

The procedure in our example can be generalized. For any experimentally determined confusion matrix, measure the average row sample variance, s^2_{eff}. "Unwrap" this measure using Equation 3.3 to determine the estimated response error, s^2, underlying the matrix. Simulate using s to obtain an estimate for Information, $I(N, m, R)$, to any desired number of trials.

One should note that averaging simulations together gives a stronger estimate of the expected value of Information $\langle I(N) \rangle$ corresponding to the underlying response error, σ. $I(N, m, R)$ resulting from a single simulation can be considered as a single experiment. The corresponding expectation, $\langle I(N) \rangle$ would represent a long term average over many experiments and has a one-to-one correspondence with σ (Carlton, 1969) [please see

see Appendix II]. The characteristic shape of $\langle I(N) \rangle$ is discussed in Section 4.1.

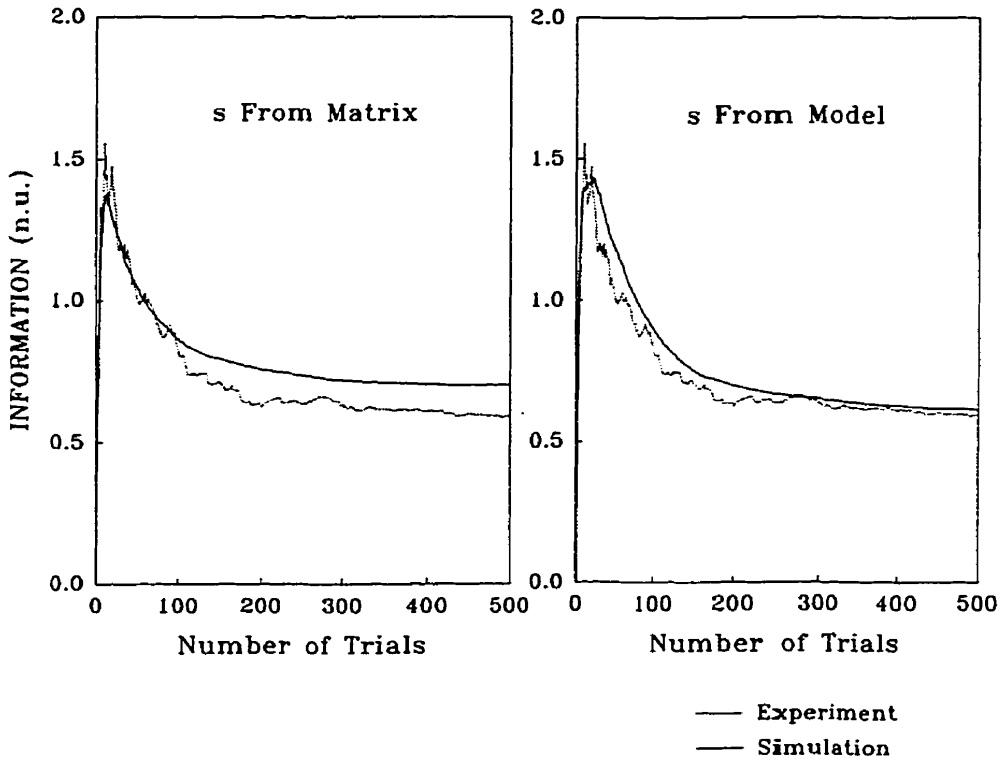

Figure 3-5: Comparison between information measured experimentally (from Subject W, $1-10$ dB) and information measured through computer simulation, both as a function of the number of trials. s^2_{eff} is the average row variance measured directly from the matrix. s^2 is the variance of the normal distribution thought to underly subject's responses and is estimated using the CV-model.

3.4 Bypassing Simulation with Approximation for Asymptotic Information

As mentioned earlier, the original purpose for simulation was to overcome the small sample bias in calculated information, $I(N)$ (short for $I(N, m, R)$), that occurs when not enough experimental trials are used. $I(N)$ becomes transmitted information, I_t, only after a significant number of trials. One of the main advantages of the CV model is that it provides a trial independent, *a priori* description of the row distributions found in the confusion matrix; namely, $p(y|x)$. This can be substituted into the equation that describes I_t.

$$\begin{aligned} I_t &= I(Y|X) = H(Y) - H(Y|X) \\ &= -\sum_{k=1}^{R} p(y_k) \log p(y_k) + \sum_{j=1}^{R} \sum_{k=1}^{R} p(x_j) p(y_k|x_j) \log p(y_k|x_j) \end{aligned} \tag{3.4}$$

This equation is independent of N and describes I_t exactly, but is too difficult to handle analytically. We can, however make some approximations. First, let us move over from the discrete realm to the continuous whereby sums are converted into integral signs. This is analogous to considering infinitesimally small category widths.

Notice that the sums in our expression span the integers $\{1, 2, ..., R\}$. In this case, the number of categories, m, used to measure information is a discrete representation of the stimulus range. We are not confined to this situation. In fact, increasing the number of categories over a fixed range tends to increase I_t. This increase, however, is bounded. That is, for a large enough number of categories, I_t is unchanging. Please recall that

$$m \to \infty \Longrightarrow I_t(m, R) \to I_t(R).$$

At this point, it will be taken as a conjecture that going from $m = R$ dB to $m \rightarrow \infty$ in 3.4 has little or no effect on I_t. That is, we shall assume that $I_t(m = R \text{ dB}, R) \simeq I_t(R)$. The issue of varying the available number of categories over a fixed range will be addressed more rigorously in section 4.2. Hence,

$$\begin{aligned} I_t &= H(Y) - H(Y|X) \\ &= -\int_0^R p(y) \log p(y) dy + \int_0^R \int_0^R p(x)p(y|x) \log p(y|x) dy dx \end{aligned}$$

Symmetry in a confusion matrix would imply that $p(y) = p(x) = \frac{1}{R}$:

$$I_t = \log R + \frac{1}{R} \int_0^R \int_0^R p(y|x) \log p(y|x) dy dx$$

Using the expression for the row distributions derived earlier in the CV model, $p(y|x) = \frac{e^{-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2}}{I(x)\sqrt{2\pi\sigma^2}}$, and expressing the transmitted information in natural units, we have (please refer to Appendix I)

$$I_t = \ln R - \frac{1}{2} \ln(2\pi e\sigma^2) - \frac{1}{R} \int_0^R \left(\ln I(x) + \sigma^2 \frac{I''(x)}{2I(x)} \right) dx$$

or,

$$I_t = \ln R - \frac{1}{2} \ln(2\pi e\sigma^2) + \Phi(R, \sigma) \tag{3.5}$$

The right most term, $\Phi(R, \sigma) \geq 0$ must be solved numerically, but vanishes for large R. One can consider $\Phi(R, \sigma)$ as an information gain due to the presence of edges. Notice how no assumptions were necessary regarding mechanisms for anchoring (please see section 2.6).

Shown below is a table depicting how well the asymptotic approximation, I_t^{as}, compares with the simulator. For estimates of I_t using the simulator, i.e. I_t^{sim}, 100,000 trials are used to overcome small sample bias. Also, 20 simulations are averaged for each estimate to ensure the simulator approximates $\langle I(N) \rangle$. The values used for s^2 are "unwrapped" from the s_{eff}^2's measured experimentally from one subject over several ranges.

Finally, to illustrate the contribution of edge effects to estimates of I_t, the asymptotic formula is also depicted without the contribution of the $\Phi(R, \sigma)$ term. This will fall under the heading of $I_t^* = \ln R - \frac{1}{2}\ln(2\pi e\sigma^2)$.

Range (dB)	s_{eff}	s	I_t^*	I_t^{as}	I_t^{sim}
$1-10$	1.49	1.75	0.325	0.563	0.557
$1-30$	2.50	2.71	0.987	1.107	1.110
$1-50$	3.72	3.99	1.110	1.215	1.224
$1-70$	4.41	4.66	1.290	1.376	1.391
$1-90$	5.58	5.90	1.306	1.391	1.427

Please notice that Equation 3.5 is sufficient to describe the results of averaged simulations over large trials. Hereafter, all estimates of I_t will use Equation 3.5.

3.4.1 Confidence Interval for estimate of I_t

One of the advantages of the assumption that a Normal distribution with a constant variance underlies responses for a given range is that sample estimates for variance, s^2, will follow a Chi-squared distribution. Specifically,

> "If $\bar{x}$ and s^2 are the mean and the variance of a random sample of size N from a normal population with the mean μ and the variance σ^2, then ... the random variable $\frac{(N-1)s^2}{\sigma^2}$ has a chi-square distribution with $N-1$ degrees of freedom (Freund and Walpole (1980), page 268)."

Please recall that the CV model uses s_{eff}^2 from the matrix to obtain an estimate for response error. Subsequently, s^2 is determined using Equation 3.3. Essentially, the CV model allows us to treat all the rows of the confusion matrix as originating from one distribution. Hence, all N trials used in measuring s_{eff}^2 can be used to measure s^2 and chi-square confidence intervals can be applied to σ^2 with $N-1$ degrees of freedom. For degrees of freedom in excess of about 30, we may use a normal approximation to the

chi-square distribution (Milton, 1992, page 223) where α represents our confidence value, υ the degrees of freedom and z the standard-normal score.

$$\chi^2_{\upsilon,\alpha} \simeq \frac{1}{2}\left[z_\alpha + \sqrt{2\upsilon - 1}\right]^2$$

Hence, the 90% Chi-square confidence interval for 499 degrees of freedom would be:

$$P\left(\chi^2_{499,.05} \leq \frac{499s^2}{\sigma^2} \leq \chi^2_{499,.95}\right) = 0.90$$

where

$$\chi^2_{499,.05} \simeq \frac{1}{2}\left[-1.65 + \sqrt{997}\right]^2 = 447.76$$

and

$$\chi^2_{499,.95} \simeq \frac{1}{2}\left[1.65 + \sqrt{997}\right]^2 = 551.96$$

Therefore, the confidence interval becomes:

$$P\left(447.76 \leq \frac{499s^2}{\sigma^2} \leq 551.96\right) = 0.90$$

Solving out the inequality for σ gives:

$$P(0.951s \leq \sigma \leq 1.056s) = 0.90$$

We can interpret this as saying that 90% of the time, σ will be found in the interval $[0.951s, 1.056s]$. This only holds true if the assumption of underlying normality is correct.

We would now like to know the amount of error that can be expected in estimating the transmitted information, I_t. Estimates of I_t are completely dependant on estimates of σ. The most amount of error will be found in the limiting case where the stimulus range is large. This can be understood from Equation 3.5. The only two terms that depend on σ subtract from each other; hence, any error incurred in the use of this Equation will also subtract. However, the term corresponding to the added information due to edge

effects, $\Phi(R, \sigma)$, diminishes as the range increases. For very large ranges, Equation 3.5 becomes

$$I_t = \ln R - \frac{1}{2} \ln \left(2\pi e \sigma^2 \right)$$

and the right hand term dominates errors incurred from estimates of σ. Substituting our 90% confidence interval gives

$$I_t = \ln R - \frac{1}{2} \ln \left(2\pi e s^2 \right) \pm 0.05 \text{ n.u.}$$

a maximum error in I_t of about ± 0.05 natural units of information.

Chapter 4

Results and Analysis

Typically, absolute identification experiments are conducted over some fixed range R that has been discretized into m stimulus/response categories. Information is calculated from the resulting m x m confusion matrix, but is sensitive to the number of trials, N, used for experimentation. Let us represent the information measure symbolically as $I(N, m, R)$.

Three kinds of "asymptotes" arise with the information measure. First, the information measure has a characteristic behaviour associated with the number of stimulus trials, N in the form of a small sample bias. Given enough trials, the information measure approaches the information transmitted to the subject for that absolute identification experiment; i.e. $I(N, m, R) \rightarrow I_t(m, R)$. Next, there is the increase of information with the number of stimulus/response categories, m. The information measure artifactually increases with increasing m, approaching an upper limit. Assuming that small sample bias has been overcome, reaching the upper categorical limit can be represented as $I_t(m, R) \rightarrow I_t(R)$. This is the situation where the resolving power of $I_t(m, R)$ reflects the informational properties of the range, $I_t(R)$. Finally, one finds that $I_t(R)$ increases with increasing R up to a "channel capacity", I_∞; the upper limit for information processing over the entire range of auditory intensity. In contrast to $I_t(m, R) \rightarrow I_t(R)$, this phenomenon, $I_t(R) \rightarrow I_\infty$, depends on the sensory properties of the individual.

As described in Chapter 3, a confusion matrix can be "recreated" using a single

constant value for σ over that range; i.e. $\sigma(R)$. It is hypothesized that $\sigma(R)$ determines the results of all AI experiments conducted over R regardless of the number of categories used. From experiments conducted in our laboratory, it may be found that $\sigma(R)$ increases with R. The increase is well described by a linear relationship $\sigma(R) = aR + b$. This relationship essentially determines $I_t(N, m, R)$!

4.1 Information and Number of Experimental Trials; $I(N)$

In Chapter 2 above, the bias in information estimates that results from using too few experimental trials was discussed. Shown in Figure 4.1 is a graph of the information measured as a function of the number of experimental trials over five ranges for subject W. One can see that $I(N, m, R)$ is characterized by an early rise to a peak followed by a decline towards an asymptote. The asymptote is the information transmitted to the subject for that absolute identification experiment. That is, $I(N, m, R) \rightarrow I_t(m, R)$ as $N \rightarrow \infty$. Hence, the bias takes the form of an overestimation of $I_t(m, R)$.

The characteristic shape of the bias can be understood using Carlton's approximation for the information measure, $\langle I(N) \rangle$ (please refer to Appendix II). We recall that,

$$I(N, m, R) = H(Y) - H(Y|X).$$

and therefore,

$$\langle I(N) \rangle = \langle H(Y) \rangle - \langle H(Y|X) \rangle .$$

In Figure 4.2, $\langle I(N) \rangle$, $\langle H(Y) \rangle$, and $\langle H(Y|X) \rangle$ are plotted with the number of trials using the value for subject error measured for subject W in the $1 - 10$ dB experiment. Notice that $\langle H(Y) \rangle$ and $\langle H(Y|X) \rangle$ increase monotonically with N, approaching separate asymptotes; however, $\langle H(Y) \rangle$ saturates more quickly than $\langle H(Y|X) \rangle$. Upon subtracting the two terms, the quicker rise in $\langle H(Y) \rangle$ causes the overshoot in $\langle I(N) \rangle$. As $\langle H(Y) \rangle$

saturates, $\langle H(Y|X)\rangle$ still increases, causing $\langle I(N)\rangle$ to fall and approach $I_t(m,R)$.

We can attempt a mechanical explanation for why $\langle H(Y)\rangle$ rises more quickly than $\langle H(Y|X)\rangle$ by looking at how the m x m confusion matrix fills progressively with N. $\langle H(Y)\rangle$ and $\langle H(Y|X)\rangle$ depend explicitly on the response distribution $p(y)$ and the row distribution $p(y|x)$ respectively. With increasing trials, $p(y)$ matures over m cells while $p(y|x)$ matures over m^2 cells. Hence, $p(y)$ matures more quickly than $p(y|x)$. Also, the monotonic increase in $\langle H(Y)\rangle$ and $\langle H(Y|X)\rangle$ reflects the maturation in $p(y)$ and $p(y|x)$ respectively. Therefore, $\langle H(Y)\rangle$ increases more quickly than $\langle H(Y|X)\rangle$.

Upon further inspection of Figure 4.1, increasing the range of experimentation has two effects on the peak value for the bias in $I(N,m,R)$. Essentially, peak bias values increase and occur at later trials for larger ranges. Since, in our experiments, $m = R$ dB, larger ranges required larger m x m confusion matrices. With larger confusion matrices, the characteristic shape of $I(N)$ will prolong over N. Notice, however, that the location of the peak bias tends to equal the number of categories used. Looking at Figure 4.2 we see that the peak bias value in $\langle I(N)\rangle$ occurred after approximately 10 trials. Generally, $\langle I(N)\rangle$ will peak after approximately m trials, but this only occurs for values of σ measured experimentally. That is, the location of the peak bias in the information measure depends on σ and typically occurs after m trials for human subjects.

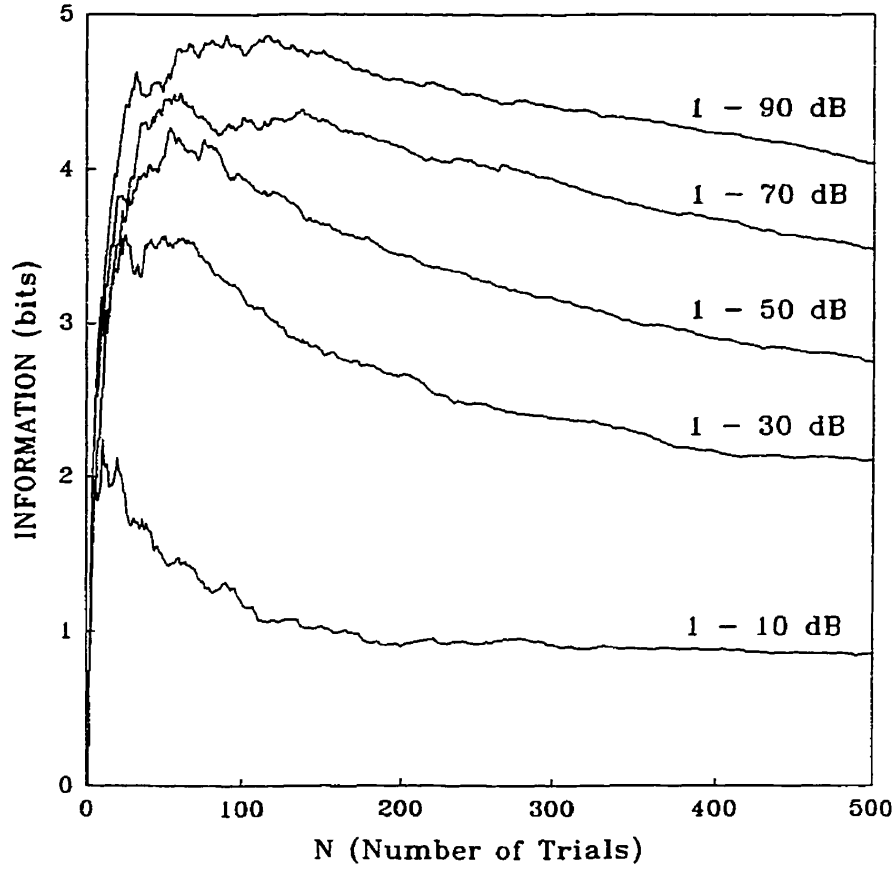

Figure 4-1: Information measured (in bits) as a function of the number of experimental trials over five stimulus ranges for subject W. In each range, the number of categories used equalled the range in dB.

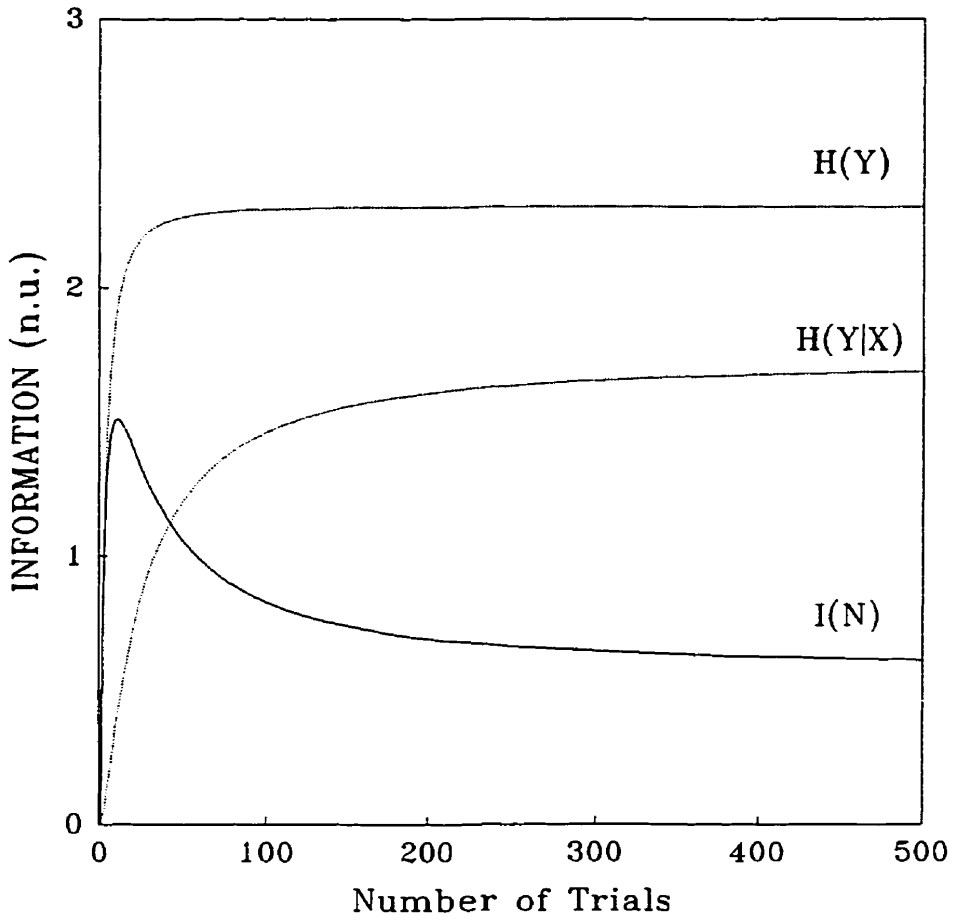

Figure 4-2: Expected value of information (measured in natural units) as a function of the number of experimental trials, $\langle I(N) \rangle$, using Carlton's approximation.

4.2 Information and Number of Categories Over a Fixed Range: $I_t(m, R_0)$

4.2.1 Mathematical Basis for Increase of Information With The Number of Stimulus Categories, $I_t(m, R_0)$

As explained earlier, each confusion matrix represents an AI experiment conducted over some fixed range $R = R_0$ using a specified number of categories, m. It has been known for nearly fifty years that increasing m in this situation tends to increase $I_t(m, R_0)$. Shown in Figure 4.3 is a typical representation of how $I_t(m, R_0)$ tends to increase with increasing m. Notice how the increase is bounded from above. A mathematical argument can explain the increase as an artifact of the information measure and independent of the perceiver. The upper bound is, however, determined by the sensory properties of the subject.

Consider an arbitrary confusion matrix expressed as follows:

	Y_1	...	Y_k	Y_{k+1}	...	$Y_{m'}$	X_j^{total}
X_1	n_{11}	...	n_{1k}	n_{1k+1}	...	$n_{1m'}$	$n_{1\cdot}$
:	:		:	:		:	:
X_j	n_{j1}	...	n_{jk}	n_{jk+1}	...	$n_{jm'}$	$n_{j\cdot}$
:	:		:	:		:	:
X_m	n_{m1}	...	n_{mk}	n_{mk+1}	...	$n_{mm'}$	$n_{m\cdot}$
Y_k^{total}	$n_{\cdot 1}$	...	$n_{\cdot k}$	$n_{\cdot k+1}$	...	$n_{\cdot m'}$	N

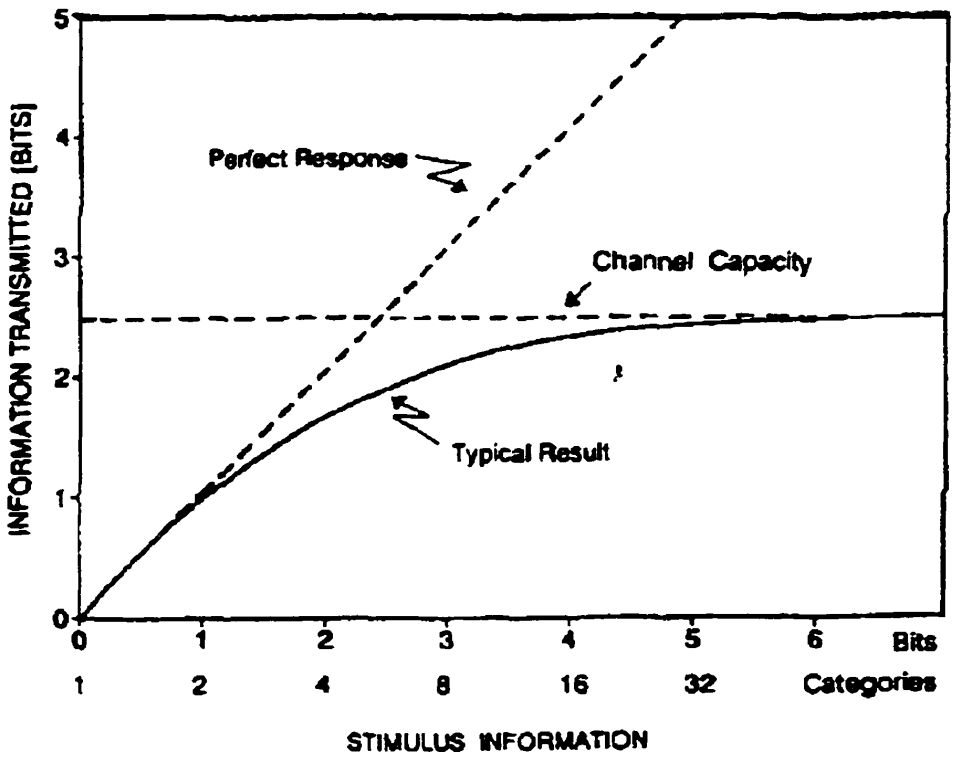

Figure 4-3: A schematic graph of how $I_t(m, R_0)$ increases with a progressively increasing number of categories m over a fixed range R_0. For a small enough number of categories, the information transmitted to the subject resembles noise-free transmission. As the number of categories increases, $I_t(m, R_0)$ saturates towards its channel capacity. One should note that STIMULUS INFORMATION $= \log_2(m)$. From Norwich (1993, pg. 82).

The information measure (using the shorthand $I(N)$) can be expressed in terms of the *a posteriori* values from the matrix.

$$I(N) = \log N + \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{k=1}^{m'} n_{jk}\log n_{jk} - \sum_{j=1}^{m} n_{j\cdot}\log n_{j\cdot} - \sum_{k=1}^{m'} n_{\cdot k}\log n_{\cdot k}\right) \tag{4.1}$$

Consider merging columns k and $k+1$ as follows:

n_{1k}	n_{1k+1}		$n_{1k}+n_{1k+1}$
:	:		:
n_{jk}	n_{jk+1}	$\longrightarrow$	$n_{jk}+n_{jk+1}$
:	:		:
n_{mk}	n_{mk+1}		$n_{mk}+n_{mk+1}$
$n_{\cdot k}$	$n_{\cdot k+1}$		$n_{\cdot k}+n_{\cdot k+1}$

How does this affect the information measured for that matrix. This affect can be described in terms of the information difference, or $\Delta I = I(N)_{unmerged} - I(N)_{merged}$. If we can show that $\Delta I \geq 0$, this would imply that arbitrarily merging any two columns always decreases the information. By the symmetry of $I(N)$, the same should be true when we merge any two rows. Hence, showing that $\Delta I \geq 0$ would imply that using a larger number of categories to represent the same data set gives information values that increase or remain the same.

The proof that $\Delta I \geq 0$ will be described here briefly and is outlined more rigorously in Appendix III. For simplicity, let $a_j = n_{jk}$ and let $b_j = n_{jk+1}$ in the arbitrary matrix outlined above. Making the necessary substitutions into 4.1 we find after some algebra that

$$\Delta I \;=\; I(N)_{unmerged} - I(N)_{merged}$$

$$= \sum_{j=1}^{m} \phi(a_j, b_j) - \phi\left(\sum_{j=1}^{m} a_j, \sum_{j=1}^{m} b_j\right)$$

where

$$\phi(a, b) = a \log a + b \log b - (a + b) \log (a + b).$$

In Appendix III, it is demonstrated that $\phi(a, b)$ is convex implying, from the theory of convex functions, that $\Delta I \geq 0$ (Hardy et. al. (1964)).

So we now see, at least theoretically, that the increase of $I(N, m, R)$ with m (R fixed), is a mathematical artifact of the information measure. Please recall that $I(N, m, R) \rightarrow I_t(m, R)$ as $N \rightarrow \infty$ implying that the artifactual increase in the information measure with increasing categories will also apply to $I_t(m, R)$. As demonstrated in Figure 4.3, the increase in $I_t(m, R)$ with m is bounded from above. Also, $I_t(m, R)$ does not increase significantly beyond approximately 20 categories. The absolute upper bound is the theoretical case obtained by representing the data set using an infinite number of categories. Referring back to Section 3.3 above, we see that the approximation of an infinite number of categories is what allowed us to jump from the discrete case to the continuous case in developing a formula for the asymptotic information, i.e. $I_t(R)$. Hence, the upper bound can be considered as the asymptotic information associated with that stimulus range, or $I_t(R)$.

4.2.2 $\sigma(R_0)$ Determines $I_t(m, R_0)$

As a demonstration of the mathematical basis for the increase of $I_t(m, R_0)$ with m, consider the experiment of Garner (1953). Absolute identification experiments were conducted over a fixed range of 15 to 110 dB using $4, 5, 6, 7, 10, 20$ stimulus categories. To overcome small sample bias, data were pooled over several subjects such that the number of trials used corresponded to $N = 600m$. Garner's paradigm was recreated using the computer simulator. A "best value" was found for $\sigma(95)$ such that the resulting simulated value of information corresponded to Garner's $I_t(20, 95)$ using $N = 12,000$.

Setting $\sigma(95) \simeq 4.8$ dB and using the same seed for pseudorandom number generation, values of $I_t(m, 95)$ were found by reorganizing the simulated data set into the appropriate number of categories. Each simulated data set corresponded to $N = 600m$ and $m = 4, 5, 6, 7, 10, 20$. This is shown in Figure 4.4. Filled circles represent Garner's data while open circles represent simulated data.

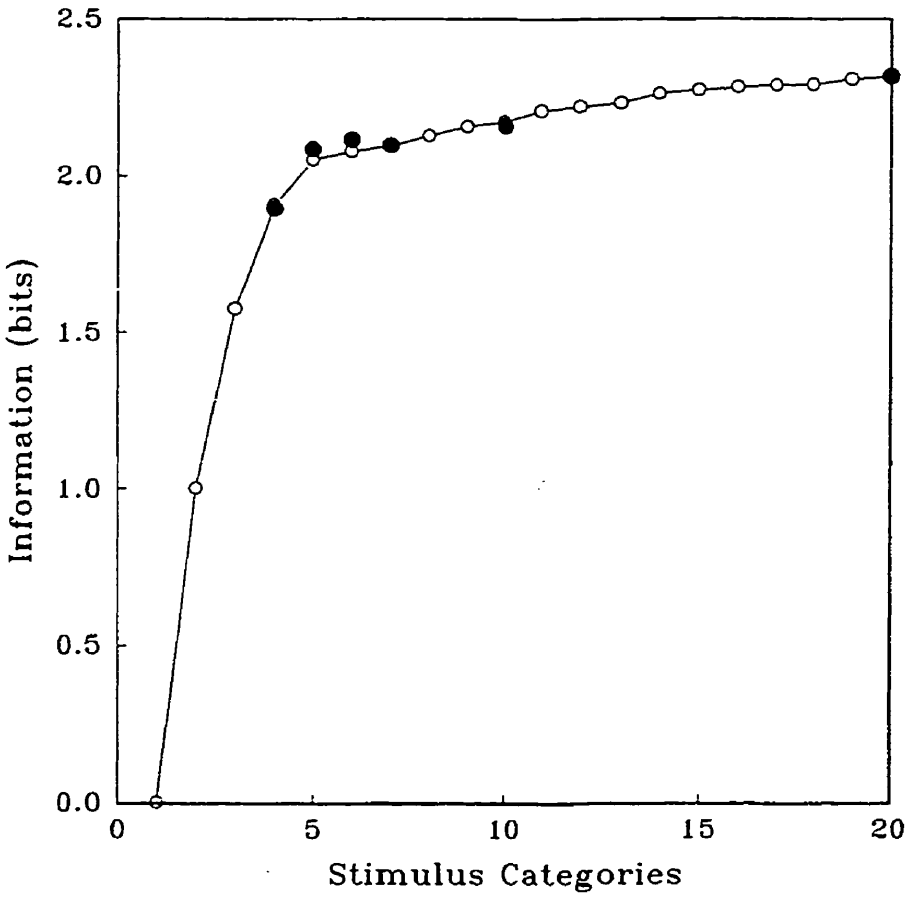

Figure 4-4: Filled circles represent data of Garner (1953). Transmitted information measured (in bits) for varying m over a fixed range spanning 95 dB. Open circles represent simulation of Garner's data using a single value of $\sigma(95) \simeq 4.8$ dB.

4.3 Transmitted Information and Stimulus Range; $I_t(R)$

4.3.1 The Increase of Information with Stimulus Range $I_t(R)$

Absolute Identification experiments were conducted in our laboratory on six subjects over several ranges; for example, $1-10$ dB, $1-30$ dB, $1-50$ dB, $1-70$ dB and $1-90$ dB. In each experiment, the number of categories used equalled the range in dB. That is, ten stimulus and response categories were used in the absolute identification experiment conducted over $1-10$ dB. For most subjects, 500 trials were used for a given range. The average row variance, s^2_{eff}, was measured from the matrix and s^2 was then calculated using the CV model. Estimates for I_t were obtained by substituting R and s into the asymptotic formula 3.5 above. Results are presented for each subject in the tables below. Notice that I_t increases with increasing range in accordance with previous investigators (Braida and Durlach (1972); Luce, Green and Weber (1976); Norwich, Wong and Sagi (1998) [NWS]). One should note that the results presented herein differ from NWS in that the estimates for I_t in NWS were obtained using s^2_{eff}.

Range (dB) [B 52]	N	s_{eff}	s	I_t
$11-20$	160	1.57	1.88	0.511
$11-30$	480	3.32	4.03	0.463
$11-40$	160	3.46	3.88	0.803
$11-50$	467	4.78	5.40	0.768
$11-60$	160	4.95	5.45	0.945
$11-70$	160	5.47	5.96	1.024
$11-90$	478	6.56	7.08	1.124

Range (dB) [C 19]	N	s_{eff}	s	I_t
$1-10$	500	1.10	1.23	0.841
$1-30$	500	2.56	2.77	1.086
$1-90$	500	6.46	6.89	1.252

Range (dB) [E 22]	N	s_{eff}	s	I_t
$1-10$	500	1.38	1.60	0.632
$1-30$	500	2.46	2.65	1.125
$1-50$	500	3.68	3.94	1.226
$1-90$	500	5.35	5.64	1.432

Range (dB) [J 18]	N	s_{eff}	s	I_t
$1-10$	500	1.44	1.68	0.595
$1-30$	500	2.87	3.14	0.979
$1-50$	500	4.43	4.82	1.049
$1-70$	500	5.12	5.48	1.232
$1-90$	500	5.08	5.34	1.482

Range (dB) [R 19]	N	s_{eff}	s	I_t
$1-10$	500	1.16	1.31	0.790
$1-30$	500	2.25	2.41	1.208
$1-90$	500	4.99	5.24	1.499

Range (dB) [W 25]	N	s_{eff}	s	I_t
$1-10$	500	1.49	1.75	0.563
$1-30$	500	2.50	2.71	1.107
$1-50$	500	3.72	3.99	1.215
$1-70$	500	4.41	4.66	1.376
$1-90$	500	5.58	5.90	1.391

4.3.2 The Increase of σ With Stimulus Range; $\sigma(R) = aR + b$

From the tables above, one finds that s, and subsequently σ, tends to increase with R. Shown below in Figure 4.5 and Figure 4.6 are graphs of s plotted as a function of R. One can see that the increase is well described by a linear relationship. That is,

$$\begin{aligned} s(R) &= aR + b \\ &\quad \text{a,b constants greater than zero.} \\ &\Rightarrow \sigma(R) = aR + b \end{aligned} \tag{4.2}$$

In Chapter 5, a theoretical basis is proposed for the form of $\sigma(R)$ with R. Essentially, by imposing the criteria that I_t reaches a channel capacity for extremely large ranges, we find that $\frac{d\sigma}{dR} = const$ in the limit of a large range. Although Equation 4.2 does satisfy this condition, it is probably a special case of a more general relationship. However, to first order, $\sigma(R) = aR + b$ does account for the increase of I_t with R.

Please recall that any estimate for $I_t(R)$ depends on the estimate for $\sigma(R)$. Hence, the $\sigma(R)$ relationship can be used as an input to the asymptotic approximation resulting in the smooth rise of I_t with R. Shown in Figure 4.7 is a graph of $I_t(R)$ with stimulus range for subject W. The smooth curve of $I_t(R)$ was calculated using the best-fit line $s(R) = aR + b$ from Figure 4.5. The smooth curve is compared with I_t values estimated from experiments.

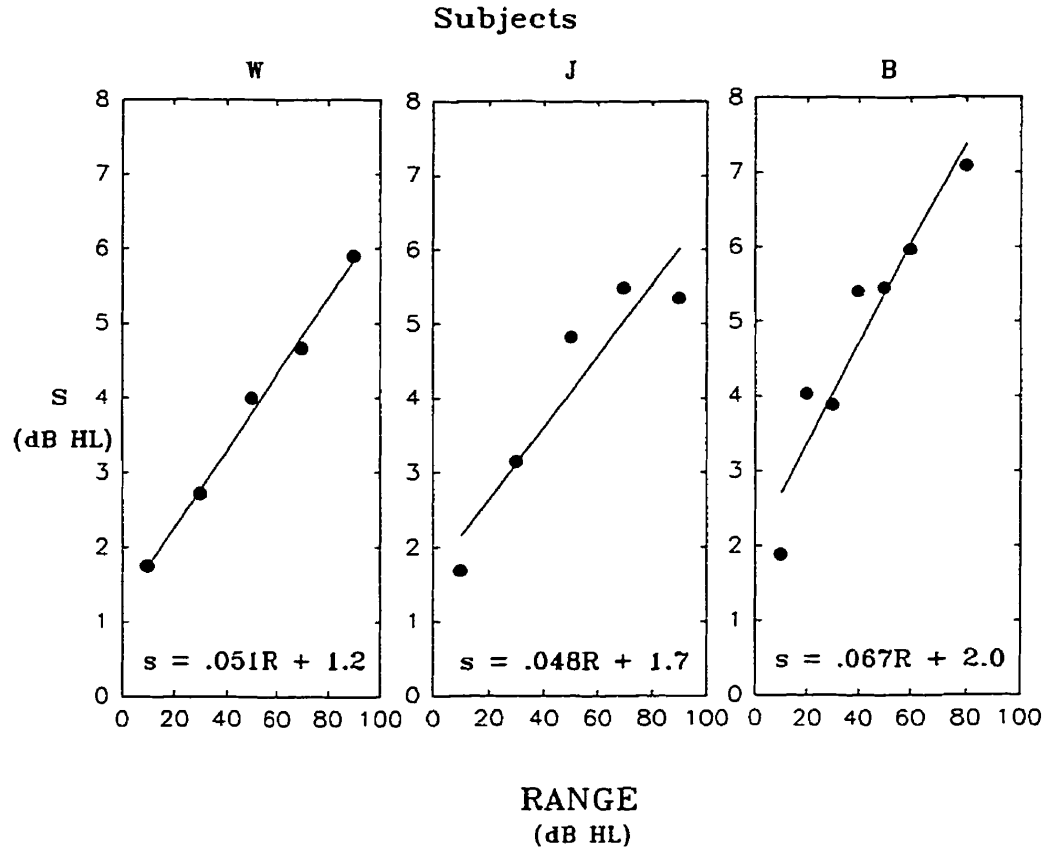

Figure 4-5: The increase of s with stimulus range R for subjects W, J and B.

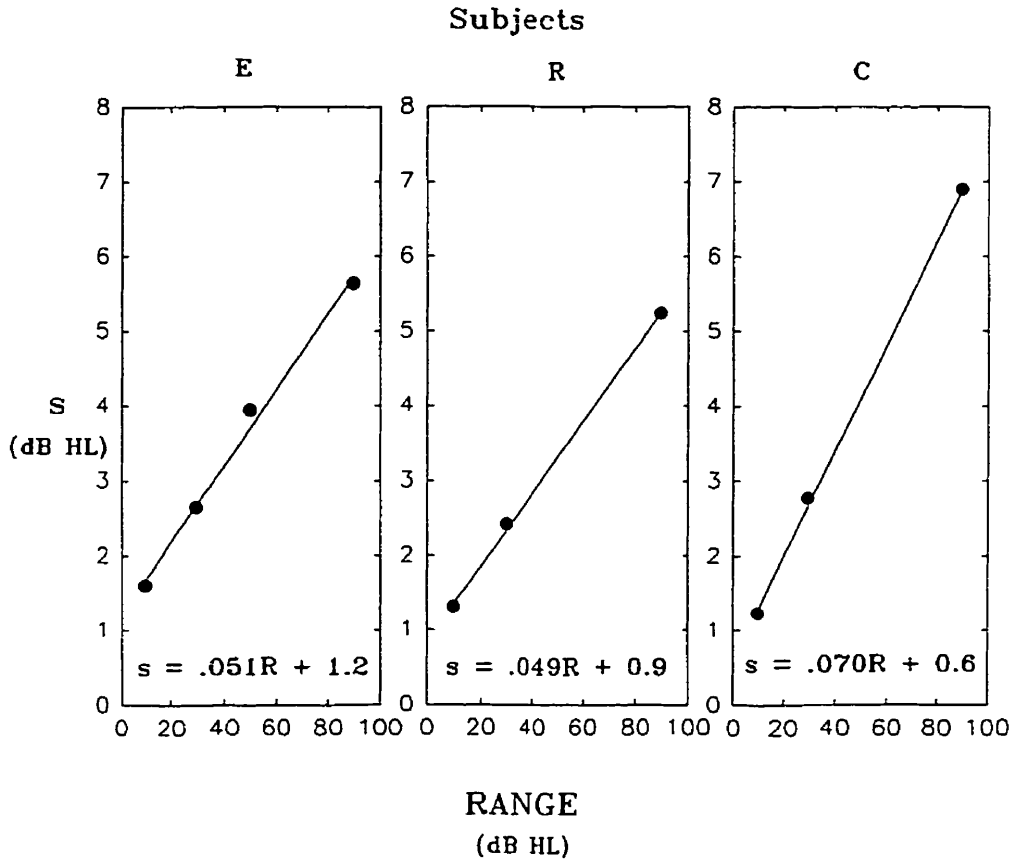

Figure 4-6: The increase of s with stimulus range R for subjects E, R and C.

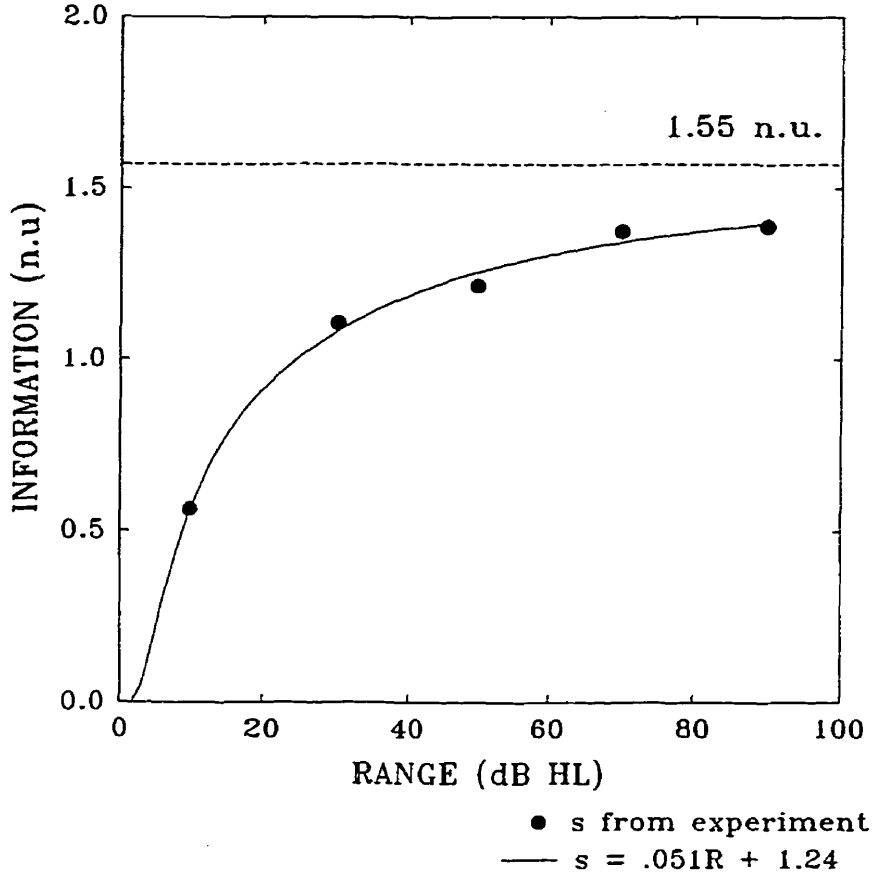

Figure 4-7: Plot of transmitted information, $I_t(R)$ (in natural units), as a function of stimulus range R for subject W. Filled circles represent estimates of $I_t(R)$ using s values estimated from experiments. Solid line is obtained using $s(R) = .051R + 1.2$. $s(R)$ determines both the rise in $I_t(R)$ and its subsequent saturation towards $I_\infty = 1.55$ natural units.

For a qualitative description of how $\sigma(R)$ determines $I_t(R)$, one finds that the constant b determines the increase or "gain" of I_t while the constant a determines the saturation or "channel capacity" of I_t. This can be seen by substituting Equation 4.2 into Equation 3.5 in the limiting case of a large stimulus range where the contribution of edge effects becomes minimized.

$$\begin{aligned} I_t &= \ln R - \frac{1}{2}\ln(2\pi e\sigma^2) \\ &= \ln\left(\frac{R}{\sqrt{2\pi e}(aR+b)}\right) \\ &= \ln\left(\frac{1}{\sqrt{2\pi e}(a+\frac{b}{R})}\right) \end{aligned}$$

Notice how the intercept term will disappear as $R \rightarrow \infty$ leaving us with a channel capacity of

$$I_\infty = \ln\left(\frac{1}{\sqrt{2\pi e a}}\right)$$

A typical value found for the slope in the above graphs is $a = 0.05$ reporting a channel capacity of $I_\infty = 1.58$ natural units. Miller's upper limit was reported at around 1.75 natural units, giving a slope value of approximately $a = 0.04$. Hence the slope of the line that describes the linear increase of $\sigma(R)$ with stimulus range directly indicates subject performance. The lower the slope, the higher the subject's informational channel capacity.

In the table below, the channel capacities for the six subjects who participated in our experiments are listed.

Subject	B	C	E	J	R	W	average
a	0.067	0.070	0.051	0.048	0.049	.051	0.056
I_∞ [n.u.]	1.28	1.24	1.56	1.62	1.60	1.56	1.46

We now see that the $\sigma(R)$ that underlies a subject's response error for the absolute

identification experiment over range R is sufficient to describe the informational properties of that subject. Namely, $\sigma(R)$ determines $I_t(R)$.

Chapter 5

Implications

5.1 Contribution to Absolute Identification Theory

The development of the CV-model is significant in the general theory of absolute identification. First, the method of overcoming small sample bias through computer simulation developed by Wong and Norwich (1997) can be improved. The improvement occurs because we are able to extract an estimate of σ directly from the confusion matrix. σ is the constant variance of the normal distribution thought to underlie a subject's response error and can be used as an input for the simulator to generate pseudo-stimulus response pairs. In turn, $I(N, m, R)$ as measured from the simulated matrices conforms to experiment more closely.

One of the results of more accurate simulation through estimation of σ, is that the effect of the number of categories on the transmitted information, $I_t(m, R)$, can be demonstrated (please see Figure 4.4). Using a single value for $\sigma(R_0)$ and simply reorganizing the data into the appropriate number of categories, one can account for the increase in $I_t(m, R_0)$ due to increasing m as demonstrated with the data of Garner (1953). In conjunction, we have provided a mathematical argument which states that organizing the same data into larger confusion matrices yields information estimates that increase or stay the same, demonstrating that the categorical affect is a mathematical property

of the information measure itself and not a function of the perceiver.

The ability to extract σ from the confusion matrix lies in the relationship between σ and σ^2_{eff} (Equation 3.3), the arithmetic mean row variance as measured from the matrix. The CV-model predicts this relationship and provides a model for the probability distributions found along the rows of the confusion matrix, i.e. $p(y|x)$. Utilizing both σ and $p(y|x)$, estimates of $I_t(R)$ are achievable for each subject with a limited number of experimental trials. Furthermore, the empirical finding that $\sigma(R) = aR + b$ accounts for the increase of $I_t(R)$ with R found by several investigators (Sections 2.5 and 4.3). The slope value, a, allows for estimation of I_∞. One should note that $\sigma(R)$ depends solely on the sensory properties of the perceiver; consequently, $I_t(R)$ and I_∞ are determined by the sensory properties of the perceiver.

One of the natural outcomes of the CV-model is the prediction of edge-effects known to occur in absolute identification experiments. By confining the underlying normal distribution of constant variance to the fixed stimulus range, we found that $p(y|x)$, especially along the first and last rows, displays skewing similar to that found in experiment. Please recall that Berliner et al. (1978), gave evidence for the existence of intrinsic anchors. The CV-model would hold that these anchors naturally develop as a result of requiring the subject to respond to stimuli that lie on a fixed stimulus range.

5.2 A Criterion For Channel Capacity

Hitherto, it was explained how the results of every experiment on absolute identification can be summarized by $\sigma(R)$. Also, it was demonstrated empirically that σ is governed by the following function of R:

$$\sigma(R) = aR + b \text{ where } a, b > 0. \tag{5.1}$$

No basis was given for the increase in subject response variance with stimulus range. To begin, let us look at why $\sigma(R)$ should increase rather than decrease with the stimulus

range, by exploring the concept of channel capacity.

In Section 2.5 and Section 4.3, it was described how information tends to increase with increasing stimulus range. One can observe this phenomenon by conducting absolute identification experiments for several ranges. In each range, the small sample bias must be carefully overcome, say, through computer simulation. One must also remove any artifactual increases in information due to increasing the number of categories used for experimentation. This can be done either by fixing the number of categories used for all experiments (eg. Braida & Durlach set $m = 10$) or by using a large enough number of categories such that the artifactual increase is negligible, eg. set $m = R$ dB. Upon measuring the information for each experiment under these conditions, the information measure reflects $I(N, m, R) \rightarrow I_t(R)$.

In any case, one finds that $I_t(R)$ increases monotonically with R, but eventually saturates at a constant value, I_∞. That is,

$$I_t(R) \rightarrow I_\infty \text{ as } R \rightarrow \infty.$$

We can consider this as the criterion for "Channel Capacity" or the upper limit on information processing for human intensity perception. For the case of a large stimulus range, we found that the information transmitted to a subject is well described by the asymptotic formula in Equation 3.5. Let us now suppose that the stimulus range is large enough such that $I_t(R)$ no longer increases significantly and has reached channel capacity, I_∞, to some degree of significance. Hence we have,

$$I_\infty = \lim_{R\rightarrow\infty} I_t(R) = \lim_{R\rightarrow\infty} \left(\ln \frac{R}{\sqrt{2\pi e \sigma}} \right) \text{ where } \sigma = \sigma(R).$$

Now, I_∞ will not be constant unless $\sigma(R)$ grows with R such that $\lim_{R\rightarrow\infty} \sigma(R) = \infty$. A little calculus indicates that

$$\exp(I_\infty) = \frac{1}{\sqrt{2\pi e}} \lim_{R\rightarrow\infty} \left(\frac{R}{\sigma} \right) \stackrel{H}{=} \frac{1}{\sqrt{2\pi e}} \lim_{R\rightarrow\infty} \left(\frac{1}{\sigma'(R)} \right).$$

where the "H" over the equal sign represents the use of *l'Hospital*'s rule and $\sigma'(R)$ represents the derivative of σ with respect to R. Some more rearranging gives us

$$\lim_{R\to\infty} \sigma'(R) = \frac{1}{\sqrt{2\pi e}\exp(I_\infty)} \tag{5.2}$$

Equation 5.2 tells us that in the limit of a large range, the criterion of channel capacity would require that the rate of change of a subject's response error with stimulus range remains constant. In other words, a constant channel capacity would require that $\sigma(R)$ grows with the stimulus range such that $\sigma'(R) = const$ in the limit of a large range. As it stands, Equation 5.1 satisfies Equation 5.2, but is not a unique solution.

Please recall that Durlach and Braida (1969) had postulated that a subject's error in response to a fixed stimulus in an absolute identification experiment should increase with stimulus range in the following manner:

$$\sigma^2 = \alpha^2 R^2 + \beta^2 \text{ where } \alpha, \beta > 0. \tag{5.3}$$

In this case we find that $\lim_{R\to\infty} \sigma'(R) = \alpha = const.$ Hence, the form predicted by Durlach and Braida also satisfies Equation 5.2. The preference of one form over another will be discussed in the next section. Regardless, a criterion of channel capacity would require any model of the increase of $\sigma(R)$ with range to satisfy Equation 5.2 which, in turn, provides a way to predict a subject's channel capacity. For example, we found in our experiments that a typical value for $\sigma'(R)$ was measured at $\sigma'(R) = a = .05$. By Equation 5.2, $a = .05$ would imply that $I_\infty = 1.58$ natural units. Also in the experiments of Braida & Durlach (1972), a value of $\alpha \simeq 0.07$ was found giving $I_\infty = 1.24$ natural units.

5.3 Wherefore $\sigma = aR + b$

A criterion of constant channel capacity will require $\sigma(R)$ to increase with R such that $\sigma'(R) = const.$ As we have seen, Equation 5.1 is not the only form that satisfies this

criterion. In the model proposed by Durlach and Braida, Equation 5.3 does have some psychological appeal. If one considers two independent processes whose variances are σ_1^2 and σ_2^2 respectively, then the total variance can be expressed as $\sigma^2 = \sigma_1^2 + \sigma_2^2$. Durlach and Braida hypothesized that the subject's response error was governed by two independent "noise" compartments. The first compartment was a kind of analogue noise resulting from the physiological circuits that mediate the loudness of a signal to the brain and was assumed constant, i.e. $\sigma_1 = \beta$. The second compartment was a kind of memory noise that results in the process of relating the intensity of the signal to the internal context or representation of stimulus range. This noise was postulated to increase linearly with the context, i.e. $\sigma_2 = \alpha R$. Combining the two compartments gives Equation 5.3.

A physiological basis for Equation 5.1 will now be discussed. Upon closer inspection, we understand that both σ and R are measured in dB, i.e. σ_{dB} and R_{dB}. The decibel is, however, a convenient way of describing large quantities in logarithmic form, relative to some fixed quantity. Linear or absolute sound intensity, I_{lin}, varies directly with the square amplitude of a sound pressure wave. If we consider the absolute threshold of hearing for a stimulus tone of fixed frequency to be some constant pressure, p_0, we can define the corresponding fixed intensity, $I_0 = kp_0^2$ (k constant). Intensity measured in dB, i.e. I_{dB}, would be measured as follows:

$$I_{dB} = 10 \log_{10} \frac{I_{lin}}{I_0}.$$

If σ_{dB} and R_{dB} are measured using the same reference intensity, I_0, then Equation 5.1 becomes

$$\begin{aligned} 10 \log_{10} \frac{\sigma_{lin}}{I_0} &= a 10 \log_{10} \frac{R_{lin}}{I_0} + b \\ \frac{\sigma_{lin}}{I_0} &= 10^{\frac{b}{10}} \left(\frac{R_{lin}}{I_0} \right)^a \\ \sigma_{lin} &= \delta' R_{lin}^a \end{aligned} \tag{5.4}$$

where $\delta' = 10^{\frac{b}{10}} I_0^{1-a} = const.$

Please recall that in our absolute identification experiments, subjects are presented with $m = R$ dB possible stimulus intensities over the range $1 - R$ dB. The intensities are distributed uniformly over the range. This means that on a given trial, each intensity has an equal likelihood of presentation. In the long run, the average stimulus intensity presented to the subject would be

$$\begin{aligned} I_{dB} &= \frac{R_{dB}}{2} \\ 10\log_{10}\frac{I_{lin}}{I_0} &= 10\log_{10}\left(\frac{R_{lin}}{I_0}\right)^{\frac{1}{2}} \\ \frac{I_{lin}}{I_0} &= \left(\frac{R_{lin}}{I_0}\right)^{\frac{1}{2}} \\ R_{lin} &= \frac{I_{lin}^2}{I_0}. \end{aligned}$$

Substituting this into Equation 5.4 gives

$$\begin{aligned} \sigma_{lin} &= \delta I_{lin}^{2a} \\ \sigma_{lin}^2 &= \delta^2 \left(I_{lin}\right)^{4a} \end{aligned} \tag{5.5}$$

where $\delta = 10^{\frac{b}{10}} I_0^{-a} = const.$

Hence, in absolute terms, Equation 5.1 gives a variance-mean relationship of the form

$$\sigma_{lin}^2 \propto I_{lin}^n. \tag{5.6}$$

This relationship is very common and arises in both the psychophysical and neurophysiological realms. Before any further digressions are made, the context in which this relationship occurs in our situation needs clarification. Essentially, we are stating that a subject's performance in an absolute identification experiment is governed by the subject's response error that depends explicitly on the average intensity presented to the subject in the long term. That is, a subject's response error does not depend on the

intensities of the individual stimulus tones presented at any given trial. If this were the case, σ would not be constant for a fixed stimulus range as described by the CV-model. Instead, a subject's response error is governed by the stimulus range. The subject perceives the stimulus range after a relatively long-term exposure to the intensities which underlie it.

5.4 Psychophysical $\sigma^2_{lin} \propto I^n_{lin}$

S.S. Stevens (1956) developed a methodology known as "magnitude estimation" that can be used to quantify how subjective magnitude, F, varies with stimulus intensity. As with the absolute identification paradigm, magnitude estimation can be used for several sensory modalities. These two methodologies differ, however, in how subject performance is quantified. In a typical magnitude estimation experiment for loudness, a subject would be familiarized with a stimulus tone that is fixed in frequency and set to a standard loudness. Subsequently, tones of similar frequency, but varying in loudness, would be presented in conjunction with the standard. The standard tone was assigned a subjective value, say "100", and subjects would be required to estimate the loudness of stimulus tones relative to the standard. For example, if a stimulus tone was thought to be twice as loud as the standard, the subject would respond "200". Similarly, if a tone was thought to be half as loud as the standard, the subject would respond "50". Responses would then be normalized onto a scale such that the standard would be assigned a subjective magnitude value of "$F = 1$ sone". Upon plotting the logarithm of subjective magnitude as a function of stimulus intensity (in dB), most of the data conformed to a straight line. Hence, Stevens proposed that subjective magnitude was related to absolute intensity by way of a power function of the form

$$F = \theta I^n_{lin} \text{ where } \theta, n = const.$$

As of yet, no theoretical basis has been proposed for Stevens' "power law of sensation"

rendering it a purely empirical formulation. Probably one of the most fascinating aspects of the power law is the value of the exponent, n. The power law has been applied by many laboratories to many sensory modalities. Apparently, n assumes characteristic values (or range of values) for different sensory modalities. For example, a typical exponent value for the loudness of a 1-kHz tone is $n \simeq 0.3$.

Notably, Stevens power law has dominated the psychology literature since its inception, but is not the only "law of sensation". In the latter half of the nineteenth century, Ernst Heinrich Weber (1795-1878) and Gustav Theodor Fechner (1801-1887) together proposed that subjective magnitude should vary with the logarithm of the absolute stimulus intensity. Hence,

$$F = C \log I_{lin} + D \text{ where } C, D = const.$$

One finds that magnitude estimation data conforms equally well to this "logarithmic law of sensation" as it does to the power law. As an example, Stevens (1969) conducted magnitude estimation experiments on the degree of saltiness for a sodium chloride solution. In a full-logarithmic plot of subjective magnitude and concentration of stimulus intensity, the data conformed to a straight line with the exception of the last two or three points. That is, the subjective magnitude corresponding to the most concentrated solutions fell well below Stevens' linear relationship. Norwich (1993) plotted the same data, but on a semi-logarithmic plot. Subjective magnitude was plotted against the logarithm of concentration and the data also conformed well to a straight line; however, the first two or three points (i.e. least concentrated solutions) fell above the straight line. Hence, the power law has a tendency to predict the subjective magnitude for lower intensities while the logarithmic law has a tendency to predict that of higher intensities.

Norwich (1993) describes how the different forms of the law of sensation can be described as two realizations of the same underlying phenomenon. To begin, he postulated that the subjective magnitude was directly proportional to the uncertainty, H, as reflected in the *perceptual unit* which is to be defined as, "the smallest and simplest configuration

of anatomical structures required to mediate the process of perception in some modality" (Norwich, 1993 [pg. 281]). Hence,

$$F = kH \text{ where } k = const.$$

The mathematical form of the uncertainty function, H, was modelled after Shannon's information theoretical approach outlined in Section 2.1 above.

In the case of loudness perception, the perceptual unit is required to obtain an estimate of stimulus intensity, I_{lin}. This is probably achieved through a sampling process at the neuronal level that works on a time scale on the order of milliseconds. In each sampling, consider the formation of a distribution of intensity estimates with mean $\mu(I_{lin})$ and variance σ_S^2. After m' samples, the Central Limit Theorem of statistics (Freund & Walpole, pg. 256) would state that a collection of sample means would follow a normal distribution of mean I_{lin} and variance $\frac{\sigma^2}{m'}$. Since sampling estimates are limited by the neuronal capabilities of the perceptual unit, consider the process to be limited by a gaussian white noise of constant variance σ_R^2. From Equation 2.1,

$$I(Y|X) = H(Y) - H(Y|X).$$

In modelling H after $I(Y|X)$, the "response entropy", i.e. $H(Y)$, would correspond to the uncertainty in the overall estimate for signal intensity and includes the noise distribution. Also, the "response equivocation", i.e. $H(Y|X)$, would correspond to the uncertainty associated with the noise distribution. Hence, $H(Y) \equiv H_{S+R}$ and $H(Y|X) \equiv H_R$. We therefore have,

$$\begin{aligned} H &= H_{S+R} - H_R \\ &= \frac{1}{2}\ln(2\pi e[\frac{\sigma_S^2}{m'} + \sigma_R^2]) - \frac{1}{2}\ln(2\pi e\sigma_R^2) \\ &= \frac{1}{2}\ln\left(1 + \frac{\sigma_S^2}{m'\sigma_R^2}\right). \end{aligned}$$

So, subjective magnitude would take the form

$$F = \frac{k}{2} \ln \left(1 + \frac{\sigma_S^2}{m' \sigma_R^2} \right).$$

Essentially, Norwich postulated that subjective magnitude should relate to the overall uncertainty in the sampling process of the perceptual unit. That is, a macroscopic observation, F, is related to a combination of microscopic sampling events. To apply this relationship in the laboratory, the microscopic components need to be converted into measurable variables. In the case of loudness perception, we are interested in how F relates to I_{lin}. Since stimulus tones are of constant duration, the total number of samples, m', should remain constant. For simplicity, let the variance of the reference noise inherent to the system remain fixed. We shall now be concerned with how the variance of a sampling distribution, σ_S^2, should relate to I_{lin}. Norwich empirically selected the relationship

$$\sigma_S^2 \propto I_{lin}^n \tag{5.7}$$

which commonly arises in statistical physics. Applying this relationship to our previous equation gives

$$F = \frac{k}{2} \ln \left(1 + \gamma I_{lin}^n \right) \text{ where } \gamma = const. \tag{5.8}$$

One should take note of the following approximations:

$$\begin{aligned} \gamma I_{lin}^n &\ll 1 \Rightarrow F \propto I_{lin}^n \text{ (Power Law)} \\ \gamma I_{lin}^n &\gg 1 \Rightarrow F \propto \ln(I_{lin}^n) \text{ (Logarithmic Law).} \end{aligned}$$

Hence, Equation 5.8 gives rise to both forms of the law of sensation. Also, at the heart of Equation 5.8 is a sampling process described by the variance-mean relationship in Equation 5.7.

Let us now cast the fine thread that connects this digression to the world of absolute identification by way of example. Figure 5.2 is taken from Norwich (1993, pg. 161)

whereby magnitude estimation data for loudness of a 1-kHz tone is fit to Equation 5.8. Notice that the exponent value is approximately $n \simeq 0.29$. In our absolute identification experiments, the slope value for $\sigma(R)$ fell within the range $a = 0.048$ to $a = 0.070$. Substituting these values into Equation 5.5 gives exponent values that would range from $n = 0.19$ to $n = 0.28$. With the aid of the variance-mean relationship in Equation 5.6, we find that exponent values that arise from absolute identification are comparable, but slightly lower than that of magnitude estimation. To account for the difference, we recall that the variance-mean relationship for magnitude estimation is directly related to stimulus intensity while that of absolute identification is related to the average stimulus intensity as governed by the overall stimulus range.

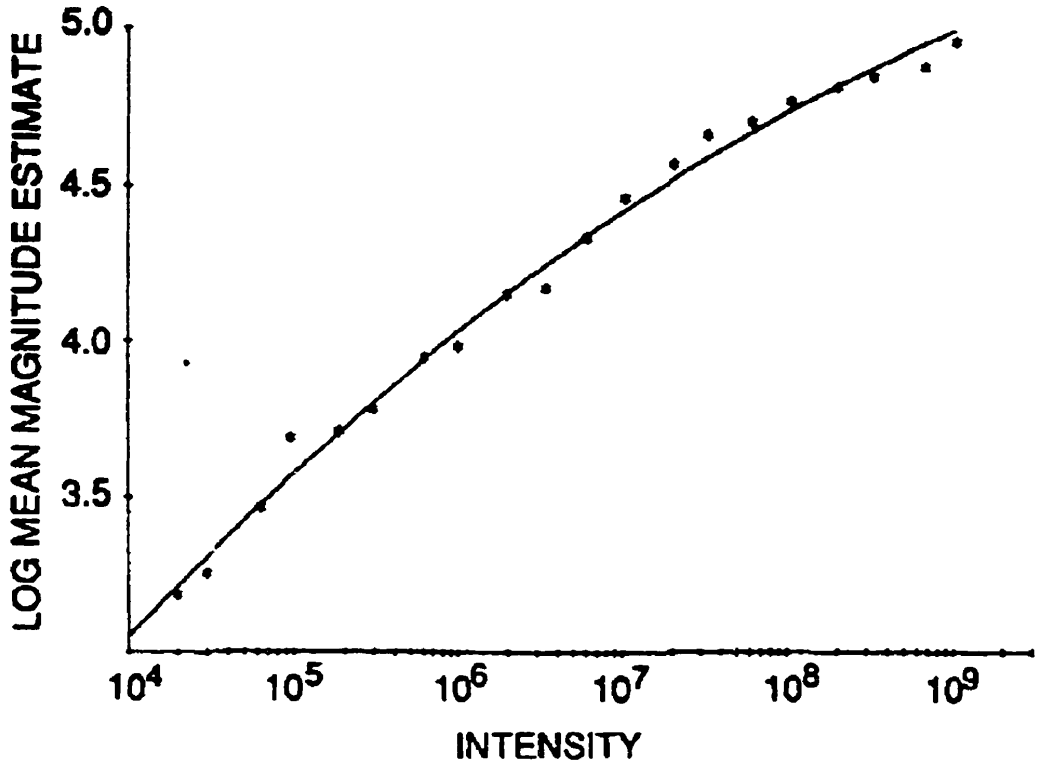

Figure 5-1: Data of Luce and Mo (1965). Natural log of mean magnitude estimate of intensity of 1000-Hz tone (subject 9) plotted against log of sound intensity. The data is fit to the entropy equation: $F = (\frac{113.1}{2}) \ln (1 + .03131 I^{.2896})$. From Norwich (1993, pg. 161).

5.5 Neurophysiological $\sigma^2_{lin} \propto I^n_{lin}$

To extend the concept of sensory performance being regulated by the sampling properties of a perceptual unit, let's consider response properties at the neurophysiological level. To begin, Borg et. al. (1967) performed magnitude estimation experiments on taste perception in a psychophysical and neurophysiological setting. In humans, the sensory fibers that mediate taste from the anterior two-thirds of the surface of the tongue extend backward toward the brain in a nerve called the *chorda tympani.* This nerve passes through the middle ear and is surgically accessible. Two days before surgery, magnitude estimation experiments on taste perception of citric acid (sour), sodium chloride (salty) and sucrose (sweet) were carried out. During surgery, several concentrations (in units of molarity) for each solution type were applied to the surface of the tongue and resulting firing rates were recorded from the *chorda tympani.* When subjective magnitude and firing rate were plotted on a log-log plot against molarity of solution, the data conformed fairly well to a linear relationship. Furthermore, the slope of the line was comparable for both psychophysical and neurophysiological responses. The results of Borg et. al. are displayed for the citric acid and sucrose solutions in Figure 5.2 [their Figure 7].

Hence, psychophysical and neurophysiological responses for the modality of taste tend to share a common exponent value. Although Stevens' power law was used to describe the responses, we understand from Equation 5.8 that the exponent value describes the variance-mean relationship governing the sampling properties of the perceptual unit. It doesn't seem fair to assume that the finding of Borg et. al. should extend to all modalities. In fact, the subjective magnitude corresponding to the loudness of a stimulus tone does not equate with neuronal responses under the same conditions. For example, Relkin and Doucet (1997) were able to empirically measure the growth in auditory nerve spike count with stimulus intensity for a 1-kHz tone. They developed a technique for

recording a compound action potential described as the perstimulus compound action potential (PCAP) from the chinchilla auditory nerve. Upon plotting the PCAP spike count against stimulus intensity on log-log plot, the data conformed well to a linear relationship with a slope value of $n \simeq 0.20$ which is two-thirds of the typical value for the subjective magnitude corresponding to the loudness of a 1-kHz tone burst.

Although the results of Relkin and Doucet show that subjective magnitude doesn't necessarily correspond with neural firing rate, their data did conform to a power function relationship. This supports the notion that sensory performance is regulated by the sampling properties of a perceptual unit. To account for the discrepancy in exponent values, we must examine the biological composition of a perceptual unit. Does a neuron, in itself, constitute a perceptual unit? To restate, a perceptual unit can be defined as, "the smallest and simplest configuration of anatomical structures required to mediate the process of perception in some modality" (Norwich, 1993). For example, a Paramecium (a single cell organism) is able to detect and move along a concentration gradient in the search for food. In the context of our definition, a Paramecium would be considered a perceptual unit. The key idea is that a perceptual unit is capable of mediating the process of perception. As discussed in the Introduction, perception involves the attribute of choice such that a selection can be made from a set of alternatives (Norwich, 1993). If a neuron is capable of mediating the full set of alternatives inherent in a precept, then the neuron is a perceptual unit. A similar argument would hold for a complex array of neurons or for the brain as a whole. I therefore make the conjecture that when sensory performance is measured at different levels (ex. psychophysical vs. neurophysiological), common exponent values would imply that the measurements were taken from a common perceptual unit. This arises from the proposition that the exponent n-value corresponds to the sampling properties of that perceptual unit as described by the variance-mean relationship in Equation 5.6 and Equation 5.7.

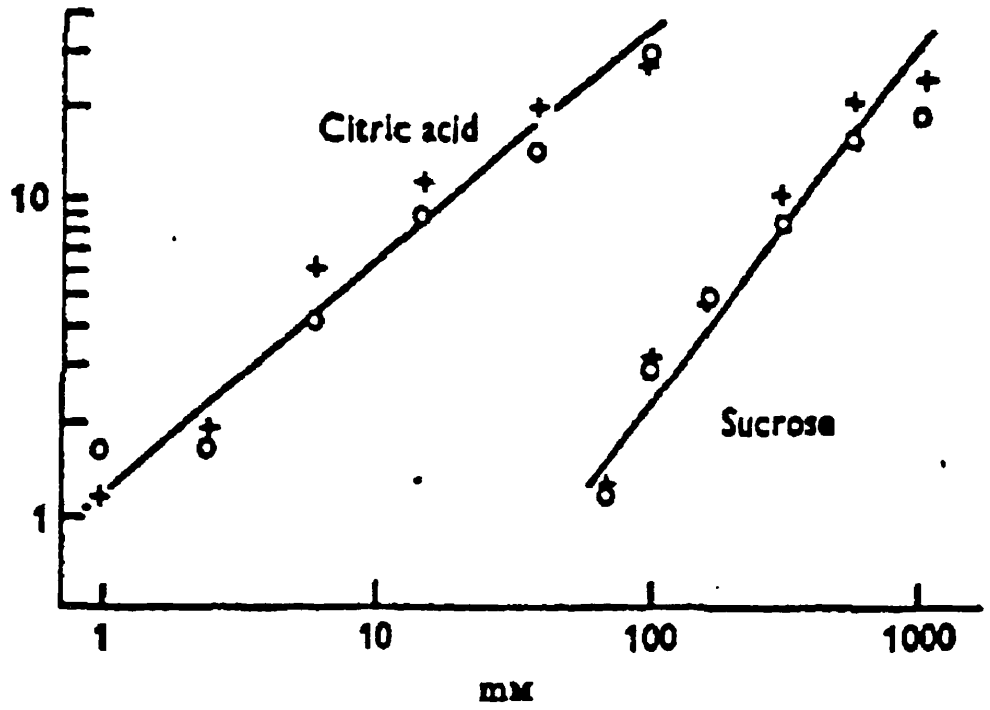

Figure 5-2: Base ten log of mean values of neural response (open circles) and of subjective response from two patients plotted against molarity of citric acid and sucrose solution. Form Borg et. al. (1967, Fig. 7).

Chapter 6

Conclusion

A typical experiment on absolute identification of loudness involves the selection of a range, R, fixed over the continuum of stimulus intensity. The range is discretized into m stimulus/response categories. Upon the presentation of a stimulus, selected randomly from one of the stimulus categories, a subject is required to estimate the intensity of the stimulus tone to the nearest response category. If the pairing of a stimulus category with a response category is considered as one trial, N such trials can be compiled into an m x m confusion matrix. A subject's ability to match stimulus categories with response categories can be described in terms of the amount of information transmitted to the subject and can be estimated from the matrix in the form of the information measure $I(N, m, R)$. The three variables, N, m and R, each have a distinct effect on the information measure.

For an insufficiently large number of trials, $I(N, m, R)$ overestimates the true value of the information transmitted to the subject for that absolute identification experiment, $I_t(m, R)$. Overcoming small sample bias can be represented as

$$N \to \infty \Longrightarrow I(N, m, R) \to I_t(m, R)$$

and is achievable through a process of computer simulation as described in Norwich and Wong (1997). The average variance, σ^2_{eff}, as measured from the rows of a confusion

matrix can be measured experimentally in the form of the sample estimate, s^2_{eff}. The sample estimate is a measure of the underlying error governing a subject's response and can be used as an input for simulating stimulus/response pairs. However, an accurate estimate of subject response error must account for apparent edge effects. That is, subjects tend to respond with less error to stimulus tones with intensities more to the extreme values of the stimulus range. Hence, edge effects result in the extreme rows having an apparently smaller measurable error. This reduction of error is incorporated in s^2_{eff} which does not, therefore, accurately convey the underlying error governing a subject's response.

This thesis proposed a model (CV model) for the error, σ, underlying a subject's response in an absolute identification experiment. Using the assumption that the underlying distribution is normal with a constant variance, σ^2, we were able to account for all the distributions that appear along the row of the confusion matrix, including the extreme rows where edge effects are mostly apparent. By confining the underlying distribution to the fixed stimulus range, R, it was possible to find a relationship between σ^2 and σ^2_{eff} (Equation 3.3). This relationship can be used to obtain a sample estimate of σ, i.e. s, from s^2_{eff}, thereby allowing for a more accurate matrix simulation. Hence, a subject's response, and therefore his/her performance, in an absolute identification experiment can be described by one parameter; the underlying variance σ^2.

For nearly fifty years, it has been well known that increasing the number of categories, m, used to discretize the fixed stimulus range, R_0, in an absolute identification experiment tends to increase $I_t(m, R_0)$. The increase, however, saturates at an upper bound corresponding to the channel capacity of the information transmitted to the subject for that absolute identification experiment, $I_t(R)$. This category effect can be represented as

$$m \to \infty \Longrightarrow I_t(m, R) \to I_t(R).$$

One of the advantages of the CV-model is that a single value, $\sigma(R_0)$, can be used to recreate this effect. For example, using the data of Garner (1953), a "best value" for

$\sigma(R_0)$ was found using one reported value of $I_t(m, R_0)$. Subsequently, $\sigma(R_0)$ was used to simulate data. By simply reorganizing simulated data into the appropriate number of categories, the resulting simulated values for $I_t(m, R_0)$ conformed closely to Garner's measured values. Also, this thesis has provided a mathematical argument which states that organizing stimulus/response pairs into larger confusion matrices causes $I(N, m, R)$ to increase or stay the same. Coupling this argument with the ability to recreate data such as Garner's using $\sigma(R_0)$ indicates that the effect of increasing $I_t(m, R_0)$ with m is a purely mathematical property of the information measure.

An effect that parallels the category effect is the range effect. That is, increasing the size of the stimulus range, R, that is fixed for an absolute identification experiment tends to increase $I_t(R)$. The increase saturates at an upper limit corresponding to a physiological channel capacity of the information transmitted to the subject, I_∞. The range effect can be represented as

$$R \to \infty \Longrightarrow I_t(R) \to I_\infty.$$

We have found, empirically, that a subject's response error, σ, depends solely on the fixed stimulus range used throughout an absolute identification experiment. Specifically, the increase of σ with R is well described by the relationship

$$\sigma(R) = aR + b \text{ where } a, b > 0. \tag{6.1}$$

Using the CV-model, we have been able to construct a theoretical description of the *a priori* probabilities that govern the distribution of responses found along the rows of the confusion matrix. These probabilities were then used to obtain estimates of $I_t(R)$. Also, the stated relationship of $\sigma(R)$ with R was used to account for the increase in $I_t(R)$ with R as well as the upper limit, I_∞. We make the conjecture that Equation 6.1 is determined by the sensory properties of the individual. Hence, the range effect is also determined by the sensory properties of the individual.

The relationship between a subject's response error and the stimulus range used for experiment (Equation 6.1) bears significance to the general area of audition. First, the form of the relationship satisfies a criterion for channel capacity. Second, when transformed into absolute (linear) terms of intensity, Equation 6.1 (measured in decibels) takes the form

$$\sigma_{lin}^2 \propto I_{lin}^n.$$

In this case, intensity was taken as the average stimulus intensity presented to the subject in the long term. Specifically, the absolute identification experiments conducted in our laboratory yielded an average exponent value of $n \simeq 0.22$. This compares well to exponent values found in other psychophysical experiments as well as neurophysiological experiments in hearing.

Chapter 7

APPENDIX

7.1 APPENDIX I: Supplement to CV-Model Calculations

Definition 1 *Consider $x, y \in [0, R]$. Let the probability distribution that underlies the occurrence of y given x be defined such that:*

$$p(y|x) = \frac{1}{I(x)\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right]$$

where

$$I(x) = \frac{1}{2}\left[\text{erf}\left(\frac{R-x}{\sqrt{2}\sigma}\right) + \text{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)\right]$$

and

$$\int_0^R p(y|x)dy = 1.$$

Remark 1 *The first two derivatives of the normalizing factor, $I(x)$, are a useful shorthand in the calculations that will follow. The function $\text{erf}(x)$ is defined as follows:*

$$\text{erf}(x) = \int_0^x \frac{2}{\sqrt{\pi}} \exp(-t^2)dt.$$

1. *1) $I(x)$*

$$I(x) = \int_0^{\frac{R-x}{\sqrt{2}\sigma}} \frac{1}{\sqrt{\pi}} \exp(-t^2)dt + \int_0^{\frac{x}{\sqrt{2}\sigma}} \frac{1}{\sqrt{\pi}} \exp(-t^2)dt \tag{7.1}$$

2. *2) $I'(x)$; from Equation 7.1,*

$$\frac{dI(x)}{dx} = \frac{1}{\sqrt{\pi}} \frac{d}{dx}\left(\frac{R-x}{\sqrt{2}\sigma}\right) \exp\left[-\left(\frac{R-x}{\sqrt{2}\sigma}\right)^2\right] + \frac{1}{\sqrt{\pi}} \frac{d}{dx}\left(\frac{x}{\sqrt{2}\sigma}\right) \exp\left[-\left(\frac{x}{\sqrt{2}\sigma}\right)^2\right]$$

$$= \frac{\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{R-x}{\sigma}\right)^2\right)}{\sqrt{2\pi\sigma^2}} \tag{7.2}$$

$$= \frac{\sqrt{2}}{\sigma}\left(\frac{-1}{2\sqrt{\pi}} \exp(-t^2)|_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}}\right) \tag{7.3}$$

3. *3) $I''(x)$; from Equation 7.2,*

$$\frac{d^2I(x)}{dx^2} = \frac{\left(\frac{-x}{\sigma}\right)\left(\frac{1}{\sigma}\right)\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right) - \left(-\frac{R-x}{\sigma}\right)\left(\frac{1}{\sigma}\right)\exp\left(-\frac{1}{2}\left(\frac{R-x}{\sigma}\right)^2\right)}{\sqrt{2\pi\sigma^2}}$$

$$= \frac{-1}{\sigma^2\sqrt{\pi}} \frac{x\exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right) + (R-x)\exp\left(-\frac{1}{2}\left(\frac{R-x}{\sigma}\right)^2\right)}{\sqrt{2}\sigma}$$

$$= \frac{1}{\sigma^2\sqrt{\pi}}\left(-t\exp(-t^2)|_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}}\right) \tag{7.4}$$

Exercise 7.1.1 *1) Calculate the variance of $p(y|x)$.*

First, consider the expectation of $p(y|x)$; namely $\langle y \rangle$ or $E(y)$,.

$$\begin{aligned}\langle y\rangle &= \int_0^R yp(y|x)dy \\ &= \int_0^R \frac{y}{I(x)\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right]dy.\end{aligned}$$

By making the following substitution:

$$t = \frac{y-x}{\sqrt{2}\sigma} \text{ or } y = \sqrt{2}\sigma t + x, \tag{7.5}$$

we can simplify the calculation for the variance of $p(y|x)$ as follows:

$$\begin{aligned}var[p(y|x)] &\equiv \left\langle y^2\right\rangle - \langle y\rangle^2 \\ &= E\left[\left(\sqrt{2}\sigma t + x\right)^2\right] - \left(\sqrt{2}\sigma E(t) + x\right)^2 \\ &= 2\sigma^2(\left\langle t^2\right\rangle - \langle t\rangle^2).\end{aligned}$$

Hence, we have reduced the problem of calculating $var[p(y|x)]$ to finding $\langle t\rangle$ and $\langle t^2\rangle$. To find $\langle t\rangle$, we shall substitute 7.5 into $\langle y\rangle$ as follows:

$$\begin{aligned}\langle y\rangle &= \int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{\left(\sqrt{2}\sigma t + x\right)}{I(x)\sqrt{2\pi\sigma^2}}\exp\left[-t^2\right]\left(\sqrt{2}\sigma dt\right) \\ &= \sqrt{2}\sigma\int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t\exp\left(-t^2\right)}{I(x)\sqrt{\pi}} + x\int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t\exp\left(-t^2\right)}{I(x)\sqrt{\pi}} \\ &= \sqrt{2}\sigma\int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t\exp\left(-t^2\right)}{I(x)\sqrt{\pi}} + x.\end{aligned}$$

Since

$$\langle y\rangle = \sqrt{2}\sigma\langle t\rangle + x,$$

we find that

$$\langle t\rangle = \int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t\exp\left(-t^2\right)}{I(x)\sqrt{\pi}} \tag{7.6}$$

$$= \frac{1}{I(x)}\left(\frac{-1}{2\sqrt{\pi}}\exp(-t^2)\Big|_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}}\right).$$

Using the shorthand from Equation 7.3,

$$\langle t\rangle = \frac{\sigma}{I(x)\sqrt{2}}\frac{dI(x)}{dx}.$$

We can evaluate $\langle t^2\rangle$ from Equation 7.6 as follows:

$$\begin{aligned}\left\langle t^2\right\rangle &= \int_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}} \frac{t^2\exp(-t^2)}{I(x)\sqrt{\pi}}dt \\ &= \frac{1}{I(x)2\sqrt{\pi}}\left(-t\exp(-t^2)\Big|_{\frac{-x}{\sqrt{2}\sigma}}^{\frac{R-x}{\sqrt{2}\sigma}}\right)+\frac{1}{2}\end{aligned}$$

using the shorthand from Equation 7.4,

$$\left\langle t^2\right\rangle = \frac{\sigma^2}{2I(x)}\frac{d^2I(x)}{dx^2}+\frac{1}{2}$$

we now can express the variance of $p(y|x)$:

$$\begin{aligned}var[p(y|x)] &= 2\sigma^2\left(\left\langle t^2\right\rangle - \langle t\rangle^2\right) \\ &= \frac{\sigma^4}{I(x)}\frac{d^2I(x)}{dx^2}+\sigma^2-\frac{\sigma^4}{(I(x))^2}\left(\frac{dI(x)}{dx}\right)^2.\end{aligned}$$

Exercise 7.1.2 *2) Evaluate the following integral:*

$$H(Y|X) = -\frac{1}{R}\int_0^R\int_0^R p(y|x)\ln p(y|x)dydx.$$

Using the definition of $p(y|x)$ and the substitution in 7.5,

$$H(Y|X) = -\frac{1}{R}\int_0^R \int_0^R \left[\frac{e^{-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2}}{I(x)\sqrt{2\pi\sigma^2}}\right] \ln\left[\frac{e^{-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2}}{I(x)\sqrt{2\pi\sigma^2}}\right] dydx$$

$$\begin{aligned} &= \frac{1}{R}\int_0^R \int_0^R \frac{\left[\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right] e^{-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2}}{I(x)\sqrt{2\pi\sigma^2}} dydx + \\ &\quad \frac{1}{R}\int_0^R \ln\left(I(x)\sqrt{2\pi\sigma^2}\right)\left[\int_0^R \frac{e^{-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2}}{I(x)\sqrt{2\pi\sigma^2}} dy\right] dx \\ &= \frac{1}{R}\int_0^R \left\langle t^2 \right\rangle dx + \frac{1}{R}\int_0^R \ln(I(x))dx + \frac{1}{2}\ln(2\pi\sigma^2) \end{aligned}$$

where $\langle t^2 \rangle$ is described in Exercise 1. The full expression therefore becomes

$$H(Y|X) = \frac{1}{R}\int_0^R \left[\left(\frac{\sigma^2}{2I(x)}\frac{d^2I(x)}{dx^2} + \frac{1}{2}\right) + \ln I(x)\right] dx + \frac{1}{2}\ln(2\pi\sigma^2),$$

or finally,

$$H(Y|X) = \frac{1}{2}\ln(2\pi e\sigma^2) + \frac{1}{R}\int_0^R \left(\frac{\sigma^2}{2I(x)}\frac{d^2I(x)}{dx^2} + \ln I(x)\right) dx.$$

7.2 APPENDIX II: Extending Carlton's Approximation to <I(N)>

Carlton (1969) developed a strong approximation for the bias in information estimates. Theoretically, the information associated with a set of m possible events whereby the probability of each event is $p_i > 0$, $i = 1, 2, ..., m$, is

$$H \triangleq -\sum_{i=1}^{m} p_i \log p_i.$$

After N observations, the maximum likelihood estimate of H, i.e. $\hat{H}$, is

$$\hat{H} \triangleq -\sum_{i=1}^{m} \frac{n_i}{N} \log \frac{n_i}{N}$$

where n_i represents the number of times event i was observed. We are able to monitor the bias in how $\hat{H} \rightarrow H$ by evaluating the expectation in $\hat{H}$, i.e. $\left\langle \hat{H} \right\rangle$. Carlton's approximation for $\left\langle \hat{H} \right\rangle$ (or $E(H)$), in natural units, is as follows:

$$\begin{aligned}
\left\langle \hat{H} \right\rangle &= -\frac{1}{N} \sum_{i=1}^{m} E\left(n_i \log \frac{n_i}{N} | N, p_i \right) \\
\left\langle \hat{H} \right\rangle &\cong \sum_{i=1}^{m} p_i f(N, p_i) \\
&\text{where} \\
f(N, p_i) &= -\ln \left[p_i + \frac{1 - p_i}{N} \right] + \frac{(N-1)(1-p_i)p_i}{(Np_i + q_i)^2}.
\end{aligned}$$

We shall now extend Carlton's approximation to the expectation of the estimate for the transmitted information, I_t, i.e. $\left\langle \hat{I}_t \right\rangle$ or $\langle I(N) \rangle$. From Section 2.2, using m stimulus and response categories,

$$\begin{aligned}
I(N) &= -\sum_{k=1}^{m} \frac{n_{.k}}{N} \log \frac{n_{.k}}{N} + \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{n_{jk}}{N} \log \frac{n_{jk}}{n_{j.}} \\
&= -\frac{1}{N} \sum_{k=1}^{m} n_{.k} \log \frac{n_{.k}}{N} + \frac{1}{N} \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk} \log \frac{n_{jk}}{n_{j.}\left(\frac{N}{N}\right)} \\
&= -\frac{1}{N} \sum_{k=1}^{m} n_{.k} \log \frac{n_{.k}}{N} - \frac{1}{N} \sum_{j=1}^{m} n_{j.} \log \frac{n_{j.}}{N} + \frac{1}{N} \sum_{j=1}^{m} \sum_{k=1}^{m} n_{jk} \log \frac{n_{jk}}{N}.
\end{aligned}$$

The expected value of the estimate for the transmitted information is, therefore,

$$\begin{aligned}
\langle I(N) \rangle &= -\frac{1}{N} \sum_{k=1}^{m} E\left(n_{.k} \log \frac{n_{.k}}{N} | N, p_k \right) - \frac{1}{N} \sum_{j=1}^{m} E\left(n_{j.} \log \frac{n_{j.}}{N} | N, p_j \right) \\
&\quad + \frac{1}{N} \sum_{j=1}^{m} \sum_{k=1}^{m} E\left(n_{jk} \log \frac{n_{jk}}{N} | N, p_{jk} \right).
\end{aligned}$$

Applying Carlton's approximation,

$$\langle I(N)\rangle \cong \sum_{k=1}^{m} p_k f(N,p_k) + \sum_{j=1}^{m} p_j f(N,p_j) - \sum_{j=1}^{m}\sum_{k=1}^{m} p_{jk} f(N,p_{jk}).$$

The probabilities p_j, p_k and p_{jk} are shorthands for the "response probability" $p(Y_j)$, the "stimulus probability" $p(X_k)$ and the "joint probability" $p(Y_k, X_j)$ respectively. In our absolute identification experiments, the probability distribution underlying the presentation of a stimulus tone was uniform, hence

$$p(X_j) = \frac{1}{m}.$$

From the CV-model, the probability distribution governing the response y to a stimulus x for the continuous random variables $x, y \in [0, R]$ was

$$p(y|x) = \frac{1}{I(x)\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-x}{\sigma}\right)^2\right].$$

Extending this to the discrete random variables X_j, Y_k, whereby $X_j \in X = \{X_1, \ldots, X_m\}$ and $Y_k \in Y = \{Y_1, \ldots, Y_m\}$,

$$p(Y_k|X_j) = \frac{1}{I(X_j - \frac{1}{2})\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_k - \left(X_j - \frac{1}{2}\right)}{\sigma}\right)^2\right].$$

From this, we can evaluate $p(Y_k)$ and $p(Y_k, X_j)$ as follows:

$$\begin{aligned} p(Y_k) &= \frac{1}{m}\sum_{j=1}^{m} p(Y_k|X_j) \\ p(Y_k, X_j) &= p(Y_k|X_j)p(X_j). \end{aligned}$$

7.3 APPENDIX III: Increase in $I(N, m, R)$ with m

Theorem

If an m x n confusion matrix is reformatted by merging two columns into one single column, the transmitted information calculated from the resulting m x (n − 1) matrix will be equal to or less than the transmitted information calculated from the original matrix.

Shannon (1948) outlined a proof for this theorem using a priori probabilities for signal transmission and receipt. That is, it was assumed that the probabilities of transmission for each signal were known beforehand, as properties, say, of the English alphabet. Here we offer a proof in the language of stimulus-response matrices, where the probabilities must be obtained a posteriori. The proof we offer differs from that put forward by Shannon in that it follows from the theory of continuous convex functions. The geometric interpretation of a convex function is that if every point on a chord lies above the surface which the chord spans, the mathematical function, ϕ, that defines the curve is convex.

From Hardy, Littlewood & Polya (1964), a continuous function of two variables, Φ is convex if

$$\Phi(\sum_{i=1}^{p} q_i x_i, \sum_{i=1}^{p} q_i y_i) \leq \sum_{i=1}^{p} q_i \Phi(x_i, y_i)$$

where $q_i \geq 0$ and $\sum q_i = 1$. When all the q_i's are equal, we have

$$\Phi(\sum x_i, \sum y_i) \leq \sum \Phi(x_i, y_i) \tag{7.7}$$

From Equation 4.1 in Section 4.3.1, the information, $I(N)$, calculated from an m x m' confusion matrix with the sum of all entries equal to N can be written in a simple form consisting of only 4 terms:

$$N \cdot I(N) = N \log N + \sum_{j=1}^{m} \sum_{k=1}^{m'} n_{jk} \log n_{jk} - \sum_{j=1}^{m} \left(\sum_{k=1}^{m'} n_{jk} \right) \log \left(\sum_{k=1}^{m'} n_{jk} \right) \tag{7.8}$$

$$-\sum_{k=1}^{m'}\left(\sum_{j=1}^{m} n_{jk}\right)\log\left(\sum_{j=1}^{m} n_{jk}\right),$$

where n_{jk} is the value of the element at row j, column k. Expressed simply

$$N \cdot I(N) = N\log N + \text{element entropy} - \text{row entropy} - \text{column entropy}.$$

We consider now two calculated quantities of informations: $I(N)_{merged}$ is the information obtained from the matrix in which a single column has been formed by the merging of 2 adjacent columns, α and $\alpha+1$ of an original matrix. $I(N)_{unmerged}$ is the information obtained from the original matrix.

The quantity of relevance to our theorem is $\Delta I = I(N)_{unmerged} - I(N)_{merged}$. Note that combining two adjacent columns does not affect the row entropies (third term on the right-hand side of Eq. (7.8)). This is because the row entropy depends only on the sum of the entries along the row and doesn't change by merging two adjacent columns. Therefore,

$$\begin{aligned} N \cdot \Delta I \;=\; & [(\text{element entropy})_{unmerged} - (\text{element entropy})_{merged}] \\ & - [(\text{column entropy})_{unmerged} - (\text{column entropy})_{merged}]\,. \end{aligned} \tag{7.9}$$

The difference between the element entropies can be written fully as

$$\sum_{j=1}^{m}\left[\sum_{k=1}^{\alpha-1} n_{jk}\log n_{jk} + n_{j\alpha}\log n_{j\alpha} + n_{j(\alpha+1)}\log n_{j(\alpha+1)} + \sum_{k=\alpha+2}^{m'} n_{jk}\log n_{jk}\right]$$

$$-\sum_{j=1}^{m}\left[\sum_{k=1}^{\alpha-1} n_{jk}\log n_{jk} + (n_{j\alpha}+n_{j(\alpha+1)})\log\,(n_{j\alpha}+n_{j(\alpha+1)}) + \sum_{k=\alpha+2}^{m'} n_{jk}\log n_{jk}\right]$$

$$= \sum_{j=1}^{m} \left[n_{j\alpha} \log n_{j\alpha} + n_{j(\alpha+1)} \log n_{j(\alpha+1)} - (n_{j\alpha} + n_{j(\alpha+1)}) \log\, (n_{j\alpha} + n_{j(\alpha+1)}) \right] .$$

For ease of notation we let $a_j = n_{j\alpha}$ and $b_j = n_{j(\alpha+1)}$. Hence the difference between element entropies can be written succinctly in the form

$$\sum_{j=1}^{m} \left[a_j \log a_j + b_j \log b_j - (a_j + b_j) \log\, (a_j + b_j) \right] . \tag{7.10}$$

The difference between the column entropies can be written fully as

$$\sum_{k=1}^{\alpha-1} \left[\left(\sum_{j=1}^{m} n_{jk} \right) \log \left(\sum_{j=1}^{m} n_{jk} \right) \right] + \left(\sum_{j=1}^{m} n_{j\alpha} \right) \log \left(\sum_{j=1}^{m} n_{j\alpha} \right) + \left(\sum_{j=1}^{m} n_{j(\alpha+1)} \right) \log \left(\sum_{j=1}^{m} n_{j(\alpha+1)} \right)$$

$$+ \sum_{k=\alpha+2}^{m'} \left[\left(\sum_{j=1}^{m} n_{jk} \right) \log \left(\sum_{j=1}^{m} n_{jk} \right) \right]$$

$$- \sum_{k=1}^{\alpha-1} \left[\left(\sum_{j=1}^{m} n_{jk} \right) \log \left(\sum_{j=1}^{m} n_{jk} \right) \right] - \left(\sum_{j=1}^{m} (n_{j\alpha} + n_{j(\alpha+1)}) \right) \log \left(\sum_{j=1}^{m} (n_{j\alpha} + (n_{j(\alpha+1)} \right)$$

$$- \sum_{k=\alpha+2}^{m'} \left(\sum_{j=1}^{m} n_{jk} \right) \log \left(\sum_{j=1}^{m} n_{jk} \right)$$

$$= \left(\sum_{j=1}^{m} n_{j\alpha} \right) \log \left(\sum_{j=1}^{m} n_{j\alpha} \right) + \left(\sum_{j=1}^{m} n_{j(\alpha+1)} \right) \log \left(\sum_{j=1}^{m} n_{j(\alpha+1)} \right)$$

$$- \left(\sum_{j=1}^{m} (n_{j\alpha} + n_{j(\alpha+1)}) \right) \log \left(\sum_{j=1}^{m} (n_{j\alpha} + (n_{j(\alpha+1)}) \right)$$

Making the same substitutions as before, the difference between the column entropies is equal to

$$\left(\sum_{j=1}^{m} a_j\right) \log \left(\sum_{j=1}^{m} a_j\right) + \left(\sum_{j=1}^{m} b_j\right) \log \left(\sum_{j=1}^{m} b_j\right) - \left(\sum_{j=1}^{m} (a_j + b_j)\right) \log \left(\sum_{j=1}^{m} (a_j + b_j)\right) \tag{7.11}$$

We now consider the function

$$\phi(a, b) = a \log a + b \log b - (a + b) \log (a + b) \tag{7.12}$$

We see from equations (7.9), (7.10), (7.11) and (7.12) above that

$$N \cdot \Delta I = \sum_{j=1}^{m} \phi(a_j, b_j) - \phi \left(\sum_{j=1}^{m} a_j, \sum_{j=1}^{m} b_j\right) . \tag{7.13}$$

If we could show that $N \cdot \Delta I \geq 0$, we would prove our proposition. However, to do so, we must first prove a lemma.

Lemma 1

If $Q \equiv \phi_{aa}\, u^2 + 2\phi_{ab}\, uw + \phi_{bb}\, w^2$, then $Q \geq 0$, where $\phi_{ij} = \frac{\partial^2 \phi}{\partial i \partial j}$.

From the defining function (7.12),

$\phi_{aa} = b/\,[a(a+b)]; \qquad \phi_{ab} = -1/(a+b)\ ; \qquad \phi_{bb} = a/[b(a+b)]$

Therefore,

$$Q = \frac{1}{a+b} \left[\sqrt{\frac{b}{a}} u - \sqrt{\frac{a}{b}} w\right]^2 \geq 0$$

since a and b are greater than zero by definition, which proves the lemma.

Now Hardy et al (1964) have shown that a necessary and sufficient condition for the convexity of the function ϕ on an open domain is that $Q \geq 0$, which we have now proved.

Therefore we know from Equation (7.7) defining convex functions that

$$\sum_{j=1}^{m} \phi(a_j, b_j) - \phi\left(\sum_{j=1}^{m} a_j,\ \sum_{j=1}^{m} b_j\right) \geq 0\ .$$

We also know, however, from Eq. (7.13) that the latter inequality is the condition for which $N \cdot \Delta I \geq 0$ and hence $\Delta I \geq 0$. Thus the theorem is proved.

In evaluating the change in transmitted information by merging two columns into one, row entropies cancel (Equation (7.9)). It follows that column entropies would cancel in the merging of two rows into one. Hence, merging two rows will produce the same effect on the transmitted information as merging two columns. It will also follow from the proof of this theorem that any increase in the number of rows or columns of the confusion matrix produced by a process of dividing the set of rows and columns of a previous matrix will always lead to an increase in the calculated information.

Bibliography

[1] Berliner, J. E., Durlach, N. I., & Braida, L. D. (1978). Intensity perception. IX. Effect of a fixed standard on resolution in identification. *Journal of the Acoustical Society of America*, **64**, 687-689.

[2] Borg, G., Diamant, H., Strom, L., & Zotterman, Y. (1967). The relation between neural and perceptual intensity: A comparative study on the neural and psychophysical response to taste stimuli. *Journal of Physiology*, **192**, 13-20.

[3] Braida, L. D., & Durlach, N. I. (1972). Intensity perception. II. Resolution in one-interval paradigms. *Journal of the Acoustical Society of America*, **51**, 483-502.

[4] Braida, L. D., & Durlach, N. I. (1988). Peripheral and central factors in intensity perception. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Auditory function* (pp. 559-583). New York: Wiley.

[5] Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception. XIII. Perceptual anchor model of context-coding. *Journal of the Acoustical Society of America*, **76**, 722-731.

[6] Carlton, A. G. (1969). On the bias of information estimates. *Psychological Bulletin*, **71**(2), 108-109.

[7] Coren, S., & Ward, L. M. (1989). *Sensation and Perception.* San Diego, CA: Harcourt Brace Jovanovich, Inc.

[8] Doucet, J. R., & Relkin, E. M. (1997). Neural contributions to the perstimulus compound action potential: implications for measuring the growth of the auditory nerve spike count as a function of stimulus intensity. *Journal of the Acoustical Society of America*, **101**(5), 2720-2734.

[9] Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, **46**, 372-383.

[10] Erikson, C. W., & Hake, H. W. (1955). Absolute judgments as a function of stimulus range and number of stimulus and response categories. *Journal of Experimental Psychology*, **49**, 323-332.

[11] Freund, J. E., & Walpole, R. E. (1980). *Mathematical Statistics. 3rd ed.* Englewood Cliffs, N. J.: Prentice-Hall, Inc.

[12] Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology*, **46**, 373-380.

[13] Garner, W. R., & Hake, H. W. (1951). The amount of information in absolute judgments. *Psychological Review*, **58**, 446-459.

[14] Hardy, W. R., Littlewood, J. E., & Polya, G. (1964). *Inequalities.* Cambridge University Press.

[15] Houtsma, A. J. M. (1983). Estimation of mutual information from limited experimental data. *Journal of the Acoustical Society of America*, **74**, 1626-1629.

[16] Luce, R. D. (1977). Thurstone's discriminal processes fifty years later. *Psychometrika*, **42**, 461-489.

[17] Luce, R. D., Green, D. M., & Weber, D. L. (1976). Attention bands in absolute identification. *Perception & Psychophysics*, **20**, 49-54.

[18] Luce, R. D., & Mo, S. S. (1965). Magnitude estimation of heaviness and loudness by individual subjects: A test of a probabilistic response theory. *The British Journal of Mathematical and Statistical Psychology*, **18** (Part 2), 159-174.

[19] MacRae, A. W. (1970). Channel capacity in absolute judgment tasks: an artifact of information bias? *Psychological Bulletin*, **73**, 112-121.

[20] Miller, G. A. (1955). Note on the bias of information estimates. In H. Quastler (Ed.), *Information theory in psychology: problems and methods* (pp.95-100). Glencoe, IL: Free Press.

[21] Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **65**, 81-97.

[22] Milton, S. J. (1992). *Statistical Methods in the Biological and Health Sciences. 2nd ed.* New York: McGraw-Hill, Inc.

[23] Norwich, K. H. (1993). *Information, sensation, and perception.* San Diego, CA: Academic Press.

[24] Norwich, K. H., Wong, W., & Sagi, E. (1998). Range as a factor in determining the information of loudness judgments: overcoming small sample bias. *Canadian Journal of Experimental Psychology*, in press.

[25] Relkin, E. M., & Doucet, J. R. (1997). Is loudness simply proportional to the auditory nerve spike count? *Journal of the Acoustical Society of America*, **101**(5), 2735-2740.

[26] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**, 379-423.

[27] Stevens, S. S. (1956). The direct estimation of sensory magnitudes-loudness. *American Journal of Psychology*, **69**, 1-25.

[28] Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273-286.

[29] Wiener, N. (1948). *Cybernetics*. Cambridge, MA: The Technology Press, MIT.

[30] Wong, W., & Norwich, K. H. (1997). Simulation of human sensory performance. *Biosystems*, **43**, 189-197.