CHAPTER 1

1.1 Introduction

Perception is the information processing of sensory data in the nervous system that leads to a conscious experience reflecting the outside world. We cannot, of course, measure conscious experience directly but we can define a perceptual task and record the observers response to the task. The response can take many forms: it can be a verbal response, it can be in the form of reaching one's arm towards an object or it could be an eye-movement. However, here we will consider only categorical responses, where the observer selects a discrete response from a set of one or more response options. A single stimulus-response pair is called an experimental *trial*.

The perceptual process is noisy. This may seem surprising as our perceptual experience is often clear and vivid. When you look at the letters on this page you see them clearly and perception does not seem noisy. That is because the signal-to-noise ratio of these letters are far above your *perceptual threshold*, the minimal signal-to-noise ratio or stimulus intensity, at which you can identify the letters. If you saw the letters at low contrast or only for a very brief amount of time then the signal-to-noise ratio could be near your perceptual threshold and you would make errors. These errors are stochastic. Even when the stimulus is carefully controlled so that it is identical across multiple trials, the observer's response may not be the same.

We want to quantify an observer's performance in a perceptual task. We will start by calculating the response proportion $\frac{n_r}{N}$ where n_r is the number of trials where the observer chose response category r and N is the total number of trials for a particular stimulus. The response proportion for the correct response option is called the *proportion correct*. At first, we might think that the proportion correct is a reasonable measure of the observer's perceptual sensitivity. After all, we will expect that an observer that performs well will have a high perceptual sensitivity. However, even this simple measure is not just influenced by perceptual sensitivity but also by the observer's *response bias*.

Pure tone audiometry will serve as an example to illustrate the problem in using the proportion correct as a measure of perceptual sensitivity. In this test sinusoidal (pure) tones at various frequencies are played at a certain sound intensity level. You have probably experienced a similar test as it is used as a screening method for frequency specific hearing loss. Every time a sound is played the observer can answer 'yes' to indicate that she could hear the sound or 'no' to indicate that she could not hear the sound. Now assume that an observer answers 'yes' in 100% of the trials in which a tone was played at a certain sound intensity. We may think that this means that the observer has a high perceptual sensitivity but we might well be mistaken: perhaps the observer did not really care to perform well and just responded 'yes' all the time. In order to check whether this was the case we should look at trials in which no sound was played: if the observer also responds 'yes' to all of those trials then we know that the observer was strongly biased towards 'yes'-responses and we have learned nothing about the observers perceptual sensitivity. In order to estimate the observer's perceptual sensitivity we need, in other words, to compare the rate of *true negative* (TN) responses (responding 'no' when no sound was played) *and* the rate of *true positive* (TP) responses (responding 'yes' when a sound was played).

1.2 Signal detection theory

Signal detection theory is a theoretical framework commonly used in cognitive science to separate the cognitive processes of perception and response selection. Perception is modelled as the *encoding* of the stimulus onto a scalar *internal representation*, x. Response selection is modeled as decoding the internal representation, x, to a response category, r.

The encoding process is noisy so that noise, typically assumed to be Gaussian, is added to the encoding x of the stimulus s, so that

$$p\left(x \mid s\right) = f\left(x \mid \mu_s, \sigma^2\right) = \phi\left(\frac{x-\mu_s}{\sigma}\right)$$

where s denotes the stimulus, f denotes the normal probability density and ϕ denotes the *standard* normal probability function for which $\mu = 0$ and $\sigma = 1$.

Importantly, the noise is also present when the stimulus, or signal, is not present so that

$$p\left(x \mid s_0\right) = f\left(x \mid \mu_0 = 0, \sigma_0 = 1\right) = \phi\left(x\right)$$

where $\sigma = 0$ denotes the null stimulus.

The choice of setting $\mu_0 = 0$ and $\sigma_0 = 1$ is made to ensure that the model is identifiable. We need to define the origin (zero) of the internal representation, x. Choosing $\mu_0 = 0$ as the origin seems reasonable: It means that when no stimulus, or signal, is presented the distribution of x is, logically, centered around zero. Setting the $\sigma_0 = 1$ sets the scale, or unit, of the model.

Decoding the internal representation, x, to a response option, is modeled by a threshold function, so that observers respond 'no' if $x < c$ and 'yes' if $x > c$ where c is the threshold, or, *response criterion*. We can thus calculate the response probability, $P\left(r = no \mid s\right)$, of a false negative response as the probability mass of $p\left(x \mid s\right)$ that lies below the criterion c as

$$P\left(r = no \mid s\right) = P\left(x < c \mid s\right) = \int_{-\infty}^{c} p\left(x \mid s\right) dx = \Phi\left(\frac{c-\mu_s}{\sigma}\right) \tag{1.1}$$

where Φ denotes the standard normal cumulative distribution function.

Likewise, we can calculate the response probability, $P\left(r = no \mid s_0\right)$, of a true negative response as the probability mass of $p\left(x \mid s_0\right)$ that lies below the criterion c as

$$P\left(r = no \mid s_0\right) = P\left(x < c \mid s_0\right) = \int_{-\infty}^{c} p\left(x \mid s_0\right) = \Phi\left(c\right) \tag{1.2}$$

The true negative and false negative responses are both 'no'-responses but whereas the true negative response are correct 'no'-responses when no stimulus was presented, the false negative responses are incorrect 'no'-responses when a stimulus *was* presented. We can, of course, also calculate the 'yes'-response probabilities by integration but it is simpler to note that if 'yes' and 'no' are the only response options then

$$P\left(r = no \mid s\right) + P\left(r = yes \mid s\right) = 1$$
$$P\left(r = no \mid s_0\right) + P\left(r = yes \mid s_0\right) = 1$$

From this we can derive

$$P\left(r = yes \mid s\right) = 1 - P\left(r = no \mid s\right) = 1 - \Phi\left(\frac{c-\mu_s}{\sigma}\right) = \Phi\left(\frac{\mu_s - c}{\sigma}\right) \tag{1.3}$$

$$P\left(r = yes \mid s_0\right) = 1 - P\left(r = no \mid s_0\right) = 1 - \Phi\left(c\right) = \Phi\left(-c\right) \tag{1.4}$$

where we have used the identity $1 - \Phi\left(x\right) = \Phi\left(-x\right)$.

1.2.1 The equal variance model

An additional simplifying assumption commonly used in signal detection theory is the *equal variance model*, which assumes that $\sigma = \sigma_0 = 1$. This model is illustrated in Figure 1.1 where the overlap between the two Gaussian probability distributions $p\left(x \mid s_0\right)$ and $p\left(x \mid s\right)$ have been filled in two shades of grey. These grey areas represent the probabilities of making an error: The dark grey area indicates the

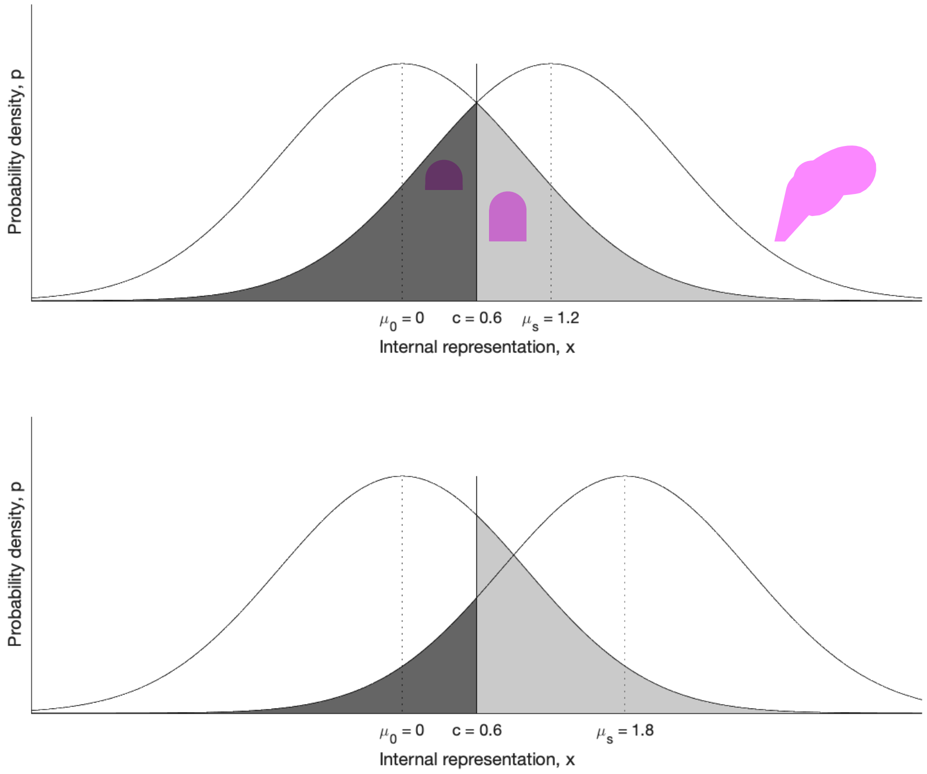


Figure 1.1: The equal variance model from signal detection theory. In the top panel the criterion of the observer, c, is set to the optimal value $c = \frac{\mu_s}{2}$. In the lower panel the criterion, c, is set to bias the observer towards 'yes'-responses. This criterion is suboptimal because it increases the total probability of making an error. The area shaded in dark represents the probability of making a false negative error. The area shaded in lighter grey represents the probability of making a false positive error..

probability, $P(r = no \mid s)$, of a false negative response and the lighter grey area indicates the probability, $P(r = yes \mid s_0)$, of a false positive response.

In the model illustrated in the top panel of Figure 1.1 the criterion, c, is set to $c = \frac{\mu_s}{2} = 0.6$ mid between to two distributions. This is the optimal criterion in terms of maximising the probability, P_{corr}, of a correct response *if* the probability that a trial contains a stimulus is equal to the probability that a trial does not contain a stimulus. This can be shown by calculating the probability, P_{corr}, of a correct response as

$$P_{corr} = P(r = yes \mid s)\, P(s) + P(r = no \mid s_0)\, P(s_0) = \Phi(mu_s - c) + \Phi(c)$$

and finding

$$\arg\max_c (P_{corr}) = \frac{\mu_s}{2}$$

using differentiation.

We can also see that $c = \frac{\mu_s}{2}$ is the optimal criterion by inspecting the lower panel of Figure 1.1, which illustrates the effect of shifting the criterion to a value lower than $c = \frac{\mu_s}{2}$. This increases the probability of a 'yes'-response, which increases the probability of a false positive error. It also decreases the probability of a false negative error but this decrease is less than the increase in the probability of a false positive because $p(x \mid s_0) > p(x \mid s)$ when $x < \frac{\mu_s}{2}$.

For the unbiased observer model in the top panel of Fig. 1, the amount of overlap of the two distributions, and hence the probability of making an error, depends only on the distance μ between the distributions and this distance is thus a measures of *perceptual sensitivity*. Therefore the distance μ is often denoted as $d' = \mu$ only for the equal variance model. Using d' as a measure of perceptual sensitivity is commonly used in the cognitive science literature.

To further illustrate the role of d' as a measure of perceptual sensitivity we will look at some extreme case. In the extreme case when $\mu = 0$, the two distributions overlap completely, so that $p(x \mid s) = p(x \mid s_0)$. In this case the response probabilities for each response category will be the same, so that $P(r = no \mid s) = P(r = no \mid s_0)$ and $P(r = yes \mid s) = P(r = yes \mid s_0)$. This extreme case of $d' = \mu_s = 0$ means that the observer was unable to perceive the stimulus. In the other extreme case when $d' >> 1$ the overlap will be small and the probability of making an error similarly small. This, confirms that $d' = \mu_s$ is a reasonable measure of sensitivity under the assumptions of the model.

The validity of using $d' = \mu_s$ as a measure of sensitivity depends on the validity of the equal variance assumption. Nevertheless, d' is often used as a measure of sensitivity in the cognitive science literature without testing whether the the equal variance assumption holds.

Note that there is an unfortunate confusion of terminology between the signal detection theory literature and the machine learning literature. In the machine learning literature sensitivity typically refers to the true positive probability $P(r = yes \mid s)$ or an estimate thereof. It is a reasonable way the word in this way as a very sensitive sensor would detect the stimulus with a high chance of success possibly at the cost of a high chance of making false alarms. Hence, sensitivity in the machine learning literature is influenced by response bias as a high probability of 'yes'-responses lead to a high sensitivity in this sense of the word. Here, I have used the term *perceptual sensitivity* to denote sensitivity in terms of signal detection theory even though the term used in the cognitive science literature is often simply 'sensitivity'.

1.2.2 Parameter estimation for the equal variance model

So far, we have described how to calculate response probabilities from the model parameters d' and c. In practice, we would face the inverse problem of estimating the model parameters from data. In that case we can estimate the underlying response probabilities from response proportions

$$\hat{P}(r = yes \mid s) = \frac{n_{tp}}{N_s}$$

$$\hat{P}(r = yes \mid s_0) = \frac{n_{fp}}{N_{s_0}}$$

where n_{tp} is the number of true positive responses, n_{fp} is the number of false positive responses, N_s is the number of trials in which a stimulus was presented and N_{s_0} is the number of trials in which no stimulus was presented.

We can now isolate c and μ_s from Equations 1.3 and 1.4 using the inverse standard normal cumulative distribution function, also known as the *probit* function, $\Phi^{-1}(P) = x$ for the equal variance model in which $\sigma = 1$

$$\hat{\mu_s} - \hat{c} = \Phi^{-1}\left(\hat{P}(r = yes \mid s)\right) = \Phi^{-1}\left(\frac{n_{tp}}{N_s}\right) \tag{1.5}$$

$$-\hat{c} = \Phi^{-1}\left(\hat{P}(r = yes \mid s_0)\right) = \Phi^{-1}\left(\frac{n_{fp}}{N_{s_0}}\right) \tag{1.6}$$

By multiplying Equation 1.6 by -1 gives us the expression for $\hat{c}$

$$\hat{c} = -\Phi^{-1}\left(\frac{n_{fp}}{N_{s_0}}\right)$$

Adding this equation expression to Equation 1.5 gives us the expression for d'

$$d' = \hat{\mu_s} = \Phi^{-1}\left(\frac{n_{tp}}{N_s}\right) - \Phi^{-1}\left(\frac{n_{fp}}{N_{s_0}}\right)$$

Note that this last expression is undefined when either of the following conditions are met

$$\begin{aligned} n_{tp} &= 0 \\ n_{fp} &= 0 \\ n_{tp} &= N_s \\ n_{fp} &= N_{s_0} \end{aligned}$$

In order to estimate the model parameters, d', and c, the observer must, in other words, use both response categories, for trials containing a stimulus and for trials containing no stimulus. The typical problem is that the detection task is too easy so that the observer makes no errors. This means that the perceptual sensitivity, d', is large but we cannot estimate how large it is. The literature contains poor heuristic approaches in which an arbitrary small number is added or subtracted to to n_{tp} or n_{fp} to solve this problem but the better solutions is to discard the data. If more data is needed then the experiment should be repeated using a stimulus that has been designed to ensure that the observer use both response categories.

Note that there is no closed form solution to the probit function, $\Phi^{-1}(P)$ but most analysis software contain a function for calculating it.

1.2.3 Equal variance model exercise

Simulate responses from 100 experiments with 3 observers each completing 50 trials containing a stimulus and 50 trials containing no stimulus. All three observers behave according to the equal variance model and have a perceptual sensitivity of $d' = 1$ but they have different response criteria: One observer is biased towards ‘yes’-responses, one is biased towards ‘no’-responses, and one is not very strongly biased towards ‘yes’- or ‘no’-responses.

Estimate the perceptual sensitivity for each experiment and each observer. Plot the distribution of the perceptual sensitivity across trials for each of the three observers. Are the distributions centered around the correct estimate for d' each of the three simulated data sets?

Simulate responses from 100 experiments with 3 observers each completing 50 trials. All three observers behave according to a model $\mu_s = 1$ and $\sigma = 0.8$ but they have different response criteria: One observer is biased towards ‘yes’-responses, one is biased towards ‘no’-responses, and one is not very strongly biased towards ‘yes’- or ‘no’-responses.

Assume that you did not know that the data came from an observer for which the equal variance assumption $\sigma = 1$ does not hold and estimate $d' = \mu_s$ using the equal variance model. Plot the distribution of the perceptual sensitivity across trials for each of the three observers. Do you get the correct estimate for each of the three simulated data sets?

What are the implications of your results?

1.2.4 Unequal variance model

Releasing the constraint, $\sigma = 1$, of the equal variance model leads to the more complex unequal variance model, which is depicted in Figure 1.2. For this model, the degree of overlap between the probability densities $p(x \mid s_0)$ and $p(x \mid s)$ and hence the total probability of an error depends not only on the mean, μ_s, but also on the standard deviation, σ, of $p(x \mid s)$. This can be seen in Figure 1.2. In the top panel, $p(x \mid s)$ is more narrow than for the equal variance model depicted in Figure 1.1 because $\sigma = 0.85$ and the overlap between the distributions is therefore smaller. In the lower panel, $p(x \mid s)$ is more wide than for the equal variance model because $\sigma = 1.8$ and the overlap between the distributions is therefore

Adding this equation expression to Equation 1.5 gives us the expression for d'

$$d' = \hat{\mu_s} = \Phi^{-1}\left(\frac{n_{tp}}{N_s}\right) - \Phi^{-1}\left(\frac{n_{fp}}{N_{s_0}}\right)$$

Note that this last expression is undefined when either of the following conditions are met

$$n_{tp} = 0$$
$$n_{fp} = 0$$
$$n_{tp} = N_s$$
$$n_{fp} = N_{s_0}$$

In order to estimate the model parameters, d', and c, the observer must, in other words, use both response categories, for trials containing a stimulus and for trials containing no stimulus. The typical problem is that the detection task is too easy so that the observer makes no errors. This means that the perceptual sensitivity, d', is large but we cannot estimate how large it is. The literature contains poor heuristic approaches in which an arbitrary small number is added or subtracted to to n_{tp} or n_{fp} to solve this problem but the better solutions is to discard the data. If more data is needed then the experiment should be repeated using a stimulus that has been designed to ensure that the observer use both response categories.

Note that there is no closed form solution to the probit function, $\Phi^{-1}(P)$ but most analysis software contain a function for calculating it.

1.2.3 Equal variance model exercise

Simulate responses from 100 experiments with 3 observers each completing 50 trials containing a stimulus and 50 trials containing no stimulus. All three observers behave according to the equal variance model and have a perceptual sensitivity of $d' = 1$ but they have different response criteria: One observer is biased towards 'yes'-responses, one is biased towards 'no'-responses, and one is not very strongly biased towards 'yes'- or 'no'-responses.

Estimate the perceptual sensitivity for each experiment and each observer. Plot the distribution of the perceptual sensitivity across trials for each of the three observers. Are the distributions centered around the correct estimate for d' each of the three simulated data sets?

Simulate responses from 100 experiments with 3 observers each completing 50 trials. All three observers behave according to a model $\mu_s = 1$ and $\sigma = 0.8$ but they have different response criteria: One observer is biased towards 'yes'-responses, one is biased towards 'no'-responses, and one is not very strongly biased towards 'yes'- or 'no'-responses.

Assume that you did not know that the data came from an observer for which the equal variance assumption $\sigma = 1$ does not hold and estimate $d' = \mu_s$ using the equal variance model. Plot the distribution of the perceptual sensitivity across trials for each of the three observers. Do you get the correct estimate for each of the three simulated data sets?

What are the implications of your results?

1.2.4 Unequal variance model

Releasing the constraint, $\sigma = 1$, of the equal variance model leads to the more complex unequal variance model, which is depicted in Figure 1.2. For this model, the degree of overlap between the probability densities $p(x \mid s_0)$ and $p(x \mid s)$ and hence the total probability of an error depends not only on the mean, μ_s, but also on the standard deviation, σ, of $p(x \mid s)$. This can be seen in Figure 1.2. In the top panel, $p(x \mid s)$ is more narrow than for the equal variance model depicted in Figure 1.1 because $\sigma = 0.85$ and the overlap between the distributions is therefore smaller. In the lower panel, $p(x \mid s)$ is more wide than for the equal variance model because $\sigma = 1.8$ and the overlap between the distributions is therefore

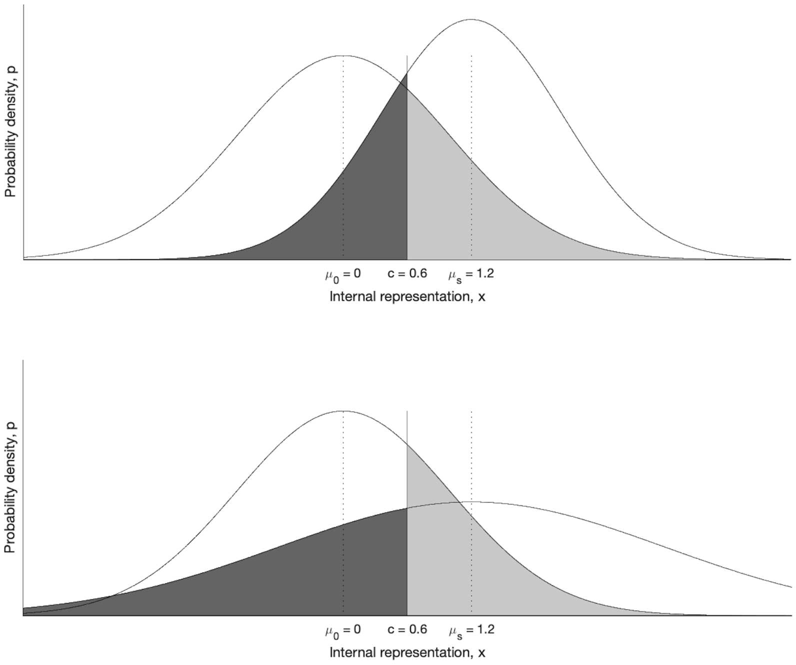


Figure 1.2: The unequal variance model from signal detection theory. The areas shaded in dark represents the probability of making a false negative error. The areas shaded in lighter grey represents the probability of making a false positive error. In the top panel the standard deviation of $p\,(x \mid s)$ is $\sigma = 0.85 < 1$ and therefore the probability, $p\,(r = no \mid s)$, of a false negative is smaller than than for the equal variance model depicted in the top panel of Figure 1.1. In the lower panel the standard deviation of $p\,(x \mid s)$ is $\sigma = 1.8 > 1$ and therefore the probability, $p\,(r = no \mid s)$, of a false negative is greater than than for the equal variance model depicted in the top panel of Figure 1.1. Note that the criterion is the same for the models depicted here and that depicted in the top panel of Figure 1.1. The probability, $p\,(r = yes \mid s_0)$, of false positive is therefore the same for all three models..

greater. Perceptual sensitivity is therefore not specified only by $d' = \mu_s$ but also depends on the standard deviation, σ, for the unequal variance model. Note, that decoding the internal representation, x, onto a response category depends only on the criterion, c, for both the equal and unequal variance model.

The unequal variance model is specified by three parameters: μ_s, and σ but the yes/no-detection paradigm only provides two independent equations, Equation 1.3 and 1.4, relating response probabilities to parameter values. Equation 1.4 can be solved for the response criterion, c, but we are left with one equation (1.3) for determining the two parameters, μ_s and σ, that specifies perceptual sensitivity. The unequal variance model is thus under-determined for the yes/no-paradigm and we need at least one more equation to solve it.

1.2.5 The Receiver Operating Characteristics (ROC) curve

We can view Equations 1.3-1.4 in a different way by using the probit transform on both sides and rearrange to

$$\Phi^{-1}\left(P\left(r = yes \mid s\right)\right) = -\frac{1}{\sigma}c + \frac{\mu_s}{\sigma} \tag{1.7}$$

$$\Phi^{-1}\left(P\left(r = yes \mid s_0\right)\right) = -c \tag{1.8}$$

Inserting Equation 1.8 into Equation 1.7 presents the the problem as a single linear equation

$$\Phi^{-1}\left(P\left(r = yes \mid s\right)\right) = \frac{1}{\sigma}\Phi^{-1}\left(P\left(r = yes \mid s_0\right)\right) + \frac{\mu_s}{\sigma} \tag{1.9}$$

Equation 1.9 is the probit transformed *Receiver Operating Characteristics* (ROC) curve. Examples of probit transformed ROC curves are shown in the upper panels of Figure 1.3.

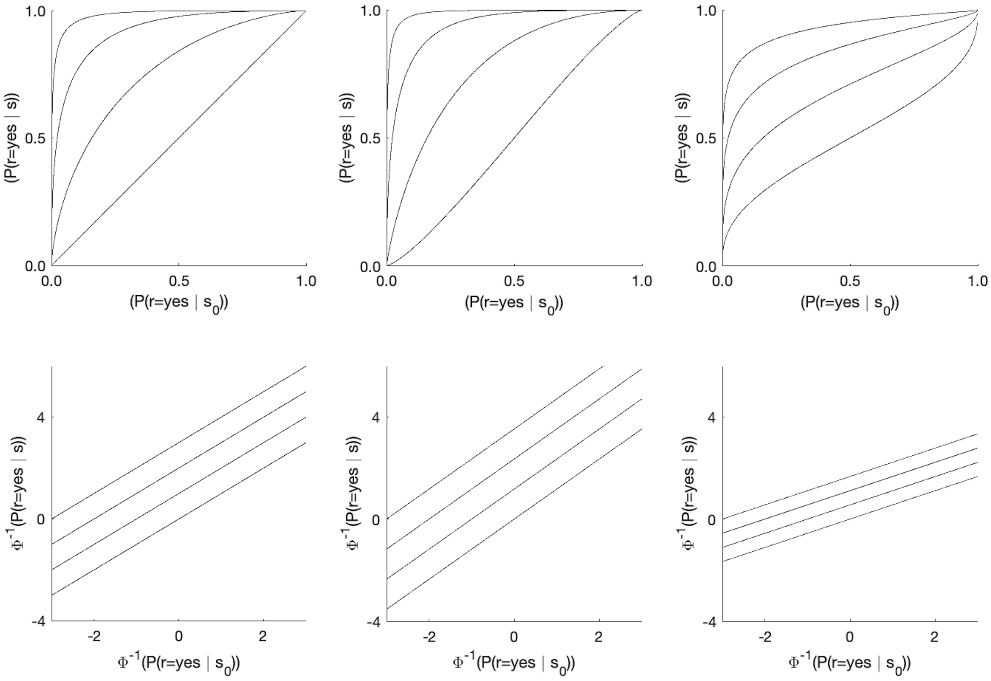


Figure 1.3: The Receiver Operating Characteristics (ROC). The top panels depicts ROC curves for values of $\sigma = 1$ (left), $\sigma = 0.85$ (middle) and $\sigma = 1.8$ (left). The lower panels show the same ROC curves for probit transformed probabilities as described in Equation 1.9. From bottom to top, each curve in every panel shows the ROC curve for values of $\mu_s = 0, 1, 2,$ and 3.

The lower panels of Figure 1.3 shows ROC curves that are not probit transformed. This is the typical way of depicting ROC curves. The free parameters, μ_s and σ of the ROC curve are related only to perceptual sensitivity while the independent variable $\Phi^{-1}\left(P\left(r = yes \mid s_0\right)\right) = -c$ is related only to the response criterion, c. The shape of ROC curves thus reflects the perceptual sensitivity of the observer while a single point on a specific ROC curve reflects the response criterion, c, of the observer.

The yes/no-paradigm provides us with a single point on the probit transformed ROC curve, not enough to fit a line. This is only possible under the equal variance assumption for which the slope of the probit transformed ROC curve is fixed to $\frac{1}{\sigma} = 1$. For the unequal variance model, we need another point, i.e. another response criterion in order to fit the curve. The most efficient way of obtaining another point on the ROC curve is to offer an additional response category such as 'maybe' to the observer. In this paradigm the observer will have to use two response criteria. One criterion separates the 'yes'- and 'maybe'-categories. This gives us the point on the ROC transformed ROC curve described by Equation 1.9. The other criterion separates the 'no'- and 'maybe'-categories. For this criterion both 'yes' and 'maybe' responses count as positives. Hence the additional point on the gives us the additional point on the probit transformed ROC curve described by Equation 1.10,

$$\Phi^{-1}\left(P\left(r = yes \vee r = maybe \mid s\right)\right) = \frac{1}{\sigma}\Phi^{-1}\left(P\left(r = yes \vee r = maybe \mid s_0\right)\right) + \frac{\mu_s}{\sigma} \tag{1.10}$$

which can be used with Equation 1.9 to fit the unequal variance model to observed response proportions.

The relation between the shape of the ROC curve and percpetual sensitivity is well captured by the *Area Under the Curve* (AUC) of the ROC curve. It can be shown that the AUC equals the proportion correct of an unbiased observer in a two-alternative forced choice task in which the observer must choose the one display, out of two, that contains the stimulus, s. This relationship is particularly clear in the equal variance ROC curve displayed in the top left panel of Figure 1.3. For $d' = \mu_s = 0$ the ROC curve is the line $P\left(r = yes \mid s\right) = P\left(r = yes \mid s_0\right)$, which divides the plane in half so that the proportion correct, P_c =AUC=0.5. This means that the observer is at chance level of a correct answer, which is what we should expect for an unbiased observer with zero perceptual sensitivity, $d' = 0$. As d' increases the ROC curve bulges further and further towards the top left corner where $P\left(r = yes \mid s_0\right) = 0$ and $P\left(r = yes \mid s\right) = 1$), so that the AUC contains the entire plane. This means that the proportion correct, P_c, increases towards $P_c = 1$, which is what we should expect for an observer with great perceptual sensitivity, $d' >> 1$.

The equation P_c=AUC holds not only for the Gaussian models we have described here but also for models based on arbitrary probability densities. The AUC is therefore a better measure of perceptual sensitivity than e.g. d' because it is more general measure. Note however, that in order to estimate the AUC we need to know the full shape of the ROC, which require some assumptions on the underlying probability densities. We can, of course, interpolate the ROC curve between empirically obtained points but the choice of interpolation method translates to making assumptions on the underlying probability densities.

The AUC is often used to quantify the performance of a classifier in the machine learning literature. For some classification algorithms, it is possible to vary the bias very precisely and at low compuational cost and thereby obtain many points on the ROC. This, in turn, means that the distributions $P\left(r = yes \mid s\right)$ and $P\left(r = yes \mid s_0\right)$ are well described empirically and the effect of the choice of interpolation method is negligible. However, in cognitive science, we cannot ask the observer to vary the bias in a very fine-grained manner. As an example, imagine if you had to rate your confidence in hearing a sound on a 1-100 scale: you would probably struggle with distinguishing between, e.g. confidence levels 77 and 78 and you might choose to use only some anchor points on the scale. Even if you could distinguish meaningfully between 100 confidence levels it would require many trials before you would actually use all the levels. Therefore, in the cognitive science literature, the observer is rarely offered more than seven response categories.

1.2.6 Parameter estimation for the unequal variance model

Equations 1.9 and 1.10 gives us to points on the linear probit transformed ROC curve. In order to fit the line we must calculate the corresponding response proportions. For estimating the point on the line given by Equation 1.9 we can estimate the response probabilities as

$$\hat{P}\left(r = yes \mid s\right) = \frac{n_{yes}}{N_s}$$
$$\hat{P}\left(r = yes \mid s_0\right) = \frac{n_{yes}}{N_{s_0}}$$

For estimating the point on the line given by Equation 1.10 we can estimate the response probabilities as

$$\hat{P}\left(r = yes \vee r = maybe \mid s\right) = \frac{n_{yes} + n_{maybe}}{N_s}$$
$$\hat{P}\left(r = yes \vee r = maybe \mid s_0\right) = \frac{n_{yes} + n_{maybe}}{N_{s_0}}$$

where n_{maybe} is the number of ‘maybe’-responses.

By fitting the linear probit transformed ROC curve we can estimate its slope $\frac{1}{\sigma}$ and its intercept $\frac{\mu_s}{\sigma}$. The final stage is to isolate σ and μ_s from these expressions.

1.2.7 Unequal variance model exercise

In a confidence rating task the observer can indicate her confidence as ’high’ or ’low’, in addition to answering ‘yes’ or ‘no’. Simulate responses from 100 experiments with one observers completing 50 trials containing a stimulus and 50 trials containing no stimulus. The observer behave according to the unequal variance model $\mu_s = 1$ and $\sigma = 0.8$.

Estimate the parameters of the model for each experiment. Plot the distribution of the parameters across experiments. Are the distributions of the parameters centered around the correct estimate?

Compare your results to the equal variance model exercise in 1.2.3.

1.3 The psychometric function

In the signal detection tasks discussed in the previous sections, only two stimuli, s and s_0, are presented to the observer. We can extend these tasks to include multiple stimuli. The psychometric function $\Psi(I_s)$, where I_s is the physical stimulus intensity can be used to quantify an observer’s perceptual ability in this type of tasks. When the signal intensity is encoded onto a one-dimensional internal representation, x, the psychometric function returns the probability of a ‘yes’-response for a certain stimulus intensity so that $\Psi(I_s) = P(r = yes \mid I_s)$. In this case, the equal variance signal detection model can be extended to a model of the psychometric function as illustrated in Figure 1.4.

In the lower panel of Figure 1.4 we see the internal representation of the perceived stimulus intensity. Just as for the equal variance signal detection model, the value of the internal representation varies randomly due to Gaussian noise added in the encoding process with the mean, μ_s, of the Gaussian noise determined by the true, physical, stimulus intensity. Also, as in signal detection theory, the observer responds ‘yes’ when the perceived stimulus intensity exceeds a criterion. The psychometric function can thus be derived from Equation 1.4

$$\Psi(I_s) = P(r = yes \mid I_s) = \Phi\left(\frac{I_s - c}{\sigma}\right) \tag{1.11}$$

The psychometric function is illustrated in the top panel of Figure 1.4. Note that it takes the physical stimulus intensity, I_s, as input. The parameters, the criterion, c and the standard deviation, σ are thus also in physical units.

The criterion, c, is the stimulus intensity for which $\Psi(I_s) = P(r = yes \mid I_s) = 0.5$. This is also called the 50%-threshold. Shifting the criterion shifts the entire psychometric function without altering its shape.

The standard deviation, σ, reflects the perceptual sensitivity of the observer. Low values of σ means that little noise is added in the perceptual process, so that the encoding of the stimulus intensity is very precise. This means that the psychometric function is steep: A small increase in stimulus intensity can increase the perceived stimulus intensity above the criterion so that the observer will detect the stimulus consistently. Hence the perceptual sensitivity is high when σ is small. This might seem different from the equal variance signal detection model described in previous sections. For this model, the sensitivity was quantified entirely by the distance between the probability densities $p(r = yes \mid s_0)$ and $p(r = yes \mid s)$. However, this distance was measured relative to the standard deviation, σ, which was set to $\sigma = 1$. In the internal representation underlying the psychometric function illustrated in the lower panel of Figure 1.4, the standard deviation is in physical units to ensure correspondence with the psychometric function. Hence, decreasing the standard deviation, σ, increases the distance between $p(r = yes \mid s_0)$ and $p(r = yes \mid s)$ in units of standard deviations, which means that the perceptual sensitivity is lowered.

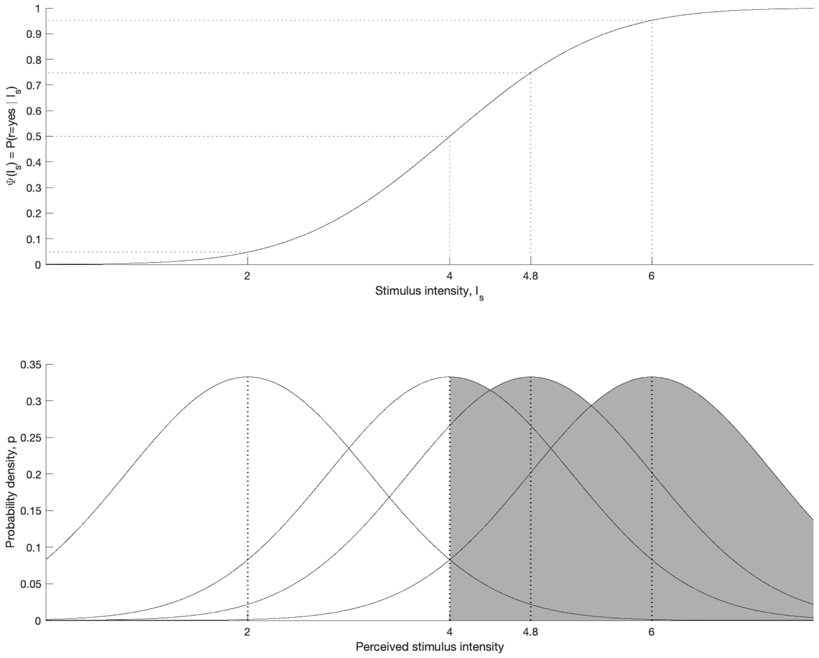


Figure 1.4: The psychometric function. The top panel depicts the cumulative Gaussian psychometric function. Note that the x-axis should be in physical units. The lower panel depict the noisy internal representation of stimulus intensity as the underlying model for the Gaussian psychometric function. The shaded area under each probability density function represents the probability of a ‘yes’-response. The criterion of the observer is set to a perceived intensity value of 4 corresponding to the 50%-threshold of the psychometric function.

Signal detection theory, including the psychometric function described above, rests on two assumptions. First, the stimulus must be encoded onto a one-dimensional internal representation, x. Second, the response categories must be ordered on the internal representation, meaning that the internal representation axis can be divided into response categories by criterion values. In many tasks these assumptions do not hold. Take for instance a letter identification task. The letters are high-dimensional stimuli and so is their internal representation. There is no one-dimensional internal representation, for which the response categories are ordered, so that response shift from one response category to another when the stimulus intensity changes. However, we can still use the psychometric function to model the proportion correct, P_c, as a function of stimulus intensity, even though the observer model illustrated in the lower panel of Figure 1.4 does not hold.

In *forced choice* paradigms, the observer must respond in one of N_r response categories. For a letter identification task, with one response category for each letter in the English alphabet, $N_r = 26$. Using the psychometric function in this task requires that we consider how the observer performs when the signal intensity, I_s, is near zero. The observer might not be able to perceive the letter but still has to respond and will have a probability $P_{guess} = \frac{1}{N_r}$ of guessing correctly. Even at higher stimulus intensities, the observer might not perceive the stimulus due to random fluctuations in the noisy encoding process. The high threshold model in Equation 1.12 is often used to model this behavior

$$P_c\left(I_s\right) = \tilde{P}_c\left(I_s\right) + \left(1 - \tilde{P}_c\left(I_s\right)\right) P_{guess} \tag{1.12}$$

where $\tilde{P}_c\left(I_s\right)$ is the proportion correct after *correcting for guessing.*

We can think of the proportion correct after correcting for guessing, $\tilde{P}_c$, as the true underlying psychometric function, which is not influenced by guessing, so that $\Psi\left(I_s\right) = \tilde{P}_c\left(I_s\right)$. We can isolate

$\tilde{P}_c(I_s)$ in Equation 1.12 to find

$$\Psi(I_s) = \tilde{P}_c(I_s) = \frac{P_c(I_s) - P_{guess}}{1 - P_{guess}} \tag{1.13}$$

In practise, we can fit Equation 1.12 to experimental data and use the parameter estimates to quantify the observer's perceptual sensitivity and response criterion.

1.3.1 Parameter estimation for the psychometric function

Theoretically, we could estimate the parameters of the psychometric function in Equation 1.11 by probit transforming the response proportions $\hat{P}(r = yes \mid I_s)$ and fitting a line to them much as we did for ROC curves. In practise, this is not a good solution since $\Phi^{-1}(P)$ is undefined for $P = 1$ and $P = 0$. Also, this approach does not work for the high threshold psychometric function in Equation 1.12.

In stead of using the probit transform, we can use the more general approach of finding the parameter values for σ and c that maximise the likelihood $\mathcal{L}(n \mid \sigma, c)$ of the observed response counts, n_s, for a specific stimulus, s. The response counts can be the number of 'yes'-responses, n_{yes} if we are using Equation 1.11, or the number correct responses, n_c, if we are using Equation 1.12. Since we have two response options, the response counts follow a binomial distribution, so that the likelihood of the response counts, n_s, for a particular stimulus, s, is given by Equation 1.14.

$$\mathcal{L}(n_s \mid \sigma, c) = \left(\begin{array}{c} N_s \\ n_s \end{array} \right) P_s^{n_s} (1 - P_s)^{N_s - n_s} \tag{1.14}$$

where P is the response probability, which can be $P = P_{yes}$ if we are using Equation 1.11, or $P = P_{correct}$ if we are using Equation 1.12.

We must, of course, fit the psychometric function to the response counts for *all* the stimuli, not just one specific stimulus. We do this based on the assumption that the response counts for each stimulus is independent of the response counts for the other stimuli so that the total likelihood is the product over stimuli, s, of the likelihood in Equation 1.14. This turns out to be problematic because the value of the likelihood can be very small, below machine precision. We can solve this problem by maximising the log likelihood (logarithm of the likelihood), $\mathcal{L}(n_s \mid \sigma, c)$ for a particular stimulus, s, as in Equation 1.15.

$$\log(\mathcal{L}(n_s \mid \sigma, c)) = \sum_{i=1}^{N_s} \log(i) - \sum_{i=1}^{n_s} \log(i) - \sum_{i=1}^{N-n_s} \log(i) + n_s \log(P_s) + (N_s - n_s) \log(1 - P_s) \tag{1.15}$$

Again, we must, of course, fit the psychometric function to the response counts for all the stimuli. Since the total likelihood is the product product over stimuli, s, of the likelihood in Equation 1.14, the total log likelihood is the sum over stimuli, s, of the log likelihood in Equation 1.15.

Note that, in order to avoid very small numerical values in the calculation of the log likelihood, $\log(\mathcal{L}(n_s \mid \sigma, c))$, it is important to take the logarithm for each term as specified in Equation 1.15 rather than first calculating the likelihood $\mathcal{L}(n_s \mid \sigma, c)$ from Equation 1.14 and then taking the logarithm.

Finally, note that, the negative log likelihood $-\log(\mathcal{L}(n_s \mid \sigma, c))$ can be thought of as a *cost function.* Per convention, model fitting is often performed by minimising the cost function. Hence we may seek to minimise the negative log likelihood rather than maximising the log likelihood.

1.3.2 Psychometric function Exercise

In a 3-alternative classification task, the observer classifies speech sounds under varying sound intensities. The experiment consists of 30 experimental trials at each sound intensity. The sound intensities and the corresponding number of correct responses are shown in the table below.

Stimulus intensity (dB)	5	10	15	20	25	30
Number of correct responses	12	11	19	27	30	30

Fit the high threshold model in Equation 1.12 and the psychometric function in Equation 1.11 to the data.

- Make one plot with the two psychometric functions with and the data. Estimate which of the two psychometric functions fit the data better
- List the value of the negative log likelihood of the two models. Which is lower?
- List the parameter values for each of the two models. Do the two models give similar estimates of the parameter values?

1.4 Magnitude estimation

In the previous section we described the psychometric function $\Psi(I_s)$ as a function of the physical stimulus, I_s. We also introduced an observer model for the psychometric function, as illustrated in the lower panel of Figure 1.4. In this observer model the internal representation represents *perceived* stimulus intensity. Hence, the model require an alignment between physical and perceived stimulus intensity. In order to achieve this alignment we should choose the unit of measurement of the physical stimulus intensity carefully so that it aligns with perceived physical stimulus intensity.

Loudness, the perceived intensity of sound, provides a good example of why it is important to choose the right measurement unit for the physical stimulus intensity. Sound is changes in air pressure and we can sound intensity in the SI-unit for pressure, pascal (Pa). However, this measure aligns poorly with perceived sound intensity. The more commonly used measure of decibels (dB) aligns better with perceived sound intensity, so that for sound, intensity is measured as

$$I_s(b) = 20 \log_{10}\left(\frac{b}{b_0}\right) dB \tag{1.16}$$

where b is the sound pressure in units of pascal. The constant, b_0, is the typical hearing threshold for humans, so that that $I_s(b_0) = 0$.

The logarithmic decibel scale stems back to what is some of the earliest work in cognitive modeling: Fechner's law of perceived magnitude. Fechner based his work on the observations of Weber who studied perceptual sensitivity in *change detection* tasks. Using an adaptive procedure in which the change in stimulus intensity, ΔI_s, is adjusted to a level where to observer can just notice the difference, Weber measured the change detection threshold and named it the *just noticeable difference* (JND). Weber found that the JND is proportional to the baseline stimulus intensity, I_s, across many perceptual modalities

$$\Delta I_s = k_w I_s \tag{1.17}$$

where k_w is called the Weber fraction.

Note that the value of Weber fraction $k_w = \frac{\Delta I_s}{I_s}$ is different for for different perceptual modalities. We can thus use the Weber fraction, k_w, as a measure that compares the perceptual sensitivity across perceptual modalities. A small Weber fraction value means that a small relative change in stimulus intensity can be detected meaning that the observer is very sensitive to this modality.

Note that we can rearrange equation to

$$\frac{1}{k_w}\frac{\Delta I_s}{I_s} = 1 \tag{1.18}$$

Fechner hypothesised that Equation 1.18 could provide a unit for the change in perceived magnitude, I_p, so that, in the limit $\Delta I_s \to 0$, perceived intensity would change as

Stimulus intensity (dB)	5	10	15	20	25	30
Number of correct responses	12	11	19	27	30	30

Fit the high threshold model in Equation 1.12 and the psychometric function in Equation 1.11 to the data.

- Make one plot with the two psychometric functions with and the data. Estimate which of the two psychometric functions fit the data better
- List the value of the negative log likelihood of the two models. Which is lower?
- List the parameter values for each of the two models. Do the two models give similar estimates of the parameter values?

1.4 Magnitude estimation

In the previous section we described the psychometric function $\Psi\left(I_s\right)$ as a function of the physical stimulus, I_s. We also introduced an observer model for the psychometric function, as illustrated in the lower panel of Figure 1.4. In this observer model the internal representation represents *perceived* stimulus intensity. Hence, the model require an alignment between physical and perceived stimulus intensity. In order to achieve this alignment we should choose the unit of measurement of the physical stimulus intensity carefully so that it aligns with perceived physical stimulus intensity.

Loudness, the perceived intensity of sound, provides a good example of why it is important to choose the right measurement unit for the physical stimulus intensity. Sound is changes in air pressure and we can sound intensity in the SI-unit for pressure, pascal (Pa). However, this measure aligns poorly with perceived sound intensity. The more commonly used measure of decibels (dB) aligns better with perceived sound intensity, so that for sound, intensity is measured as

$$I_s(b) = 20\log_{10}\left(\frac{b}{b_0}\right) dB \tag{1.16}$$

where b is the sound pressure in units of pascal. The constant, b_0, is the typical hearing threshold for humans, so that that $I_s(b_0) = 0$.

The logarithmic decibel scale stems back to what is some of the earliest work in cognitive modeling: Fechner's law of perceived magnitude. Fechner based his work on the observations of Weber who studied perceptual sensitivity in *change detection* tasks. Using an adaptive procedure in which the change in stimulus intensity, ΔI_s, is adjusted to a level where to observer can just notice the difference, Weber measured the change detection threshold and named it the *just noticeable difference* (JND). Weber found that the JND is proportional to the baseline stimulus intensity, I_s, across many perceptual modalities

$$\Delta I_s = k_w I_s \tag{1.17}$$

where k_w is called the Weber fraction.

Note that the value of Weber fraction $k_w = \frac{\Delta I_s}{I_s}$ is different for for different perceptual modalities. We can thus use the Weber fraction, k_w, as a measure that compares the perceptual sensitivity across perceptual modalities. A small Weber fraction value means that a small relative change in stimulus intensity can be detected meaning that the observer is very sensitive to this modality.

Note that we can rearrange equation to

$$\frac{1}{k_w}\frac{\Delta I_s}{I_s} = 1 \tag{1.18}$$

Fechner hypothesised that Equation 1.18 could provide a unit for the change in perceived magnitude, I_p, so that, in the limit $\Delta I_s \to 0$, perceived intensity would change as

$$dI_p = \frac{1}{k_w}\frac{dI_s}{I_s} \tag{1.19}$$

perceived magnitude, I_p.

In order to find the actual perceived stimulus intensity, I_p, Fechner found his law by integrating Equation 1.19

$$I_p = \int_{I_0}^{I_s} \frac{1}{k_w}\frac{dI_s}{I_s} = \frac{1}{k_w}\ln\left(\frac{I_s}{I_0}\right) \tag{1.20}$$

where I_0 is the absolute threshold, the minimal intensity value that can be perceived. Comparing Fechner's law in Equation 1.20 to the decibel scale in Equation 1.16 we see that they are identical except for the choice of base in the logarithm.

Fechner published his studies in 1860 and it was only seriously challenged 100 years later when Stevens showed that it does not apply generally to all sensory modalities. It fails, for example, to describe magnitude perception for color saturation, warmth and electric shock. Stevens introduced an alternative law of magnitude perception now known as Stevens' law, which describes perceived stimulus intensity as a power function of physical stimulus intensity.

$$I_p = k_s I_s^a \tag{1.21}$$

where the parameters k_s and the exponent, a varies across stimulus modalities.

Stevens' law seem to apply reasonably well to a very wide range of stimulus modalities including those, like sound intensity, for which Fechner's law applies. The reason for this is that power laws with an exponent $a < 1$ can have a shape similar to that of a logarithmic function.

Fechner and Stevens' work informs us on the choice of unit to use for the stimulus intensity when fitting a psychometric function. If previous studies have shown that perceived intensity follows one of Fechner and Stevens's laws then the physical unit should be chosen accordingly. It is therefore common to use dB as the unit of sound intensity because perceived sound intensity has been shown to be approximated well by a logarithmic function. For other units may apply for other modalities and experimental paradigms.

1.4.1 Magnitude estimation exercise

Although mathematically different, Fechner and Stevens' laws of perceptual intensity provide fairly good fits to perceived brightness as a function of luminance. This is because the exponent of Stevens' law is approximately $a = 0.33 < 1$ so that

$$I_p = 10 I_s^{0.33}$$

To see that this relationship might be mistaken for a logarithmic relationship, first calculate the perceived stimulus intensity, I_p, for physical intensities $I_s = 1, 2, \ldots, 10$ using Steven's law. This simulates an observer that rates the perceived intensity according to Stevens' law. Fit Fechner's law to the simulated data. Note that Fechner's law is linear with respect to I_s.

- List the parameter values for Fechner's law
- Plot the simulated data and curve showing Fechner's law
- Evaluate whether Fechner's law provides a reasonable fit by visual inspection of the simulated data and the model

For electric shock, Stevens found the exponent to be approximately $a = 3.3 > 1$. As before, calculate the perceived stimulus intensity for physical intensities $I_s = 1, 2, \ldots, 10$ using Stevens' law and fit Fechner's law to the simulated data.

- List the parameter values for Fechner's law
- Plot the simulated data and curve showing Fechner's law
- Evaluate whether Fechner's law provides a reasonable fit by visual inspection of the simulated data and the model

CHAPTER 2

2.1 Introduction

Our senses are able to encode high-dimensional stimuli to lower-dimensional feature spaces. Typically, these features are of particular relevance to our actions and, hence, our survival.

As an example, humans can encode images to extract relevant features such as facial expression. If we represent a monochrome image as a vector where each pixel is a dimension then the value of each dimension represents the brightness of the corresponding pixel. If we assume that an image is *linearly* encoded onto an internal representation, x, of some feature then

$$x = \boldsymbol{i}^T \boldsymbol{w} + \delta + \epsilon \tag{2.1}$$

where, $\boldsymbol{i}$, is the image column vector, $\boldsymbol{w}$ is a weight column vector and $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$ is the noise added to the encoding process.

The linear encoding model is a simplified model, which, in general is too simple. Still, it does capture some important aspects of the encoding process. It does, however, rely on some assumptions. Importantly, the linear encoding model relies on the assumption that each pixel represent the same aspect across all images. This is a reasonable assumption if the images are aligned. For images of faces, this can be achieved by scaling, shifting and rotating the images so that the two eyes are in the same pixels in all images.

In the previous chapters we assumed that the stimuli were carefully designed and delivered to have a stimulus intensity known with great precision. If we want to study how the observer perceives facial features, such as a smile or gender characteristics, we cannot control the stimulus in the same way. First, we do not know the relevant image features and even if we did, we would probably be unable to control them precisely. How should we generate a face with typically masculine traits at a certain level of intensity?

Unable to control the stimulus precisely we can instead sample from a repository or generate our own samples, i.e. we can obtain a random selection of N photographs of faces. We can then ask observers to rate the images with respect to a certain facial feature. We will assume that their rating reflect the internal representation value, x, and therefore refer to both the rating and the internal representation value as x. This experiment will give us a N-dimensional column vector, $\boldsymbol{x}$, of ratings, where each entry reflects the internal representation value for a particular image, and a matrix, $\boldsymbol{I}$, of images, where each row represents an image and each column represents a pixel that, according to Equation 2.1, are related as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_1^T \\ \vdots \\ \boldsymbol{i}_N^T \end{bmatrix} \boldsymbol{w} + \delta + \epsilon = \boldsymbol{I}\boldsymbol{w} + \delta + \epsilon \tag{2.2}$$

If the images each consist of M pixels then the image vectors $\boldsymbol{i}_1, \ldots, \boldsymbol{i}_N$ and the weight vector, $\boldsymbol{w}$, are M-dimensional and $\boldsymbol{I}$ is an N-by-M matrix. Note that a certain image can be rated multiple times. In that case the image will be repeated across multiple rows of the image matrix, $\boldsymbol{I}$. The *intercept*, denoted as δ, and the weight vector, $\boldsymbol{w}$ are free parameters that can be stacked into a parameter vector by rewriting Equation 2.2 to

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}_1^T & 1 \\ \vdots & \vdots \\ \boldsymbol{i}_N^T & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ \delta \end{bmatrix} + \epsilon = \boldsymbol{I}_\delta \boldsymbol{w} + \epsilon \tag{2.3}$$

If we have more equations than unknowns, $(N > M)$, then Equation 2.3 can be solved using the normal equations to find

$$\begin{bmatrix} \boldsymbol{w} \\ \delta \end{bmatrix} = \left(\boldsymbol{I}_{\delta}^{T} \boldsymbol{I}_{\delta} \right)^{-1} \boldsymbol{I}_{\delta}^{T} \boldsymbol{x} \tag{2.4}$$

where $\hat{\boldsymbol{w}} = \left(\boldsymbol{I}^{T} \boldsymbol{I} \right)^{-1} \boldsymbol{I}^{T}$ is called the *pseudo-inverse* matrix of $\boldsymbol{I}$. This solution is described in more detail in Chapter 8 in the work of Herlau, Schmidt and Mørup[HMS21].

Before we attempt to solve Equation 2.2 using Equation 2.4 we must consider the relationship between the number of, $M + 1$, of free parameters and the number of equations, N. The resolution of digital images is typically measured in megapixels, meaning that $M > 10^6$. This corresponds reasonably well with the number of photo-receptors in the human eye, normally tens of millions. Compare this number to the number of images that an observer, or a group of observers, can rate in a reasonable amount of time. If we assume that it takes 2 s to rate one image then an observer can rate 1800 images in half an hour. The task is tedious and we cannot expect an observer to perform well for more than an hour without breaks. Assume that we can obtain approximately 1000 reliable ratings in one session with one observer, then we will need thousands of experimental sessions before $N = M$. Fortunately, we can use far fewer ratings even though this leaves Equation 2.2 extremely *ill-posed*. In order to do so, we must *regularise* Equation 2.2. We can do this by reducing the dimensionality of the images.

The simplest steps to reduce the dimensionality of the images rated by observers is to use cropping and down-sampling. If the task of the observers is to rate facial features then we may crop the images to contain only the face and not the hair or any irrelevant background. For cropped images of faces, it is possible to reliably detect facial features in images with a resolution of only thousands of pixels, which is a significant reduction in the dimensionality, M, of the images compared to the millions of dimensions in uncropped megapixel images.

The next step in reducing the dimensionality of images is to use *Principal Component Analysis* (PCA). PCA provides a new orthogonal basis, the principal components, (PCs) for the image vector. The first PC points in the direction of the most variance in image space. The second PC points in the direction of the second most variance in image space and so forth. The total number of PCs equals the number of dimensions, M, in the original image space *if* we have more images than than the image dimensionality, i.e. if $M > N$. If, as in our case, the images have more dimensions than the number of images, i.e. $N > M$ then the number of PCs will equal the number of images, N.

Since the PCs are orthogonal, the sum of the variance across PCs equals the total amount of variance of the images. By projecting the images into PC space we will often find that most of the variance is contained in the first few PCs. These PCs define a *subspace* that has a dimensionality, which is typically much lower than the original image space. By projecting the images onto this subspace we obtain a representation, of the images, known as the *score*, which is of a much lower dimensionality than the original images. Replacing the images, $\boldsymbol{I}$, with their lower-dimensional scores, $\tilde{\boldsymbol{I}}$ in Equations 2.1–2.3 thus provides us with equations that are not ill-posed. This approach is explained in detail in Chapter 3, especially subsection 3.4.2, in the work of Herlau, Schmidt and Mørup [HMS21].

We may be able to reduce the dimensionality of the images even further. Variance in image space will be due to variance in many image features. For images of faces, the variance can be related to the angle of the light source, gender characteristics, wearing spectacles and facial expression. Selecting the PCs containing most of the variance will include the variance across all those features. If possible, we would like to select the PCs that include variance related to the image feature that we intend to capture in the weight vector, $\boldsymbol{w}$. We can use methods like *forward selection* to achieve this. Forward selection is an iterative procedure in which the model is first fitted using each of the PCs included in the subspace, $\tilde{\boldsymbol{I}}$—one at a time—so that the number of models equals the number of PCs included in the subspace, $\tilde{\boldsymbol{I}}$. The models are then evaluated for each of the PCs. The PC that gives the best model will be included in the final model. Then, the model is fitted again using each of the remaining PCs—one at a time—*and* the PC that was already selected in the first iteration. This process is repeated until the model is no longer improved by adding additional PCs. In this way a number of PCs that are relevant for improving the model is selected.

Note that in forward selection the models should not be evaluated by their goodness-of-fit. In that case, forward selection would select all PCs in the subspace more free parameters generally provide better fits. Instead the models should be evaluated using *cross-validation* where the models are fitted to a *training set* and evaluated on a *test set*. Once the PCs that will be included in the final model have been selected using forward selection, the model can be fitted to all the data to obtain an estimate of the weight vector, w. Cross-validation and the forward selection approach to feature selection is described in more detail in Chapter 10 in the work of Herlau, Schmidt and Mørup [HMS21].

The model described in Equation 2.2 describes a line in image space, or, in a subspace of image space. In order to validate the model we can visualise a point on the line corresponding to a rating, x_0, as an image, $\boldsymbol{i}_0$. In order to do so we need to solve Equation 2.1 for $\boldsymbol{i}$ given the weight vector, $\boldsymbol{w}$, and the intercept, δ, which are parameter values obtained by fitting the model. Note that there is, an $(M-1)$-dimensional, space of images that will have the same rating according to Equation 2.1. This subspace consists of images that all have the same rating but varies in image components not relevant to the rating, i.e. they are orthogonal to the weight vector, $\boldsymbol{w}$. Note that adding such components to $\boldsymbol{i}$ in Equation 2.1 do not change the rating, x. We can find a unique image, $\boldsymbol{i}_0$, containing no such irrelevant components for a given rating, x_0, by constraining the solution to be parallel to the weight vector, $\boldsymbol{w}$, so that $\boldsymbol{i}_0 = \alpha\boldsymbol{w}$ where α is a scalar.

$$x_0 = \boldsymbol{i}_0^T\boldsymbol{w} + \delta = \alpha\boldsymbol{w}^T\boldsymbol{w} + \delta = \alpha\|\boldsymbol{w}\| + \delta \tag{2.5}$$

Solving Equation 2.5 for alpha gives us

$$\alpha = \frac{x_0 - \delta}{\|\boldsymbol{w}\|} \tag{2.6}$$

and hence Fagertun, Andersen and Paulsen [FAP12] used a similar approach for creating synthetic images with varying gender characteristics. In their approach the rating, x, of gender characteristics, ranging from feminine to masculine, was calculated using binary responses and reaction times, so that the rating was deemed more extreme if for faster responses. They validated the synthetic images informally using visual inspection and found that typical masculine characteristics such as facial hair was present in images rated more masculine. This result was confirmed across three publicly available image sets that were constrained to contain cropped frontal facial images and in one image set that was obtained by collecting images with greater variability from LinkedIn™. Fagertun, Andersen, Hansen and Paulsen [Fag+13] used the same approach on high resolution three-dimensional images dividing the images into shape and texture components. They observed that synthetically generated more masculine images showed characteristic masculine head shape features such as a stronger jaw line and also characteristic masculine texture features such as beard stubble. This confirms that the approach can be used to create synthetic images varying in only in a single perceptual feature.

Bibliography

[Fag+13] Jens Fagertun et al. "3D gender recognition using cognitive modeling." English. In: *2013 International Workshop on Biometrics and Forensics (IWBF)*. 2013 International Workshop on Biometrics and Forensics (IWBF) ; Conference date: 04-04-2013 Through 05-04-2013. United States: IEEE, 2013. ISBN: 978-1-4673-4987-1. DOI: 10.1109/IWBF.2013.6547324. URL: http://www.img.lx.it.pt/iwbf2013/.

[FAP12] Jens Fagertun, Tobias Andersen, and Rasmus Reinhold Paulsen. "Gender Recognition Using Cognitive Modeling." English. In: *Computer Vision – ECCV 2012*. Lecture Notes in Computer Science. 12th European Conference on Computer Vision (ECCV 2012) ; Conference date: 07-10-2012 Through 13-10-2012. Springer, 2012, pages 300–308. ISBN: 978-3-642-33867-0. DOI: 10.1007/978-3-642-33868-7_30. URL: http://eccv2012.unifi.it/.

[HMS21] Tue Herlau, Morten Mørup, and Mikkel N. Schmidt. *Introduction to Machine Learning and Data Mining*. DTU lecture notes, 2021.

CHAPTER 2

Bayesian models of multisensory perception

2.1 The strong fusion observer model

Our multiple senses allows us to receive information about the world from multiple encodings of the same source. Take, for instance, a cat hunting for mice. The cat might both see and hear the mouse, so the location of the mouse will be encoded in the auditory modality as x_a and in the visual modality as x_v but, in order to catch the mouse, the cat needs a single integrated estimate of the location, S, of the mouse. We are now searching for a model for integrating the two internal representation values, x_a and x_v, using Bayes' rule.

First, let us consider the case where the observer (the cat) has access to only one sensory modality. If the observer can only hear the mouse, then it has access only to x_a. It seems reasonable to assume that the most probable location, i.e. the *maximum a posteriori* (MAP) estimate, $\hat{S}_{MAP}$, of the true location of the sound of the mouse, S_a, should correspond to the internal representation value, x_a. Even so, we will show this more formally as it will help us in establishing a useful mathematical framework. Using Bayes' rule, the observer model can estimate the posterior probability of an estimate $\hat{S}$, of the location, S.

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} \tag{2.1}$$

For simplicity, we will assume that the observer has no prior information on the location, S, so that the prior, $P(\hat{S})$, is constant across all locations. In this case, the prior probability does not influence the posterior. We can show this by expanding the denominator

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} = \frac{P(x_a \mid \hat{S})P(\hat{S})}{\int_{\hat{S}} P(x_a \mid S)P(S)} = \frac{P(x_a \mid \hat{S})}{\int_{\hat{S}} P(x_a \mid \hat{S})} \tag{2.2}$$

In this case, the MAP estimate is equal to the maximum likelihood estimate, which is why the strong fusion model is often referred to as the maximum likelihood estimation (MLE) model.

As in signal detection theory, the observer model assumes that the internal representation value, x_a, vary randomly due to Gaussian sensory noise added in the encoding process, so that it centered around the true location of the sound, S_a. The, observer can thus calculate the likelihood as

$$P(x_a \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2) \tag{2.3}$$

Inserting Equation 2.3 into Bayes' rule in the form of Equation 2.2 gives us

$$P(\hat{S} \mid x_a) = \frac{f(x_a \mid \hat{S}, \sigma_a^2)}{\int_{\hat{S}} f(x_a \mid \hat{S}, \sigma_a^2)} = f(\hat{S} \mid x_a, \sigma_a^2) \tag{2.4}$$

Note that the normalisation of the Gaussian likelihood, $f(x_a \mid \hat{S}, \sigma_a^2)$, with respect to the estimate, $\hat{S}$, of the location, S, results in a posterior probability density, $P(\hat{S} \mid x_a)$, which is centered around the internal representation value, x_a. Now, since the maximum of a Gaussian distribution is its mean, the MAP solution, $\hat{S}_{MAP}$, to Equation 2.4 is the mean of the posterior so that

$$\hat{S}_{MAP} = x_a \tag{2.5}$$

In order to fit the simple unisensory observer model to observable data, we will assume that the observer responds according to its MAP solution, $\hat{S}_{MAP} = x_a$. In the case of a cat chasing a mouse, the response could be a pounce or an orienting response to a sound. If the observer model holds then the distribution of the internal representation value, x_a, will centered around the true location of the sound, S_a.

$$P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid S_a, \sigma_a^2) \tag{2.6}$$

In this case the only free parameter of the model is σ_a. However, the observer model may not hold. Therefore, the true location of the sound, S_a, is typically replaced by a free parameter for the mean, μ_a, of the distribution of the observer's MAP solution.

$$P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid \mu_a, \sigma_a^2) \tag{2.7}$$

Trivially, the case in which the observer can only see the stimulus, but not hear it, is described by replacing μ_a with μ_v and σ_a with σ_v in Equations 2.1-2.6.

We have now established a framework, which will help us analyse the case of multisensory stimuli for which the observer has access to a bivariate internal representation, (x_a, x_v), but still needs a single integrated estimate, $\hat{S}$, of the location of the mouse. This is the basis of strong fusion: It only makes sense to fuse the two sensory modalities into a single estimate if they both convey information about the same location. Still assuming a uninformative prior, as in Equation 2.2, we can write Bayes' rule as

$$P(\hat{S} \mid x_a, x_v) = \frac{P(x_a, x_v \mid \hat{S})}{P(x_a, x_v)} = \frac{P(x_a, x_v \mid \hat{S})}{\int_{\hat{S}} P(x_a, x_v \mid \hat{S})} \tag{2.8}$$

Since the sensory noise is added in the encoding process, the observer model assumes that the noise added to the auditory internal representation, x_a, is independent from the noise added to the visual internal representation. This means that the internal representation values are *conditionally* independent on S, so that we can write the likelihood for the multisensory case as

$$P(x_a, x_v \mid \hat{S}) = P(x_a \mid \hat{S})P(x_v \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2) f(x_v \mid \hat{S}, \sigma_v^2) \tag{2.9}$$

Note that conditional independence, as described in Equation 2.9, does not imply unconditional independence, $P(x_a, x_v) = P(x_a)P(x_v)$. This would be a poor assumption because the internal representation values, x_a and x_v will, in general, be dependent as they are both influenced by location of co-occurring changes in sound and light. Conditional independence means that the internal representation values, x_a and x_v, are independent across multiple occurrences of the exact same stimulus because their values would, in that case, vary only due to the noise added in the encoding process.

Equation 2.9 describes the likelihood for audiovisual stimuli as a product of two Gaussian probability density distributions. In general, the product of two Gaussian distributions with random variables, x_a and x_v is proportional to a Gaussian distribution of a random variable, x_{av}, that is equal to a weighted average of x_a and x_v

$$x_{av} = w_a x_a + (1 - w_a)_v \tag{2.10}$$

where the weight, w_a, is given by

$$w_a = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{2.11}$$

so that

$$f(x_a \mid \mu_a, \sigma_a^2) f(x_v \mid \mu_v, \sigma_v^2) = A f(x_{av} \mid \mu_{av}, \sigma_{av}^2) \tag{2.12}$$

where A is a constant that will be of little importance.

Since x_{av} is a weighted mean, its mean, μ_{av}, is a weighted average of the mean, μ_a, of x_a, and the mean μ_v, of x_v

$$\mu_{av} = w_a \mu_a + (1 - w_a)\mu_v \tag{2.13}$$

and the variance, σ_{av}^2, is given by

$$\sigma_{av}^2 = w_a^2 \sigma_a^2 + (1 - w_a)^2 \sigma_v^2 = \frac{\sigma_a^2 \sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{2.14}$$

For the strong fusion observer model in Equation 2.12, the means of the distributions of x_a and x_v, are both assumed to be equal to one and the same estimate, $\hat{S}$. A weighted average of the two means is therefore also equal to $\hat{S}$. Inserting this into the observer model gives us

$$P(\hat{S} \mid x_a, x_v) = \frac{Af(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} Af(x_{av} \mid \hat{S}, \sigma_{av}^2)} = \frac{f(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} f(x_{av} \mid \hat{S}, \sigma_{av}^2)} = f(\hat{S} \mid x_{av}, \sigma_{av}^2) \tag{2.15}$$

As in Equation 2.5 we find that the MAP estimate, $\hat{S}_{MAP}$, of the location S is

$$\hat{S}_{MAP} = x_{av} = w_a x_a + (1 - w_a)x_v \tag{2.16}$$

2.1.1 Properties of the strong fusion observer model

Here we will pause and reflect on the properties of the strong fusion model. The first thing that we will note is that its implementation is actually quite simple: The auditory and visual stimuli are encoded as scalars, x_a and x_v. When the observer has access to both, the combined estimate is simply a weighted sum. The fairly complex probability theoretical considerations above are not necessary for the implementation. They only serve for us to analyse the model.

In order to better understand the weight, w_a, we rephrase it in terms of precision, $r = \frac{1}{\sigma^2}$, by dividing the numerator and denominator on the right hand side of Equation 2.11 by $\sigma_a^2 \sigma_v^2$. This gives us

$$w_a = \frac{r_a}{r_a + r_v} \qquad\qquad w_v = (1 - w_a) = \frac{r_v}{r_a + r_v} \tag{2.17}$$

Equation 2.17 shows that the strong fusion observer weighs the sensory modalities by their precision. Somewhat confusingly, this is called the *information reliability* principle, even though it should, perhaps, be called the information precision principle. Regardless of the name, the principle of weighing information according to its precision according to the makes intuitive sense.

Calculating the precision, r_{av} of the internal representation value x_{av} reveals another interesting property of the strong fusion model. We can do this by rearranging equation 2.14

$$r_{av} = \frac{1}{\sigma_{av}^2} = \frac{\sigma_a^2 + \sigma_v^2}{\sigma_a^2 \sigma_v^2} = r_a + r_v \tag{2.18}$$

Equation 2.18 reveals that the precision of the observer given audiovisual stimuli is the sum of the precision given the auditory stimulus and precision given the visual stimulus. A strong fusion observer will thus always be more precise when integrating the multisensory internal representations, x_a and x_v.

We can even show that the strong fusion observer weight, w_a, is the Bayes' optimal weight in terms of maximising the precision, r_{av}, or, equivalently, minimising the variance, σ_{av}^2, by expressing the variance as a function of the weight as in Equation 2.14 and finding the minimum by differentiation

$$\frac{\partial \sigma_{av}^2}{\partial w} = \frac{\partial w^2 \sigma_a^2 + (1 - w)^2 \sigma_v^2}{\partial w} = \frac{\partial (\sigma_a^2 + \sigma_v^2) w^2 - 2w\sigma_v^2 + \sigma_v^2}{\partial w} = 2(\sigma_a^2 + \sigma_v^2)w - 2\sigma_v^2 \tag{2.19}$$

The solution to Equation 2.19, is $w = w_a = \frac{\sigma_v^2}{\sigma_a^2+\sigma_v^2}$, which is the minimum $\arg\min_w(\sigma_{av}^2) = w_a$, which shows that the choice of weight in the strong fusion observer model is optimal for maximising the precision, r_{av}.

Finally, we will look at the case where the true location of the sound, S_a, is different from the true location of the visual stimulus, S_v. Small differences could occur in the natural world if, say, a cat sees the tail of the mouse while the sound originates from the snout of the mouse. The strong fusion estimate, $\hat{S}$, of the location would, on average, lie somewhere between the snout and the tail and in that case a pounce may well prove successful. Much larger differences occur mostly in artificial settings where the location of the sound, S_a, and the location of the visual stimulus, S_v, are designed to be different. A striking example of this is the ventriloquist illusion. When we observe a ventriloquist, we often perceive the sound as appearing from a dummy even though it does, of course, originate from the ventriloquist. A more common version of this illusion is experienced when watching a movie of a conversation with a monaural sound source. Although the sound of the voices originate from one location, observers tend to perceive that the voice of each actor originates from the location of the image of that actor. In order to test how the strong fusion accounts for the ventriloquist illusion, we may measure an observer's estimate of the location of the sound, the image and the audiovisual combination of the two. It has been found that the strong fusion model is able to account for this type of data. Since vision has a higher spatial resolution than hearing in humans, estimates of the location of the visual stimulus are typically much more precise than estimates of the location of the sound. Therefore, according to Equation 2.17 observers will weigh the location of the visual stimulus much higher when estimating the location of the audiovisual stimulus. This, according to the strong fusion model, is the basis of the ventriloquist illusion.

2.1.2 Bayesian causal inference (BCI)

Studies have shown that many factors influence the strength of audiovisual integration. The timing is, for example, a very important factor. The greater the asynchrony between the auditory and visual stimuli, the less likely they are to be integrated. The strong fusion model cannot account for such effects. Therefore the Bayesian causal inference model (BCI) has been introduced.

The BCI model posits that the brain tries to infer whether audiovisual stimuli has one or two underlying causes. In the case that there is one underlying cause, the assumption of the strong fusion model is valid and the observer will integrate the auditory and visual internal representations accordingly. In the case that there are two underlying causes the auditory and visual internal representations are not integrated. Note that this means that the response to audiovisual stimuli depends on the task. If the observer is given the task of responding according to what she heard then she will respond according to the auditory internal representation. In this case

$$P(\hat{S}_{MAP} \mid S_a, S_v) = P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)P(N_c = 1) + P(\hat{S}_{MAP} \mid S_a, N_c = 2)(1 - P(N_c = 1)) \tag{2.20}$$

where $P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)$ is probability of the strong fusion MAP estimate, $P(\hat{S}_{MAP} \mid S_a, N_c = 2)$ is probability of the estimate based only on the auditory stimulus as in Equation 2.6, and N_c is the number of underlying causes inferred by the observer. Conversely, if the observer is given the task of responding according to what she saw then she will respond according to the visual internal representation. In this case, the distribution of S_{MAP} is described by

$$P(\hat{S}_{MAP} \mid S_a, S_v) = P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)P(N_c = 1) + P(\hat{S}_{MAP} \mid S_v, N_c = 2)(1 - P(N_c = 1)) \tag{2.21}$$

The BCI model is a mixture model of two distributions. When fitting it to observed data, $P(C = 1)$, is typically a free parameter. The BCI model thus has one more free parameter than the MLE model and this must be taken into account when assessing the models.

CHAPTER 2

Bayesian models of multisensory perception

2.1 Models for continuous responses

2.1.1 The strong fusion observers model

Our multiple senses allows us to receive information about the world from multiple encodings of the same source. Take, for instance, a cat hunting for mice. The cat might both see and hear the mouse, so the location of the mouse will be encoded in the auditory modality as x_a and in the visual modality as x_v but, in order to catch the mouse, the cat needs a single integrated estimate of the location, S, of the mouse. We are now searching for a model for integrating the two internal representation values, x_a and x_v, using Bayes' rule.

First, let us consider the case where the observer (the cat) has access to only one sensory modality. If the observer can only hear the mouse, then it has access only to x_a. It seems reasonable to assume that the most probable location, i.e. the *maximum a posteriori* (MAP) estimate, $\hat{S}_{MAP}$, of the true location of the sound of the mouse, S_a, should correspond to the internal representation value, x_a. Even so, we will show this more formally as it will help us in establishing a useful mathematical framework. Using Bayes' rule, the observer model can estimate the posterior probability of an estimate $\hat{S}$, of the location, S.

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} \tag{2.1}$$

For simplicity, we will assume that the observer has no prior information on the location, S, so that the prior, $P(\hat{S})$, is constant across all locations. In this case, the prior probability does not influence the posterior. We can show this by expanding the denominator

$$P(\hat{S} \mid x_a) = \frac{P(x_a \mid \hat{S})P(\hat{S})}{P(x_a)} = \frac{P(x_a \mid \hat{S})P(\hat{S})}{\int_{\hat{S}} P(x_a \mid S)P(S)} = \frac{P(x_a \mid \hat{S})}{\int_{\hat{S}} P(x_a \mid \hat{S})} \tag{2.2}$$

In this case, the MAP estimate is equal to the maximum likelihood estimate, which is why the strong fusion model is often referred to as the maximum likelihood estimation (MLE) model.

As in signal detection theory, the observer model assumes that the internal representation value, x_a, vary randomly due to Gaussian sensory noise added in the encoding process, so that it centered around the true location of the sound, S_a. The, observer can thus calculate the likelihood as

$$P(x_a \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2) \tag{2.3}$$

Inserting Equation 2.3 into Bayes' rule in the form of Equation 2.2 gives us

$$P(\hat{S} \mid x_a) = \frac{f(x_a \mid \hat{S}, \sigma_a^2)}{\int_{\hat{S}} f(x_a \mid \hat{S}, \sigma_a^2)} = f(\hat{S} \mid x_a, \sigma_a^2) \tag{2.4}$$

Note that the normalisation of the Gaussian likelihood, $f(x_a \mid \hat{S}, \sigma_a^2)$, with respect to the estimate, $\hat{S}$, of the location, S, results in a posterior probability density, $P(\hat{S} \mid x_a)$, which is centered around the internal representation value, x_a. Now, since the maximum of a Gaussian distribution is its mean, the MAP solution, $\hat{S}_{MAP}$, to Equation 2.4 is the mean of the posterior so that

$$\hat{S}_{MAP} = x_a \tag{2.5}$$

In order to fit the simple unisensory observer model to observable data, we will assume that the observer responds according to its MAP solution, $\hat{S}_{MAP} = x_a$. In the case of a cat chasing a mouse, the response could be a pounce or an orienting response to a sound. If the observer model holds then the distribution of the internal representation value, x_a, will centered around the true location of the sound, S_a.

$$P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid S_a, \sigma_a^2) \tag{2.6}$$

In this case the only free parameter of the model is σ_a. However, the observer model may not hold. Therefore, the true location of the sound, S_a, is typically replaced by a free parameter for the mean, μ_a, of the distribution of the observer's MAP solution.

$$P(\hat{S}_{MAP} \mid S_a, \sigma_a^2) = f(x_a \mid \mu_a, \sigma_a^2) \tag{2.7}$$

Trivially, the case in which the observer can only see the stimulus, but not hear it, is described by replacing μ_a with μ_v and σ_a with σ_v in Equations 2.1-2.6.

We have now established a framework, which will help us analyse the case of multisensory stimuli for which the observer has access to a bivariate internal representation, (x_a, x_v), but still needs a single integrated estimate, $\hat{S}$, of the location of the mouse. This is the basis of strong fusion: It only makes sense to fuse the two sensory modalities into a single estimate if they both convey information about the same location. Still assuming a uninformative prior, as in Equation 2.2, we can write Bayes' rule as

$$P(\hat{S} \mid x_a, x_v) = \frac{P(x_a, x_v \mid \hat{S})}{P(x_a, x_v)} = \frac{P(x_a, x_v \mid \hat{S})}{\int_{\hat{S}} P(x_a, x_v \mid \hat{S})} \tag{2.8}$$

Since the sensory noise is added in the encoding process, the observer model assumes that the noise added to the auditory internal representation, x_a, is independent from the noise added to the visual internal representation. This means that the internal representation values are *conditionally* independent on S, so that we can write the likelihood for the multisensory case as

$$P(x_a, x_v \mid \hat{S}) = P(x_a \mid \hat{S})P(x_v \mid \hat{S}) = f(x_a \mid \hat{S}, \sigma_a^2) f(x_v \mid \hat{S}, \sigma_v^2) \tag{2.9}$$

Note that conditional independence, as described in Equation 2.9, does not imply unconditional independence, $P(x_a, x_v) = P(x_a)P(x_v)$. This would be a poor assumption because the internal representation values, x_a and x_v will, in general, be dependent as they are both influenced by location of co-occurring changes in sound and light. Conditional independence means that the internal representation values, x_a and x_v, are independent across multiple occurrences of the exact same stimulus because their values would, in that case, vary only due to the noise added in the encoding process.

Equation 2.9 describes the likelihood for audiovisual stimuli as a product of two Gaussian probability density distributions. In general, the product of two Gaussian distributions with random variables, x_a and x_v is proportional to a Gaussian distribution of a random variable, x_{av}, that is equal to a weighted average of x_a and x_v

$$x_{av} = w_a x_a + (1 - w_a) x_v \tag{2.10}$$

where the weight, w_a, is given by

$$w_a = \frac{\sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{2.11}$$

so that

$$f(x_a \mid \mu_a, \sigma_a^2) f(x_v \mid \mu_v, \sigma_v^2) = A f(x_{av} \mid \mu_{av}, \sigma_{av}^2) \tag{2.12}$$

where A is a constant that will be of little importance.

Since x_{av} is a weighted mean, its mean, μ_{av}, is a weighted average of the mean, μ_a, of x_a, and the mean μ_v, of x_v

$$\mu_{av} = w_a \mu_a + (1 - w_a)\mu_v \tag{2.13}$$

and the variance, σ_{av}^2, is given by

$$\sigma_{av}^2 = w_a^2 \sigma_a^2 + (1 - w_a)^2 \sigma_v^2 = \frac{\sigma_a^2 \sigma_v^2}{\sigma_a^2 + \sigma_v^2} \tag{2.14}$$

For the strong fusion observer model in Equation 2.12, the means of the distributions of x_a and x_v, are both assumed to be equal to one and the same estimate, $\hat{S}$. A weighted average of the two means is therefore also equal to $\hat{S}$. Inserting this into the observer model gives us

$$P(\hat{S} \mid x_a, x_v) = \frac{A f(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} A f(x_{av} \mid \hat{S}, \sigma_{av}^2)} = \frac{f(x_{av} \mid \hat{S}, \sigma_{av}^2)}{\int_{\hat{S}} f(x_{av} \mid \hat{S}, \sigma_{av}^2)} = f(\hat{S} \mid x_{av}, \sigma_{av}^2) \tag{2.15}$$

As in Equation 2.5 we find that the MAP estimate, $\hat{S}_{MAP}$, of the location S is

$$\hat{S}_{MAP} = x_{av} = w_a x_a + (1 - w_a) x_v \tag{2.16}$$

2.1.2 Properties of the strong fusion observer model

Here we will pause and reflect on the properties of the strong fusion model. The first thing that we will note is that its implementation is actually quite simple: The auditory and visual stimuli are encoded as scalars, x_a and x_v. When the observer has access to both, the combined estimate is simply a weighted sum. The fairly complex probability theoretical considerations above are not necessary for the implementation. They only serve for us to analyse the model.

In order to better understand the weight, w_a, we rephrase it in terms of precision, $r = \frac{1}{\sigma^2}$, by dividing the numerator and denominator on the right hand side of Equation 2.11 by $\sigma_a^2 \sigma_v^2$. This gives us

$$w_a = \frac{r_a}{r_a + r_v} \qquad w_v = (1 - w_a) = \frac{r_v}{r_a + r_v} \tag{2.17}$$

Equation 2.17 shows that the strong fusion observer weighs the sensory modalities by their precision. Somewhat confusingly, this is called the *information reliability* principle, even though it should, perhaps, be called the information precision principle. Regardless of the name, the principle of weighing information according to its precision according to the makes intuitive sense.

Calculating the precision, r_{av} of the internal representation value x_{av} reveals another interesting property of the strong fusion model. We can do this by rearranging equation 2.14

$$r_{av} = \frac{1}{\sigma_{av}^2} = \frac{\sigma_a^2 + \sigma_v^2}{\sigma_a^2 \sigma_v^2} = r_a + r_v \tag{2.18}$$

Equation 2.18 reveals that the precision of the observer given audiovisual stimuli is the sum of the precision given the auditory stimulus and precision given the visual stimulus. A strong fusion observer will thus always be more precise when integrating the multisensory internal representations, x_a and x_v.

We can even show that the strong fusion observer weight, w_a, is the Bayes' optimal weight in terms of maximising the precision, r_{av}, or, equivalently, minimising the variance, σ_{av}^2, by expressing the variance as a function of the weight as in Equation 2.14 and finding the minimum by differentiation

$$\frac{\partial \sigma_{av}^2}{\partial w} = \frac{\partial w^2\sigma_a^2 + (1-w)^2\sigma_v^2}{\partial w} = \frac{\partial (\sigma_a^2 + \sigma_v^2)w^2 - 2w\sigma_v^2 + \sigma_v^2}{\partial w} = 2(\sigma_a^2 + \sigma_v^2)w - 2\sigma_v^2 \tag{2.19}$$

The solution to Equation 2.19, is $w = w_a = \frac{\sigma_v^2}{\sigma_a^2+\sigma_v^2}$, which is the minimum $\arg\min_w(\sigma_{av}^2) = w_a$, which shows that the choice of weight in the strong fusion observer model is optimal for maximising the precision, r_{av}.

Finally, we will look at the case where the true location of the sound, S_a, is different from the true location of the visual stimulus, S_v. Small differences could occur in the natural world if, say, a cat sees the tail of the mouse while the sound originates from the snout of the mouse. The strong fusion estimate, $\hat{S}$, of the location would, on average, lie somewhere between the snout and the tail and in that case a pounce may well prove successful. Much larger differences occur mostly in artificial settings where the location of the sound, S_a, and the location of the visual stimulus, S_v, are designed to be different. A striking example of this is the ventriloquist illusion. When we observe a ventriloquist, we often perceive the sound as appearing from a dummy even though it does, of course, originate from the ventriloquist. A more common version of this illusion is experienced when watching a movie of a conversation with a monaural sound source. Although the sound of the voices originate from one location, observers tend to perceive that the voice of each actor originates from the location of the image of that actor. In order to test how the strong fusion accounts for the ventriloquist illusion, we may measure an observer's estimate of the location of the sound, the image and the audiovisual combination of the two. It has been found that the strong fusion model is able to account for this type of data. Since vision has a higher spatial resolution than hearing in humans, estimates of the location of the visual stimulus are typically much more precise than estimates of the location of the sound. Therefore, according to Equation 2.17 observers will weigh the location of the visual stimulus much higher when estimating the location of the audiovisual stimulus. This, according to the strong fusion model, is the basis of the ventriloquist illusion.

2.1.3 Bayesian causal inference (BCI)

Studies have shown that many factors influence the strength of audiovisual integration. The timing is, for example, a very important factor. The greater the asynchrony between the auditory and visual stimuli, the less likely they are to be integrated. The strong fusion model cannot account for such effects. Therefore the Bayesian causal inference model (BCI) has been introduced.

The BCI model posits that the brain tries to infer whether audiovisual stimuli has one or two underlying causes. In the case that there is one underlying cause, the assumption of the strong fusion model is valid and the observer will integrate the auditory and visual internal representations accordingly. In the case that there are two underlying causes the auditory and visual internal representations are not integrated. Note that this means that the response to audiovisual stimuli depends on the task. If the observer is given the task of responding according to what she heard then she will respond according to the auditory internal representation. In this case

$$P(\hat{S}_{MAP} \mid S_a, S_v) = P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)P(N_c = 1) + P(\hat{S}_{MAP} \mid S_a, N_c = 2)(1 - P(N_c = 1)) \tag{2.20}$$

where $P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)$ is probability of the strong fusion MAP estimate, $P(\hat{S}_{MAP} \mid S_a, N_c = 2)$ is probability of the estimate based only on the auditory stimulus as in Equation 2.6, and N_c is the number of underlying causes inferred by the observer. Conversely, if the observer is given the task of responding according to what she saw then she will respond according to the visual internal representation. In this case, the distribution of S_{MAP} is described by

$$P(\hat{S}_{MAP} \mid S_a, S_v) = P(\hat{S}_{MAP} \mid S_a, S_v, N_c = 1)P(N_c = 1) + P(\hat{S}_{MAP} \mid S_v, N_c = 2)(1 - P(N_c = 1)) \tag{2.21}$$

The BCI model is a mixture model of two distributions. When fitting it to observed data, $P(C = 1)$, is typically a free parameter. The BCI model thus has one more free parameter than the MLE model and this must be taken into account when assessing the models.

2.2 Models for categorical responses

2.2.1 Audiovisual speech perception

The brain integrates information across the sensory modalities for many types of stimuli. In Section 2.1 we described models of audiovisual integration for stimuli, such as the location of a mouse, that are naturally perceived on a continuum. In this section we will describe models of audiovisual integration that are naturally perceived in discrete categories.

Audiovisual speech perception serves as a good example of a stimulus that is integrated across the senses and which is perceived in categories. Speech perception is enhanced when we can see the face of the talker compared to when we can only hear the voice. This is called the *enchancement effect*. Although there could be several reasons for this effect, it indicates that perceptual precision is increased when we perceive speech from two modalities rather than just one. This is similar to the effect of audiovisual integration for spatial localisation discussed in Section 2.1. As in the ventriloquist illusion for spatial localisation, illusory percepts occur also in speech perception when the two senses convey incongruent information. The most commonly known example of this is the McGurk fusion illusion where a voice saying /ba/ is dubbed on a video of a face articulating /ga/, in which case observers typically hear /da/. In this case, auditory and visual information seem to be fused into a percept different from that contained in either modality much as in the ventriloquist illusion.

Unlike spatial location, speech is perceived in discrete categories that are called phonemes. Phonemes are often thought of as speech sounds such as vocals and consonants. Defining phonemes in terms of the acoustic properties is, however, very difficult. One reason for this is that a phoneme, such as the ‘p’ in the word ‘sport’, sounds slightly differently every time you say it. Worse is that the sound depends many characteristics of the speaker such as the speaker’s gender, age and level of inebriation. Even worse, the the ‘p’ in the word ‘sport’ is actually articulated more like the ‘b’ in the word ‘bored’ than the ‘p’ in the word ‘port’. This is an example of co-articulation meaning that the sound of phonemes depends on the phonetic context. The acoustic variability within phonetic categories is thus large compared to the variability across phonetic categories, which is why it is difficult to define phonemes acoustically.

It is much more straight forward to define phonemes by the way they are articulated. A full description of all the articulatory features is beyond the scope of this text and we will focus on just one, the *place of articulation*, since it is particularly important for audiovisual speech perception. The consonant /b/ is a bilabial phoneme as it is articulated by a brief closure of the lips. A /b/ is thus articulated at the very front of the mouth. The consonants /d/ is articulated further back, at the alveolar ridge, right behind the teeth. It is therefore an alveolar consonants. The consonant /g/ is called a velar consonants and is articulated at the very back of the mouth. Obviously, the exact place of articulation may vary depending on the same factors that influence the acoustic characteristics of phonemes as we do not close the lips exactly the same way every time we articulate a /b/. Yet, this variability in the place of articulation within phonetic categories is small compared to the variability across phonetic categories. We do not, for example, ever close our lips when articulating a /d/. This is why phonemes are better characterised by articulatory features.

The place of articulation is important in audiovisual integration of speech. The *manner-place* hypothesis has been has found much support in the study of audiovisual integration of speech. Its claim is that it is information about the place of articulation that is integrated across modalities because it can be perceived visually, as you can see if someone is articulating a /b/ or a /g/, and auditorily, as you can also hear the difference. Other phonetic features can not be seen and are therefore not integrated audiovisually.

2.2.2 The early strong fusion model of audiovisual integration of speech

Perception of phonemes in discrete categories may rely on a continuous internal representation of the place of articulation much as the discrete responses, 'yes' and 'no', relies on a continuous internal representation of signal intensity in signal detection theory. As we have already described models of audiovisual integration for continuous responses in Section 2.1 it is natural to combine these models with signal detection theory, to build models of audiovisual integration of speech. We will refer to these models as *early* models of audiovisual integration of speech because they posit that integration occurs based on the continuous internal representation before decoding the internal representation into discrete response probabilities.

As in Section 2.1, the early strong fusion model posits that the observers estimate, $\hat{S}$, of the place of articulation, S is distributed as a Gaussian distribution so that

$$P(\hat{S}_a \mid S_a) = f(\hat{S} \mid \mu_a, \sigma_a^2) \tag{2.22}$$

$$P(\hat{S}_v \mid S_v) = f(\hat{S} \mid \mu_v, \sigma_v^2) \tag{2.23}$$

$$P(\hat{S}_{av} \mid S_{av}) = f(\hat{S} \mid \mu_{av}, \sigma_{av}^2) \tag{2.24}$$

where expressions for μ_{av}, w_a and σ_{av} are given by Equations 2.13, 2.11 and 2.14 respectively.

Just as in signal detection theory, criteria values, c, are introduced to decode the continuous values for $\hat{S}$ onto discrete response categories. We will, arbitrarily, choose that a bilabial place of articulation (towards the front of the mouth) corresponds to smaller values of $\hat{S}$, so that the probability, $P(r_b)$, of a b-response, r_b, is given by

$$P(r_b) = P(\hat{S} < c_{bd}) = \int_{-\infty}^{c_{bd}} P(\hat{S} \mid S) = \Phi(c_{bd}, \mu, \sigma) \tag{2.25}$$

where c_{bd} denotes the response criteria that separates b- and d-responses. Note that the stimulus, S, should be replaced by S_a, S_v and S_av for auditory, visual and audiovisual stimuli respectively.

Likewise, the probability, $P(r_d)$, for a d-response, r_b, is given by

$$P(r_d) = P(c_{bd} < \hat{S} < c_{dg}) = \int_{c_{bd}}^{c_{dg}} P(\hat{S} \mid S) = \Phi(c_{dg}, \mu, \sigma) - \Phi(c_{bd}, \mu, \sigma) \tag{2.26}$$

where c_{dg} denotes the response criteria that separates d- and g-responses.

Finally, the probability, $P(r_g)$, for a g-response, r_g, is given by

$$P(r_g) = P(\hat{S} > c_{dg}) = \int_{c_{dg}}^{\infty} P(\hat{S} \mid S) = 1 - \Phi(c_{dg}, \mu, \sigma) \tag{2.27}$$

In general, the means μ_a and μ_v, for each stimulus are free parameters of the early strong fusion model. Also the standard deviations, σ_a and σ_v, are free parameters, which typically are assumed not to depend on the specific stimulus. Finally, the response criteria, c_{bd} and c_{dg} are also free parameters. Depending on the experimental paradigm, the strong fusion model may be over-parameterised making it a highly flexible model that is likely to over-fit. This can, of course, be tested using cross-validation as over-fitting models tend to have a high validation error despite a low training error. In the next section we will describe how we sometimes can use observer theory to constrain the model so that it will be less likely to over-fit.

2.2.3 Applying the early strong fusion model of audiovisual integration

Andersen (JASA, 2015) applied the early strong fusion model to categorical responses to auditory, visual and audiovisual speech stimuli. In a previous study, auditory speech had been generated using a speech synthesizer and videos of a face articulating speech had been generated using an animation. Five auditory and five visual stimuli had been generated so that they were perceptually evenly spaced on a continuum ranging from a clear /ba/ to a clear /da/. Additionally, 25 Audiovisual stimuli had been generated as the combination of the five auditory and five visual stimuli. Note that this continuum spans only the bilabial and alveolar (front to mid) range of the full range of place of articulation. Accordingly, the participants in the experiments could only respond with b- or d-responses. This simplifies the early strong fusion model, as the only dependent variable is the number of d-responses.

With these simplifications, the probability, $P_a(r_d)$, of a d-response given an auditory stimulus is given by

$$P_a(r_d) = P(\hat{S} > c \mid S_a) = \int_c^{\infty} P(\hat{S} \mid S_a) = 1 - F(c_a \mid \mu_a, \sigma_a) = \Phi\left(\frac{\mu_a - c_a}{\sigma_a}\right) \tag{2.28}$$

where $F(c_a \mid \mu_a, \sigma_a)$ denotes the cumulative Gaussian probability function with mean, μ_a, and standard deviation, σ_a.

Note that the last step in Equation 2.28 is based on the mathematical identity

$$1 - F(x \mid \mu, \sigma) = F(-x \mid \mu, \sigma) = \Phi\left(\frac{-x - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - x}{\sigma}\right) \tag{2.29}$$

Note also that since the auditory stimuli were perceptually evenly spaced along the /b/-/d/ continuum, we can recognise the last step in Equation 2.28 as the psychometric function from Equation 1.11 if we replace μ by the stimulus intensity, I_s. This is in line with the observer model illustrated in Figure 1.4 where each point on the psychometric function corresponds to the mean of the Gaussian distribution of internal representation values for that perceived stimulus intensity. Since the intervals of the synthetic continuum was designed to match the perceived intensity, we can replace $\mu = I_s$ by $1, \ldots, 5$ for the five auditory stimuli. Similarly the response probability, $P_v(r_d)$, of a d-response given a visual stimulus is given by replacing changing the subscript a by v.

Now that we have described the psychometric functions for the auditory and visual stimuli, we are ready to derive the response probabilities for audiovisual stimuli. The approach we will use is first to use the strong fusion model described in Equations 2.13-2.14 to derive expressions for the mean and standard deviation for audiovisual stimuli and then use Equation 2.28 to derive an expression for the response probabilities. We must, however, first align the internal representations. Recall that we have replaced the mean by the perceived signal intensity so that $\mu = I_s = 1, \ldots, 5$ for the five auditory and the five visual stimuli so that Equation 2.28 represents a psychometric function. We did this for both auditory and visual stimuli. However, Unless $c_a = c_v$, the two psychometric functions are misaligned, which will also misalign the auditory and visual internal representations. The means of the distributions for auditory and visual stimuli are, however, defined with respect to the response criterion, c. Introducing $\tilde{\mu} = \mu - c = I_s - c$ will fix this problem. This is illustrated in Figure 2.1.

To summarise, applied to the experimental paradigm used by Andersen (JASA, 2015), the early strong fusion model has four free parameters: c_a, c_v, σ_a and σ_v. From these we can directly calculate the response probabilities for auditory and visual stimuli according to Equation 2.28 by setting $\mu = I_s = 1, \ldots, 5$. In order to calculate the response probabilities for audiovisual stimuli we must first calculate $\tilde{\mu}_a = \mu_a - c_a$ and $\tilde{\mu}_v = \mu_v - c_v$. We can then calculate $\tilde{\mu}_{av} = w_a\tilde{\mu}_a + (1 - w_a)\tilde{\mu}_v$ according to Equations 2.13. Finally, we can calculate $P_{av}(r_d) = \Phi\left(\frac{\tilde{\mu}_{av}}{\sigma_{av}}\right)$ according to Equation 2.28.

With this parameterization, Andersen (JASA, 2015) could show that the early strong fusion model provided a good fit to the experimental data. Importantly, the use of the observer model to limit the number of free parameters ensured that validation error in a leave-one-stimulus cross validation test was low.

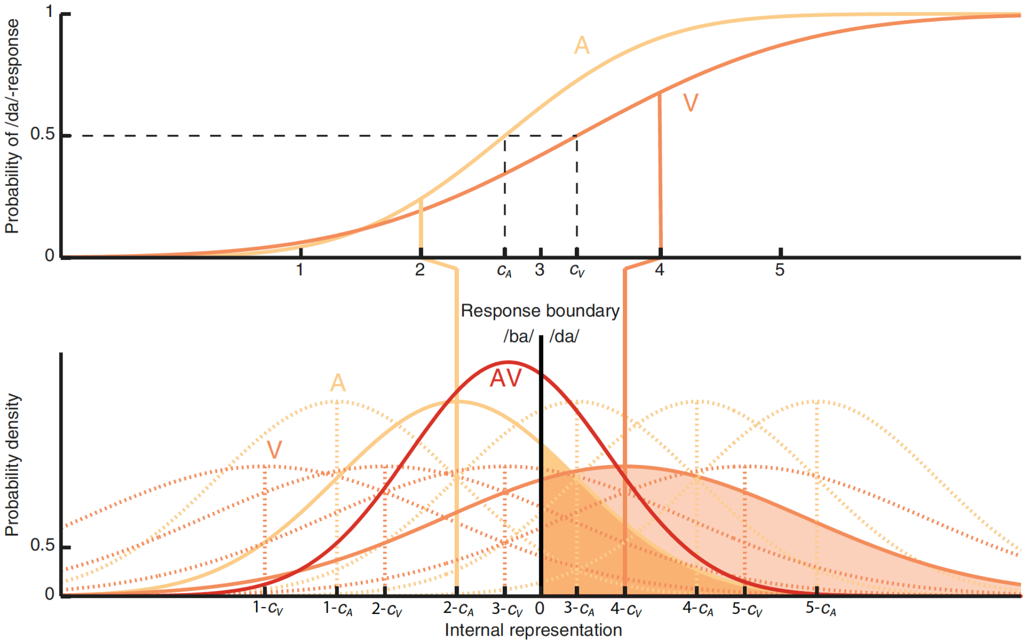


Figure 2.1: The psychometric function and observer model for audiovisual speech stimuli. The top panel depicts psychometric functions for auditory and visual stimuli. Note that they have different thresholds, c_a and c_v. The lower panel depicts the corresponding internal representation. Substituting the means, μ_a and μ_v, by $\tilde{\mu}_a = I_s - c_a$ and $\tilde{\mu}_v = I_s - c_v$ where $I_s = 1, \ldots, 5$ denotes the perceived stimulus intensity aligns the internal representations to a common threshold arbitrarily set to zero. Figure from Andersen (JASA, 2015).

The early maximum likelihood estimation model of audiovisual integration in speech perception

Tobias S. Andersen

Citation: The Journal of the Acoustical Society of America **137**, 2884 (2015); doi: 10.1121/1.4916691
View online: https://doi.org/10.1121/1.4916691
View Table of Contents: https://asa.scitation.org/toc/jas/137/5
Published by the Acoustical Society of America

CALL FOR PAPERS
JASA
THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA
Special Issue: Fish Bioacoustics: Hearing and Sound Communication

The early maximum likelihood estimation model of audiovisual integration in speech perception

Tobias S. Andersen[a)]
Section for Cognitive Systems, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark

(Received 4 July 2014; revised 12 March 2015; accepted 17 March 2015)

Speech perception is facilitated by seeing the articulatory mouth movements of the talker. This is due to perceptual audiovisual integration, which also causes the McGurk−MacDonald illusion, and for which a comprehensive computational account is still lacking. Decades of research have largely focused on the fuzzy logical model of perception (FLMP), which provides excellent fits to experimental observations but also has been criticized for being too flexible, *post hoc* and difficult to interpret. The current study introduces the early maximum likelihood estimation (MLE) model of audiovisual integration to speech perception along with three model variations. In early MLE, integration is based on a continuous internal representation before categorization, which can make the model more parsimonious by imposing constraints that reflect experimental designs. The study also shows that cross-validation can evaluate models of audiovisual integration based on typical data sets taking both goodness-of-fit and model flexibility into account. All models were tested on a published data set previously used for testing the FLMP. Cross-validation favored the early MLE while more conventional error measures favored more complex models. This difference between conventional error measures and cross-validation was found to be indicative of over-fitting in more complex models such as the FLMP. © *2015 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4916691]

[JFC] Pages: 2884–2891

I. INTRODUCTION

Speech perception is facilitated when the face of the talker is seen, as in face-to-face conversation, compared to when it is not, as in a phone conversation (Sumby and Pollack, 1954). This effect is stronger when auditory speech perception is poor and is an important aid for hearing impaired listeners (Grant *et al.,* 1998). The effect is widely believed to be caused by perceptual audiovisual integration rather than just a post-perceptual combination of information from auditory speech perception and lip-reading. The McGurk−MacDonald illusion is a striking demonstration of the perceptual nature of audiovisual integration of speech (MacDonald and McGurk, 1978; McGurk and MacDonald, 1976). In this illusion a speech sound, e.g., /ba/, is dubbed onto a video of a face articulating an incongruent phoneme, e.g., /ga/. This creates an illusory percept, in this example, of hearing /da/.

Decades of research on the computational mechanisms underlying audiovisual integration in speech perception has largely focused on the fuzzy logical model of perception (FLMP), which has been shown to provide good fits to empirical data in a number of studies (Massaro, 1998; Massaro and Cohen, 1983, 2000; Massaro *et al.,* 2011; Schwartz, 2010). The good fits of the FLMP have, however, been argued to be due to the model's flexibility rather than its ability to capture the underlying computational mechanisms (Andersen *et al.,* 2002; Cutting *et al.,* 1992; Myung and Pitt, 1997; Pitt, 1995; Pitt *et al.,* 2003; Schwartz, 2003, 2006; Vroomen and Gelder, 2000). Although much of this criticism has been addressed (Massaro, 2000, 2003; Massaro and Cohen, 1993; Massaro *et al.,* 2001) a consensus has not been reached despite the invocation of a broad spectrum of methods for model evaluation.

The current study has two main purposes. First, it introduces early maximum likelihood estimation (MLE) (Andersen *et al.,* 2005) as a new model of audiovisual integration in speech perception. Early MLE is based on the MLE model of multisensory integration of continuous representations (Ernst and Banks, 2002). By introducing a response boundary, as in signal detection theory (Green and Swets, 1966), the model can be applied to categorical responses. In this model, integration occurs before categorization, hence the name early MLE.

The idea of modeling audiovisual integration in speech perception based on a continuous internal representation is not new. The pre-labeling model introduced by Braida (1991) is also based on this idea but differs in the way it models the mechanism of audiovisual integration. In the pre-labeling model, auditory and visual internal representations are assumed to be orthogonal and integration occurs by basing the decision on the Pythagorean sum of the two. Thus, the model is inherently multidimensional. In addition to assigning separate dimensions to the perceptual modalities, the pre-labeling model can also assign multiple dimensions to the representation within modalities. A multidimensional phonetic representation is probably very realistic as speech perception relies on multiple perceptual features. This realism comes, however, at the cost of computational complexity because the multiple

[a)] Author to whom correspondence should be addressed. Electronic mail: toban@dtu.dk dimensions necessitate numerical evaluation of multidimensional integrals when fitting it. To avoid this problem, the current study is limited to models with a one-dimensional internal representation. The current study aims to show that this can be done without critical loss of realism when applying the models to data from experimental paradigms in which a single phonetic contrast is varied.

The MLE principle is also not new to models of audiovisual speech perception as it is inherent to the FLMP, which can be interpreted as MLE based on categorical representations (Massaro, 1998). Hence, integration happens after categorization in the FLMP and it can therefore be seen as a late MLE model. Although this difference between early and late MLE integration may seem subtle, the current study aims to show that there are great differences between the models in terms of parameterization and model complexity. These differences form the basis for the design of three related models all of which will be considered as alternatives to early MLE and the FLMP.

The other main purpose of this paper is to show that cross-validation effectively includes both goodness-of-fit and model flexibility in model evaluation, and provides meaningful selection of models. The development of methods for model evaluation in cognitive and perceptual science is an important field in its own right and the FLMP has had an important role in this field. Model evaluation methods that have been applied to the FLMP can be divided into three categories.

First, methods such as Akaike's information criterion (AIC; Akaike, 1974, Pitt *et al.*, 2002), Bayesian information criterion (Schwarz, 1978; Pitt *et al.*, 2002) and the root mean squared error (RMSE) corrected for the number of degrees of freedom (e.g., Massaro, 1998) depend on the goodness-of-fit penalized by a function of the number of free parameters. The problem with these methods is that since the number of free parameters is a poor measure of model complexity for non-linear models they do not always correct adequately for model complexity (Pitt *et al.*, 2002). This is problematic when evaluating the FLMP, which is non-linear (Myung and Pitt, 1997), especially in some regions of its parameter space (Andersen *et al.*, 2002; Schwartz, 2006).

Other methods, such as the Bayes factor (Massaro *et al.*, 2001; Myung and Pitt, 1997; Schwartz, 2006, 2010) and minimum description length (Pitt *et al.*, 2002) do not suffer from this problem but are algorithmically complex (Pitt *et al.*, 2002) although a simplifying assumption exists for the Bayes factor (see Schwartz, 2010).

Finally, cross-validation methods do not suffer from the same problems: They apply to all types of models and are straightforward to apply. They aim to estimate the generalization, or prediction, error, which is the expected error for new data not used in fitting the model parameters. The generalization error differs from the training error, the error for the data that were used in fitting the model parameters. Variability in data is generally due to fixed and random effects. Flexible models will, generally, fit closely to both types of variations. This is problematic because they have, so to speak, found a trend in randomness and this trend will, generally, not reappear in new data. This is called overfitting and causes flexible models to have high generalization errors. At the other end of the spectrum of complexity are models that are not sufficiently complex to capture the fixed effects. These models are said to under-fit and will have high training errors as well as high generalization errors. Somewhere between these two extremes lies the true model that fits the fixed effects perfectly. The true model will have higher training error than more flexible models because it cannot accommodate random variations in the data. This is why the training error is a poor criterion for evaluating models. The true model will, however, have the lowest possible generalization error, which is why the generalization error is the ideal criterion for evaluating models. The problem is that estimating the generalization error requires separate data for fitting the model and for evaluating the model. This increases the amount of data required. Pitt and Myung (2002) provide a good introduction to these concepts.

In cross-validation the data are split into a training set, which is used for fitting the model, and a test set, which is used for estimating the generalization error. The process of splitting, fitting, and evaluating is repeated so that all the data are used in the evaluation. In this way, cross-validation circumvents the requirement for separate training and test data. Each split is called a *fold* and the sum of the generalization error estimates across folds is called the test, or validation, error. The validation error is thus an estimate of the generalization error, which is based on the entire data set. Hastie *et al.* (2009) and MacKay (2003) provide introductions to cross-validation and compare it to other model evaluation techniques.

Data splitting can be done in several different ways: between observers, trials, conditions, or stimuli. It is important that the way that the data are split reflects how the model aims to generalize. The FLMP and other models of audiovisual integration aim at predicting the audiovisual percept based on the auditory and visual percepts, or, more generally, at generalizing perception across stimuli and modalities. Therefore, cross-validation splits should be made between stimuli within observers.

To ensure that models and methods are compared using representative data all model comparisons in the current study are based on the University of California Santa Cruz (UCSC) corpus (Massaro, 1998; Massaro *et al.*, 1995; Massaro *et al.*, 1993), which has been used extensively for comparing models of audiovisual integration of speech (Massaro, 1998; Massaro *et al.*, 2001; Schwartz, 2006, 2010; Wagenmakers *et al.*, 2004).

II. METHODS

A. Data

The data used in the current study are the UCSC corpus collected by Massaro and co-workers (Massaro, 1998; Massaro *et al.*, 1993; Massaro *et al.*, 1995) who kindly made it available online.[1] In this data set, 82 observers identified five auditory, five visual, and 25 audiovisual speech stimuli. The stimuli were synthesized using a speech synthesizer and an animated talking head. The auditory and visual stimuli were designed to fall approximately linearly on a continuum ranging from a clear /ba/ to a clear /da/. The audiovisual stimuli consisted of all the 25 possible combinations of the auditory and visual stimuli. The observers identified the stimuli as /ba/ or /da/. All data were stored as the proportion of /da/-responses. According to the reports describing the experimental procedures, each stimulus was presented 24 times. Hence, multiplying the response proportion by 24 should yield the response counts, which should be integer values. This was not the case for several response proportions indicating that there was some variability in the number of stimulus presentations. This has prevented the usage of likelihood based error measures in the current study, which therefore uses error measures based on the squared error.

B. Models

1. Gaussian model without integration

The Gaussian model without integration only introduces a psychometric function in order to impose constraints based on the experimental design. The purpose of this model is two-fold. First, it is contained in some of the models of integration described here. Hence, comparing these models with the model without integration will provide a more detailed view of whether it is their mechanisms of integration or the psychometric function that determines their performance. Second, as the model without integration has the highest number of free parameters of the models in this study it will serve to show how the number of free parameters influences model performance in terms of goodness-of-fit and validation error differently. As such it serves as a baseline model with maximal complexity.

The psychometric function, $\Phi(S;\ c,\ \sigma)$, is here the Gaussian cumulative distribution function. It returns the probability of a /da/-response as a function of the stimulus level, $S = 1,\ldots,5$, where $S = 1$ indicates a clear /ba/ and $S = 5$ indicates a clear /da/. The psychometric function has two free parameters: the threshold parameter, c, denoting the 0.5 threshold and the standard deviation, σ, which determines the slope of the function. Hence, the psychometric function models the response proportions for five data points using two free parameters. For audiovisual stimuli, the stimulus level, $S_{AV} = 1,\ldots,5$, is determined by the stimulus level of the auditory component of the stimulus while the slope and threshold depend on the visual stimulus. Technically, this model can also be constructed so that the visual stimulus component determines the stimulus level while the slope and threshold depend on the auditory stimulus but, for simplicity, this model is not included in the current study. The complete model thus consists of seven psychometric functions: one auditory, one visual and five audiovisual. As each function contains two free parameters, the model has 14 free parameters. The way in which the visual stimulus influences auditory perception in this model does not reflect a perceptual integration process, which is why the model is referred to as a model without integration.

The psychometric function can be interpreted as a model of the underlying perceptual process (Gescheider, 1997). According to this model observers base their responses on a scalar internal representation value, x, of a stimulus feature that distinguishes /ba/ from /da/. If the value of the internal representation exceeds the threshold, c, the observer responds /da/. Otherwise the observer responds /ba/. The mapping of the stimulus onto the internal representation is stochastic due to additive noise. The values of x are thus distributed according to the normal probability density function, $\varphi(x;\ \mu,\ \sigma)$ with mean $\mu = S$. The probability of responding /da/ is the probability of x exceeding the threshold, $x > c$, which is given by the integral

$$\int_c^\infty \varphi(x;\mu,\sigma) = \Phi(\mu;c,\sigma) = \Phi(S;c,\sigma).$$

The psychometric function, thus allows us to transform response probabilities to probability densities on a continuous internal representation. This is of great interest because cross-modal integration of continuous internal representations of stimulus features such as spatial location or size has been successfully model by the MLE model (Alais and Burr, 2004; Ernst and Banks, 2002).

2. Early MLE

In the MLE model, the distributions, φ_A and φ_V, of the auditory and visual internal representation values, x_A and x_V, are assumed to be independent. Therefore, the maximum likelihood estimate of the corresponding audiovisual distribution φ_{AV} is the normalized product of the auditory and visual probability densities, φ_A and φ_V. This product is also a Gaussian distribution with a mean, μ_{AV}, which is a weighted sum of the means, μ_A and μ_V, of the distributions, φ_A and φ_V. The weights, w_A and w_V, are given by the expressions $w_A = r_A/(r_A + r_V)$ and $w_V = r_V/(r_A + r_V)$. Note that the weights are mutually dependent since $w_A = (1–w_V)$. The parameter, $r = \sigma^{-2}$, denotes the precision. The more precise, or reliable, modality is thus given greater weight. This is known as the information reliability principle and is in accordance with many observations in studies of multisensory perception (Andersen *et al.*, 2004; Alais and Burr, 2004; Ernst and Banks, 2002). The precision, r_{AV}, of the audiovisual distribution is given by the (unweighted) sum of the reliabilities, r_A and r_V, of the auditory and visual distributions. Hence, integration of information always leads to a more precise estimate according to the MLE model.

Inherent to MLE is the assumption that the auditory and visual internal representations are one and the same. Hence the threshold, c, should be the same for the auditory and visual internal representations, which it is not in the model without integration. Alignment of the representations and thresholds can be achieved by noticing that $\Phi(S;\ c,\ \sigma) = \Phi(S - c;\ 0,\ \sigma)$. This transformation has no effect on the psychometric function but it implies a shift of the mean, $\mu_A = S_A - c_A$ and $\mu_V = S_V - c_V$, of the probability density functions, φ_A and φ_V. This aligns the auditory, visual, and hence also the audiovisual internal representations so that the threshold is zero for all of them. It also contains an important constraint on the early MLE model: just as the stimulus levels, S_A and S_V, are fixed at integer values from 1 to 5, so are the means of the distributions within a modality evenly distributed with a distance of 1 between them. It thus only takes two free parameters, c_A and c_V, to determine the means of the five auditory and five visual distributions. This is illustrated in Fig. 1. The early MLE model thus has four free parameters; two for the auditory and two for the visual psychometric function.

3. The weighted model

In early MLE, the reliability, r, of the auditory and visual modalities determines their weight, w. The weighted model releases this constraint and assigns a free parameter to the weight. The standard deviation of the audiovisual probability density function is given by summing of variances $\sigma_{av}^2 = w_a^2\sigma_a^2 + w_v^2\sigma_v^2$.

There are several reasons for why the weight given to each modality would not be determined (entirely) by its reliability. First, early MLE assumes that the distance between stimulus levels is identical for auditory and visual stimuli. If this assumption is violated the auditory and visual internal representations are scaled differently and it is not possible to determine the standard deviation of the auditory and visual probability densities relative to one another. This difference in scale will thus require an additional free parameter and it can be shown that adding this free parameter to the early MLE model makes it equivalent to the weighted model. Another reason is that stimuli in one modality may distract attention from the other modality. This could mean that the information in one modality is more reliable for unimodal stimuli, when attention is focused, than for bimodal stimuli, when attention is divided (Andersen *et al.*, 2005). The weighted model can take this effect into account.

4. The FLMP

In the FLMP, audiovisual integration is based on response probabilities (or, equivalently, fuzzy truth values). If P_a and P_v denote the auditory and visual response probability, respectively, then the audiovisual response probability is given by the normalized product of P_a and P_v,

$$P_{av} = \frac{P_a P_v}{P_a P_v + (1 - P_a)(1 - P_v)}.$$

Applied to the UCSC corpus, the FLMP requires 10 free parameters—five for the auditory response probabilities and five for the visual response probabilities.

Note that as the audiovisual probability distribution is based on the normalized product of the auditory and visual probability distributions, the FLMP can be interpreted as MLE based on a categorical internal representation (binomial in this case, multinomial in the general case of more than two response categories). Therefore, integration occurs after categorization and the FLMP can thus be considered as being based on late MLE.

5. Gaussian late MLE model

Early MLE has the potential advantage that the constraints imposed on the experimental design—using stimuli evenly spaced on a continuum—are incorporated into the model. The FLMP does not have this potential advantage. Any difference in the performance of the two models can thus be due to this as well as on the different ways (early vs late) they implement MLE. It is, however, possible to construct a late MLE model that contains a continuous internal representation. In this model, the auditory and visual response probabilities are calculated from the psychometric function exactly as in early MLE. The audiovisual response

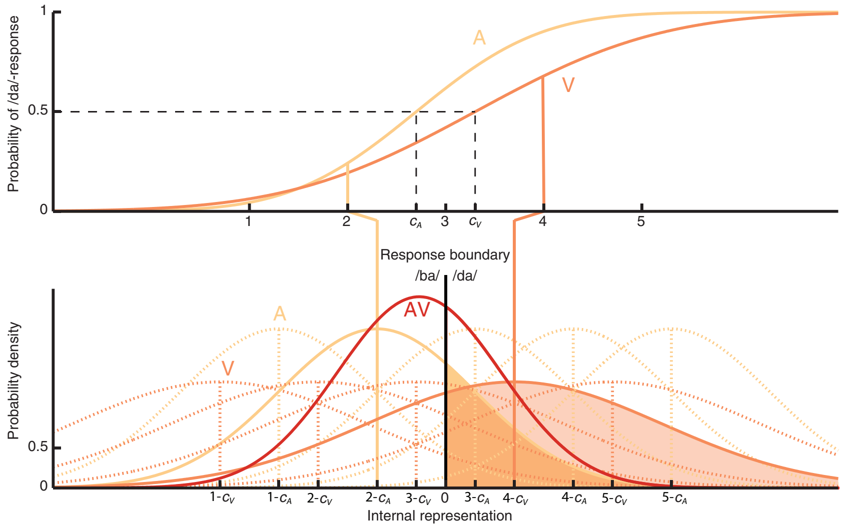


FIG. 1. (Color online) Illustration of the early MLE model. Upper axis: Example auditory (A) and visual (V) psychometric functions. Lower axis: Probability density functions of auditory (A) and visual (V) internal representation values corresponding to the psychometric functions in the upper axis. Each stimulus level in the upper axis determines the mean of a distribution in the lower axis. Examples are shown by lines connecting the axes. Note that the even spacing between stimulus levels in the upper axis is reflected in the even spacing between the distributions in the lower axis. The means of the five auditory and five visual distributions are thus determined by the auditory and visual thresholds, c_A and c_V, respectively. The example probability density function for the audiovisual internal representation values is calculated from MLE integration of the solid auditory and visual density functions. The response probability (of a /da/ response) is given by the probability mass falling above zero. Examples are shown by shaded areas.

probabilities are then calculated exactly as in the FLMP. This model is here termed the Gaussian late MLE model because it contains Gaussian noise in the early, continuous stage and late MLE as the model of integration. The parameters of the Gaussian late MLE model are the same as the parameters of the early MLE model.

6. Summary of models

The five models are summarized in Table I. The data set contains 35 data points (response proportions) for each subject (five auditory, five visual, and 25 audiovisual). In the five models there are three ways of reducing this complexity. First, the psychometric function (with no modeling of integration) reduces five degrees of freedom to two. Hence the Gaussian model without integration reduces 35 degrees of freedom to $2 \times 35/5 = 14$ free parameters. Second, modeling integration (without a psychometric function) predicts the 25 audiovisual data points from the five auditory and five visual data points. Hence the FLMP has $35 - 25 = 10$ free parameters. Finally, including both the psychometric function and a model of audiovisual integration predicts 35 data points from two psychometric functions. The early and late MLE models thus contain $2 + 2 = 4$ free parameters. The weighted model containing an additional free parameter for the weight contains five free parameters.

C. Fitting and cross-validation

The five models were all fitted to the data from each subject by minimizing the squared error between observed response proportions and the model response probabilities using the non-linear least squares solver from the Matlab$^{\text{TM}}$ Optimization Toolbox. As this is an unconstrained solver, constrained parameters were modeled as transformed unconstrained parameters. The weight, w, in the weighted model and response probabilities, P_a and P_v, in the FLMP were constrained to the range of 0 to 1 by applying a sigmoid function to unconstrained parameters. Standard deviations, σ, were constrained to be positive by applying the exponential function to unconstrained parameters.

Every model was fitted with 100 random initial conditions to minimize the chance of the optimization ending in a local minimum. The RMSE was calculated as the square root of the mean squared error for each subject. For each model, the RMSE corrected for degrees of freedom, henceforth referred to as the corrected RMSE, was calculated by dividing the RMSE by $(N_d - N_p)/N_d$, where N_d denotes the number of independent data points (35) and N_p denotes the number of free parameters.

Cross-validation was performed as a 35-fold leave-one-out procedure in which the models were fitted to the data from each subject separately. In each fold, the response proportion for one stimulus was left out from the fit. The validation squared error was then calculated between the model response probability and the observed response proportion for the stimulus left out from the fitting. The validation RMSE was then calculated as the square root of the across-fold mean squared error for each subject.

To test the significance of the differences in validation errors across models, the validation errors were subject to a one-way repeated measures analysis of variance (ANOVA). *Post hoc* tests were conducted in two ways. First, the

TABLE I. The parameters (pars.), their number (# pars.) and equations for the five models. P_A, P_V, and P_{AV} denote response probabilities for auditory, visual, and audiovisual stimuli, respectively. S_A, S_V, and S_{AV} denote stimulus level for auditory, visual, and audiovisual stimuli, respectively.

Model	Parameters	N_p	Description	Equations
Gaussian model w/o integration	C_A, σ_A C_V, σ_V C_{AV}, σ_{AV}	14	Thresholds and slopes for auditory, visual and five audiovisual psychometric functions	$P_a = \Phi(\mu_A;0, \sigma_A)$ $P_v = \Phi(\mu_V;0, \sigma_V)$ $P_{av} = \Phi(\mu_{AV};0, \sigma_{AV})$ $\mu_A = S_A - c_A$ $\mu_V = S_V - c_V$ $\mu_{AV} = S_{AV} - c_{AV}$
Early MLE	C_A, σ_A C_V, σ_V	4	Thresholds and slopes for auditory and visual psychometric functions	P_a, P_v, P_{av}, μ_A, and μ_V as in the Gaussian model w/o integration $\mu_{AV} = w_A\mu_A + w_V\mu_V$ $w_A = r_A/(r_A + r_V)$ $w_V = r_V/(r_A + r_V)$ $r_A = \sigma_A^{-2}; r_V = \sigma_V^{-2}$ $\sigma_{AV} = r_{AV}^{-0.5}; r_{AV} = r_A + r_V$
Weighted model	C_A, σ_A C_V, σ_V	5	Thresholds and slopes for auditory and visual psychometric functions	P_a, P_v, P_{av}, μ_A, and μ_V as in the Gaussian model w/o integration $\mu_{AV} = w_A\mu_A + (1-w_A)\mu_V$ $\sigma_{AV} = \sqrt{w_a^2\sigma_a^2 + w_v^2\sigma_v^2}$
	w_A		Weight parameters	
FLMP	P_a, P_v	10	Auditory and visual response probabilities	$P_{av} = \dfrac{P_aP_v}{P_aP_v + (1-P_a)(1-P_v)}$
Gaussian late MLE	C_A, σ_A C_V, σ_V	4	Thresholds and slopes for auditory and visual psychometric functions	P_a, P_v, μ_A, and μ_V as in the Gaussian model w/o integration P_{av} as in FLMP

validation error of each model was compared to every other model using a two-tailed *t*-test. Second, in another, less conventional, way, the models were ordered according to their validation error. Paired, one-sided *t*-tests were then performed between consecutive models. This was done in order to conduct *post hoc* tests with a smaller number of independent tests.

III. RESULTS

The results of the model fitting and cross-validation are displayed in Fig. 2 as the RMSE, the corrected RMSE and the validation RMSE. The models are ordered by number of free parameters so that models with more free parameters are to the left of models with fewer free parameters. The horizontal dashed line indicates the expectation value for the mean RMSE. The expectation value is calculated as the standard deviation of the response proportion assuming that the response count is distributed according to the binomial distribution with the response probability estimated by the observed response proportion.

As seen in Fig. 2, the differences in validation errors between models appear to be rather small. However, the ANOVA showed that the difference between the means of the validation errors is highly significant [$p < 0.001$, Greenhouse−Geisser corrected $F(2.6, 209.4) = 34.8$]. *Post hoc* paired two-tailed *t*-tests showed that the validation error of the early MLE model is significantly lower than the validation error of all of the other four models ($p < 0.0002$ for each comparison). The validation error of the weighted model is significantly lower than that of the late MLE, the FLMP and the early Gaussian model without integration ($p < 0.02$ for each comparison). The late MLE does not have lower validation error than the FLMP ($p > 0.9$) but both the late MLE and the FLMP has lower validation error than the Gaussian model without integration ($p < 10^{-6}$ for each comparison). When the models were ranked according to their validation error, paired one-sided *t*-tests, confirmed this pattern of significance. The *p*-values of these tests are displayed in Fig. 2.

The goodness-of-fit of over-fitting models is highly sensitive to small changes in parameter values. To test whether this is the case for the models described here, a sensitivity analysis was performed. For each subject a random number ranging from −5% to +5% of the parameter values was added to the best fitting unconstrained parameters. The RMSE was then calculated for these parameters. This procedure was repeated 1000 times and the mean difference between this RMSE and the RMSE of the best fit was calculated. This mean difference was small (<0.01) for all models compared to the difference in RMSE between models.

IV. DISCUSSION

The first purpose of the current study is to evaluate the early MLE model in comparison with the FLMP and the three other models described above. The early MLE model had the lowest validation error of the five models tested here and the difference in validation error between early MLE and the weighted model was highly significant. This finding shows that early MLE is a promising new model of audiovisual integration of speech.

However, this promise should be accompanied by words of caution. MLE models, early or late, contain a very strong constraint: the influence of each sensory modality depends only on the reliability of that modality. Audiovisual integration of speech may however vary across individuals (Magnotti and Beauchamp, 2014; Schwartz, 2010) beyond what can be explained due to variability in unimodal perception. Schwartz (2010) introduced a weighted version of the FLMP to account for this and showed that it performed better than the unweighted FLMP when applied to the UCSC corpus, the same data set as used here. The difference between Schwartz' findings and the findings in the current study may be due to differences in the type of weighted model and differences in the model evaluation methods. It is also possible that the individual differences in integration are distributed so that a majority of subjects integrate in agreement with early MLE while a significant minority integrates differently. Although early MLE performed significantly better than the weighted model in the current study, there was some variability across subjects and the weighted model was actually better for 24 out of 82 subjects.

Furthermore, several results in the literature suggest that audiovisual integration of speech can be influenced by the state of the observer without a corresponding change in unisensory perception (Alsius *et al.*, 2005; Nahorna *et al.*, 2012; Tuomainen *et al.*, 2005). These findings may require models

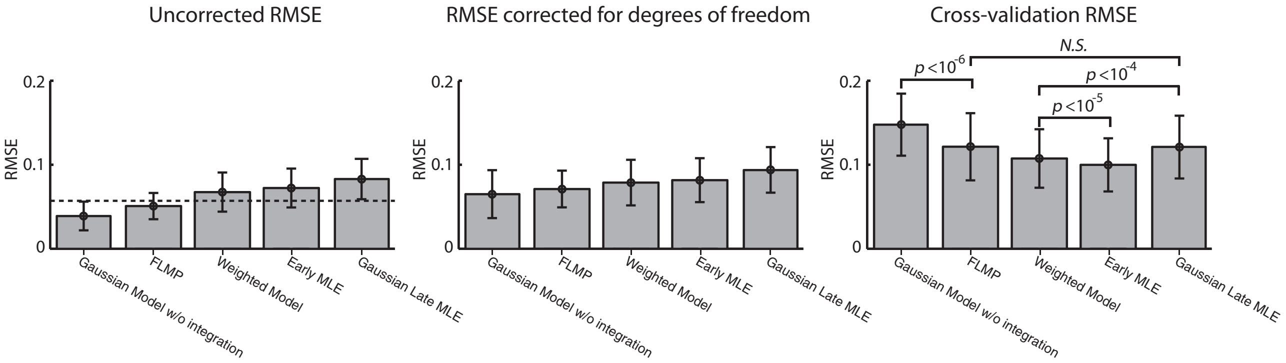


FIG. 2. The across-subject average RMSE, RMSE corrected for degrees of freedom, and validation RMSE for each of the seven models tested. Error bars represent the standard deviation (not the standard error of the mean as it would be too small to be clearly visible).

with a variable mechanism of integration such as weighted models or the Bayesian models of audiovisual integration suggested by Ernst (2006) and Shams *et al*. (2005).

So, how can the simple early MLE model perform well in the current study? The answer may lie in limitations in the data set. The single phonetic contrast, the limited number of stimuli along the stimulus continuum, the single signal-to-noise level and the lack of variation in the attentional state of the observer do not reflect the richness of everyday speech perception. Still, the data set has been influential and is not significantly smaller than data sets typically used in the literature on perceptual and cognitive models. This may indicate that the complexity of the data sets used to test models of audiovisual integration of speech so far has not matched the complexity of the models tested.

The weighted model and early MLE both had significantly lower validation error than Gaussian late MLE. This indicates that early integration reflects the mechanism of integration better than late MLE as this is the only difference between these two models.

Gaussian late MLE model did not have significantly lower validation error than the FLMP. This indicates that introducing the early continuous representation does not, in itself, lead to much improvement. This is confirmed by the FLMP having significantly lower validation error than the early Gaussian model without integration. From this we also learn that late MLE integration (Gaussian or FLMP) does seem to capture some of the underlying mechanism of integration, only not as well as early MLE.

The second purpose of the current study is to show that cross-validation effectively includes both goodness-of-fit and model flexibility in model evaluation, and provides meaningful selection of models. This is perhaps best seen by comparing model selection based on the validation error with model selection based on the corrected RMSE. Unsurprisingly, the RMSE consistently favored models with more free parameters. More importantly, this trend persisted when the RMSE was corrected for the degrees of freedom. Interestingly, this means that these measures did not favor the FLMP, in contrast to previous findings (Massaro, 1998), as the Gaussian model without integration, having the highest number of free parameters, had the lowest RMSE and corrected RMSE. This trend stands in stark contrast to the trend seen in the validation RMSE, which tends to favor the models with the fewest free parameters. The models with the more free parameters thus have low training errors and high validation errors, which is the hallmark of over-fitting. A further indication of over-fitting is that the RMSE was lower than the expectation value for the FLMP and the Gaussian model without integration. This suggests that these models fit not only to the variability due to fixed effects but also to variability due to the random effects.

The result of the sensitivity analysis indicated that all models were fairly robust to small variations in parameter values. Hence, although some models might over-fit in this study they do not do so to the extreme degree that was seen by Schwartz (2006) in a similar analysis of the FLMP. The reason for this discrepancy may be that Schwartz conducted his analysis on a different data set. This data set may have contained more response proportions close to zero for which the FLMP becomes highly non-linear and unstable.

That the early MLE is the best model of audiovisual integration of speech in terms of the cross-validation RMSE is a promising result. However, it may prove difficult to generalize it to more complex experimental designs that reflect real-life speech perception more closely. The reason for this is that whereas the continuous internal representation of speech is assumed to be one-dimensional in the current study, this is unlikely to be the case in general. Still, models with multidimensional representations do exist (Ashby, 1992) and it may be possible to insert a mechanism of integration into them. Although this may prove challenging, it also carries a promise: The inclusion of the experimental design in model design can lead to a more interpretable model with the dimensions of the model reflecting the perceptual features of audiovisual speech. Early MLE also contains a clear prediction for the effect of lowering the acoustic signal-to-noise ratio. This should lead to an increase in the variance of the Gaussian distribution in the auditory modality and increase the variability of responses across response categories as has been seen in early studies (Miller and Nicely, 1955). It should also lead to an increased visual influence in the McGurk illusion, which has also been reported (Sekiyama and Tohkura, 1991; Andersen *et al.*, 2001). The FLMP can make no such prediction, as it does not parameterize the acoustic signal-to-noise ratio.

The conclusion of the current study is that cross-validation shows that audiovisual integration of speech is best modeled by the parsimonious early MLE model in the UCSD data corpus. Whether more complex models, such as multidimensional or weighted models, are required to model audiovisual integration of speech in general will require more complex data sets and is a task is left for future studies.

ACKNOWLEDGMENT

This study was supported by the Oticon Centre of Excellence for Hearing and Speech Sciences (CHESS).

[1] http://mambo.ucsc.edu/psl/8236/ (Last viewed June 6, 2010).

Akaike, H. (**1974**). "A new look at the statistical model identification," IEEE Trans. Autom. Control **19**, 716–723.

Alais, D., and Burr, D. (**2004**). "The ventriloquist effect results from near-optimal bimodal integration," Curr. Biol. **14**, 257–262.

Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S. (**2005**). "Audiovisual integration of speech falters under high attention demands," Curr. Biol. **15**, 839–843.

Andersen, T. S., Tiippana, K., Lampinen, J., and Sams, M. (**2001**). "Modeling of audiovisual speech perception in noise," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, Aalborg, pp. 172–176.

Andersen, T. S., Tiippana, K., and Sams, M. (**2002**). "Using the fuzzy logical model of perception in measuring integration of audiovisual speech in humans," in *Proceedings of the First International NAISO Congress on Neuro Fuzzy Technologies*, Havana.

Andersen, T. S., Tiippana, K., and Sams, M. (**2004**). "Factors influencing audiovisual fission and fusion illusions," Brain Res. Cogn. Brain Res. **21**, 301–308.

Andersen, T. S., Tiippana, K., and Sams, M. (**2005**). "Maximum likelihood integration of rapid flashes and beeps," Neurosci. Lett. **380**, 155–160.

Ashby, F. G. (**1992**). "Multidimensional models of categorization," in *Multidimensional Models of Perception and Cognition*, edited by F. G. Ashby (Erlbaum, Hillsdale, NJ), pp. 449–483.

Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. A **43**, 647–677.

Cutting, J. E., Bruno, N., Brady, N. P., and Moore, C. (**1992**). "Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth," J. Exp. Psychol. **121**, 364–381.

Ernst, M. O. (**2006**). "A Bayesian view on multimodal cue integration," in *Human Body Perception From The Inside Out*, edited by G. Knoblich, I. M. Thornton, M. Grosjean, and M. Shiffrar (Oxford University Press, New York), Chap. 6, pp. 105–131.

Ernst, M. O., and Banks, M. S. (**2002**). "Humans integrate visual and haptic information in a statistically optimal fashion," Nature **415**, 429–433.

Gescheider, G. A. (**1997**). "Classical psychophysical theory," in *Psychophysics: The Fundamentals*, 3rd ed. (Psychology Press, East Sussex, UK), Chap. 4, pp. 73–103.

Grant, K. W., Walden, B. E., and Seitz, P. F. (**1998**). "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," J. Acoust. Soc. Am **103**, 2677–2690.

Green, D. M., and Swets, J. A. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).

Hastie, T., Tibshirani, R., and Friedman, J. (**2009**). "Model assessment and selection," in *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer, New York), Chap. 7, pp. 219–257.

MacDonald, J., and McGurk, H. (**1978**). "Visual influences on speech perception processes," Percept. Psychophys. **24**, 253–257.

MacKay, D. (**2003**). "Model comparison and Occam's razor," in *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK), Chap. 28, pp. 343–353.

Magnotti, J. F., and Beauchamp, M. S. (**2014**). "The noisy encoding of disparity model of the McGurk effect," Psychonom. Bull. Rev. pp. 1–9.

Massaro, D. W. (**1998**). *Perceiving Talking Faces* (MIT Press, Cambridge, MA), 507 pp.

Massaro, D. W. (**2000**). "Reply to Vroomen and de Gelder," Trends Cognit. Sci. **4**, 38–39.

Massaro, D. W. (**2003**). "Model selection in AVSP: Some old and not so old news," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, St. Jorioz, France, pp. 83–88.

Massaro, D. W., and Cohen, M. M. (**1983**). "Evaluation and integration of visual and auditory information in speech perception," J. Exp. Psychol. **9**, 753–771.

Massaro, D. W., and Cohen, M. M. (**1993**). "The paradigm and the fuzzy logical model of perception are alive and well," J. Exp. Psychol. **122**, 115–124.

Massaro, D. W., and Cohen, M. M. (**2000**). "Tests of auditory–visual integration efficiency within the framework of the fuzzy logical model of perception," J. Acoust. Soc. Am. **108**, 784–789.

Massaro, D. W., Cohen, M. M., Campbell, C. S., and Rodriguez, T. (**2001**). "Bayes factor of model selection validates FLMP," Psychonom. Bull. Rev **8**, 1–17.

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., and Tsuzaki, M. (**1993**). "Bimodal speech perception: An examination across languages," J. Phon. **21**, 445–478.

Massaro, D., Cohen, M. M., Meyer, H., Stribling, T., Sterling, C., and Vanderhyden, S. (**2011**). "Integration of facial and newly learned visual cues in speech perception," Am. J. Psychol. **124**, 341–354.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. (**1995**). "Cross-linguistic comparisons in the integration of visual and auditory speech," Mem. Cognit. **23**, 113–131.

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices," Nature **264**, 746–748.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Myung, I. J., and Pitt, M. A. (**1997**). "Applying Occam's razor in modeling cognition: A Bayesian approach," Psychonom. Bull. Rev. **4**, 79–95.

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (**2012**). "Binding and unbinding the auditory and visual streams in the McGurk effect," J. Acoust. Soc. Am. **132**, 1061–1077.

Pitt, M. A. (**1995**). "Data fitting and detection theory: Reply to Massaro and Oden," J. Exp. Psychol. **21**, 1065–1067 (1995).

Pitt, M. A., Kim, W., and Myung, I. J. (**2003**). "Flexibility versus generalizability in model selection," Psychonom. Bull. Rev. **10**, 29–44.

Pitt, M. A., and Myung, I. J. (**2002**). "When a good fit can be bad," Trends Cognit. Sci. **6**, 421–425.

Pitt, M. A., Myung, I. J., and Zhang, S. (**2002**). "Toward a method of selecting among computational models of cognition," Psychol. Rev. **109**, 472–491.

Schwarz, G. E. (**1978**). "Estimating the dimension of a model," Ann. Stat. **6**, 461–464.

Schwartz, J.-L. (**2003**). "Why the FLMP should not be applied to McGurk data or how to better compare models in the Bayesian framework," in *International Conference on Auditory–Visual Speech Processing (AVSP)*, St. Jorioz, France, pp. 77–82.

Schwartz, J.-L. (**2006**). "The 0/0 problem in the fuzzy-logical model of perception," J. Acoust. Soc. Am. **120**, 1795–1798.

Schwartz, J.-L. (**2010**). "A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent," J. Acoust. Soc. Am. **127**, 1584–1594.

Sekiyama, K., and Tohkura, Y. (**1991**). "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," J. Acoust. Soc. Am. **90**, 1797–1805.

Shams, L., Ma, W. J., and Beierholm, U. (**2005**). "Sound-induced flash illusion as an optimal percept," NeuroReport **16**, 1923–1927.

Sumby, W. H., and Pollack, I. (**1954**). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212–215.

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (**2005**). "Audiovisual speech perception is special," Cognition **96**, B13–B22.

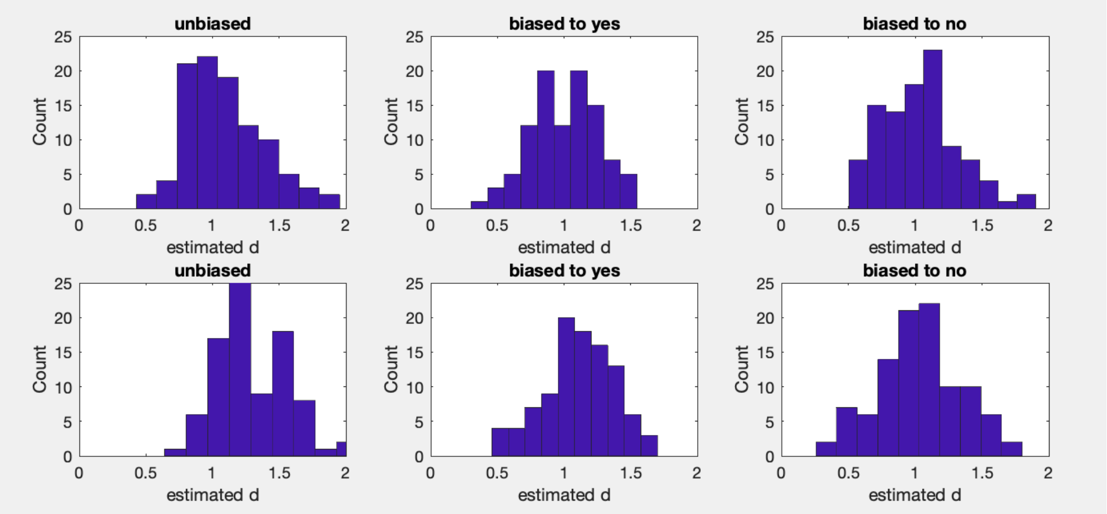
Vroomen, J., and de Gelder, B. (**2000**). "Crossmodal integration: A good fit is no criterion," Trends Cognit. Sci. **4**, 37–38.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and Iverson, G. J. (**2004**). "Assessing model mimicry using the parametric bootstrap," J. Math. Psychol. **48**, 28–50.

0.1 Week 1 - Equal variance model

We have 3 observers with criteria $c_y = 0.5$, $c_n = 0.o$ and $c_u = 1.1$ that are unbiased, biased towards yes-responses, and biased towards no-responses respectively. They all have sensitivity $d' = 1$. To sample the internal representation value, x, for $N = 50$ trials with no stimulus we draw 50 samples from $x \sim \mathcal{N}(\mu = 0, \sigma = 1)$ for each of the three observers. To obtain the number of false positives, n_{fp} we count the trials for which $x > c$ for each observer. Likewise, to sample the internal representation value, x, for $N = 50$ trials with stimulus we draw 50 samples from $x \sim \mathcal{N}(\mu = 1, \sigma = 1)$. To obtain the number of true positives, n_{tp} we count the trials for which $x > c$ for each observer. This allows us to estimate the perceptual sensitivity, d' as $d' = \Phi^{-1}\left(\frac{n_{tp}}{N}\right) - \Phi^{-1}\left(\frac{n_{fp}}{N}\right)$.

We repeat the process described above 100 times to simulate 100 experiments. A histogram of the obtained values for the estimate of d' for each observer are shown in the top plots in the figure below. Taking the average across experiments we found that the estimate for d' was 1.09, 1.01 and 1.03 for the unbiased, biased towards yes-responses, and biased towards no-responses observers respectively. This shows that we are able to retrieve the true value of d' on average for observers with different response criteria.

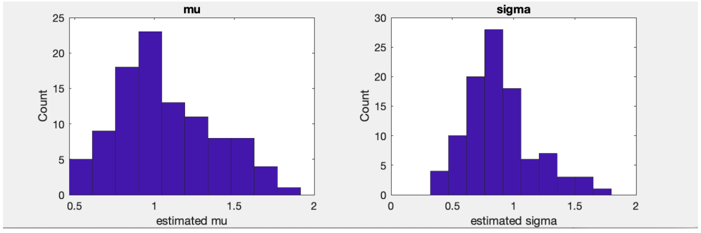


We then repeat the process but for unequal variance observers for which $\sigma = 0.8$. The process is the same as before except that we sample the internal representation value, x, for $N = 50$ trials with stimulus by drawing 50 samples from $x \sim \mathcal{N}(\mu = 1, \sigma = 0.8)$. A histogram of the obtained values for the estimate of d' for each observer are shown in the bottom plots in the figure above. Taking the average across experiments we found that the estimate for d' was 1.30, 1.12 and 1.03 for the unbiased, biased towards yes-responses, and biased towards no-responses observers respectively. This shows that d' was not estimated correctly and that the estimate depends on the criterion of the observer. Hence, d' is not a correct measure of sensitivity for an unequal variance observer.

0.2 Week 2 - ROC curves

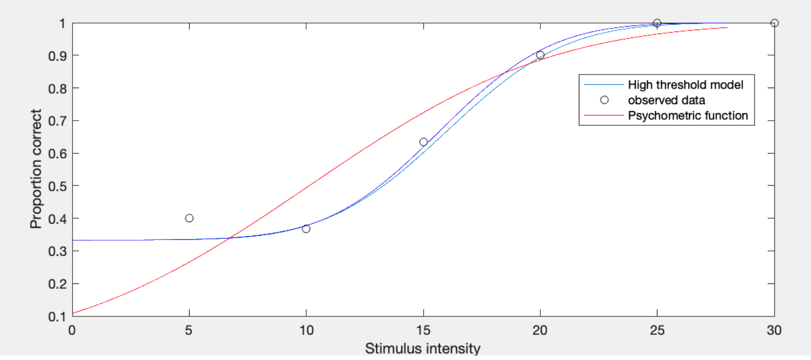
We have one observer responding in four ordered categories, so we need three response criteria, which are chosen to be the same as in the previous problem. The sampling of internal representation values, x, is done as in the unequal variance model above. Trials for which $x > 1.1$ are counted as 'yes - high confidence' responses. Trials for which $0.5 < x < 1.1$ are counted as 'yes - low confidence' responses. Trials for which $0.0 < x < 0.5$ are counted as 'no - low confidence' responses. Remaining trials are counted as 'no - high confidence responses.

We then count the number positive responses for each response criterion. For $c = 1.1$ only 'yes - high confidence' responses count as positive responses. For $c = 0.5$ all 'yes'-responses counts as positive responses. For $c = 0.0$ all 'yes'-responses *and* 'no - low confidence'-responses count as positive responses. True positive responses are responses to a stimulus trial and false positive responses are responses to a no-stimulus trials. This give us three points on the probit transformed ROC curve $\Phi^{-1}(P_{tp}) = \frac{1}{\sigma}\Phi^{-1}(P_{fp}) + \frac{\mu_s}{\sigma}$ by fitting a line to these points we can estimate the slope, a and the intercept, b, and calculate $\sigma = \frac{1}{a}$ and $\mu_s = b\sigma$. The process is repeated 100 times to simulate 100 experiments. The distribution of the estimated parameter values are shown in the histograms below. Averaging the estimates of μ_s and σ across experiments produced the values 1.08 and 0.88 respectively showing that the distributions are approximately centered around the true underlying values.



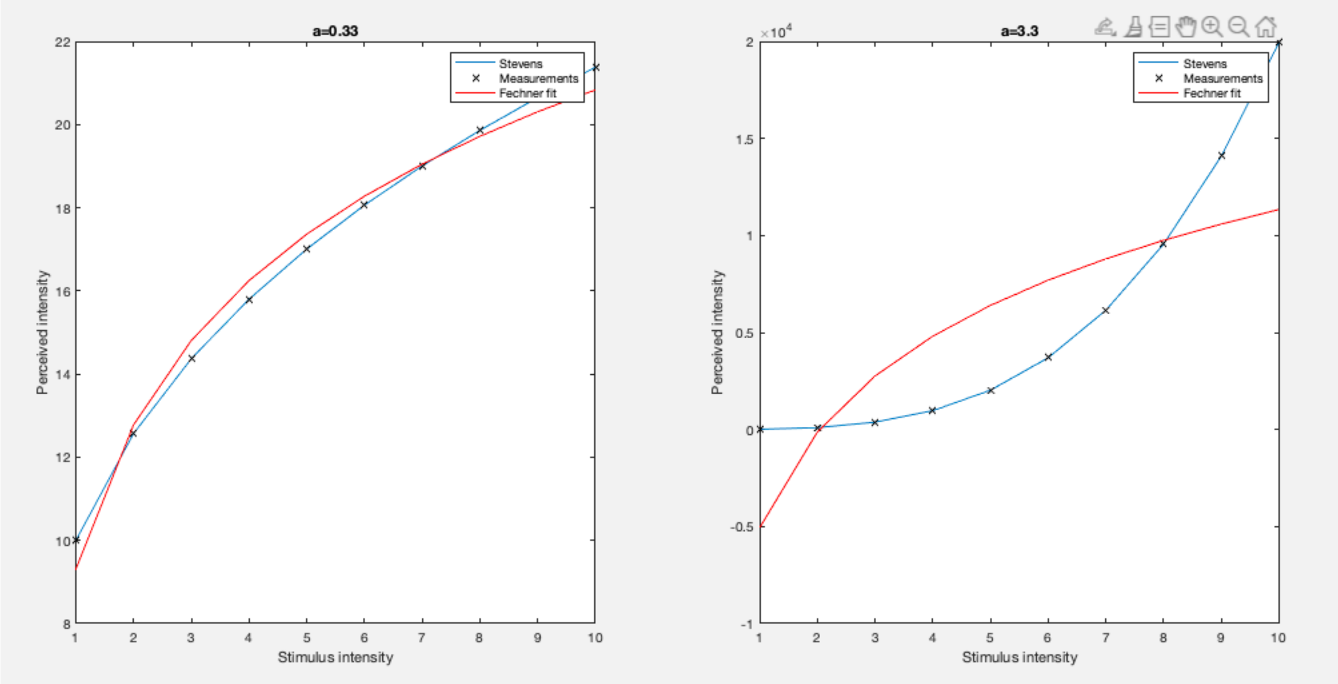
0.3 Week 2 - The psychometric function

We find the optimal parameter values for c and σ using an optimisation routine that maximises the log likelihood (Lecture notes Eq. 1.15) summed over stimulus intensities, I_s, where $P_s = \Psi(I_s)$ is given by the psychometric function (Eq. 1.11), n_s is the number of correct listed in the table and $N_s = 30$. We find that $\sigma = 8.19$ and $c = 10.13$. The negative log likelihood was found to be 11.32. We do the same for the high-threshold model using Eq. 1.12 and $P_{g}uess = \frac{1}{3}$. We find that $\sigma = 3.78$ and $c = 15.69$. These parameter estimates are very different from the parameter estimates above. The negative log likelihood was found to be 7.64. The psychometric functions and the response proportions are plotted below. From visual inspection we find that the high threshold fit the data better–especially for low stimulus intensities where the effect of guessing is greater. Accordingly, the likelihood of the high threshold model is greater than the likelihood of the psychometric function that does not take guessing into account.



0.4 Week 3 - Magnitude estimation

We first calculate the perceived intensity, I_p for stimulus intensities, $I_s = 1, 2, \ldots, 10$, using Stevens' law $I_p = 10I_s^a$ with $a = 0.33$. We then fit Fechner's law to these simulated data noticing that Fechner's law can be written as $I_p = \frac{1}{k_w} \ln I_s - \frac{1}{k_w} \ln I_0$, so that we can fit Fechner's law by linear regression with respect to $\ln I_s$ and with slope, $a = \frac{1}{k_w}$, and intercept, $b = -\frac{1}{k_w} \ln I_0$. This fit is plotted in the left plot below along with the simulated data. We observe that the fit is very good. The parameter values obtained from the fit are $k = 0.20$ and $I_0 = 0.16$. We then repeat the process but for $a = 3.3$ and fit Fechner's law again. This fit is plotted in the right plot below along with the simulated data. We observe that the fit is very poor. The parameter values obtained from the fit are $k = 0.00014$ and $I_0 = 2.04$. We conclude that Fechner's law can mimic Stevens' law for $a < 1$ but not when $a > 1$.


a=0.33
Stevens
Measurements
Fechner fit
Perceived intensity
Stimulus intensity
a=3.3
×10^4
Stevens
Measurements
Fechner fit
Perceived intensity
Stimulus intensity