The Morton–Massaro Law of Information Integration: Implications for Models of Perception

Javier R. Movellan
Department of Cognitive Science &
Institute for Neural Computation
University of California San Diego
James L. McClelland
Department of Psychology&
Center for the Neural Basis of Cognition
Carnegie Mellon University

Technical Report 2000.01

Machine Perception Laboratory
University of California, San Diego
La Jolla, CA 92093-0515

May 1, 2000

The Morton–Massaro Law of Information Integration: Implications for Models of Perception

Javier R. Movellan
Department of Cognitive Science &
Institute for Neural Computation
University of California San Diego
James L. McClelland
Department of Psychology&
Center for the Neural Basis of Cognition
Carnegie Mellon University

Abstract

Information integration may be studied by analyzing the effect of two or more sources (e.g., auditory and visual) on subjects' responses. Experiments show that ratios of response probabilities often factorize into components selectively influenced by only one source (e.g., one component affected by the acoustic source and another one affected by the visual source). We call this the Morton-Massaro law (MML). We find conditions where the law is optimal and note that it reflects an implicit assumption about the statistics of the environment. Adherence to the MML can be used to assess whether the assumption is being made, and analyses of natural stimuli can be used to determine whether the assumption is reasonable. Feedforward and interactive models subject to a channel separability constraint are consistent with the law.

Many experiments and demonstrations document that the perception of a wide range of stimuli is affected by multiple sources of information. Figure 1 shows a classic example, the Ebbinghaus' circles, in which the perceived size of an object is affected by its actual size and by the context in which it appears. Extensive empirical research has been carried out in a variety of fields including cognitive psychology, neuroscience, behavioral ethology, and machine perception, to understand how information is integrated in perception (InuiMcClelland96Stein & Meredith, 1993PartanMarler99; Clark & Yuille, 1990a).

In this paper we focus on a specific characteristic of the results that has often been observed in this research. The characteristic pattern arises in experiments examining the effects of variation of two or more different sources of information on perceptual identification responses. For example, one might investigate the effects of visual and auditory sources of information on the identification of a spoken syllable. The pattern will be described in detail below, but its essence can be stated in simple terms: The probabilities of identification responses (e.g., saying that a given item is /ba/) given to combinations of inputs from the different sources can be accounted for with a set of stimulus-response factors, where each factor represents the support provided by one of the sources to one of the possible identification responses. The pattern indicates that a given input from one source makes the same contribution in support of each identification response, regardless of the input from the other sources.

Two important and related theories have been developed that predict this empirical pattern, one by Morton (1969) and one by Massaro and colleagues (e.g. Oden & Massaro, 1978). Morton (1969) reviewed a wide range of experiments and proposed a model, called the *logogen model*, that described reasonably well the effects of stimulus and context information on word identification responses. Massaro and colleagues (Oden & Massaro, 1978; Massaro, 1987a, 1989a) developed a powerful experimental paradigm for the study of information integration in which several levels of each independent variable or information source are factorially combined (see Figure 2 for an example) and the dependent variables are the probabilities of a given set of response alternatives. Massaro and colleagues (Oden & Massaro, 1978) proposed a model of information integration named the *fuzzy logical model of perception* (FLMP) to explain the results obtained using this experimental paradigm.

An important feature of the logogen model and the FLMP is the fact that

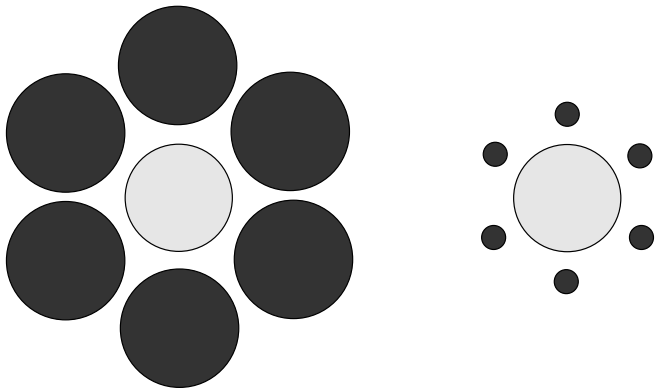


Figure 1: The Ebbinghaus circles are classic example of the effects of context on perception. The perception of the size of the inner circle is affected by the context.

they use autonomous modules that selectively process each of the different information sources, and are not influenced by the other sources or by later processing stages (see Figure 3). These modules map low-level inputs into quantities that represent the strength of support for high-level perceptual hypotheses (e.g., whether the stimulus is a "G" or a "Q"). As we will see later in more detail, the fact that these models are modular leads them to predict that the strength of support for each alternative will factorize as described above.

Both Morton and Massaro have pointed out many cases where the effects of different information sources can be factored in this way. Morton (1969) showed that his model provided a good fit to a number of experiments examining the effects of context and stimulus information on word identification, and Massaro and colleagues have shown that, as predicted by the FLMP, factorability provides good approximations for response probabilities obtained in a remarkable range of experiments in domains such as word and letter perception, object identification, depth perception, audio-visual speech recognition, memory retrieval and recognition of emotions—See Chapter 6 of Massaro (1989a) for a review.

The range of cases in which effects of different sources factor is extensive, and we refer to this fact as the *Morton–Massaro law*. However, it is important to note that the extent of applicability of this law is controversial. Many investigators have claimed to find violations (Samuel81Samuel96Bülthoff &

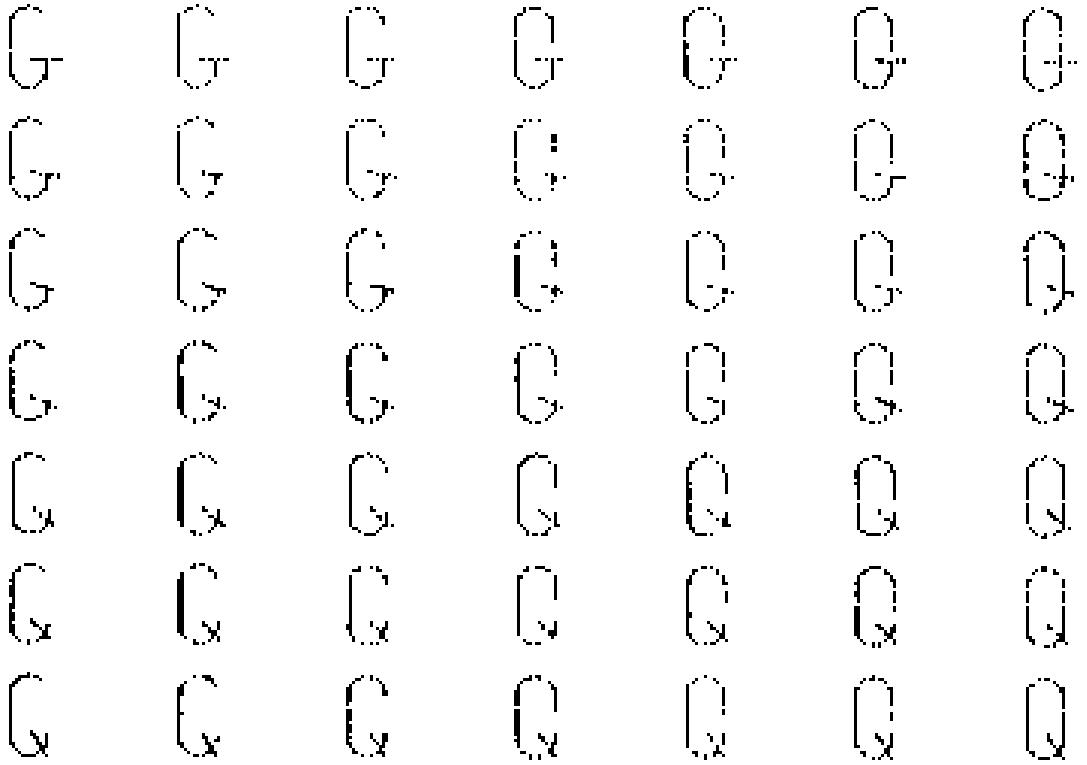


Figure 2: Stimuli illustrating Massaro's factorial experimental paradigm. On each trial subjects see one of the 49 stimuli for 400 msecs and are asked to categorize it as G or Q. The two sources of information are the angle of the straight line and the openness of the oval gap. Redrawn from Figure 6, p. 329 of Massaro89book) Permission Pending.

Yuille, 1996). Some such violations can be disputed (see, e.g., Massaro, 1996), leaving the exact extent of the applicability of the law in doubt. It is not our intention to take a stand on these particular cases, and we do not use the term "law" to endorse the view (Massaro, 1989a, ch. 4) that factorability is a universal principle of human information processing that spans all domains in which information integration is necessary. Instead, we use the term to suggest only that the empirical pattern of factorability is worth taking note of, in view of the wide range of cases in which it does hold, and also to suggest that the implications of factorability for theories of perception must be seriously considered. Thus, we treat the Morton-Massaro law as a very useful relationship whose confirmation or refutation in particular experiments may inform us about the underlying structure of the perceptual system. Indeed, we present a case of a violation of the law later in this article, and use it to illustrate how such a violation may reflect important characteristics of the underlying perceptual system that produces it. This way of using factorability of response probability ratios is similar to the way Sternberg used additivity of reaction times to understand the organization of underlying cognitive processes (Sternberg69).

An important issue raised by the Morton–Massaro law is its potential incompatibility with interactive models of perception, i.e, models in which all the processing units may be coupled via feed-back and lateral connections. According to the FLMP, perceptual processes are feed-forward and modular. They can be decomposed into three stages (see Figure 3): First there is an evaluation stage in which information sources are evaluated in a modular manner. Next there is an integration stage in which the outputs of the evaluation modules are integrated. Finally there is a decision stage in which a response is chosen based on the output of the integration stage. The outputs of the different stages may overlap in time and may be organized hierarchically, but in all such cases information processing is modularized and proceeds in a strictly feed-forward manner. (Note that when the FLMP is implemented as a process model, propagation is thought of as continuous, so that information made available to the integration and to the response stages can change gradually as a function of processing time.)

In neuroscience the feed-forward characterization reflects traditional views about the early stages of visual processing. For example, classic studies on the behavior of single neurons at different levels of the visual pathway showed that, in anesthetized cats and monkeys, neurons in the primary visual cortex have small receptive fields and selectively respond to simple visual stimuli

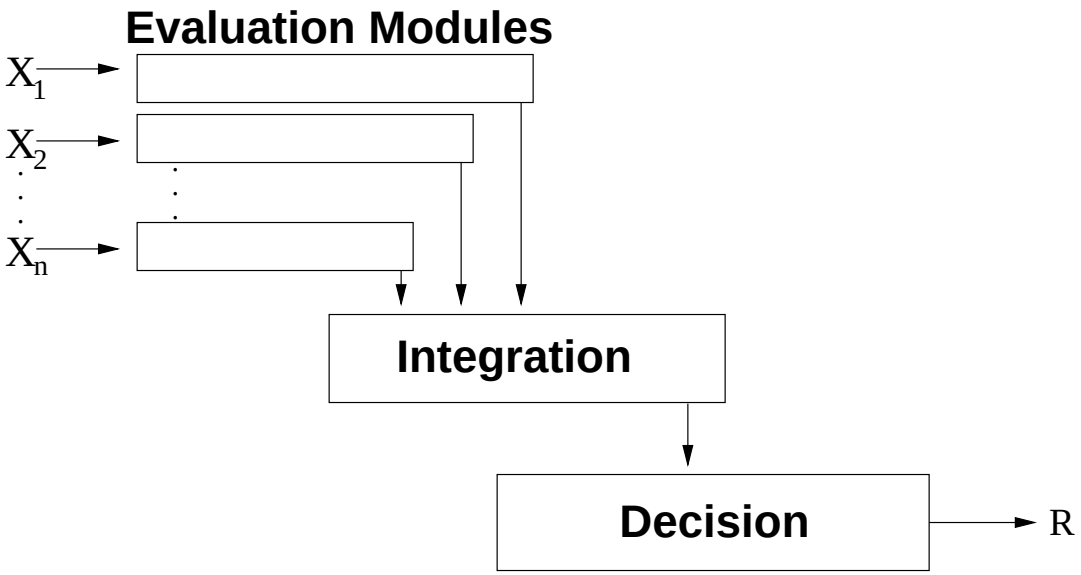


Figure 3: Schematic representation of the FLMP. The information sources are represented with the uppercase letter X. Each source of information is selectively processed by a different evaluation module. The output of each module represents the degree of support of an information source for the different response alternatives. These outputs are integrated at a later stage. The response stage operates on the output of the integration stage to produce an external response such as discrete decision or a continuous rating. Note that propagation of information may be cascaded, so that information available to the response stage may change as a function of processing time.

(e.g., gratings of a particular spatial frequency and orientation) while neurons in more anterior areas have larger receptive fields and respond to more complex stimuli (Hubel & Weisel, 1962, 1968). The notion that neurons in primary cortex have small local receptive fields is consistent with feed-forward information processing models. If feed-back and lateral connections had a functional influence on simple neurons in primary visual cortex, their receptive fields should be larger, reflecting the influence of neighboring neurons and of anterior neurons. Feed-forward models, like the logogen model or the FLMP, provide a nice information processing metaphor for this picture of the neurophysiology.

An alternative approach views perception as arising from a bi-directional propagation of information. In this approach, a perceptual processing sys-

tem may be subdivided into levels, but there are bi-directional influences between and within levels, so that the outcome of processing at every level is potentially subject to influence from higher and lower levels as well as from other units at the same level. McClelland and Rumelhart's (1981) interactive activation model of word perception is an early example of a model of this type. In this model a letter unit may influence a word-level unit through feed-forward connections and conversely the word-level unit can return the influence back to the letter level via feed-back connections (see Figure 4). The model was implemented as a connectionist network in which features, letters, and words were represented as local units that interacted according to simple processing principles. Top-down and bottom-up combination of information was achieved through bi-directional connections between the feature and letter units and between the letter and word units, as the figure illustrates. These bi-directional connections allowed the network to exhibit a wide variety of word perception phenomena, leading McClelland and Rumelhart (1981) to suggest that such connectionist models might serve as a general paradigm for perception, and these and other authors have applied these ideas to a range of topics in perception, attention, memory, word reading, language production, and conceptual knowledge representation (e.g., Dell, 1986; Hinton & Shallice, 1991; McClelland, 1981; McClelland & Elman, 1986; McClelland & Rumelhart, 1986; Plaut, McClelland, Seidenberg, & Patterson, 1996; Phaf, van der Heijden, & Hudson, 1990; Rumelhart, Smolensky, McClelland, & Hinton, 1986).

Interactive concepts also have a long history in neuroscience. Ramon y Cajal (1892) emphasized the existence of feed-back connections in his early studies of the brain, and Lorente de No (1922), a pioneer of theoretical ideas about neural computation, proposed "reciprocidad" as a design principle of the brain. Recent anatomical studies confirm this principle by showing that when there is a projection from one brain region to another, there is almost always a reciprocal return projection (Felleman & Essen, 1991). A number of neuroscientists (Peon, 1961; Sperry, 1969; Szentagothai & Arbib, 1974) have cautioned against modular and unidirectional views of information processing. Mumford92) points out that the majority of cortical cells have inter-area projections as opposed to exclusively intra-area projections and that this fact runs against modular computing metaphors. More recently, physiological studies on alert behaving animals have shown that contextual attributes far beyond the classical receptive field modulate the response of single neurons in primary visual cortex (Allman, Miezin, & McGuiness, 1985; Lamme,

1995; Lee, Mumford, Romero, & Lamme, 1998). These contextual effects are interpreted as due to feed-back connections from down-stream areas and collateral interactions among neurons within the same cortical region. There are now direct demonstrations that contextual influences seen in somatosensory thalamic nuclei arise from feed-back connections from somatosensory cortex (Ghazanfar & Nicolelis, 1997). Modular feed-forward models offer little help to understand this picture of the neurophysiology.

The interactive character of connectionist models with feed-back and lateral connections appears to be in direct contradiction with the modular character of models such as the FLMP. In interactive models all units may potentially affect and be affected by the state of all the other units and all the external information sources. It seems counterintuitive that in such models one could possibly have the contribution of a stimulus on the response probabilities being independent of all the other stimuli. Indeed Massaro (1989b) found that the original version of the interactive activation model did not adhere to the Morton–Massaro law. This raised some interesting questions: Are interactive models intrinsically inconsistent with the Morton–Massaro law? More generally, what, if anything, does the Morton–Massaro law tell us about the organization of perceptual processes? Does it make sense from the point of view of the rationality or optimality of the perceptual process? Is it a consequence of the statistical structure of the environment? Should we expect this law to be violated under some conditions? The goal of this paper is to address these and other related questions. Some of these questions have been previously considered by other authors and are included here to provide a more complete picture of the Morton–Massaro law. In other cases, the answers arise from new theoretical and empirical results that will presented for the first time in this article.

The paper proceeds as follows. We consider the phenomenon of factorability in detail, relating it to classical models including Morton's (1969) logogen model, Oden and Massaro's (1978) Fuzzy Logical Model of Perception (FLMP), random utility models, and to a statistical technique called logistic regression. Then we identify the conditions under which this phenomenon may represent an optimal strategy of perceptual inference. The analysis suggests that in situations where the Morton–Massaro law holds, the environment should have a particular statistical structure. It also suggests that when that statistical structure does not hold, the Morton–Massaro law may not hold either. An experiment and a statistical analysis of audio-visual speech signals support these suggestions. Next we examine whether

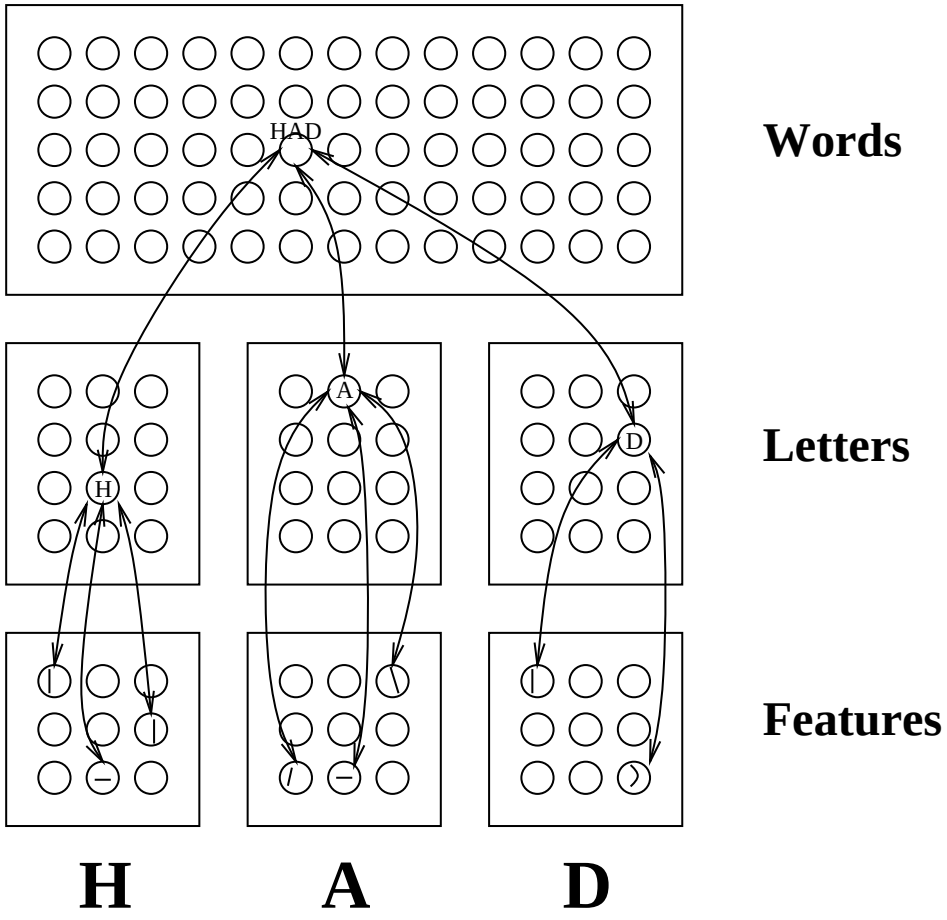


Figure 4: The interactive activation architecture (McClelland & Rumelhart, 1981). The version shown is applicable to three-letter words such as those used in the experiment described later in the text. There is a unit for every three letter word in English, a unit for each letter in each position, and a set of feature units for each position. Not all units are shown. Bi-directional, excitatory connections for the word HAD are illustrated. Between level connections are excitatory and symmetric. Within each pool of units there are symmetric, mutually inhibitory connections between incompatible units (not shown). In some versions of the interactive activation architecture, between level inhibitory connections have been used but these are not included in this version.

interactive models may exhibit factorable effects, introduce the concept of *channel separability*, and consider the relations of this idea to other notions of channel separability and independence.

In the next section we describe typical experiments in which factorability of information sources has been observed. We formalize the notion of factorability and describe its relationship to the classical models mentioned above.

1 Notation and Experimental Paradigm

In the experiments in question, subjects are presented with inputs which are combinations of several treatment variables and are asked to report on some property of these inputs by choosing among a set of mutually exclusive response alternatives. In some cases subjects are asked about a perceptual property of one of the treatment variables (e.g., identify the first letter of a three letter string). That variable is usually known as *the stimulus* and the other variables are known as *the context*. In other cases, subjects are asked about the entire input combination. For example, subjects may be asked about their combined audio-visual experience when a visual presentation of a person saying a word is synchronized with an accompanying acoustic signal. In such cases the classification of input components as stimulus or context is arbitrary and only indicates that they are different sources of information.

Let $X_1, \cdots, X_m$ represent the different treatment factors (sources of information) and R represent the response obtained for combinations of these treatments. Let the $\mathcal{R} = \{1, \cdots, r\}$ represent the set of response alternatives, and $\mathcal{X}_i$ the set of treatment levels of factor X_i. For any set $\mathcal{S}$, the term $|\mathcal{S}|$ will represent the number of elements in that set, e.g., $|\mathcal{R}| = r$. Subjects are repeatedly presented with all possible combinations of the different treatment factors and the response distribution of each *individual subject* given each combination of treatments is recorded. If there are only two experimental factors, the response distributions can be organized as a matrix in which the rows represent the treatments of factor X_1 and the columns the treatments of factor X_2. Each cell in this matrix would contain the probability distribution of response alternatives given a particular combination of treatments.

For example, consider the following hypothetical letter recognition experiment (which we have actually carried out and will report in a later section). On each trial, a three letter stimulus is presented, followed by a mask, and the subject must decide which of two letters was presented in a particular position (for example, the second). In this case there are three sources of information: The "stimulus" X_2 is the letter presented at the second position, and the other two letters X_1 and X_3 act as the context . The stimulus letters might be "E" and "U", in which case $\mathcal{X}_2 = \{\text{"E"}, \text{"U"}\}$. The contexts might be M_N, R_N, M_D, and R_D, in which case $\mathcal{X}_1 = \{\text{"M"}, \text{"R"}\}$ and $\mathcal{X}_3 = \{\text{"D"}, \text{"N"}\}$. In this case there would be 8 input combinations: MEN, MUN, REN, RUN, MED, MUD, RED, RUD. In our notation the responses

are encoded as numbers, so $\mathcal{R} = \{1, 2\}$, with 1 representing the subject choosing the alternative "E", and 2 representing the alternative "U". For each combination of stimulus and context we would obtain the probability of response 1 and response 2. Thus there would be a total of $2\times2\times2\times(2-1) = 8$ independent probability estimates per subject.

In the example just given, there was a one-to-one correspondence between stimuli and response alternatives but this does not need to be the case. Consider for example Experiment 2 of Massaro and Cohen (1983a) in which 7 subjects had to identify synthetic consonant sounds presented in the context of other phonemes. There were two response alternatives, $|\mathcal{R}| = 2$, seven stimulus conditions, $|\mathcal{X}_1| = 7$, and four context conditions, $|\mathcal{X}_2| = 4$. The response alternatives were /l/ and /r/, the stimuli were synthetic sounds generated by varying the onset frequency of the third formant, followed by the vowel /i/. Each of the 7 stimuli was placed after each of four different context consonants, /v/, /s/, /p/, and /t/. Each stimulus by context combination was tested 40 times per subject thus providing $7\times4\times(2-1) = 28$ independent probability estimates per subject.

Similar experiments can also be done with stimulus and context operating in different modalities. For example, Repp, Healy, and Crowder (1983) presented subjects with all possible combinations of auditory signals of the articulations */ba/*, */va/*, /Da/ and */da/* and synchronized visual depictions of articulations of the same alternatives, for a total of 16 different audio-visual inputs. Subjects were presented with these inputs and asked to report whether they heard */ba/*, */va/*, */Da/* or */da/*. The experiment provided $4 \times 4 \times 3 = 48$ independent probability estimates per subject (see Massaro (1987c) for a description of the methods used in Repp et al.'s (1983) study).

It should be noted that all these experiment are designed to obtain mappings between external input conditions and response probabilities. The experiments do not inform us about the extent to which these mappings reflect perceptual or post-perceptual processes. For example, it is theoretically possible that in all these experiments perception itself is not affected by context but that the post-perceptual[1] response choices made by subjects are. The experiments, and the models presented in this paper are silent about this issue. What is at hand in these experiments is the nature of the proper characterization of the mappings between information sources and responses,

[1]By post-perceptual we mean processes that map internal perceptions to external responses.

not whether these mappings were perceptual or post-perceptual.

2 Factorability

In this section we describe more precisely the concept of factorability and its relationship to models of perception. Factorability refers to a particular way in which multiple sources of information affect response probabilities. Guided by the logogen model Morton (1969) proposed that the influence of information sources on perception of words can be accounted for with a factorized version of Luce's (1959) strength model. In Luce's model a non-negative strength parameter, denoted with the Greek symbol η (eta), is assigned to each response alternative (indexed here by k) in every combination of information sources. The probability of the response is given by

$$P(R = k \mid X_1 = x_1, \cdots, X_m = x_m) \quad = \quad \frac{\eta(x_1, \cdots, x_m, k)}{\sum_{l=1}^{r} \eta(x_1, \cdots, x_m, l)}, \qquad (1)$$

where $x_1, \cdots, x_m$ represent specific values of the information sources and k, l specific response alternatives. Morton's claim was that these strength parameters can be factorized into non-negative terms here called η_{X_i} which are selectively influenced by one factor so that

$$\eta(x_1, \cdots, x_m, , k) = \eta_{X_1}(x_1, k) \cdots \eta_{X_m}(x_m, k)\,, \qquad (2)$$

and

$$P(R = k \mid X_1 = x_1, \cdots, X_m = x_m) \quad = \quad \frac{\eta_{X_1}(x_1, k) \cdots \eta_{X_m}(x_m, k)}{\sum_{l=1}^{r} \eta_{X_1}(x_1, l) \cdots \eta_{X_m}(x_m, l)}\,. \qquad (3)$$

Here $\eta_{X_i}(x_i, k)$ represents the support provided to the response k when the source X_i takes the value x_i. It is important to note that (3) describes ideal probability values one would obtain with an infinite number of trials, not empirical probability estimates based on a finite number of observations.

Hereafter we will assume that no strength parameter is exactly zero, i.e., we assume that given an arbitrarily large number of trials each response alternative will appear at least once for each input combination. Under this assumption (3) is equivalent to

$$\frac{P(R = k \mid X_1 = x_1, \cdots, X_m = x_m)}{P(R = l \mid X_1 = x_1, \cdots, X_m = x_m)} = \left(\frac{\eta_{X_1}(x_1, k)}{\eta_{X_1}(x_1, l)}\right) \cdots \left(\frac{\eta_{X_m}(x_m, k)}{\eta_{X_m}(x_m, l)}\right)\,. \qquad (4)$$

This says that response probability ratios can be factorized into components selectively influenced by one and only one information source. The term $\eta_{X_i}(x_i, k)/\eta_{X_i}(x_i, l)$ is the *relative support* for responses k and l provided by the source X_i when it takes the specific value x_i. Response biases may be included but they do not add generality since they can be considered as part of the η_{X_i}terms. Hereafter we refer to (3) as a *factorized strength model*[2].

A formulation similar to Morton's logogen model arises within the Fuzzy Logical Model of Perception (Oden & Massaro, 1978; Massaro, 1987c). The FLMP is a factorized strength model in which the strength parameters are constrained to be in the interval $(0, 1)$ and to add up to one, i.e., $\Sigma_k \eta_{X_i}(x_i, k) = 1$. These provisos occasion no loss of generality. To see why let $\tilde{\eta}_{X_i}$ represent parameters of a factorized strength model and let η_{X_i}

$$\eta_{X_i}(x_i, k) = \frac{\tilde{\eta}_{X_i}(x_i, k)}{\sum_{l=1}^{r} \tilde{\eta}_{X_i}(x_i, l)}. \tag{5}$$

It is easy to see that the model with parameters $\eta_{X_1}, \cdots, \eta_{X_m}$ results in the same response probabilities as the model with parameters $\tilde{\eta}_{X_1}, \cdots, \tilde{\eta}_{X_2}$. In addition the η_{X_i} parameters follow the constraints prescribed by the FLMP, i.e., they add up to 1 are are in the (0,1) range. This shows that *response probability matrices conform to a factorized strength model if and only if they conform to the FLMP.*

It should be noted that not all factorized models foster the view of human information processing advocated by the FLMP. In particular Massaro (1987c) interprets the FLMP parameters in terms of fuzzy logic (Zadeh, 1988), and views information integration as a deterministic feed-forward process, involving separate stages of independent evaluation and integration. Evaluation consists of assigning values to the η_{X_i} terms, and integration consists of computing the product of these terms for each response alternative. The only probabilistic part of the process is the decision, which consists of randomly choosing a response based on (3). In addition there are extensions of the FLMP that make predictions about the changes in choice probabilities as a function of the ISI between the input and a backward mask (p. 256 Massaro, 1989a). Such extensions are not the focus of this paper.

[2]Factorized strength models appear in domains other than perception. For example, the sampling equations of the SAM model of information retrieval (ShiffrinRaaijmakers92GillundShiffrin84) conform to our definition of factorized strength models.

2.1 Strength Models and Random Utility Models

Strength models, like the ones we have seen so far, assume strength values are assigned deterministically to each response alternative. Response variability is achieved by choosing randomly with probability proportional to the strength of each response. On the other hand, random utility models assume that choices are made by picking deterministically response alternatives which maximize some internal utility value. In these models response variability is explained by adding a random component to the utility of each alternative. On each trial random utility values $Y_1, \cdots, Y_r$ are assigned to each response alternative. These random utilities are the sum of an utility function μ_k, which is determined by the external input, and a random noise component N_k which is independent of the stimulus and context

$$Y_k = \mu_k(X_1, \cdots, X_m) + \beta\, N_k, \tag{6}$$

where $\beta > 0$ is a parameter scaling the effect of noise. On each trial the alternative with maximum random utility value is chosen. To simplify the presentation in this section we consider experiments which manipulate only two treatment factors: X_1 and X_2. The generalization to more than two factors is transparent.

In general it is difficult to find whether random utility models factorize. However, special cases exist for which analytical results are possible. In particular, McFadden (1978) showed that if the joint cumulative distribution function of the noise is as follows

$$F_N(u_1, \cdots, u_r) = e^{-G(e^{-u_1}, \cdots, e^{-u_r})}, \quad \text{for } (u_1, \cdots, u_r) \in \mathbb{R}^r, \tag{7}$$

where G is a (generator) function satisfying some regularity conditions[3], then

$$P(R = k \mid X_1 = x_1, X_2 = x_2) = \frac{\partial \log G(e^{\mu_1(x_1,x_2)/\beta}, \cdots, e^{\mu_r(x_1,x_2)/\beta})}{\partial \mu_k(x_1, x_2)/\beta}. \tag{8}$$

From this it is possible to show that for some noise generator function G and some utility function μ random utility models factorize. In particular if we

[3]The regularity conditions are that: (1) $G(u_1, \cdots, u_r)$ be nonnegative homogeneous of degree one for $u_1 \geq 0, \cdots, u_r \geq 0$; (2) $\lim_{u_i \to \infty} G(u_1, \cdots, u_r) = \infty$ for $i = 1, \cdots, r$; and (3) for any $k \in \mathbb{N}$ and $(i_1, \cdots, i_k) \in \{1, \cdots, r\}^k$ the partial derivatives $\partial^k G(u_1, \cdots, u_r)/\partial u_{i_1} \cdots \partial u_{i_k}$ are nonnegative if k is odd and non-positive if k is even.

let

$$G(u_1, \cdots, u_r) = \left(\sum_{i=1}^{r} u_i^{1/(1-\rho)} \right)^{1-\rho}, \quad \text{for } (u_1, \cdots, u_r) \in \mathbb{R}^r, \tag{9}$$

then the noise is a multivariate version of the standard Gumbel noise (Oliveira, 1961) . Here $0 \leq \rho < 1$ is a parameter controlling the overall correlation between the noise components $N_1, \cdots, N_r$, i.e., $\rho = 0$ when the noise components are uncorrelated and $\rho = 1$ when they are perfectly correlated. Applying McFadden's equation (8) to (9) and after some simple derivations it follows that

$$P(R = k \mid X_1 = i, X_2 = j) = \frac{e^{\mu_k(x_1,x_2)/(\beta(1-\rho))}}{\sum_{l=1}^{r} e^{\mu_l(x_1,x_2)/(\beta(1-\rho))}}. \tag{10}$$

Thus if the utility is additively controlled by the different information sources

$$\mu_k(x_1, x_2) = \mu_k^1(x_1) + \mu_k^2(x_2), \tag{11}$$

then response probabilities factorize as required by the Morton–Massaro law. Moreover the deterministic utility values and strength parameters are related as follows

$$\eta_{X_1}(x_1, k) = e^{\mu_k^1(x_1)/(\beta(1-\rho))}, \tag{12}$$

$$\eta_{X_2}(x_2, k) = e^{\mu_k^2(x_2)/(\beta(1-\rho))}. \tag{13}$$

The noise scale parameter β affects the importance of the utility functions on the response probabilities. As β goes to zero, the response with maximum deterministic utility is chosen with probability one. As β goes to infinity, all responses are chosen with equal probability. In addition, as the correlation parameter ρ goes to one, the response with maximum deterministic utility is chosen with probability one. This makes sense since if the noise components are perfectly correlated they just add a constant to the utility values of all the alternatives and thus have no effect on the choice process.

Note that in random utility models factorability of information sources corresponds to the decomposition of the overall utility into additive terms selectively controlled by different sources of information. Morton himself presented the logogen model as a close relative of a random utility model. Each logogen unit produced an activation value, i.e., a utility function, which was

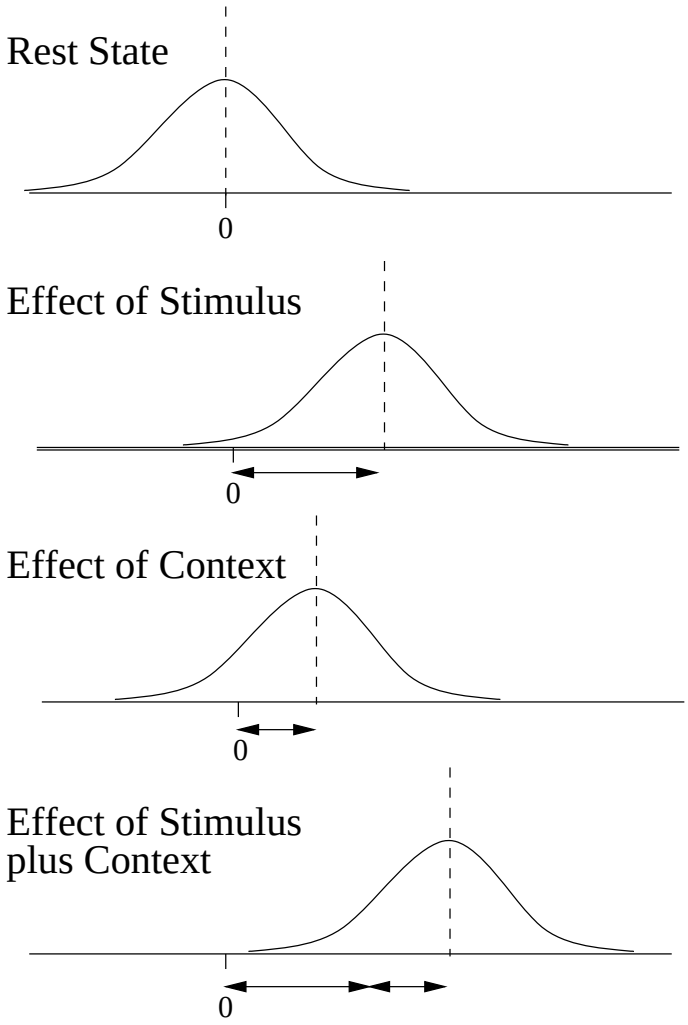


Figure 5: In Morton's logogen model stimulus and context have additive influence on the state of logogen units. Redrawn from Figure 2 in "Interaction of information in word recognition", by J. Morton, 1969, *Psychological Review, 76*, p. 167. Copyright 1969 by the American Psychological Association. Permission Pending.

the sum of terms selectively controlled by different sources of information (in his case, stimulus and context) plus independent random noise. He assumed that each logogen had a threshold, and that if any logogen reached threshold the corresponding word became available as a response (if more than one became available, the first one determined the response). This response selection policy approximates the policy used in random utility models, in which the item with the largest activation is selected.[4] Unfortunately Morton's (1969) paper did not clarify the relationship between his formulation of the model and the factorized strength model used to describe the data. The relationship between strength models and random utility models appeared in work by Holman and Marley, cited in Luce & Suppes (1965) and was extensively studied in Yellot (1977) and McFadden (1978).

3 Complexity Analysis

The Morton–Massaro law is the statement that factorized strength models provide good approximations to the choice probabilities obtained in a wide variety of perceptual experiments. One potential explanation for why factorized models fit empirical data so well is that they impose no constraints on response probability matrices, i.e., they can fit any arbitrary probability matrix. For example, it is well known that given a large number of hidden units, multilayer perceptrons can fit arbitrary input-output functions (Hornik, Stinchombe, & White, 1989). Thus it would not be of interest to observe that response probability matrices are well approximated by such models. In page 84 of the Appendix we show that this is not a good explanation.

The Appendix shows that in experiments with factors $X_1, \cdots, X_m$ and response alternatives $\{1, \cdots, r\}$ factorized strength models are fully defined using just $(r-1)(|\mathcal{X}_1| + \cdots + |\mathcal{X}_m| - m + 1)$ parameters, where $|\mathcal{X}_i|$ is the number of treatments for factor X_i. Note that arbitrary response probability matrices have $(r-1)(|\mathcal{X}_1|) \cdots (|\mathcal{X}_m|)$ degrees of freedom. Thus, the complexity[5], arbitrary response probability matrices grows exponentially with the number of information factors, while the complexity of factorized models

[4]It should be noted that Morton assumed that the threshold of a logogen was reduced by prior use, and thus was lower for frequent and recent words. In the corresponding random utility model this corresponds to assuming that the activation is increased by prior use, so that the item tends to be relatively more active for frequent and recent words.

[5]We informally use the term complexity to represent the number of non-redundant parameters in a model. This notion is adequate for our work since we operate within the

Table 1: A non-factorable probability matrix. There are two response alternatives, two stimuli and two context conditions. The number within each cell represents the probability of the first response alternative given the specific combinations of information sources indicated by the row and column of that cell. In this case the best mean square approximation achievable with factorized models gives probability 0.5 to each cell, resulting in a root mean square error of 0.4.

	$X_2 = 1$	$X_2 = 2$
$X_1 = 1$	0.1	0.9
$X_1 = 2$	0.9	0.1

grows linearly. The smallest experimental design that allows falsifying factorized models is a 2×2 factorial design with 2 response alternatives, i.e., $m = 2$, $r = 2$, and $|\mathcal{X}_1| = |\mathcal{X}_2| = 2$. In this simple case factorized strength models use $(2-1)(2+2-2+1) = 3$ degrees of freedom whereas arbitrary data matrices can have $(2-1)(2)(2) = 4$ degrees of freedom. Table 1 shows an example of a 2×2 probability matrix for which the best approximation achievable with factorized models is grossly inaccurate.

3.1 Two Response Alternatives

If there are only two response alternatives (i.e., $r = 2$), then the data in the response matrix are amenable to an analytic technique known as factorial logistic analysis (Myers, 1990). Logistic analysis is based on a simple and quite commonly-used statistical model in which treatment variables $X_1, \cdots, X_m$ are used to model a dependent variable bounded in the $(0, 1)$ range (e.g., a probability value). To simplify the presentation we will default to the case with 2 information sources, which we call stimulus (X_1) and context (X_2). The generalization to more than two sources is transparent.

The name "logistic" comes from the fact that the dependent variable is assumed to be related to the treatment variables by the so called, logistic

framework of Luce strength models. For other notions of model complexity see Vapnik (1995).

function

$$P(R = 1 \mid X_1 = x_1, X_2 = x_2) = \frac{1}{1 + \exp\left(-(\mu + \delta_{X_1,X_2}(x_1, x_2))\right)}, \tag{14}$$

where $1/(1 + \exp(-(\cdot)))$ is the logistic function, μ is a parameter constant for all the treatments, and $\delta_{X_1,X_2}(x_1, x_2)$ is a parameter known as the *logistic joint effect* of stimulus x_1 and context x_2. These parameters are commonly estimated using standard maximum-likelihood approaches. As in ANOVA the logistic joint effects are constrained to add up to zero and are decomposed into *logistic main effects*, $\delta_{X_1}(x_1)$, $\delta_{X_2}(x_2)$, and *logistic interaction effects*, $\delta_{X_1 \times X_2}(x_1, x_2)$

$$\delta_{X_1,X_2}(x_1, x_2) = \delta_{X_1}(x_1) + \delta_{X_2}(x_2) + \delta_{X_1 \times X_2}(x_1, x_2)\ , \tag{15}$$

where

$$\delta_{X_1}(x_1) = \frac{1}{|\mathcal{X}_2|} \sum_{x_2 \in \mathcal{X}_2} \delta_{X_1,X_2}(x_1, x_2)\ , \tag{16}$$

$$\delta_{X_2}(x_2) = \frac{1}{|\mathcal{X}_1|} \sum_{x_1 \in \mathcal{X}_1} \delta_{X_1,X_2}(x_1, x_2)\ , \tag{17}$$

$$\delta_{X_1 \times X_2}(x_1, x_2) = \delta_{X_1,X_2}(x_1, x_2) - \delta_{X_1}(x_1) - \delta_{X_2}(x_2)\ . \tag{18}$$

The number of free parameters divides up as follows: 1 parameter for the constant μ, $|\mathcal{X}_1|-1$ parameters for the main effects of X_1, $|\mathcal{X}_2|-1$ parameters for the main effects of X_2, and $(|\mathcal{X}_1| - 1) \times (|\mathcal{X}_2| - 1)$ parameters for the interaction effects. As in ANOVA, this partition follows from the fact that the joint effects are constrained to add up to 0.

As pointed out by MassaroFriedman90) and by crowther95) in experiments with only two response alternatives, factorized strength models are equivalent to logistic models with no interaction effects. To see why let

$$p(x_1, x_2) = P(R = 1 \mid X_1 = x_1, X_2 = x_2), \tag{19}$$

then if factorability holds,

$$p(x_1, x_2) = \frac{\eta_{X_1}(x_1, 1)\eta_{X_2}(x_2, 1)}{\eta_{X_1}(x_1, 1)\eta_{X_2}(x_2, 1) + \eta_{X_1}(x_1, 2)\eta_{X_2}(x_2, 2)}\ , \tag{20}$$

assuming none of the η_{X_1} and η_{X_2} terms are zero, this can be written as

$$p(x_1, x_2) = \frac{1}{1 + \exp\left(-(\mu + \delta_{X_1}(x_1) + \delta_{X_2}(x_2))\right)}, \tag{21}$$

where

$$\mu + \delta_{X_1}(x_1) = \log \frac{\eta_{X_1}(x_1, 1)}{\eta_{X_1}(x_1, 2)}, \tag{22}$$

$$\delta_{X_2}(x_2) = \log \frac{\eta_{X_2}(x_2, 1)}{\eta_{X_2}(x_2, 2)}, \tag{23}$$

showing that the stimulus by context interaction effects are zero, thus saving $(|\mathcal{X}_1| - 1) \times (|\mathcal{X}_2| - 1)$ degrees of freedom.

3.2 Modularity and Statistical Interaction

It should be noted that the computational notion of interactivity, as an antithesis to modularity, has very little to do with the statistical notion of interaction in analysis of variance. In particular factorized strength models can be modular (like the FLMP) yet they can produce some interaction effects. Consider for example the case in which there are only two experimental factors and two response alternatives. If the untransformed probability values predicted by factorized models are plotted against the treatments of one factor for each of the treatments of the other factor, this results in a family of curves spanning a region somewhat shaped like an American football, or a portion of one. This shape is a characteristic signature of factorized strength models and the Morton–Massaro law (see Massaro (1989a), page 107). Such patterns can yield significant statistical interaction effects when their raw response probabilities are subjected to analysis of variance.

Although factorized strength models can produce some interaction effects, *when there are only two response alternatives*, they cannot produce cross over interactions. To see why, fix the context to x_2 and examine the difference in logit probabilities for two different stimuli x_1 and $\tilde{x}_1$. Using (21) we get that

$$\operatorname{logit} p(x_1, x_2) - \operatorname{logit} p(\tilde{x}_1, x_2) = \delta_{X_1}(x_1) - \delta_{X_1}(\tilde{x}_1), \tag{24}$$

where the logit function $\operatorname{logit}(p) = \log(p/(1-p))$, is just the inverse of the logistic, so that if $p = \operatorname{logistic}(x)$, then $\operatorname{logit}(p) = x$. Note how the difference of logit probabilities is not dependent on the context x_2. The transformed

data (i.e., $\text{logit}(p(x_1, x_2))$ can be plotted against the stimulus indices x_1, for each of the different values of x_2); if the Morton–Massaro law holds, the curves should all be parallel if it were not for deviations due to sampling error. Moreover, since the logit function is an strictly increasing transformation it follows that

$$\text{sign}\Big(p(x_1, x_2) - p(\tilde{x}_1, x_2)\Big) = \text{sign}\Big(\delta_{X_1}(x_1) - \delta_{X_1}(\tilde{x}_1)\Big), \tag{25}$$

which is also independent of the context (x_2). Thus, if $\delta_{X_1}(x_1) - \delta_{X_i}(\tilde{x}_1)$ is positive, then switching from stimulus x_1 to stimulus $\tilde{x}_1$ can only increase the probability of response 1, regardless of the context (x_2), and similarly the probability can only decrease if $\delta_{X_1}(x_1) - \delta_{X_i}(\tilde{x}_1)$ is negative. In other words: *when there are only two response alternatives, factorized models cannot produce cross over interactions on the choice probabilities.* This property is useful for quick verification of gross violations of the Morton–Massaro law. If there are only two response alternatives then significant cross-over interactions indicate violations of the Morton–Massaro law. It should be pointed out that when there are more than two response alternatives cross-over interactions do not necessarily violate the law.

4 Functional Analysis

As mentioned earlier in this document we do not use the term "Morton–Massaro law" to imply that it applies in all perceptual problems at all levels of analysis. Indeed if the law applied at all levels of the perceptual process, the problem of machine perception might have been solved long ago. One of the aspects that makes machine perception so difficult is the fact that information must sometimes be integrated in a non-factorable manner. It is thus important to ask under what conditions the Morton–Massaro law reflects a sensible way to combine sources of information. Functional questions like this are an essential component of scientific enquiry in disciplines such as biology. Unfortunately, cognitive psychology has tended to focus on structural questions (e.g., the modularity debate) while paying little attention to functional questions—See Marr (1982), Anderson (1990) and Knill, Kersten, and Yuille (1996) for notable exceptions.

In this section, we adopt a Bayesian framework to formulate a notion of optimal perceptual inference, and examine under what conditions and in what sense the Morton–Massaro law conforms to this definition of optimality. The goal is to gain an understanding of the function of the Morton-Massaro law and its relationship to the statistics of the environment in which it develops. For other perspectives on the optimality of the Morton–Massaro law see MassaroFriedman90), Movellan and Chadderdon (1996), FriedmanMassaro98) and page 272 of Massaro (1989a).

4.1 An Overview of Bayesian Decision Theory

From a Bayesian perspective, human perception is viewed as a form of probabilistic inference (Knill et al., 1996; Clark & Yuille, 1990b) which is describable in terms of subjective probabilities. Bayesian approaches do not assume that such probabilities are explicitly represented by the perceptual system (e.g., as they would be in a classic Bayesian expert system). Subjective probabilities can be simply seen as descriptive variables used by Bayesian analysts to help understand the logic of the outcome of perceptual processes. Within this framework one can typically analyze in what sense and under what conditions these subjective probabilities and the inferences they afford might be seen as optimal.

We consider this question in the problem of choosing from among a set of alternatives (perceptual hypotheses or categories) when given multiple

sources of information about those alternatives. The approach will help us understand the ubiquity of the Morton–Massaro law. As in previous sections we use the random variables $X_1, \cdots, X_m$ to represent sources of information, and the random variable R to represent the subject's responses. The variables $X_1, \cdots, X_m$ provide information about a variable A which represents the alternatives (e.g., letters, words, phonemes, etc.) the subject is required to consider, in the sense that he must designate one of these as his response.

It is assumed that the perceptual system has no direct access to the value of A and its role is to infer it from the available information sources. At this level of analysis perceptual systems are just probabilistic mappings whose inputs are the information sources, $X_1, \cdots, X_m$, and whose outputs are probability distributions of responses (i.e., two perceptual systems with identical mappings are considered equivalent regardless of internal mechanisms). We denote these mappings as $p_{R|X_1 \cdots X_m}$, referring to the distribution function of responses conditional on all possible values of the information sources. The notation $p_{R|X_1 \cdots X_m}(k|x_1, \cdots, x_m)$ corresponds to the probability that R takes the value k given that X_1 takes the value x_1 ... and X_m takes the value x_m. In this context we investigate in what sense and under what conditions perceptual systems—that is, distribution functions $p_{R|X_1 \cdots X_m}$ that adhere to the Morton–Massaro law—are optimal.

At this point it is important to make a distinction between true environmental probabilities and subjective estimates of these probabilities, which are internal to the mechanism that is making perceptual decisions. The idea is that human observers do not have direct access to the probability values given by the environment. Instead the human perceptual system must work (explicitly or implicitly) with quantities that are best seen as internal or subjective estimates of the relevant probabilities. To make this distinction clear we mark such internal subjective probability estimates with a tilde.

A common criterion for optimality is minimization of the subjective error rate, i.e., maximization of the subjective probability that the response equals the correct alternative. It should be apparent that policies that satisfy the following rule are optimal with respect to that criterion (Duda & Hart, 1973)

$$\begin{aligned} &p_{R|X_1 \cdots X_m}(r \mid x_1, \cdots, x_m) = 0 \text{ if there is } r' \in \mathcal{R} \text{ such that} \\ &\tilde{p}_{A|X_1 \cdots X_m}(r' \mid x_1, \cdots, x_m) > \tilde{p}_{A|X_1 \cdots X_m}(r \mid x_1, \cdots, x_m) \ . \end{aligned} \tag{26}$$

In words, this rule says that we should never choose a response r if some other response r' is estimated to be more likely to correspond to the correct

alternative. So, if one response is more likely to correspond to the correct alternative than any other, we should always choose that response. If there are two or more responses tied for being the most likely alternative, then any distribution of responses among these alternatives are equally optimal. In the pattern recognition literature this policy is known as "maximum a posteriori" or MAP because it maximizes the subjective estimate of the posterior probability that the response is correct, given the data available to the perceiver.

Maximizing the posterior probability is equivalent to maximizing a statistic known as the expected discriminant value (Duda & Hart, 1973), which can be expressed as follows

$$E^P(\log \tilde{p}_{X_1 \cdots X_m A}(X_1, \cdots, X_m, R)) = \sum_{x_1} \cdots \sum_{x_m} \sum_{r} p_{X_1 \cdots X_m R}(x_1, \cdots, x_m, r) \log \tilde{p}_{X_1 \cdots X_m A}(x_1, \cdots, x_m, r). \tag{27}$$

In words, the expected discriminant value is the average, over all combinations of inputs and responses of the log of the subjective estimate of the probability that the given input has occurred and the response r actually corresponds to the alternative that generated the input. Large values of the expected discriminant value indicate that the policy is effectively exploiting the subjective information about contained in $X_1, \cdots, X_m$. MAP policies achieve the largest possible expected discriminant value.

Having introduced the basic ideas of Bayesian decision theory, let us return now to the optimality of the Morton–Massaro law. The first thing that we can note is that the MAP policy is in general a deterministic function of the information sources. If there is a single alternative that is subjectively most likely to be correct for a given combination of $X_1, \cdots, X_m$ then the MAP policy would always choose that response when that particular combination of information sources is presented. However in experiments in which subjects believe that one alternative is more probable than the others they consistently still choose the less probable alternatives a substantial number of times (FriedmanMassaro98).

One way to explain this deviation from MAP is to construe subjects as maximizing a combination of entropy and subjective discriminant value

$$E^P(\log \tilde{p}_{X_1 \cdots X_m A}(X_1, \cdots, X_m, R)) + \beta \, H^P(X_1, \cdots, X_m, R) \, , \tag{28}$$

where

$$H^P(X_1, \cdots, X_m, R) = \tag{29}$$
$$-\sum_{x_1} \cdots \sum_{x_m} \sum_{r} p_{X_1 \cdots X_m R}(x_1, \cdots, x_m, r) \log p_{X_1 \cdots X_2 R}(x_1, \cdots, x_m, r),$$

is the joint entropy of the information sources and the response. The joint entropy is an information theoretic measure of the dispersion of the joint distribution of information sources and responses (Cover & Thomas, 1991). The parameter $\beta > 0$ is fixed and controls the relative importance of the entropy versus the subjective discriminant value. We can now introduce a theorem and a corollary for the Morton–Massaro law that suggest in what sense and under what conditions the law is optimal. The proof for the theorem and its corollary are on page 85 of the Appendix. Here we focus on their heuristic interpretation.

OPTIMALITY THEOREM: *A Bayesian system that satisfies the two following conditions conforms to the Morton-Massaro law:*

1. *It maximizes a weighted sum of subjective discriminant value and entropy*

$$E^P(\log \tilde{p}_{X_1 \cdots X_m A}(X_1, \cdots, X_m, R)) + \beta \, H^P(X_1, \cdots, X_m, R), \tag{30}$$

   where E^P, H^P respectively symbolize expected values and entropy with respect to the probability measure P, and $\beta > 0$ is a fixed parameter.

2. *It assumes a distribution of information sources which factorizes as follows:*

$$\tilde{p}_{X_1 \cdots X_m | A}(x_1, \cdots, x_m \mid a) = \phi_A(a)\phi_X(x_1, \cdots, x_m)\phi_1(x_1, a) \cdots \phi_m(x_m, a), \tag{31}$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$. Here ϕ_R is a term selectively influenced by A, not X, ϕ_X is selectively influenced by X, not A, and the terms ϕ_i are selectively influenced by the i^{th} element of X and by A.

COROLLARY: *If $p_{R|X_1 \cdots X_m}(a \,|\, x_1, \cdots, x_m)$ satisfies the Morton–Massaro law for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$, then there is a Bayesian system that satisfies Condition 1, that assumes class conditionally independent sources (a special version of Condition 2) and whose response probabilities equal $p_{R|X_1 \cdots X_m}(a \mid x_1, \cdots, x_m)$ for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$.*

Examining Condition 1: The first condition required by the optimality theorem is that the perceptual system should maximize not just the expected value of the subjective probability of being correct, as a pure Bayesian system would do, but a combination of this and response entropy. A deterministic policy would yield no dispersion (all the probability would be associated with a single response), and distributing the probability of choice evenly across the available alternatives would yield the greatest degree of dispersion. Maximization of entropy, not just discriminant value, is beneficial in situations where the subjective internal probability estimates are not certain or the environment is non-stationary. In such cases some entropy in responding helps sampling the environment for the purpose of improving the subjective probability estimates (Dayan & Sejnowski, 1996). This positive value of entropy is well known in the pattern recognition literature where it is presented in terms of a tradeoff between "exploration" and "exploitation" (Sutton & Barto, 1988).

To illustrate this tradeoff consider the following example from a real life application of Bayesian inference (Mineiro, 1998). A Web server has a set of banners that may be displayed for advertisement purposes. The server gets paid every time a client clicks on the banners. To avoid overwhelming the client with advertisements only one banner is displayed per page. The banner chosen to be displayed plays a role analogous to the response R in our analysis. Each time a client visits the Web site the server can obtain partial information about the client (e.g., zip code, age, gender). These sources of information are analogous to $X_1, \cdots, X_m$ in our analysis. Sophisticated Web servers typically have a Bayesian model which provides estimates of the posterior probability that the client clicks on each of the banners given the information known about the client. These estimates are the equivalent of the subjective probabilities $\tilde{p}_{A|X_1 \cdots X_m}$ in our analysis.

The goal of the server is to maximize the number of clicks. If the probability estimates were perfect then the best strategy would be to display the banner which has the maximum probability of being clicked given the information at hand. However, typically the probabilities estimates are not perfect. One reason for this is the fact that people's preferences change (i.e., the environment is non-stationary) and estimates based on past experience may need recalibration. If the system displayed only the most probable banner, it would lose the opportunity to learn that now other banners may be more effective. Thus it is important for the system to display banners that have low subjective probability of success. At the same time, the goal is to

make money, and thus it is also important to give more priority to those banners that have a higher subjective probability of success. This contradiction is at the heart of the exploration–exploitation dilemma.

From a Bayesian point of view, the problem faced by the Web server is not unlike that faced by perceptual systems. Both systems need to balance exploitation of internal knowledge with the need of calibrating that knowledge. Equation (30) provides a way of formalizing a common heuristic used in the pattern recognition literature to balance exploration and exploitation (Sutton & Barto, 1988). The parameter β in (30) can be seen as an "uncertainty parameter" which modulates the importance of exploration versus exploitation. In the limit, as $\beta \to 0$ the obtained policy is that of a pure Bayesian observer which bases its decision on the subjective probability measures. As β goes to infinity all responses are chosen with equal probability, regardless of their subjective value, as might be appropriate if the probability estimation process is viewed as completely uncertain. If $\beta = 1$ then the objective probability of choosing an alternative equals the subjective probability of that alternative (a form of probability matching). The optimal value of β depends on the statistical properties of the environment and the validity of the subjective probability estimates.

The entropy term in (28) can also be viewed as an undesirable but unavoidable force that is influencing processing (i.e., noise). Under this interpretation, the subject maximizes the weighted sum by attempting to maximize the first term, which is the only one over which he actually has control. The noise can still be viewed as nature's way of forcing a sampling of alternatives, which may lead to new information allowing updating of subjective probabilities.

In our consideration of Condition 1, we have acted as though correctness of the response choice is equally important for all stimuli and all responses. However, it is possible to take into account differences in the benefit of making particular responses when they are correct and in the costs of making particular responses when they are incorrect. The optimal policy can be derived if estimates of these costs and benefits are available, and can be incorporated into the representation of the subjective discriminant value.

Examining Condition 2: The second condition of the optimality theorem amounts to assuming a particular statistical structure of the world. The assumption only needs to hold for the specific values of the information sources and response alternatives probed in the experiments in which the Morton–Massaro law is tested. An important special case that satisfies Condition 2,

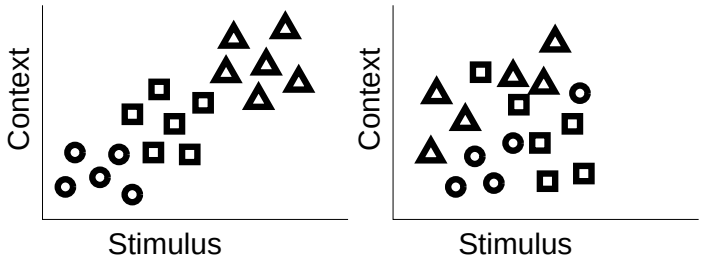


Figure 6: An illustration of the difference between conditional and unconditional independence. Each point represents a stimulus by context combination. There are three different perceptual categories, represented as circles, squares or triangles. On the left stimulus and context are class conditionally independent, since the 2 dimensions are roughly independent within each category. However, overall, stimulus and context are not independent. On the right stimulus and context are not conditionally independent but are roughly independent overall.

occurs when the information sources are assumed to be independent within each perceptual class or category. This condition is called *class-conditional independence* or just *conditional independence* for short. (For a consideration of other cases in which Condition 2 holds, see the Appendix). Figure 6 illustrates the concept of conditional independence (see also Movellan and Chadderdon (1996) and Massaro (1989a)). Each point in the figure represents an instance of a stimulus-context combination, with the symbol indicating which response category each combination comes from. On the left, the stimulus and context are roughly independent within perceptual categories (class conditional independence) but non-independent when collapsed across categories. For example, knowing that an item is an example of the square category tells us that the value of the context will be intermediate, and knowing the value of the stimulus adds no further information. On the right, stimulus and context are roughly independent when collapsing across categories even though they are non-independent within categories. For example, in this case, if we know that an item is an example of the square category, it is hard to specify the value of the context, but if we also know that the stimulus value is high we can predict that the context value will be low.

Conditional independence is sufficient to satisfy Condition 2, but it is

not necessary. Figure 7 shows distributions consistent and inconsistent with Condition 2. The horizontal and vertical axes represent two sources of information. The underlying perceptual category is depicted using different point patterns. Each figure has 100 points sampled from a theoretical probability distribution. In A the distribution is class conditionally independent, i.e., independent within each perceptual category. In B the two perceptual categories are Gaussian with equal correlation structure but different means. In C the samples were taken from Gaussian distributions with equal correlation structure and then passed through a monotonic transformation, the exponential function. In D the perceptual categories are Gaussian with different correlation structure. Case D is incompatible with the Morton–Massaro law. In this case, discrimination of the categories requires a cross-over interaction of response probabilities which, as shown in page 22, is incompatible with the law. In page 88 of the Appendix, we show that cases A, B, and C satisfy Condition 2 and thus they can be optimally discriminated using the Morton-Massaro law.

Condition 2 is well known in the pattern recognition literature, where is often called the "*naive Bayesian assumption*". This assumption has been shown to work surprisingly well in a very wide variety of pattern recognition problems (Domingos & Pazzani, 1997). It should be noted that Condition 2 involves subjective probabilities and thus it is internal to the perceptual system, not a statistical property of the environment. However, since human perception works quite well under many conditions, it is reasonable to infer that, when the Morton–Massaro law holds, the naive Bayesian assumption provides an adequate approximation to the statistical structure of the environment. Indeed one might propose a functional principle that the human perceptual system has evolved to conform to the statistical structure of the environment[6]. This principle suggests that the ubiquity of the Morton–Massaro law may be ultimately due to the fact that, for the type of cases in which the Morton–Massaro law holds, Condition 2 provides a good working approximation to the actual structure of the environment. This principle further suggests that when Condition 2 does not provide a good approximation to the environment we should expect violations of the Morton–Massaro law. Of course, it is an empirical matter to determine just how adequately

[6]For the moment, it is irrelevant whether this conformity arises through direct specification or through an adaptive process that shapes the perceptual system to the characteristics of its environment through experience.

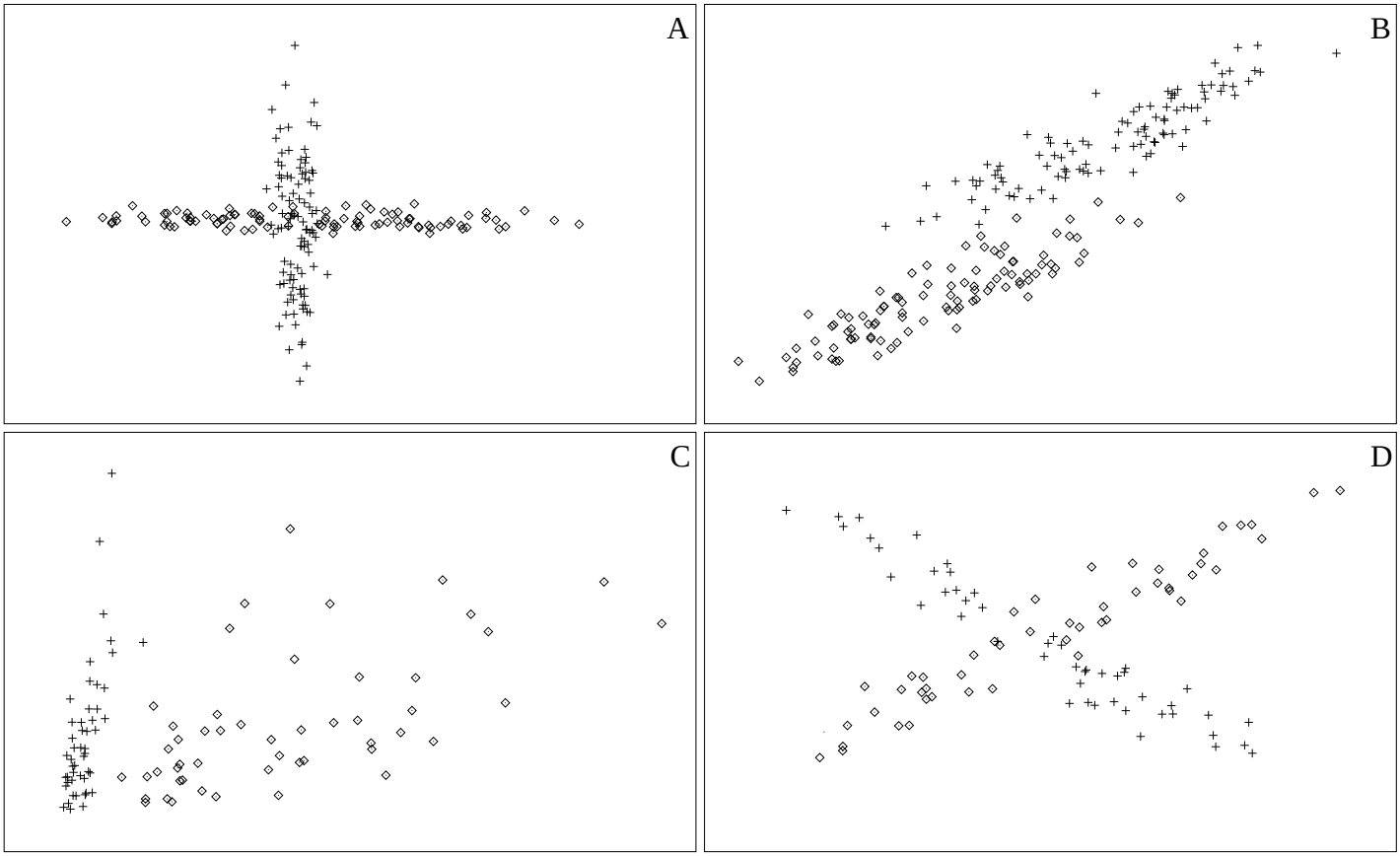


Figure 7: The horizontal and vertical axes represent two sources of information. The type of point (cross vs. diamond) represents the underlying perceptual category. In cases A, B, and C, the Morton-Massaro law provides an optimal way of combining information. In case D the law would not be an adequate way of combining information.

the human perceptual system does reflect the statistics of the environment. In the next two sections we present evidence relevant to this issue. First we present a statistical analysis of audio-visual speech signals and show that conditional independence of audio and video inputs given (within) words is a reasonable assumption. This helps explain why the Morton–Massaro law has been found to hold so well in experiments involving the integration of auditory and visual speech (Massaro, 1989a). Second, we present a case in which Condition 2 grossly misrepresents the environment. In such case the optimality theorem suggests violation of the Morton–Massaro law may occur. An experiment consistent with this prediction is presented.

5 Statistical Analysis of Audio-Visual Speech

Visual information about articulation is an important source of information in face to face speech perception. Sensitivity to correspondences in auditory and visual information for speech events has been shown in 4 month old infants (Kuhl & Meltzoff, 1982). By 6 years of age, humans consistently use audio visual contingencies to understand speech (Massaro, 1987b) and by adulthood visual articulation automatically modulates perception of the acoustic signal (McGurk & MacDonald, 1976). Massaro and colleagues (Massaro, 1989a) investigated how acoustic and visual sources are integrated in speech perception and found that the FLMP does a good job at describing their experiments. According to the Corollary of the Optimality Theorem, this suggests that acoustic and visual speech signals may indeed be treated in a factorized manner without much loss of information.

We investigated the plausibility of this hypothesis by analyzing a database of audio-visual speech signals using statistical machine perception techniques. The main point of the exercise was to illustrate how one may go about testing whether conditional independence provides reasonable approximations to the posterior probabilities of perceptual categories. Our approach involved testing whether assuming conditional independence hinders the performance of automatic audio-visual speech recognition systems which are trained to extract the statistical structure of audio-visual speech. The idea is to test a general class of speech recognition systems that do not assume conditional independence against a restricted version of that class in which we impose the assumption of conditional independence. These two classes are then optimized using a database of audio-visual speech and if the best recognizer

from the constrained class is not worse than the best recognizer from the unconstrained class, we have an indication that conditional independence is a reasonable assumption.

5.1 Database

We collected a database of audio-visual speech from 9 male and 3 female undergraduate students from the Cognitive Science Department at the University of California, San Diego. Each subject was asked to utter the digits "one" through "four" twice. Thus, the total database consists of 96 utterances. The audio sampling rate was 11.1 kHz with 8-bits per sample. The video signal was digitized at 30 frames per second, grey-scale, 100× 75 pixel image, 8 bits per pixel. The video signals show only the subjects' lips. Subjects were asked to center and align their lips in the camera during the sampling. The database, which we named "Tulips1" can be found at *http://cogsci.ucsd.edu* following the links to technical reports and software.

5.2 Visual Front-End

Automatic recognition of visual speech is still an emerging field and a great deal of effort is being done to find appropriate image processing techniques (Prasad, Stork, & Wolff, 1993; Hennecke, Stork, & Ventakesh Prasad, 1996; Luettin, Thacker, & Beet, 1996; Gray, Movellan, & Sejnowski, 1997). The approach we used here is one of the most successful in the field. First each image frame was symmetrized by averaging pixel by pixel the left and right side of each image, using the vertical mid-line as the axis of symmetry. The benefit of this transformation is resistance to changes in illumination along the horizontal axis. For each image frame an associated *delta image* was obtained by subtracting the previous image from the current image. Delta images provide resistance to changes in illumination and emphasize moving regions of the lips.

Each image and its associated *delta* image were low pass filtered and subsampled at 10×15 equidistant points. The resulting pixel values were soft-thresholded and scaled in the (0,1) range using a logistic function matched to the statistics of the images

$$y = f(\frac{\pi}{\sqrt{3}\sigma}(x - \mu)) \tag{32}$$

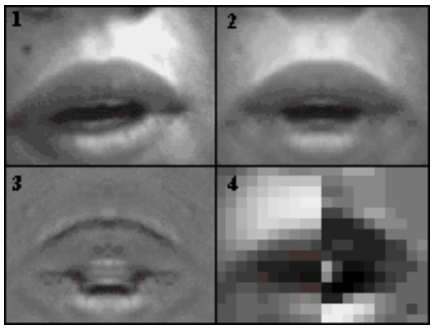


Figure 8: Image processing steps: (1) raw image, (2) symmetrized image, (3) delta image, (4) final composite.

where f is the logistic function, and μ, σ are respectively the average and standard deviation of the distribution of image intensities across an entire word. This transformation is a fast approximation to histogram-equalization, a standard image enhancement technique.

Composites of the relevant portions of the low-pass filtered images and their associated delta images were fed to the statistical recognition engine. The number of variables per processed image was 300 (150 pixel values from the original images and 150 from the delta images). Figure 8 shows the effect of the different preprocessing stages.

5.3 Acoustic Front-End

The acoustic signals were pre-processed using LPC/cepstral analysis. This is a standard technique in the speech recognition literature that parameterizes an estimate of the human vocal tract's transfer function (Rabiner & Juang, 1993). First the high frequency components of the audio track were enhanced using a first order linear filter. Then the audio signal was divided into non-overlapping frames synchronized with the visual frames (i.e., each audio frame lasted 1/30 of a second). Each frame was multiplied by a Hamming window to reduce artifacts due to the blocking procedure (Rabiner & Juang, 1993). LPC analysis was performed on each frame, yielding 8 LPC coefficients per frame. These coefficients, were transformed into a log-power coefficient and 12 cepstral coefficients, i.e., Fourier coefficients of the LPC estimates of the log power spectrum (Rabiner & Juang, 1993). The cepstral coefficients were weighted to minimize the effect of the highest and lowest

order cepstrals. Then, for each frame, a differential log-power coefficient and 12 differential cepstral coefficients were calculated using polynomial approximations with 3 frames before and after the desired frame. This resulted on a total of 26 coefficients per audio frame (12 cepstrals, 12 differential cepstral, 1 log-power and 1 differential log-power). Each of these coefficients was linearly scaled to assume values in the (0,1) range.

5.4 Recognition Engine

The recognition engine was a bank of hidden-Markov models (HMMs). Hidden Markov models, which are the dominant approach in audio and visual speech recognition, are stochastic models whose parameters can be tuned to approximate the statistical structure of time-varying sequences (Rabiner & Juang, 1993). Each of the HMMs was trained using the Expectation-Maximization (EM) algorithm (Rabiner & Juang, 1993) to estimate the log-likelihood of the speech signal given a perceptual hypothesis (one HMM was trained with examples of the word "one", another with examples of the word "two", etc). Since in our database each of the four perceptual alternatives had equal prior probability, the log-likelihood estimates provide the "subjective" discriminant values upon which the MAP rule operates. Classification of the speech signals proceeds by calculating, for each model, the log-likelihood of the observed sequence given that model and choosing the model with maximum log-likelihood.

We tested two different types of architectures, one of which was a restricted version of the other. The restricted version consisted of two independent banks of HMMs, one visual, and the other auditory (see Figure 9). These were trained on their respective feature data sets, and during testing, the resulting responses of each bank were integrated to obtain an audiovisual classification. Under the assumption of conditional independence this amounts to summing the auditory and visual log-likelihood responses for each word category and picking the category which has the highest sum as the winning classification. Hereafter we refer to this architecture as *factorized*.

We also tested an *unfactorized* architecture that did not assume conditional independence. The models used in this architecture consisted of a single bank of audiovisual HMMs. During training, visual and auditory feature data were mixed together into a single observation sequence, where each time-frame contained the concatenation of the visual and auditory feature vectors (i.e., 326 coefficients per frame).

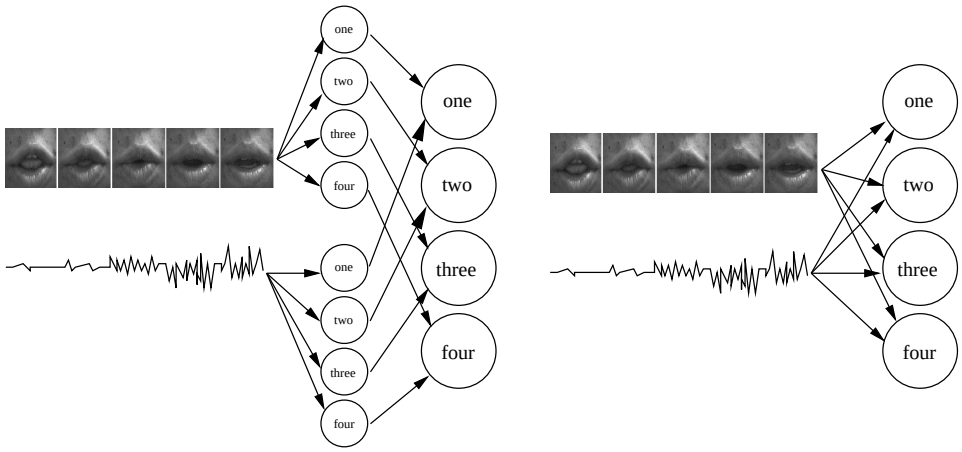


Figure 9: Two approaches to automatic audio visual speech recognition. The factorized approach (on the left side) uses separate banks of HMMs for the audio and visual signal. This approach saves a large number of parameters but is justified only if the audio and video signals are conditionally independent given the categories to be recognized. The unfactorized approach (right side) is more general and does not assume conditional independence. The first approach can be shown to be a special case of the second approach.

5.5 Testing Procedure

The two architectures were evaluated in terms of their generalization performance, i.e., their performance when tested with subjects different from those used for training the system. Tukey's jack-knife (Efron, 1982) method was used to obtain generalization estimates: the two architectures were trained leaving out the utterances of one of the 12 subjects, and testing was done on the utterances of the excluded subject. This process was repeated 12 times, leaving out a different subject each time. Jack-knife estimates were based on the average generalization obtained with these 12 samples.

Hidden Markov models differ in the number of hidden states and Gaussian mixtures per state. For a fixed number of hidden states and mixtures, the parameters trained using the EM algorithm were: (1) The hidden state transition probability matrix, this matrix was constrained to be upper triangular; (2) the centroids of each Gaussian mixture, and (3) the covariance matrix of the mixtures. These covariances were constrained to be of the form σI, where $\sigma > 0$ is an adaptive parameter, and I is the identity matrix. Note that this structure allows to capture high-order dependencies since multiple Gaussian mixtures are used. To give a fair chance to the factorized and unfactorized architectures, we optimized them by testing 45 different variations of the two architectures with different number of states (2,3,4,5,6) and Gaussian centroids per state (2,3,4,5,6,7,8,9,10). Pilot work had shown that optimal architectures were located within this region. This optimization was very costly computationally but was a crucial part of this exercise. By selecting a particular value of Gaussian centroids and number of states one could find that the factorized architecture performed better or worse than the unfactorized architecture. The crucial point here is whether the best models within the factorized architecture are as good as the best models within the unfactorized architecture. We chose the best 2 models within each architecture (factorized and unfactorized) as representative of the entire architecture and report average generalization performance with those 2 best models. For the factorized architecture, the best auditory models had 5 states/7 Gaussians, and 5 states/8 Gaussians. These models respectively had 949 and 1,084 free parameters. The best visual model (which was used with the two best auditory models) had 3 states/3 Gaussians for a total of 2,711 free parameters. For the unconstrained architecture, the best models had 3 states/8 Gaussians and 5 states/7 Gaussians; they respectively had 7,850 and 11,449 free parameters.

5.6 Results

When performance is measured in terms of generalization to new data it is actually possible for restricted versions of a model to perform better than more general versions, as long as the restricted version provides a good working approximation to the statistics of the environment. This has to do with a well known trade-off between bias and variance of statistical models (Geman, Bienenstock, & Doursat, 1992). This is precisely what happened in our study: the best 2 factorized architectures gave an average generalization performance of 98 % correct, slightly better than the average performance of the best 2 unfactorized architectures (95.5 %). Thus, conditional independence provided a close enough approximation to the actual statistical structure of the signals so that the bias introduced by the assumption was well compensated by the reduction in the number of free parameters (the best unfactorized models had about 3 times as many parameters as the best factorized models).

5.7 Discussion

The main point of this exercise was to illustrate how machine perception techniques may be used to test for conditional independence of natural, high dimensional, signals. We know that the Morton–Massaro law holds well when humans integrate audio and visual speech (Massaro, 1989a). According to the optimality theorem, this suggests that conditional independence of acoustic and visual signals approximates well the posterior probabilities of perceptual categories in audio-visual speech. Therefore assuming conditional independence should not result in a significant loss of information when the assumption is applied to artificial audio-visual speech recognition systems. We tested this prediction by building state-of-the-art audio-visual speech recognition systems with and without the assumption of conditional independence. We found that the assumption of conditional independence did not hinder performance. This suggests that indeed audio and visual speech signals may be treated as if they were conditionally independent without significant loss of information about speech categories.

The experiment we have reported illustrates an approach to these issues, and is certainly consistent with the idea that independent treatment of audio and visual speech is reasonable. The main limitation of our analysis is the fact that our database was relatively small. Thus our work should only

be seen as exploratory research. We had to use a small database because the convincingness of the analysis depended on the fact that the factorized and unfactorized architectures were as good as they could possibly be. This required optimizing over the space of HMMs, which is a very time consuming process. As computing power increases it should be possible to perform this type of studies with larger databases.

6 The Conjunctive Context Effect

The optimality theorem suggests that violations of the Morton–Massaro law may occur when objective probability ratios of perceptual categories do not factorize. A particularly obvious violation of factorability occurs when there are only two categories and optimal information integration requires cross-over interactions (see page 22). Previous research has shown that subjects have no difficulty learning category structures that require such interactions (Kruschke96CS). In this section we present a case of a cross over interaction, i.e., a violation of the Morton–Massaro law, in a letter perception task.

Consider the task of recognizing 3-letter words in English. The goal is to choose between two alternatives E and U for the identity of the middle letter in the following contexts: M_N, R_D, M_D, R_N. In this case we have three sources of information, corresponding to the three letters in the word. The stimulus is whatever letter appears in the second position. The other two letters (which are now part of the context) provide additional information about the second letter because, in written English the strings MEN and RED are high-frequency words but MUN and RUD are not; while MUD and RUN are high-frequency words but MED and REN are not. If we look at any context letter (let's say R), the constraint it places on the identity of the target letter reverses as we manipulate the identity of the other letter in the context. Taken by itself, R is consistent with both E or U, since it occurs with E in RED and with U in RUN. When taken together with one of the context letters, however, R unequivocally supports only one of the two alternatives. When the final letter is N, R supports middle U; and when the final letter is D, M supports middle E.

Consider how a factorized model may work here: (1) Three autonomous modules, one per letter location, would selectively evaluate the letter presented in a particular location and would output quantities that would indicate the support of that location for "E" and "U" on the second location; (2)

the output of these three modules would be integrated using a multiplication rule; (3) an alternative would be chosen with probability proportional to the product of the 3 evaluations.

In English, we have that

$$p(R\_N \mid \_U\_) \approx 1, \tag{33}$$

$$p(R\_N \mid \_E\_) \approx 0, \tag{34}$$

and thus the context R_N strongly supports the alternative "U" for the middle letter. However, if we assume conditional independence, an analysis of the Kucera and Francis (1967) corpus restricted to three-letter words tells us that

$$p(R\_\_ \mid \_U\_)P(\_\_N \mid \_U\_) = \frac{244}{9555}\frac{493}{9555} = \frac{132}{10000}, \tag{35}$$

$$p(R\_\_ \mid \_E\_)P(\_\_N \mid \_E\_) = \frac{256}{11178}\frac{1095}{11178} = \frac{432}{10000}. \tag{36}$$

$$\tag{37}$$

Thus, if we assume conditional independence, R_N would strongly support "E" instead of "U" for the middle letter. While the assumption of conditional independence will not always support the non-word alternative, it is apparent that the a model with such an assumption may have problems producing a strong word-superiority-effect when the context letters combine in a conjunctive manner.

6.1 Upper Bound on the Predicted Effect

Consider how the factorized model would be parameterized when the inputs are the letter combinations MEN, MUN, REN, RUN, MED, MUD, RED, RUD, the task is to recognize the middle letter and the response alternatives are E, and U. In this case we would need 2 parameters to describe the support of M__ and R__ for the E alternative, 2 parameters for the support of _E_ and _U_ for the E alternative and two parameters for the support of __D and __N for the E alternative. Thus the factorized model would use a total of 6 parameters (in fact, as we saw in a previous section two of these parameters would be redundant). Similar models could be used when the task at hand is to recognize the first or the last letters.

If our dependent variable were logit(p), where p is the probability of correct identification, then this model would predict no difference between

the word and non-word conditions. One problem with the logit function is that it is undefined for extreme values of p and thus it is desirable to work with raw probability estimates rather than their logit transformation. Unfortunately, due to the fact that the logit transformation is non-linear, it is possible to find parameter values of the factorized model that produce a small word superiority effect when the dependent variable is the raw probability of correct identification. To take care of this problem we specified the largest possible word-superiority-effect achievable by the factorized model. First we performed a grid search looking for combinations of parameter values that maximized the difference in error rate between word and non-word stimuli for groups of 3 letters of the type described above. The search was constrained by the following conditions:

1. A parameter set was not used if for any of the words in the word-set, both the stimulus and context support the incorrect alternatives. For example, consider the case of the item MEN with the E in the second letter position taken to be the stimulus. A parameter set would not be used if for this item: (a) The stimulus E supports the E response less than it supports the U response, and (b) at the same time the context M_N supports the E response less than it supports the U response.

2. The expected error rate for all conditions could not be smaller than 1%. This constraint is realistic in the experiment we will describe later, which produced an overall error rate of 22%.

The grid search was followed by Powell's derivative-free optimization algorithm (Press, Flannery, Teukolsky, & Vetterling, 1989) using the best condition found by the grid search as its starting point. The result of this extensive search was that the maximum average word superiority effect achievable by the factorized model was 2.26%. Thus, word superiority effects stronger than this are inconsistent with the factorized model.

6.2 Stimuli and Materials

Using a computer program we searched for cases in the Kucera and Francis (1967) corpus for sets of 3 letter words organized in a conjunctive manner. The condition was that for the three positions, switching one letter with the letter in the same position of another word in the set should produce a non word. For example, in the set MEN, RED, MUD, RUN letter switching

Table 2: Conjunctive word sets.

words								*non-words*							
had	5133	war	464	her	3038	wed	2	har	0	wad	0	hed	0	wer	0
aid	130	oil	93	all	3001	old	660	ail	0	oid	0	ald	0	oll	0
few	601	hex	4	fox	13	how	834	fex	0	hew	0	fow	0	hox	0
men	763	red	197	mud	32	run	212	med	0	ren	0	mun	0	rud	0
age	227	ugh	1	ash	11	use	589	agh	0	uge	0	ase	0	ush	0
nae	1	war	464	nor	195	woe	5	nar	0	wae	0	noe	0	wor	0
bad	142	gap	17	bop	3	god	318	bap	0	gad	0	bod	0	gop	0
gal	5	sat	150	got	482	sol	3	gat	0	sal	0	gol	0	sot	0
dry	68	gre	1	due	142	guy	51	dre	0	gry	0	duy	0	gue	0
sam	79	tar	12	sir	95	tim	25	sar	0	tam	0	sim	0	tir	0

generates a matched set of non-words MUN, RUD, MED, REN. We chose 10 sets of words that had this conjunctive property and that maximized median frequency of word occurrence while minimizing the difference in frequency of occurrence within the words in the same set. The 10 sets of words, matched non-words, and the frequency of occurrence of each in Kucera and Francis (1967) is displayed in Table 2. We treated membership in the lexicon as a discrete binary variable: Words are letter strings that occur in the Kucera-Francis word list, and all other letter strings are treated as nonwords. However, in our stimuli some of the "non-words" are in fact low frequency words, and others, such as MED, are used as abbreviations. It is likely that such items act as words as least to some subjects but they are generally much lower in frequency than the stimuli categorized as words (as attested by their frequency of 0 in the KF word count).

The 10 sets of words have the characteristic that any of the three letter positions can serve as target in each set. For instance, if the letter to be reported is the last instead of the middle letter, the conjunctive effect is still there. In the MEN-RED-MUD-RUN example, middle E favors N when the first letter is M but it favors D when the first letter is R. Therefore, from 10 sets of 8 3-letter stimuli there are 240 possible test items, in which each 3-letter stimulus occurs three times with each letter serving as the target once. The stimuli were presented using a 486-Dx33 based computer with

a 15" NEC4FGx color monitor running at 72Hz refresh rate and standard VGA resolution (640x480 pixels, 16 colors). The display parameters were adjusted using a pilot study so that the average error rate was about 25%.

6.3 Subjects

The subjects were 30 UCSD undergraduate students. They received partial course credit in introductory cognitive science courses for participating in the experiment. All subjects were tested individually.

6.4 Procedure

Each subject received the following instructions: "You will be presented letters for brief periods of time. Your goal is to identify these letters. There will be 24 practice trials followed by 240 testing trials. Press the space bar to get a new set of letters. Press the up or down arrow key to choose the correct alternative. 50% of the trials show letters that make a word. 50% of the trials show letters that do not make a word. The experiment is self-paced. You can take as much time as you want between stimulus presentations."

After the instructions, subjects received 4 demonstration trials and were given the opportunity to ask questions. After answering any questions the subject had at this point, the experimenter left the room. The demonstration trials were followed by 24 practice trials chosen from an additional set of stimuli organized in a conjunctive fashion. After 24 practice trials, subject received the 240 test trials. The sequence of test trials was randomized for each subject. Each trial was initiated by the subject pressing the space-bar key on the computer key-board, which triggered presentation of a fixation point for 1 sec. The fixation point was immediately followed by the 3 letter display, which in turn was followed by a mask made of 3 # characters. The SOA between the stimulus display and the mask was 28 msec (2 refresh cycles). After 200 msec of mask presentation, a probe appeared indicating the target position and the two possible response alternatives.

Each subject was run under one of three possible feed-back conditions. These conditions were included to explore the generality of our results across situations that could potentially create different response biases. In feed-back condition I, the subject received no feed-back. In feed-back condition II, the response was followed for 1 second with a presentation of the correct letter in the appropriate position. In addition, a beep occurred if the incorrect

choice had been made. Feed-back condition III was the same except that now both the stimulus and context were displayed after the response was made. Subjects in conditions II and III were told at the outset that after the response, the computer would display the correct alternative and beep if the response was incorrect.

6.5 Design

There were three independent factors: Word vs. Non-word (W); Position (P) of the tested letter (first, second, third); and Feed-back condition (F). Factors W and P were within subjects, and F was between subjects. The subject's identity (S) was treated as a random factor. The dependent variable was the probability of correct response averaged across the 10 stimulus sets. An additional analysis treating stimulus set as a factor to check on the consistency of the results across stimuli are described below.

6.6 Results

The analysis of variance indicated a significant word superiority effect. No other main effect or interaction was significant. Table 3 shows the analysis results using Keppel's (1993) standard notation.

The difference in accuracy between the word and non-word conditions was 4.472%,
$P(\text{correct} \mid \text{word}) = 0.799$, $P(\text{correct} \mid \text{non-word}) = 0.754$, $\text{F}(1,27) = 29.82$, $p < 0.001$. The word superiority effect is somewhat smaller than the difference found in the only other experiment we know of in which only one of the forced-choice alternatives makes a word with the context (Estes, 1975). The relatively small difference in our case may be explained in part by the fact our stimuli are only three letters, and the fact that our "non-words" were all pronounceable pseudo-words, while those used by Estes were not necessarily word-like.

An additional analysis using the 10 word sets as an additional factor showed a significant effect of the stimulus set ($\text{F}(9,243) = 3.29$, $p < 0.01$) but no set by word interaction ($\text{F}(9,243) = 1.14$, $p \geq 0.05$). Thus, some sets were easier than others over all, but there is no evidence that the word advantage differed across sets.

The obtained word superiority effect was significantly larger than 2.26%, the upper bound for factorized models: $t(27) = (4.47 - 2.26)/\sqrt{(2)(30.17)/90}, p <$

Table 3: ANOVA table of conjunctive context experiment.

Source	df	MS	F	p
w	1	900.0347	29.829	$p < 0.001$
w× s/f	27	30.1736		
p	2	39.6181	0.478	$p \geq 0.05$
p× s/f	54	82.9398		
w× p	2	30.4514	1.269	$p \geq 0.05$
w× p× s/f	54	24.0046		
f	2	313.3681	0.333	$p \geq 0.05$
s/f	27	942.2569		
w× f	2	22.3264	0.740	$p \geq 0.05$
w × s/f	27	30.1736		
p × f	4	35.5035	0.428	$p \geq 0.05$
p × s/f	54	82.9398		
w × p × f	4	16.9618	0.707	$p \geq 0.05$
w × p × s/f	54	24.0046		

0.01 , where the pooled estimate of the standard deviation of the effect is derived from the $w \times s/f$ error term of the ANOVA. Thus simple factorized models can be rejected, indicating a violation of the Morton–Massaro law. Equivalent results were obtained using the MANOVA approach to within subjects analysis.

6.7 Discussion

It should be noted that, as is the case in the prototypical experiments described in previous sections, our experiment does not distinguish whether the effect of context is perceptual or post-perceptual. In particular it is possible that subjects' perceptions of the stimulus were not affected by context but their choices were. The quantitative formulations of information integration that are considered in this paper (e.g., the logogen model and the FLMP) were developed to account for data obtained in experiments like ours, in which it is not possible to distinguish between perceptual and post-perceptual processes. What is an issue in our experiment is the nature of the proper characterization of the mappings between information sources and

responses, not whether these mappings were perceptual or post-perceptual. In some ways the experiment may seem an obvious one, in that it predicts a word advantage over non-words in a situation where a response bias could account for the effect. Many experiments have already reported such findings (Estes, 1975). What makes this experiment different is that it rules out the possibility that the word advantage is due to guessing based on pair-wise letter constraints and requires that it be based on whole-word constraints.

A model with three autonomous modules, each selectively evaluating the support of a letter position for each of the different response categories is in direct contradiction with our experimental results. However our results are not incompatible with all forms of factorization. In particular the effect of the stimulus may factorize from the effect of context as a whole. For example, our results are compatible with an architecture that uses a module for the stimulus and a separate module for the context as a whole. In fact this is the approach used by Massaro and Cohen (1991) to account for the word superiority effect (they use a version of the FLMP with factors selectively influenced by the stimulus and by the context as a whole). This points out an important issue: The pattern of results predicted by the Morton–Massaro law depends on how we choose to group our information sources. In our experiment the Morton–Massaro law does not hold if each letter is considered as an experimental factor but the law still holds if the two context letters are jointly considered as a single factor. Unfortunately such a version of the Morton–Massaro law does not tell us how the two sources that make up the context are combined.

Thus, our results indicate the need for further development of models such as the current version of the FLMP. It may be useful to consider what such a model might look like, since this will help bring a second important lesson of the experiment. To begin, let's imagine the case where the displays are three letters as in our experiment, but only one display position, e.g., the first, is designated as the target position; the other letter positions serve only to provide context. In this case, we could construct a FLMP model to account for the data for first position target letters. To do so, we would need a pathway for evaluation of the support for each alternative provided by the features of the target letter, and a separate pathway for evaluation of the support for each alternative provided by the context. This latter pathway would have to combine the two context letters conjunctively, while maintaining mutual independence with the pathway that processes the features of the target letter; the outputs of these two pathways would then be combined at

the integration stage, in accordance with the FLMP equation.

While this approach offers a straightforward method for accounting for the results obtained when the target position is known in advance, it does not provide a complete account of performance in our actual experiment, since the subjects did not know which position would contain the target until after the stimulus had been followed by a mask. The task was sometimes to identify the first letter, sometimes the middle letter, and sometimes the last letter, and subjects were not informed in advance of each trial which letter position would be tested. To deal with the possibility of targets in the second position, the stimulus and context pathways would have to be configured differently, and a third configuration would be needed for third-position targets. To cope with the fact that subjects have no prior information about which letter position will be tested, it would appear that the three different configurations would have to be used in parallel on each trial; otherwise the perceptual data would have to be buffered for post-processing, after the alternatives have been interpreted and the appropriate network chosen or configured.

Massaro and his colleagues have not taken a stand on these issues, and for us this reinforces our suggestion that the FLMP model (even as elaborated by Massaro & Cohen, 1991 to address the word superiority effect) may be viewed as a descriptive and functional framework for capturing aspects of experimental data (See JacobsGrainger94 for a similar way of viewing the FLMP). However, Massaro clearly views the FLMP as describing structural aspects of processing itself (e.g., its feed-forward nature) not just function. Our suggestion is that the descriptive, functional, and structural aspects of the FLMP should be consider separately. We suggest that the FLMP equations provide a good description of certain aspects of the experimental facts, that the fuzzy-logic framework provides a reasonable framework for the functionality of the equations, but that it is necessary to carry out further analyses to understand whether they provide guidance about the structure of the perceptual system.

There are models that offer detailed mechanisms explaining how information from individual context letters is combined. Our results are incompatible with models that only keep information about pair-wise letter statistics, like Golden's (1986) word recognition model. However, it is likely that our findings are compatible with models that use information about frequencies of higher-order conjunctions of letters in English, not just pair-wise letter statistics. These models include Rumelhart and Siple's (1974) model, Mc-

Clelland and Rumelhart's (1981) interactive activation model, Paap, Newsome, McDonald, and Schvaneveldt's (1982) activation verification model, and GraingerJacobs94) dual read-out model. In all of these models, the co-occurrences of three of four-letters together to form words are used in determining the likelihood of a letter in a given context. Thus, these models take conjunctions of context letters into account, thereby making it likely that they could account for the conjunctive context effect.

While these models may account for the conjunctive context effect, it is unclear whether they can also account for Massaro and Cohen's (1991) findings, i.e., the fact that the Morton–Massaro law holds when the sources of information are grouped into stimulus and context as a whole. In the next section, we consider this issue for the case of interactive models of perception. In particular, we show that a stochastic version of McClelland and Rumelhart's (1981) interactive activation model can: (1) account for the conjunctive context effect, and (2) account for the fact that the effects of the target letter and of the context-as-a-whole factorize.

7 Implementation analysis

The Morton–Massaro law is a statement about an empirical regularity: The fact that the joint effects of information sources on response probabilities often factorize. As such it can be used as a descriptive characterization of data without reference to underlying processing mechanisms. The optimality theorem helps us understand the possible rational basis of the law but it does not tell us anything about internal processing mechanisms. Are there implications of the Morton–Massaro law for the structure of the underlying processing mechanisms? In this section we consider such questions.

In particular we address potential incompatibility of the Morton–Massaro law with interactive models of perception. As mentioned earlier, the Morton–Massaro law tells us that the joint support of information sources for different response alternatives can be factorized into terms selectively influenced by each information source. This fact is consistent with modular approaches, like the FLMP or Morton's logogen model, in which information processing is decomposed into autonomous modules. But it is not clear whether the law is consistent with interactive architectures in which feed-back and lateral connections allow every part of the system to be affected by every other part. Massaro (1989b) noted that the probabilities of responses generated by the original version of the interactive activation model did not factorize, and he conjectured that this was an inescapable consequence of the interactive character of the model. He thus attempted to conclude that adherence to the law proved that the mechanism is feed-forward.

McClelland (1991) refuted Massaro's conjecture by showing that stochastic versions of the interactive activation model can be factorized if their connectivity is constrained in certain ways. The analysis presented there assumed networks of units with discrete binary states and local representations. Simulations suggested that interactive networks with continuous states may also exhibit factorability but no mathematical analysis was presented for the continuous case. Massaro and Cohen (1991) reported simulations in which they could not find good fits to actual empirical data using the stochastic interactive activation model, but Movellan and McClelland (1995) showed that their failure was due to inappropriate limitations on parameters. By removing these constraints they obtained excellent fits.

In this section we expand the mathematical analysis of interactive models to networks that work in continuous time with continuous states. These networks relate to diffusion models, which have a rich tradition in the human

information processing literature (Ratcliff, 1978TownsendAshby83Ashby89) and take us a step closer to understanding how perceptual models may be implemented in electric circuits and neural hardware.

First, we formally introduce the class of models we are concerned with and present the concept of *stochastic equilibrium*, which plays a central role in the analysis. Second, we introduce the concept of *channel separability* and show its relationship to the Morton–Massaro law.

7.1 Diffusion Networks

McClelland (1993) emphasized the principles of graded, random and interactive propagation of information to explain human cognition and to bridge gaps between the study of the mind and the brain. Following that idea here we focus our analysis on network models which are interactive (i.e., in the general case they may have lateral and feed–back connections), graded (i.e., operate in continuous time with continuous representational states) and intrinsically stochastic (i.e, they are defined via stochastic differential equations instead of ordinary differential equations). We call these models diffusion networks for they are stochastic diffusion processes defined by adding Brownian motion to the standard connectionist network dynamics (Movellan & McClelland, 1993; Movellan, 1994; Mineiro, Movellan, & Williams, 1998; Movellan, 1998). Diffusion networks are vary similar to the continuous stochastic interactive activation networks presented in McClelland (1991). We modified the networks to give them a useful physical interpretation as models of low-power electronic circuits and to facilitate their analysis using tools from the theory of stochastic differential equations and diffusion processes (Karatzas & Shreve, 1988; Kloeden & Platen, 1992; Oksendal, 1992).

Diffusion networks consist of a set of n units whose activation values change through time due to influences from other units and from external inputs. In addition the network dynamics are probabilistic due to an intrinsic noise component. Diffusion networks are defined via stochastic differential equations of the following form[7]

$$dY_i(t) = \mu_i(Y(t), X)\, dt + \sigma\, dB_i(t) \quad \text{for } i \in \{1, \cdots, n\}, \tag{38}$$

where $Y_i(t)$ is a random variable representing the *internal state* at time t of the i^{th} unit, $Y(t) = (Y_1(t), \cdots, Y_n(t))'$ is the vector of activation states at

[7]We do not include the initial distribution since it is irrelevant for the equilibrium solution, which is the focus of our paper.

time t, $X = (X_1, \cdots, X_m)'$ is the external input, and B_i is Brownian motion, a mathematical model of the random noise present in many natural systems. The constant $\sigma > 0$, known as the *dispersion*, controls the amount of noise injected into each unit. The function μ_i, known as the *drift*, determines the average instantaneous change of activation. The drift is modulated by a matrix w of connections between units, and a matrix v that controls the influence of the external inputs onto each unit,

$$\mu_i(Y_i(t), X) = \frac{1}{\kappa_i(Y_i(t))}(\bar{Y}_i(t) - \frac{1}{\rho_i} Y_i(t)) - \gamma Y_i(t), \quad \text{for } i \in \{1, \cdots, n\}, \tag{39}$$

where κ_i is a positive function which will be defined in (43), $\rho_i > 0$ is a constant parameter whose function is explained below,

$$\bar{Y}_i(t) = \sum_j w_{i,j}\ Z_j(t) + \sum_k v_{i,k} X_k, \quad \text{for } i \in \{1, \cdots, n\}, \tag{40}$$

is the *net-input* of unit i, $w_{i,j}$ is the weight from unit j into unit i, $v_{i,k}$ is a parameter modulating the effect of input X_k on unit i, and

$$Z_j(t) = \varphi_i(Y_j(t)) = \varphi(\alpha_i\, Y_j(t)) \tag{41}$$

is the *activation* of unit j. The term φ represents the activation function, which we assume to be differentiable and bounded. In practice we use the logistic function but other options are also possible:

$$Z_i = \frac{1}{1 + e^{-\alpha_i\, Y_i}}. \tag{42}$$

The term $\gamma > 0$ is a *decay* parameter, and the $\alpha_i \geq 0$ terms are *gain* parameters that control the sharpness of the activation functions. For large values of α_i the activation function of unit i converges to a step function (see Figure 10). Finally, the function κ_i is defined as follows

$$\frac{1}{\kappa_i(Y)} = \frac{d\varphi_i(Y_i)}{dY_i}. \tag{43}$$

This particular form of κ_i was chosen to facilitate the mathematical analysis of the equilibrium behavior of these networks (Movellan, 1998). In practice,

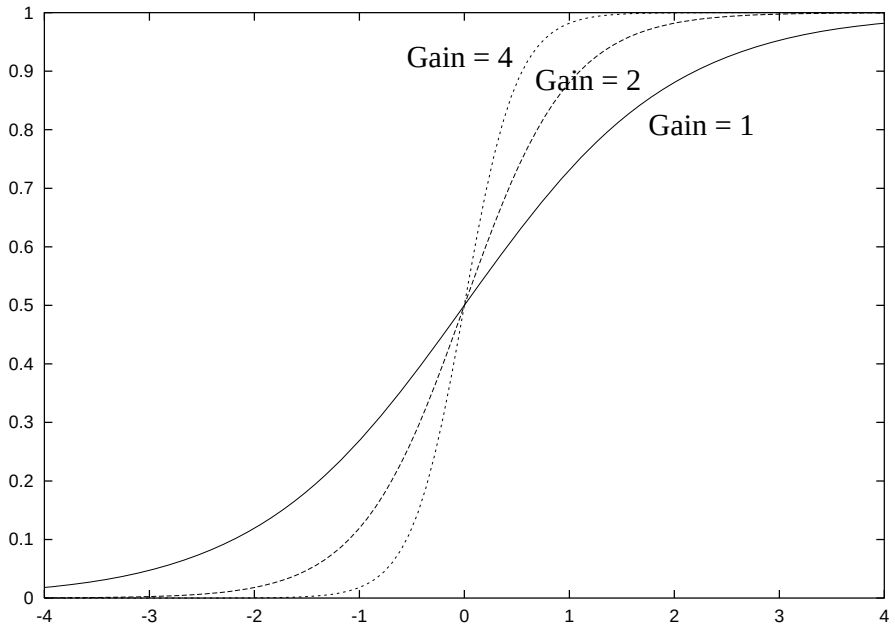


Figure 10: Logistic activation function for several values of the gain parameter. As the gain goes to infinity the logistic function converges to a step function.

if the activation function is logistic,

$$\frac{1}{\kappa_i(Y)} = \frac{d\varphi_i(Y_i)}{dY_i} = \alpha_i \varphi(Y_i)(1 - \varphi(Y_i)). \tag{44}$$

Note that the drift function in (39) changes the state of the units so that the state follows the net-input. If the net-input times ρ_i is larger than the internal state, the drift is positive, and the internal state has a tendency to increase. If the net input times ρ_i is smaller than the internal state the drift is negative and the state has a tendency to decrease. The function κ_i controls the speed with which the internal state follows the net input. The speed decreases as the activation Z_i gets closer to the extremes.

Intuition for (39) can be achieved by thinking of it as a the limit of a discrete time difference equation, in such case

$$Y_i(t + \Delta t) = Y_i(t) + \mu_i(Y_i(t), X)\Delta t + \sigma\sqrt{\Delta t}\, N_i(t), \tag{45}$$

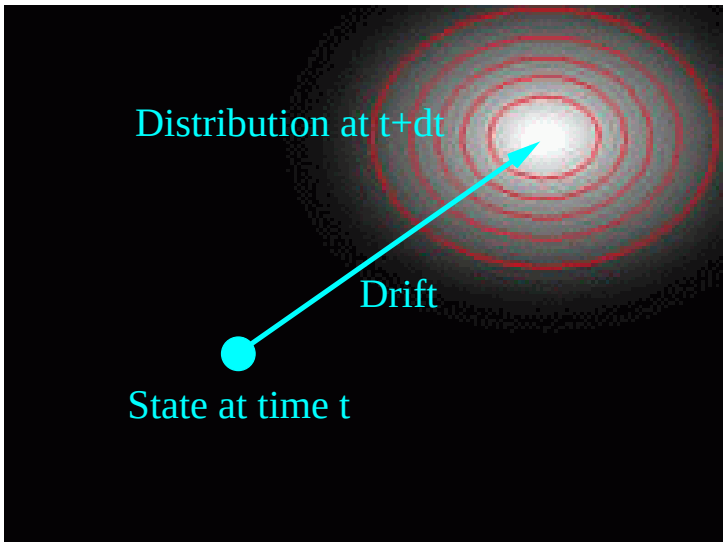


Figure 11: The drift and the dispersion combine to produce a distribution of states at time t+dt, given the state at time t.

where the $N_i(t)$ are independent standard Gaussian random variables. For a fixed state at time t there are two forces controlling the change of state: the drift, which is deterministic, and the dispersion, which is stochastic (see Figure 11). This results in a distribution of states at time $t + \Delta t$. As Δt goes to zero, the solution to the difference equation (45) converges in distribution to the diffusion process defined in (39).

Diffusion networks can be seen as a generalization of the linear diffusion model that has been used by Ratcliff's (1978) to model reaction time in memory experiments and other paradigms. In Ratcliff's model, the drift is constant within an experimental trial, and the units are not coupled. In our diffusion networks the units, are coupled via connection weights and the drift is a function of the state of activation. Diffusion networks are closely related to the coupled diffusion processes investigated by Ashby99) but our approach is more general since we allow non-linear activation functions and feed-back, not just lateral connections between processing units. Diffusion networks have a physical interpretation as models of low-power electronic circuits in which the random motion of the electrons plays a non-negligible role. In addition diffusion networks are reasonable models of biological neural networks with an intrinsic noise component. In such case each unit represents

a point neuron, the weight values represent synaptic conductances, κ_i is a capacitance function, ρ_i a transmembrane resistance, the internal states Y are pre-synaptic potentials, the elements of $\bar{Y}$ are net input currents, and the elements of Z are post-synaptic short-term firing rates (see p. 3089, Hopfield, 1984).

7.2 Stochastic Equilibrium

Figure 12 shows an example of a diffusion network with 2 units responding to a transient input. The two units were coupled to each other via bidirectional excitatory connections, the external input had an excitatory effect on unit 1. The "noisy curves" show the evolution of activation in a particular run of the network. Consider an ensemble of identical clones of the same diffusion network, what is sometimes known as a *canonical ensemble*. We can think of members of the canonical ensemble as replicas of the same network running in parallel, or as distinct runs of the same network running at different times. For each unit and for each time step we could compute expected activation values by averaging across the entire ensemble. Figure 12 displays the evolution of such averages (the two smooth curves). Note how some time after the presentation of the stimulus these averages equilibrate. We say that a network is at *stochastic equilibrium* when the probability distribution of states across the ensemble equilibrates, i.e., at stochastic equilibrium all the ensemble statistics (mean, variance, covariance, kurtosis, etc) equilibrate. Note that due to the effects of noise the activations obtained in individual runs never equilibrate. What equilibrates is the distribution of states across an entire ensemble of runs. Figure 13 shows a simulation of a single unit diffusion network. The network was run 50 million times with internal states starting at 0. The activation histogram was collected after 10, 40 and 600 cycles. The figure displays the theoretical equilibrium distribution, which was almost perfectly achieved after 600 cycles. Simulation details are in the Appendix.

A system like this can be used to do statistical inference in the following way: An input x is introduced that disturbs the established equilibrium. Eventually, a new equilibrium distribution is reached. This stable distribution represents the response of the system to the input x.

Consider now an experiment involving repeated trials each displaying particular combinations of information sources. Each such trial with a particular combination of information sources can be thought of as another sample from

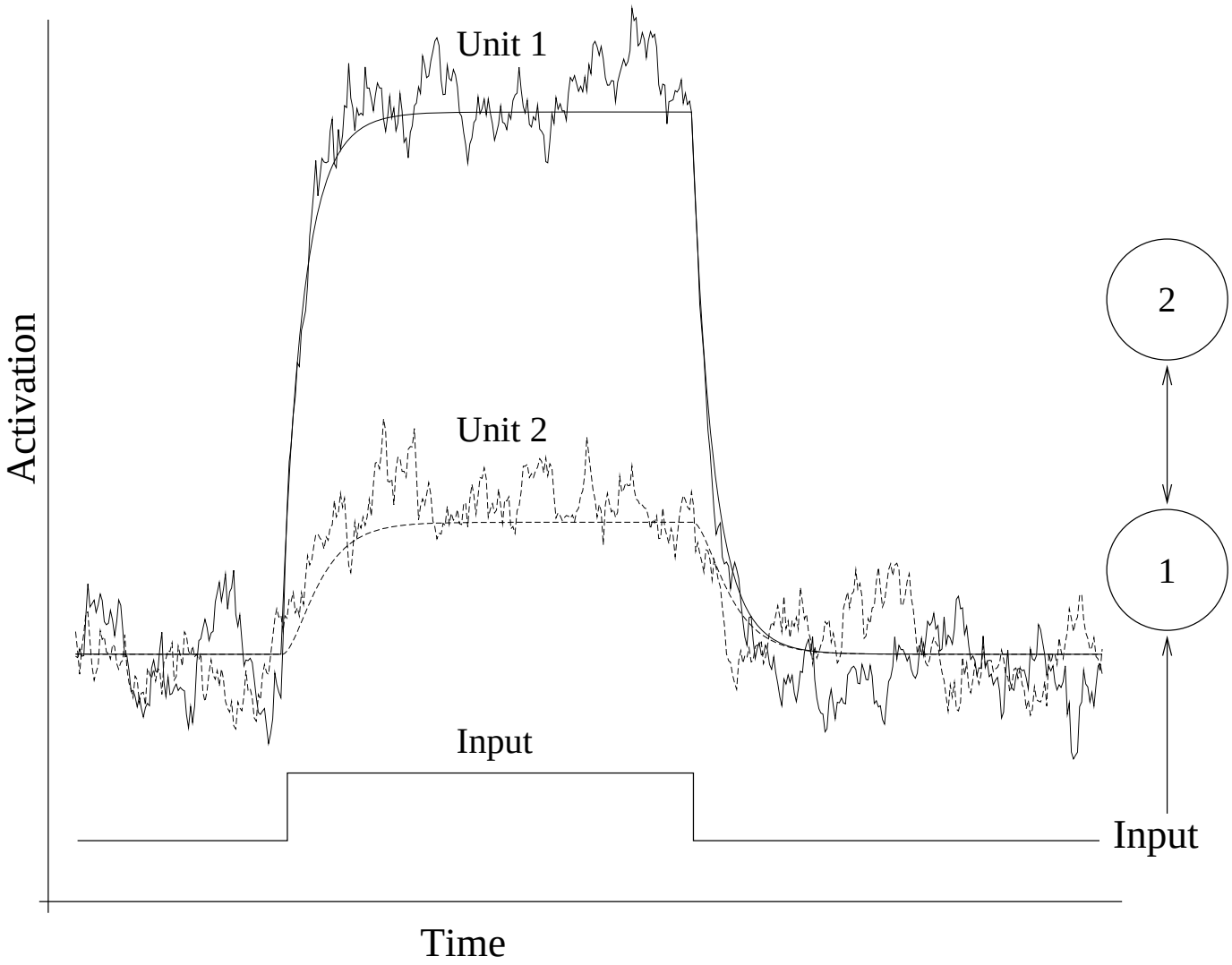


Figure 12: Simulation of a 2 unit diffusion network, illustrated at right. The input was a rectangular signal (bottom function) . The Figure shows a sample activation path, one per unit in the network, and the average activation path.

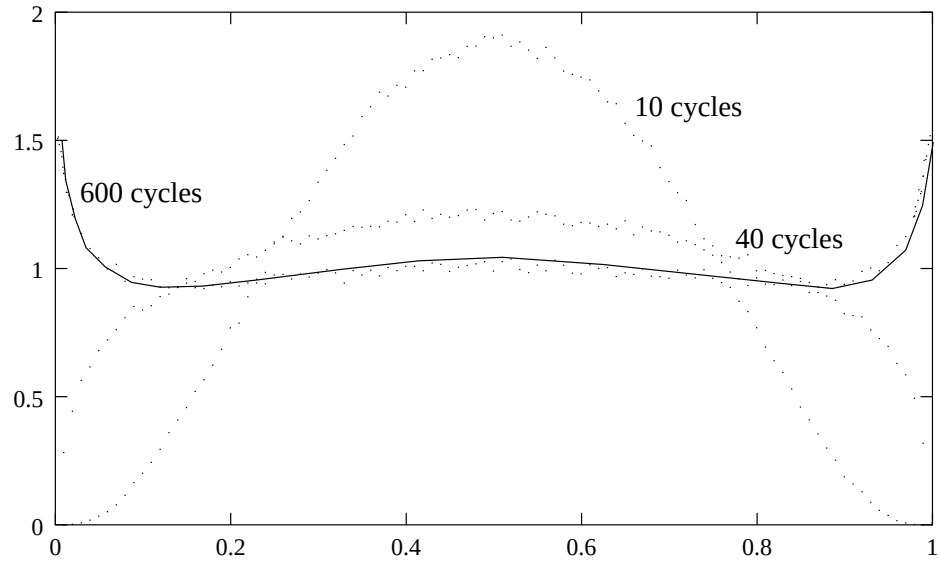


Figure 13: The figure shows the empirical activation histograms obtained by simulating a 1-unit network 50 million times. Histograms were collected after 10, 40 and 600 cycles. The theoretical equilibrium distribution is displayed with a continuous line. Simulation details are in the Appendix.

the canonical ensemble. On each trial we let the network run for a period of time sufficient to approximate stochastic equilibrium and count the proportion of trials that the state of the network is in a particular region of the state space. As the number of trials increases, this proportion converges into the equilibrium probability of being in the region. Each region is associated with one of the possible responses, so that the probability of being in a region corresponds to the probability of making a particular response. This is precisely the approach we use in the next section to analyze the Morton–Massaro law in diffusion networks.

7.3 Channel Separability

To simplify the analysis we assume the network connections are symmetric, i.e., the connection $w_{i,j}$ from unit j to unit i equals the connection $w_{j,i}$ from unit i to unit j. Later we discuss more general cases in which this assumption is relaxed. Symmetric weights allow for each state of the network to have an associated *goodness* value. Here goodness represents how well the

activation of the units in the network conforms to the constraints imposed by the external inputs and the network architecture. The goodness function has the following form

$$G(z \mid x) = \frac{1}{2}\Sigma_i\Sigma_j z_i w_{i,j} z_j + \Sigma_i\Sigma_k z_i v_{i,k} x_k - \Sigma_i S_i(z_i) \ , \tag{46}$$

where $G(z \mid x)$ is the goodness of activation state z given the external input x. The state of activation z is an n-dimensional vector, z_i is the i^{th} element of z (i.e., the activation of the i^{th} unit), x is an input vector containing the state of the external sources of information, and x_i is the j^{th} component of that vector. The expression $\Sigma_i\Sigma_j z_i w_{i,j} z_j$ is a measure of the consistency of the pattern of activation z in the network with the matrix of connection weights w. One can see that if $w_{i,j}$ is positive, then goodness increases monotonically with either z_i or z_j as long as they have the same sign. When $w_{i,j}$ is negative, goodness increases monotonically with either z_i or z_j as long as they have opposite signs. The next part of the expression, $\Sigma_i\Sigma_k z_i v_{i,k} x_k$, is a measure of the consistency of the pattern of activation z with the external input x. The first two terms in equation (46) favor states with extreme values. The function S_i in the final term penalizes extreme activations of the units. The exact form of this function is on equation (87) of the Appendix but its most important characteristic is that it rapidly increases as the activations approach extremes. Thus overall, the goodness of a state represents a balance between a tendency to maximize activations of units with positive weights from other active elements (units or inputs) while at the same time avoiding extreme activations. A fuller discussion of goodness, and how networks tend to settle toward good states, may be found in Rumelhart et al. (1986). In Movellan (1998) it is shown that if the weights are symmetric then, at equilibrium, the probability of an activation state z given an external input x is an exponential function of the goodness

$$p(z \mid x) = \frac{1}{K(x)} \exp((2/\sigma^2)\ G(z \mid x)) \ , \tag{47}$$

where $K(x)$ is a proportionally term that makes probability densities integrate to 1. Note how the "better" a state is, i.e., the better it satisfies the network constraints, the more probable it is to visit it at equilibrium.

In the Appendix we show that for networks of this type there is a condition, which we named *channel separability*, which is sufficient to exhibit

the Morton–Massaro law. Our formalization of channel separability entails separating units into the following sets:

1. **Response Specification Units**: A unit belongs to the response specification set, if, when the state of all the other units in the network is fixed, changing the state of this unit can affect the probability distribution of external responses.

2. **Information Channels**: A unit belongs to the channel $\mathcal{C}_i$ of information source X_i if: (1) It is not a response specification unit, and (2) when the state of the response units is fixed, the probability distribution of the activations of this unit can be affected by X_i.

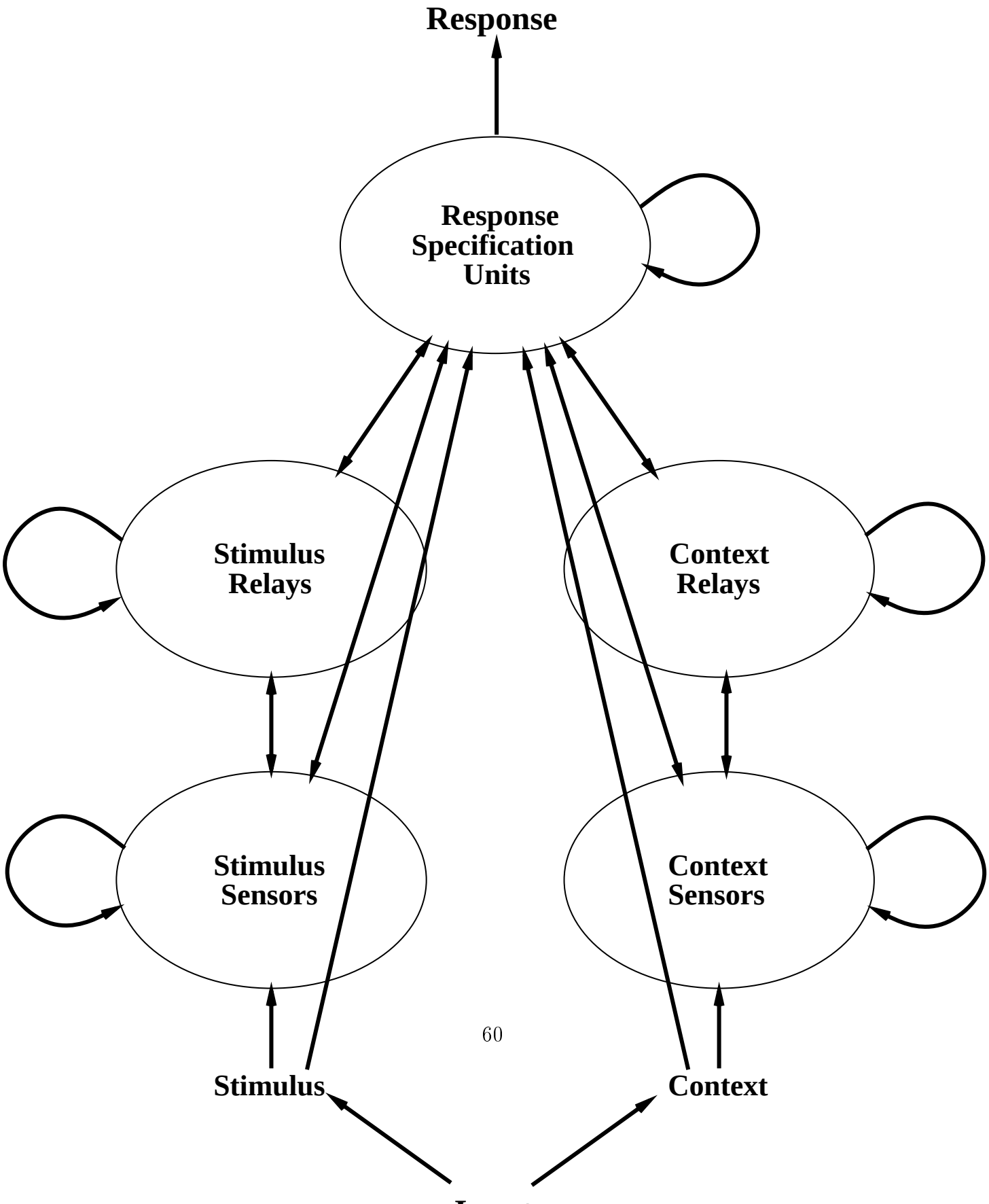
Channel Separability: We say that two channels are separable if they are disjoint, i.e., if they have no units in common. Units within a channel can be sensors or relays:

1. **Sensors:** A unit is a sensor for source X_i if when the state of all the other units in the network is fixed, the probability distribution of activations of this unit is affected by X_i. If a unit is a sensor for more than one source we say that it is a mixed sensor.

2. **Relays:** A unit is a relay if it is not a sensor and it is not a response unit.

Channel separability requires that there be no mixed sensors and that there be no direct connectivity between relays belonging to different channels, as illustrated in Figure 14. Separability would be violated if, for example, the external stimulus is visual, the external context is auditory, and the auditory stimulation is so loud that it has a physical effect on the retina strong enough to influence neural activity there. In such a extreme case the retina would be a mixed sensor, thus violating separability.

Separability implies that the influences of an information source upon the channel of another information source should be mediated via the response specification units. Since the connections in the network may be bidirectional throughout, such indirect influences are possible. Direct connections between primary visual cortex and primary auditory cortex would violate channel separability. However, if responses are read out of a downstream cortical area, bi-directional connections linking the downstream area from


Response
Response Specification Units
Stimulus Relays
Context Relays
Stimulus Sensors
Context Sensors
Stimulus
Context

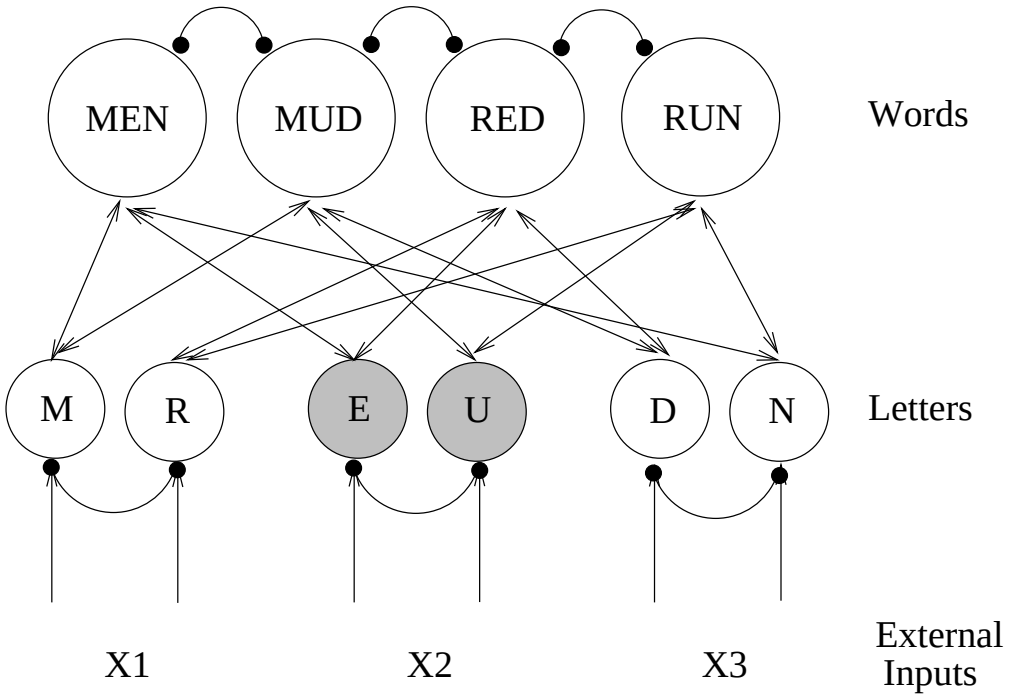


Figure 15: A simplified version of McClelland and Rumelhart's (1981) word reading model. Negative connections end with a circle. Not all the negative connections are displayed.

which responses are read out to each of the two primary cortices would not violate the separability constraint.

To analyze the channels in a diffusion network first we isolate the response specification units and disconnect them from the rest of the units in the network (i.e., we create a virtual network with all the connections in and out the response specification units set to zero). The channel $\mathcal{C}_i$ for information source X_i is the set of units in the virtual network whose activation is modulated by X_i. Two channels $\mathcal{C}_i$ and $\mathcal{C}_j$ are separable if they have no units in common. To see how this works in practice, consider the simplified version of McClelland and Rumelhart's (1981) word reading model shown in Figure 15. Suppose the task is to identify the second letter of a three letter word, and the external response is determined by the states of the letter units in the second location. Thus there are two response specification units, the units labeled "E" and "U" (See Figure 15). The stimulus is X_2, and the context is

composed of two sources: X_1 and X_3. When the response specification units are disconnected from the rest of the network, there are no units, other than the response specification units, modulated by the stimulus X_2. Thus, in this case the stimulus channel is empty, $\mathcal{C}_2 = \emptyset$. When the response specification units are disconnected, the set of units affected by X_1 is as follows

$$\mathcal{C}_1 = \{MEN, MUD, RED, RUN, M, R, D, N\}. \tag{48}$$

Finally, $\mathcal{C}_3 = \mathcal{C}_1$, since all the units affected by X_1 are also affected by X_3. We can conclude that: (1) The stimulus channel is separable from the context channel as a whole, since $\mathcal{C}_2 \cap (\mathcal{C}_1 \cup \mathcal{C}_3) = \emptyset$; and (2) The two context channels are not separable, since $\mathcal{C}_1 \cap \mathcal{C}_3 \neq \emptyset$. Note how our definition of information channel depends on the choice of response specification units. If the external responses were based on the letter units in the left-most position then the units assigned to each information channel would be different, i.e., we would get $\mathcal{C}_1 = \emptyset$, and $C_2 = C_3 = \{MEN, MUD, RED, RUN, E, U, D, N\}$. We now state a theorem relating channel separability and response factorability in diffusion networks.

Implementation Theorem: *Consider a diffusion network as defined in* (38) *through* (43) *with symmetric weights and external input* $X = (X_1, \cdots, X_m)'$. *Let* $\mathcal{C}_j$ *be the channel for* X_j, $j = 1, \cdots, m$ *as defined on page 59. If* $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ *for* $j = 1, \cdots m$, $j \neq i$ *then the equilibrium density of responses factorizes into a term controlled by* X_i *and a term jointly controlled by all the other inputs.*

The theorem is proved on page 89 of the Appendix. In words, the theorem tells us that if we can find a source X_i with a channel separable from all the other sources, then response probability ratios factorize into a term controlled by source X_i and a term jointly controlled by all the other sources. While the details of the proof are technical, the reason why separable networks factorize is easy to understand. Call X_i the stimulus, and all the other input variables the context. Separability of the stimulus and context channels means that there are no direct connections between the stimulus and context units and thus that there are no terms in the goodness function (46) that depend jointly on the stimulus and context. Because of this, the goodness can be separated into two additive terms, one that depends on the stimulus and one that depends on the context. Since probability densities are proportional to the exponential of the goodness (47), if follows that if the stimulus and context channels are separable then the probability density of the response specification units factorizes into a term selectively influenced by the stimulus

and a term selectively influenced by the context (see the Appendix). We now study how the implementation theorem can be used to analyze experiments with continuous and discrete responses.

7.4 Continuous Responses

In some cases it is of interest to model continuous responses rather than discrete categorical judgments. This situation might arise, for example, in choosing a number or the position of a dial that represents the perceived size of an object. Such cases can be modeled with diffusion networks using one to one mappings between the states of a response unit and the external responses. The implementation theorem tells us that if the weights are symmetric and the stimulus and context channels are separable then response probability densities factorize into terms selectively controlled by the stimulus and by the context as a whole.

7.5 Discrete Responses

McClelland (1991) showed that for networks of units that take on only discrete binary states, channel separability is sufficient to exhibit the Morton–Massaro law. The situation is more complex with diffusion networks since we need to map continuous states into discrete responses. To do so we associate overt responses with entire regions of the response specification space, rather than with individual points in that region. To get the probability of a particular response we need to integrate the response probability density over the region associated with that response.

The problem is that integrals of probability densities do not necessarily factorize even though the densities factorize at every point. Figure 16 illustrates a case in which probability densities factorize yet, careful choice of response regions results in a violation of the Morton–Massaro law. The figure shows the response distributions of a simple diffusion network to 4 different combinations of stimulus and context. The figure also shows a particular mapping from continuous activations into two discrete responses. This mapping produces a violation of factorability of the discrete response probabilities even though the underlying distribution of continuous responses factorizes. Note how when Context = 1, the effect of the stimulus is to increase the probability of Response 2. However when Context = 0, the stimulus decreases

the probability of Response 2. Since there are only two response alternatives this cross-over interaction indicates a clear violation of factorability.

Fortunately there are two important cases for which factorability holds, at least as a good approximation. The first case is when the response regions are relatively small and thus we can approximate the integral over that region by the density at a point times the volume of the region. In such a case the ratio of the integrals can be approximated by the ratio of the probability densities of those two points times the ratio of the two volumes. Since probability densities of individual states factorize, probability ratios of small regions should approximately factorize, with the approximation getting arbitrarily better as we make the response regions smaller.

The second case applies to models, like McClelland and Rumelhart's (1981) word reading model, in which each response is associated with a distinct response unit. These models typically have negative connections amongst the response units so that at equilibrium one unit tends to be active while the others are inactive. In such a case a common response policy chooses the response corresponding to the most active unit. It turns out that when the response units inhibit each other, this response policy can be closely approximated by another policy, the *active-wins* policy, which is easier to analyze mathematically (see Figure 17). In the active-wins approach a response is chosen when a unit is active beyond a given threshold and all the other units are below threshold. A similar policy has been used by Grainger and Jacobs in several models of visual word identification (e.g. GraingerJacobs94GraingerJacobs96). In the Appendix we show that if the network has separable channels such a policy can approximate the Morton–Massaro law to an arbitrary level of precision as the gain parameter of the response units is increased (see Figure 10). In practice we have observed that excellent approximations to the Morton–Massaro law can be achieved with small values of the gain parameters (see details of the simulation displayed in Figure 19).

7.6 Summary and Discussion

- In discrete networks with binary states, symmetric weights, separable channels and an active-wins response rule, the response probabilities factorize (McClelland, 1991).

- In diffusion networks with symmetric weights and separable channels,

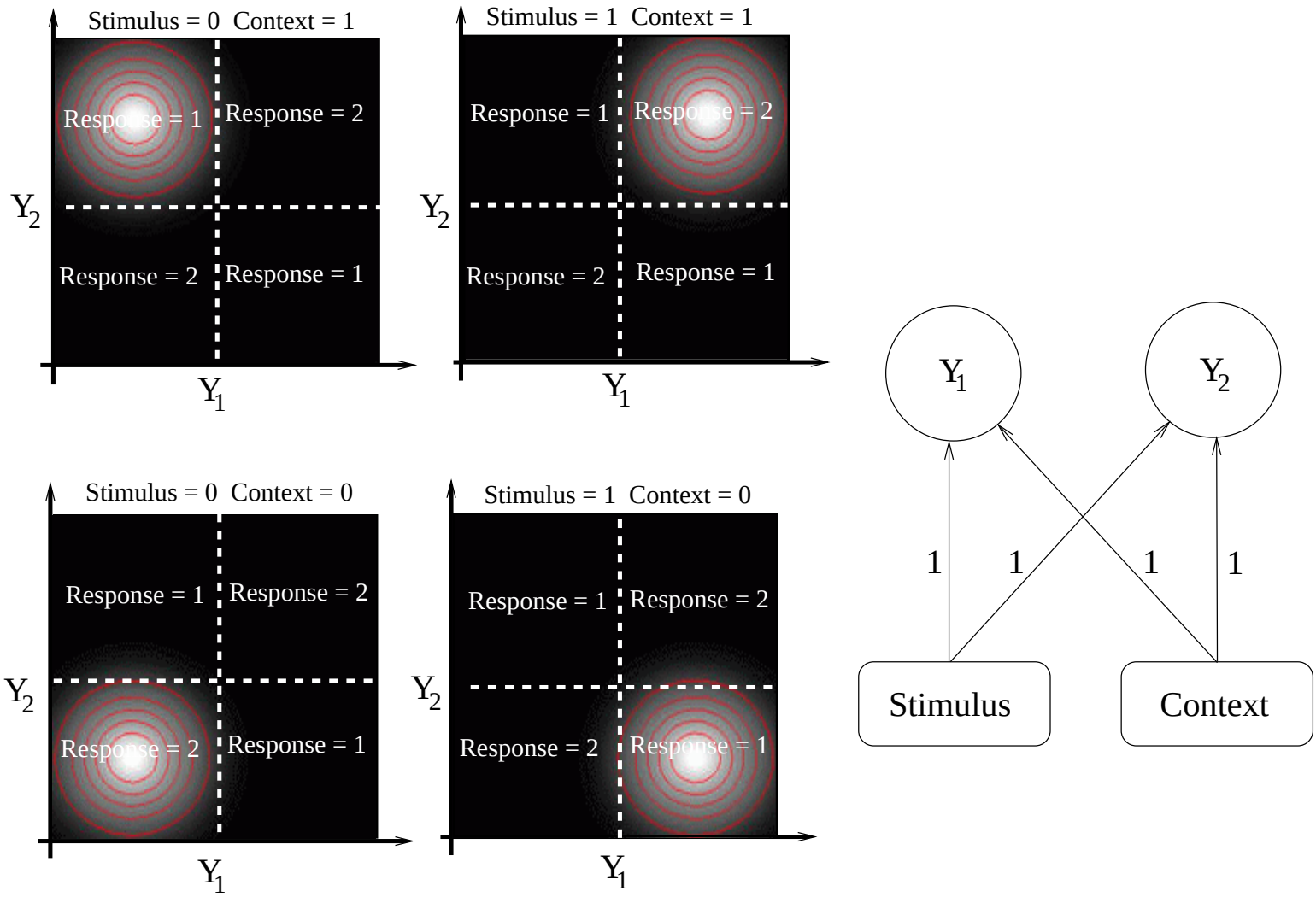


Figure 16: The four figures show the distribution of activation states in response to 4 different combinations of stimulus and context. The distribution of continuous activation states factorizes. However, for the particular mapping between continuous states and discrete responses used in this figure the distribution of discrete responses does not factorize.

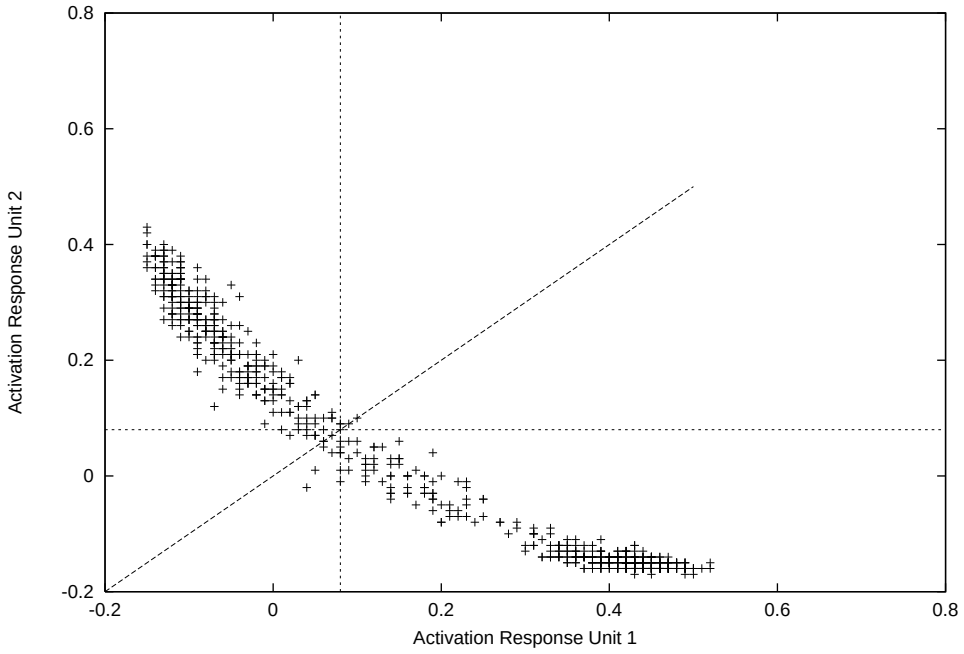


Figure 17: This figure shows the equilibrium distribution of two response specification units. Due to inhibitory connections between these units, the activations are negatively correlated. The diagonal line separates the response regions produced by a response policy that chooses the most active unit. These regions can be closely approximated using two thresholds (the horizontal and vertical lines) and an active-wins response rule.

response probability densities factorize.

- Diffusion networks with symmetric weights, separable channels, and an active-wins response rule, can arbitrarily approximate factorability by increasing the sharpness of the response units.

The main lesson of the analysis is that feed-back is not incompatible with the Morton–Massaro law. Reasons in favor or against interactive models must be found outside of this law. Even though stimulus and context units are interdependent in diffusion networks with feed-back connections, it is still possible to factorize their effect on response probabilities. The analysis presented in the Appendix takes us a long way towards understanding mathematically why this is the case.

The implementation theorem can be applied to a wide range of parallel-distributed processing models. For example, consider the schematic structure of the word perception model shown in Figure 15. As we saw earlier, the stimulus channel $\mathcal{C}_2$ is separable from the context channel as a whole, i.e., $\mathcal{C}_2 \cap (\mathcal{C}_1 \cup \mathcal{C}_3) = \emptyset$. Thus in such model response probabilities should factorize into a term controlled by the stimulus and a term controlled by the context as a whole. However, note that the channel for each context letter is not separable from the other two channels thus it is possible for this model to produce a conjunctive context effect. Moreover, the same property holds when the response units are in the first or third position, not just the second, so that responses based on activations of the letter units in any of the positions should factorize into a term controlled by the stimulus and another term controlled by the context as a whole.

To test this prediction we implemented a diffusion network version of the word perception model shown in Figure 15. There were 6 stimulus conditions, corresponding to different levels of external input to the unit "E" in the second position. There were 4 context conditions: M_N, R_D, M_D, R_N. For each of these 24 conditions 40,000 trials were simulated. On each trial stimulus and context were presented and after a settling period, the activations of the "E" and "U" units were collected. The largest activation determined the external response. The weight parameters were set by hand to roughly approximate the results we obtained in our experiment on the conjunctive context effect, i.e. when the percent correct identification of letters in the context of non-words was about 75 % the correct identification of letters in the context of words was about 80%. The specific parameters

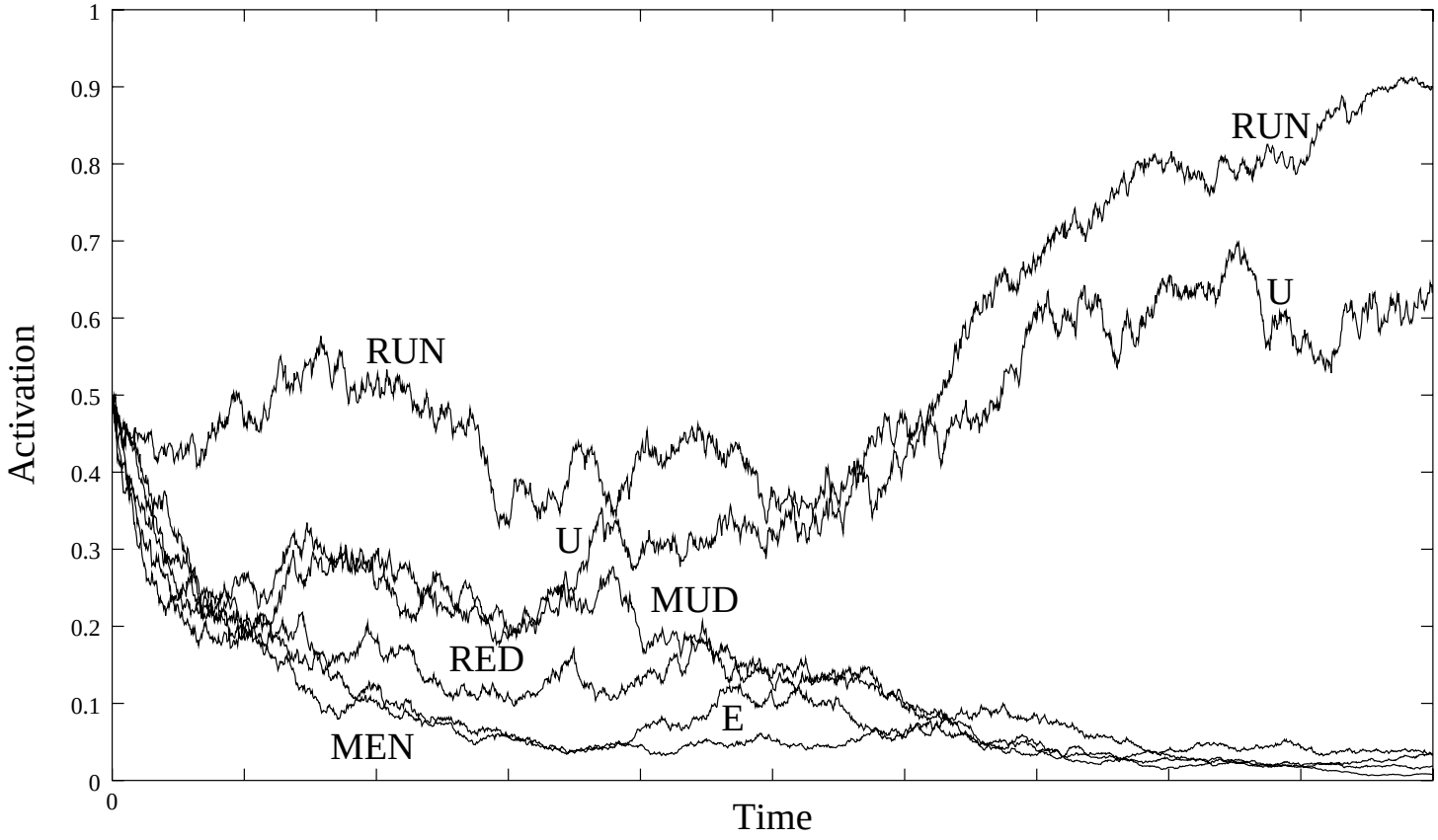


Figure 18: A simulation trial of the diffusion network displayed in Figure 15. In this trial an ambiguous stimulus in the second position was presented in the context of "R_N".

used in this simulation are presented in the Appendix (page 93). Figure 18 shows a single trial in which the network was presented with a stimulus ambiguous between "E" and "U" in the context of "R_N". Figure 19 shows, the simulation results over the entire ensemble of trials. The ordinate shows the probability of response "E" using a logit scale, i.e., $log(p/(1-p))$. As shown in a previous section, factorized effects on probability ratios are additive on the logit transform of these probabilities. As predicted the stimulus and context as a whole factorized. This can be seen in Figure 19 by the fact that all the lines are parallel. Moreover, the two context letters were processed in a conjunctive manner. This can be seen by the fact that the line corresponding to the context "R_N" is above the line corresponding to the context "R_D", while the line for the context "M_N" is below the line for the context "M_D". Thus, in the model, the support of the letter presented in the third position for the letter in the second position depends on the letter presented in the first position.

As we previously noted, interactive models are not the only ones that can produce this pattern of results. For example the FLMP can describe these results by assigning separate parameters to the stimulus and to all the possible context combinations. However such a model would not provide a mechanism to explain how the two letters that make up the context are combined. Models consistent with the conjunctive context effect, such as Rumelhart and Siple's (1974) model, Paap et al.'s (1982) activation verification model, and GraingerJacobs94) dual read-out model still need to be evaluated for their consistency with factorability of stimulus and context as a whole. Other approaches certainly cannot be ruled out but our work shows that a diffusion version of the interactive activation architecture nicely solves the problem of: (1) satisfying factorability of stimulus and context as a whole while (2) exhibiting conjunctive context effects, and (3) providing a mechanism explaining how such contextual conjunctions may emerge.

The optimality theorem assumes symmetric weights. The non-symmetric case is difficult to analyze mathematically but fortunately results can be found in some special cases. For example, it is possible to show that in diffusion networks with separable channels, if the noise is low, or if the activation functions φ_i are linear, then response probability densities factorize regardless of weight symmetry (Movellan & McClelland, 1995). These results suggest that weight symmetry is not a key property for factorability to hold. Figure 19 shows simulations in which the feed-forward connections were three times larger or three times smaller than the feed-back connections. The effect

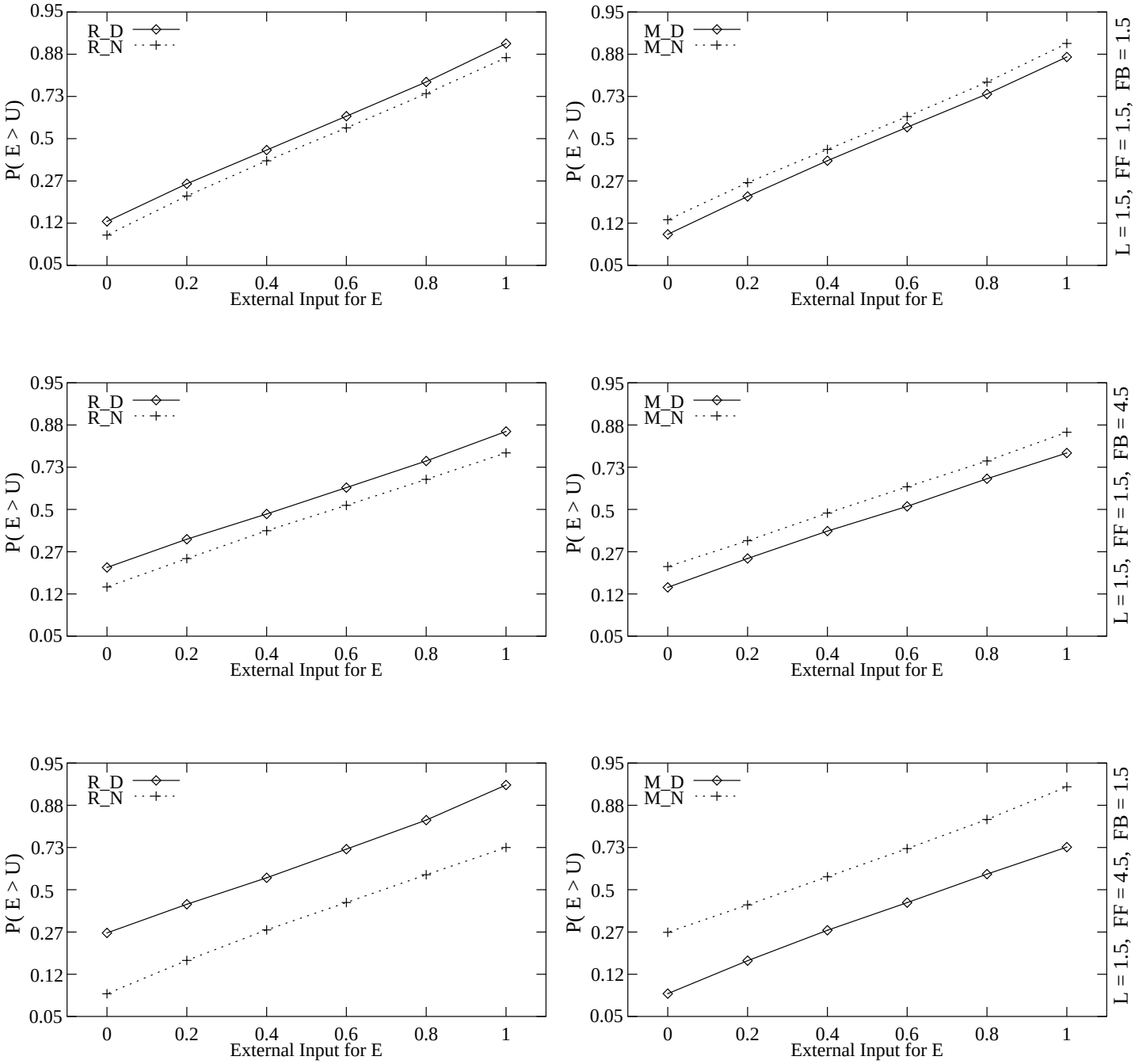


Figure 19: Simulation of the conjunctive context effect with the network displayed in Figure 15. The ordinate shows response probabilities using a logit scale. The first row shows results for a network with lateral weights (L), feed-forward weights (FF) and feedback weights (FB) set to 1.5. The second row shows results with the same network when the FB weights were set to 4.5. The last row shows results when the FF weights were set to 4.5. In all cases the network showed a conjunctive context effect while adhering well to the Morton–Massaro law for the stimulus and context as a whole.

of this weight asymmetry was to change the intercept of the lines but the overall pattern of results was the same regardless of weight asymmetry: the stimulus factorized from the context as a whole but the two context letters combined in a conjunctive manner.

When the activation values are interpreted as spiking rates, diffusion networks are reasonable models of physical neural networks. Thus, the analysis presented here makes novel predictions at the neural systems level. For example, it is now known that the response of neurons in primary visual cortex is modulated by the background stimuli outside their classical receptive fields. These contextual modulations are due to the effect of lateral and feed-back connections (see Figure 20). Suppose we present a neuron in primary visual cortex with combinations of inputs inside its classical receptive field (the stimulus) and outside its classical receptive field (the context). For each combination of stimulus and context we can collect the distribution of spike rates of the neuron under consideration. As Figure 20 makes clear the stimulus channel is separable from the background channel as a whole. Thus, the prediction is made that in this case the Morton–Massaro law should hold, i.e., at equilibrium the spike rate histograms should factorize into terms selectively controlled by the stimulus and the background. The analysis presented here may also be used to infer the functional connectivity of the brain. We may excite two different regions, and observe the activity of a third region. If the observed distribution of spike rates in the third region does not follow the Morton–Massaro law then it is likely that the channels from the two input areas to the response area are not separable.

7.7 Relation to Other Notions of Separability and Independence

The notion of perceptual independence has played a central role in the cognitive psychology literature. Unfortunately this concept has many interpretations leading to a large and confusing proliferation of terms, e.g., orthogonal channels (Green, Weber, & Duncan, 1977), stimulus separability (Shepard, 1964), sampling independence (Townsend & Hu, G. G. And Ashby, 1981), response parity (Garner & Morton, 1969), and uncorrelated pathways among others.

Ashby and Townsend's (1986) General Recognition Theory (GRT) is arguably the most thorough and successful attempt to formalize and unify the

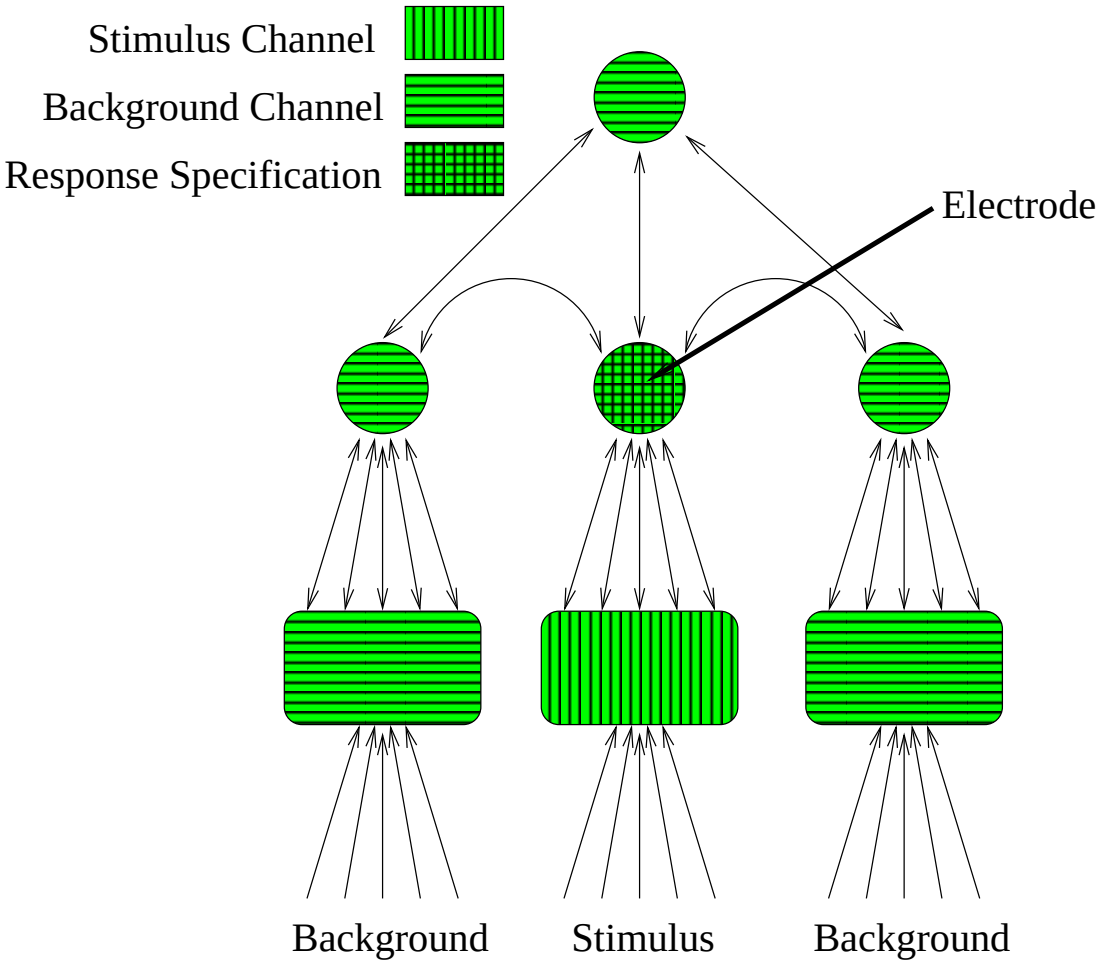


Figure 20: The arrows connecting the stimulus to the unit in the center represent the classical receptive field of that unit. External inputs affecting the classical receptive field are called "stimuli" and all the other inputs are called "background". The center unit determines the responses under consideration (e.g., spiking rates of that neuron). In this preparation the stimulus and background channels are separable.

different notions of independence present in the psychological literature. In this section we develop a framework inspired on an interpretation of GRT from the point of view of information theory. This framework helps clarify that our notion of channel separability in diffusion networks is unrelated to previous notions of independence and separability in the psychological literature. To begin the analysis it is helpful to differentiate four types of variables relevant to the description of perceptual processes:

1. External sources of information: $(X_1, \cdots, X_m)$, the external inputs.

2. Hidden variables: $(H_1, \cdots, H_m)$, pre-perceptual variables internal to the subject that are influenced by the external sources and that have an influence on the perceptual experience (e.g., the state of the ganglion cells in the retina). Here H_i is a vector of hidden variables influenced by X_i.

3. Perceptual variables: $(P_1, \cdots, P_m)$, internal variables that represent the perceptual representations of sources $X_1, \cdots, X_m$.

4. Output variables: $(O_1, \cdots, O_m)$, they represent observable responses to inquiries about the sources $X_1, \cdots, X_m$.

Different models emphasize different variables. For example, in our analysis of diffusion networks we gloss over the difference between internal perceptions and overt responses (i.e., since the focus of the Morton–Massaro law is the relationship between external inputs and responses we treat perceptual variables as relays between inputs and responses). Other models, like the GRT, focus on the relationship between inputs, perceptions and responses without using hidden variables (i.e., pre-perceptual states are ignored).

Most terms found in the perceptual independence literature can be reduced to two different notions of stochastic independence applied at different levels of the perceptual process. We call these two notions *classic independence* and *classic separability*.

Classic perceptual independence (Ashby & Townsend, 1986) is the lack of contingency between the perceptual interpretation of two sources once we partial out the effect of these sources. For example, we would say that X_i and X_j are perceptually independent if

$$I(P_i, P_j \mid X_i, X_j) = 0 \ , \tag{49}$$

where $I(P_i, P_j|X_i, X_j)$ stands for mutual information, between the perceptual representation of X_i and X_j when the effects of these two external sources are partialed out [8]. To get a better understanding of this notion, consider the following simple linear model in which

$$P_i = \sum_j \alpha_{i,j} X_j + \beta_{i,j} N_j, \tag{50}$$

where $N_1, \cdots, N_m$ are noise components independent of the external sources and of each other. The $\alpha_{i,j} \geq 0$ terms represent the effect of input source X_j on the perceptual interpretation of X_i. The $\beta_{i,j} \geq 0$ terms represent the effect of noise source N_j on the perceptual representation of X_i. In this case perceptual independence would be satisfied if there is no cross talk between the noise terms, i.e., $\beta_{i,j} = 0$ when $i \neq j$. Note how it is possible for X_i and X_j to be perceptually independent even though X_j may have an effect on the perception of X_i by virtue of having $\alpha_{i,j} > 0$.

The same mathematical notion of independence can be applied between external inputs and outputs

$$I(O_i, O_j \mid X_i, X_j) = 0 \ , \tag{51}$$

and appears in the literature with names such as "sampling independence" (p. 159, Townsend & Hu, G. G. And Ashby, 1981).

The classic notion of separability relates to a different form of stochastic independence. This notion formalizes the idea that processes are selectively influenced by different sources. Garner and Morton (1969) were first to formalize this notion mathematically as follows

$$I(O_i, \bar{X}_i \mid X_i) = 0, \tag{52}$$

where $\bar{X}_i$ represents all the external sources with the exclusion of X_i. This simply says that once we know the source X_i, then the other sources give no further information about the response O_i. In the simple linear model (50) this notion of separability would be satisfied if there are no cross terms between the external sources, i.e., $\alpha_{i,j} = 0$ when $i \neq j$. Note how X_i and X_k may be perceptually separable even though their perceptual representations may not be independent, by virtue of having $\beta_{i,j} > 0$ for $i \neq j$.

[8] Mutual information is a generalized measure of contingency between random variables. Two random variables are independent if and only if their mutual information is zero.

The same notion of separability can be applied at other levels of the perceptual process. For example when applied between external inputs and perception,

$$I(P_i, \bar{X}_i \mid X_i) = 0 \ , \tag{53}$$

it appears with the name "perceptual separability" (Ashby & Townsend, 1986). When applied between perceptions and responses

$$I(O_i, \bar{P}_i \mid P_i) = 0 \ , \tag{54}$$

it is known as "response separability" (Ashby & Townsend, 1986). Here $\bar{P}_i$ represents the perceptual representations of all external sources, except for X_i. When applied between external inputs and hidden pre-perceptual processes such as the "the output of auditory detectors"

$$I(H_i, \bar{X}_i \mid X_i) = 0 \ , \tag{55}$$

it receives names such as "independent detection" and "orthogonality" (Green et al., 1977).

Now that we have a framework in which to classify notions of independence and separability we will address where our notion of channel separability falls within this framework.

First note that in the experimental paradigm of the Morton–Massaro law we assume that more than one sources of information will have an influence on the external responses. What is at stake in these experiments is whether the external sources influence the external responses in a manner consistent with the Morton–Massaro law. Morton points out that "In a logogen, evidence relevant to a particular response is collected indifferently with respect to the source of information" (Garner & Morton, 1969). Massaro emphasizes that a central aspect of the FLMP is the fact that different sources of information are "independently" evaluated before their influences are integrated. Morton's idea of indifferent evaluation and Massaro's notion of independent evaluation is in fact a case of classic separability which can be formulated as follows

$$I(H_i, \bar{X}_i \mid X_i) = 0, \tag{56}$$

where $H_i = (H_{i,1}, \cdots, H_{i,m})$ is a vector of pre-perceptual evaluation terms. Here $H_{i,j}$ represents the support of source X_i for response alternative j.

This is simply classic separability applied between external sources and pre-perceptual terms controlled by these sources (e.g., the evaluation values in the FLMP). Both the logogen model and the FLMP exhibit this classic notion of separability.

Ashby (in press) recently developed a continuous time version of the GRT which is closely related to our work. The model is defined via a stochastic differential equation similar to the ones used in diffusion networks. The main difference between his work and ours is that we allow feed-back connections and non-linear activation functions, while the networks he studied are linear and do not have feed-back connections. Ashby (in press) showed how the classic notions of separability and independence relate to architectural constraints in feed-forward linear networks. The situation is more complex in our work, due to the presence of non-linearities and feed-back connections. As far as we can tell diffusion networks with separable channels and feed-back connections do not exhibit any of the classic notions of independence and separability presented in this section yet they exhibit the Morton–Massaro law.

8 Summary and Conclusions

First we summarize the main arguments and implications of this paper, in the form of a set of questions and answers[9]:

1. *What is the Morton-Massaro law?* A frequently observed relationship between inputs and responses in tasks that involve the integration of multiple sources of information. A prototypical example is the combination of acoustic and visual sources in speech perception.

2. *What is the behavior that defines the Morton-Massaro law?* Ratios of response probabilities factorize into components selectively influenced by a single information source.

3. *Is the law determined by the environment?* No, it is determined by the behavior of the subject.

4. *Can humans violate the law?* Yes. Many violations have been reported. It is also possible to train subjects to learn category structures that require violations of the Morton-Massaro law.

5. *Why is this relationship called a law?* Because it provides a good description of results in a very wide range of perceptual experiments and because it informs us about about important characteristics of the perceptual process. A good example of this kind of law in the physical sciences is Ohm's law. Knowing whether Ohm's law holds or does not hold in particular circumstances informs us about the characteristics of the material under investigation. Similarly, knowing whether the Morton–Massaro law holds or does not hold in particular experiments informs us about functional and structural characteristics of the perceptual process.

6. *Why is it called the Morton-Massaro law?* Morton was first to point out this behavior in the context of word and letter perception. Massaro and his colleagues have shown that it holds in many experiments in a wide range of domains.

7. *What can cases of adherence (or failure to adhere) to the Morton-Massaro law tell us about structural issues?.* Failures of adherence suggest that the information sources in question influence responses via non-separable channels.

[9]We thank Ken Paap for providing an earlier version of this summary, which we have adapted and expanded, with his permission.

The law tells us nothing about feed-forward vs. interactive mechanisms. Both feedforward and interactive architectures can give rise to performance that adheres or does not adhere to the law.

8. *What does the Morton-Massaro law tells us about functional issues?* If observers are indeed approximately optimal, adherence to the law suggests that the objective class likelihood ratios should factorize. By the same reasoning, if observers do not follow the law, this suggests that the objective class likelihood ratios do not factorize.

9. *Why is performance that adheres to the Morton-Massaro law observed so often?* The behavior is optimal for environments in which objective class likelihood ratios factorize. This condition appears to occur often in nature.

10. *How can we test whether objective class likelihoods ratios factorize?* By building pattern recognition systems that do not assume factorized class likelihoods and comparing them to restricted versions that make the assumption. If the restricted versions perform in generalization tasks as well as the unrestricted versions then the assumption is a reasonable one. This approach is more convincing the larger the training/testing database.

11. *What does this work teach us about the study of cognition?* It suggests that structural questions, which focus on internal processing mechanisms, should be complemented by functional questions, which focus on the statistics of the environment, what is optimal in that environment, and what processing mechanisms achieve optimality in that environment.

Many researchers have emphasized the importance of studying perception and cognition at multiple levels of analysis, and several different taxonomies of levels have been proposed (Marr, 1982; Anderson, 1990Newel82AI). Here we distinguish between descriptive, functional, and structural levels, and within structural levels, we further distinguish between information-processing mechanisms and physical implementations. At a descriptive level, the goal is to find succinct ways of describing patterns of results. The functional level favored by "rational" and Bayesian paradigms (Anderson, 1990; Knill et al., 1996) deals with the optimality of such patterns. Many psychologists adopt a structural approach at the information processing level of analysis for characterizing mechanisms that may account for the obtained patterns of results. At this level of analysis no effort is made to map information processing mechanisms to actual physical processes. Neuroscientists on the other

hand look for the actual underlying physical mechanisms. While it is useful to treat these different levels of analysis somewhat independently it is also important to show how they inform and constrain each other. As cognitive neuroscientists, we tend to seek models that can be viewed as characterizing the mechanisms at an information processing level while at the same time suggesting possible physical implementations that might be realizable in the brain.

At the descriptive level, the Morton–Massaro law is a mathematical expression that describes a relationship between external sources of information and the response probabilities produced by perceptual systems in response to these sources. As we have noted repeatedly this is a law that is not always observed in data, but which very often does provide at least a good approximate account. Such an account is often of great value because of its succinctness (JacobsGrainger94): While the number of parameters required to describe arbitrary response probability matrices grows exponentially with the number of sources, the number of parameters required to describe the Morton–Massaro law grows linearly, thus providing a succinct representation of the response probability matrices typically found in perceptual experiments.

At the functional level we showed that the Morton–Massaro law reflects an "assumption" made by the perceptual system about the statistical structure of the environment: factorization of likelihood ratios given (within) response categories. This assumption is adequate when it provides good approximations to the actual posterior probabilities of the perceptual categories under consideration (e.g., letters). Thus one would expect the law to hold when objective likelihood ratios factorize in the world and to be violated when they do not. We analyzed a database of audiovisual speech signals and showed evidence suggesting that conditional independence of audio and video signals is indeed a reasonable assumption. This helps explain why the Morton–Massaro law has been found to hold so well in perceptual experiments in which humans combine information from acoustic and visual speech sources (Massaro, 1989a). We then presented a case in which factorized objective likelihoods was not a good assumption and observed that in such a case subjects violated the Morton–Massaro law.

At the functional level of analysis the Morton-Massaro law is a relationship of great value to tell us whether perceptual systems behave consistently with the factorability assumption. Ultimately the ubiquity of the Morton-Massaro law may be explained by the fact that the assumption provides good

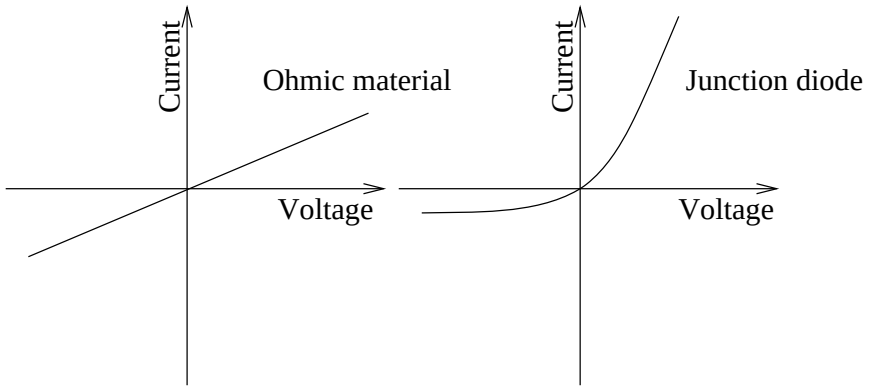


Figure 21: Ohm's law states that the electrical current flowing through a material is proportional to the voltage drop. The law does not reflect a universal principle but a relationship that applies to some devices.

approximations to the statistical structure of the environment in a great number of cases—See Domingos and Pazzani (1997) for a review of pattern recognition problems for which the assumption works very well. However violations should be expected. This kind of law is not unheard of in the physical sciences. For example, Ohm's law states that the relationship between voltage and current is linear. While Ohm's law was initially thought to be a universal principle, nowadays we know that many materials violate it. Ultimately Ohm's law is a criterion to classify materials into ohmic (e.g., conductors) and non-ohmic (e.g., ionized gas and junction diodes). This criterion is treated as a law because of its tremendous practical value. We believe the Morton–Massaro law of information integration holds a status similar to that of Ohm's law in Physics.

At an information processing level, the Morton–Massaro law is also useful, even though it has different implications for our understanding of the underlying processes than some (e.g., Massaro, 1989b) originally thought. It is consistent with a series of models, including factorized strength models like the FLMP and random utility models like Morton's logogen model. Such models can be viewed either as descriptive or as mechanistic; in the latter case, a common characteristic of these models is the fact that signals are selectively evaluated by separate modules each of which processes only one information source (i.e., one of the factors in Massaro's factorial design) and is oblivious to the other sources. These autonomous modules map low-level

inputs into representations that reflect the degree of support of these inputs for high-level perceptual categories.

What we have established here is that the Morton–Massaro law is also consistent with other classes of models, that do not make the modular, feed-forward assumption characteristic of the mechanistic interpretation of FLMP. We presented a class of interactive models called diffusion networks, and we showed that in these models the Morton–Massaro law does not rule out either feed-back or lateral connections. Instead it is related to an architectural constraint that we called channel separability. We can think of channel separability as a way to implement in interactive systems the assumption of factorability of likelihood ratios. The concept of channel separability, an implementation level concept, finds its mirror image in the assumption of factorability, a functional concept. At a functional level, separable channels in an interactive model are similar to distinct, independent modules in feed-forward models, although there is a big difference: in models with separable channels influences between the channels can occur; they are simply constrained to occur in accordance with the Morton–Massaro law.

Our work illustrates some interesting points about the issue of modularity of cognitive architectures. Suppose we are given a black-box with some input sensors (e.g., a video camera and a microphone) and some output devices (e.g., a series of lights that indicate which of n words has been uttered). We test the system using Massaro's experimental paradigm and find that response probabilities factorize into components selectively controlled by audio and video signals. According to the optimality theorem we would be justified to say that this black-box "assumes" a world in which audio and video signals are conditionally independent. At this level of analysis we would also be justified to say that the black-box treats audio and video in a modular manner. However, the work presented here shows that it is possible for the black box to appear modular at this level of analysis (in that it treats sources as independent) while being completely interactive at other levels of analysis. For example, if given access to the internal circuitry we could very well find that all the circuit elements are bimodal, i.e., that they are modulated by audio and visual information. This possibility is in fact consistent with recent findings emerging from the neurosciences: While acoustic and visual speech are perceived and integrated in a manner consistent with the Morton–Massaro law, visual information from lip movements has been shown to modify the activity of the human auditory cortex (Sams91).

In our view an important property of interactive hardware is that it affords a very flexible form of computation in which many different tasks can be solved with the same circuitry. In interactive networks what counts as input and output is not at all fixed. Depending on the task at hand the same unit can act as stimulus, as context, or as a relay for influences in other portions of the system. One can send information through each unit and distribute it throughout the entire net, or one can probe a specific unit to see what the net "thinks" about the specific hypothesis it represents. This allows interactive models to propagate new top-level information into a low-level computation on the fly in a very efficient manner.

In spite of these potential advantages, it is fair to say that at this point the psychological evidence against and in favor of interactive models is inconclusive. It is in fact possible that traditional behavioral experiments based on reaction time and percent correct may never provide enough evidence to distinguish these two hypotheses. Fortunately diffusion networks can be given physical interpretations, as models of electronic circuits or biological neural networks. The fact that these models have physical interpretations may help bridge gaps between the information processing level of interest to psychologists and the physical mechanisms of interest to neuroscientists. In this context, it is perhaps worth taking note that physiological evidence presents a picture of the brain in which bidirectional connections and influences between areas are the norm. It is our experience that while some cognitive psychologists ask "why use interactive models if we can explain our data using feed-forward models", most neuroscientists ask "why do you give us feed-forward models if all we see in the brain is interactive circuitry?" At a minimum this leaves us with the task of trying to understand how psychological laws (e.g., the Morton–Massaro law) may be implemented in the noisy and interactive circuitry used by the brain. Diffusion networks represent a small step in that direction.

Significant work remains to be done and we have identified the following lines of research: (1) It would be of interest to investigate whether the Morton–Massaro law is also found at the neural level. One could in principle probe single neurons or neural pools and test whether their responses to combined stimuli adhere to the Morton–Massaro law. Diffusion networks are reasonable models of biological neural networks and thus our analysis provides a good starting point for how to proceed in this direction; (2) our analysis centered on equilibrium behavior. It would be of interest to investigate whether diffusion networks can also simulate extensions of the FLMP

which describe the time course of information propagation not just equilibrium behavior; (3) more research needs to be done to better understand the computational advantages and disadvantages of interactive architectures; (4) the statistical analysis of audio-visual signals should be extended to other domains and larger databases to explore whether factorized likelihood models provide good approximations to the statistical structure of natural signals as ubiquitously as suggested by the Morton–Massaro law.

Our work illustrates an approach to the study of cognition and perception that emphasizes mathematical trace-ability of statements, simultaneous pursue of functional and structural questions, and exploration of neural basis of cognition. Ultimately the constraints on theory imposed by the convergence of biological, computational, and psychological investigations will provide the strongest guide in the effort to bridge the gaps between the study of the brain, the mind and the computer.

9 Appendix

9.1 Complexity of Factorized Strength Models

Let $X_1, \cdots, X_m$ be the treatment factors (sources of information) and R the subject's response to these treatments. The elements of $\mathcal{R} = \{1, \cdots, r\}$ are the response alternatives, and the elements of $\mathcal{X}_i$ for $i = 1, \cdots, m$ are the treatment levels of factor X_i. For any set $\mathcal{S}$, the term $|\mathcal{S}|$ represents the number of elements in that set. Let $\tilde{\eta}_{X_i}(x_i, k)$ be the parameters of a factorized strength model, i.e., according to this model

$$P(R = k \mid X_1 = x_1, \cdots, X_m = x_m) \;=\; \frac{\prod_{i=1}^{m} \tilde{\eta}_{X_i}(x_i, k)}{\sum_{l=1}^{r} \prod_{i=1}^{m} \tilde{\eta}_{X_i}(x_i, l)}, \qquad (57)$$

for $x_i \in \mathcal{X}_i$ and $k \in \mathcal{R}$. Thus this model appears to have $(r)(|\mathcal{X}_1| + \cdots + |\mathcal{X}_m|)$ parameters. We can re-parameterize the model to make clear that in fact some of these parameters are redundant. In particular let

$$\eta_R(k) = \prod_{i=1}^{m} \Big(\prod_{x_i \in \mathcal{X}_i} \tilde{\eta}_{X_i}(x_i, k) / \tilde{\eta}_{X_i}(x_i, 1) \Big)^{1/|\mathcal{X}_i|}, \text{ for } k \in \mathcal{R}, \qquad (58)$$

$$\eta_{X_i}(x_i, k) = \frac{\tilde{\eta}_{X_i}(x_i, k)}{\tilde{\eta}_{X_i}(x_i, 1)} \Big(\prod_{u \in \mathcal{X}_i} \tilde{\eta}_{X_i}(u, 1) / \tilde{\eta}_{X_i}(u, k) \Big)^{1/|\mathcal{X}_i|} \qquad (59)$$

for $i = 1, \cdots, m$; $x_i \in \mathcal{X}_i$, and $k \in \mathcal{R}$. Note that

$$\frac{\prod_{i=1}^{m} \tilde{\eta}_{X_i}(x_i, k)}{\sum_{l=1}^{r} \prod_{i=1}^{m} \tilde{\eta}_{X_i}(x_i, l)} = \frac{\eta_R(k) \prod_{i=1}^{m} \eta_{X_i}(x_i, k)}{\sum_{l=1}^{r} \eta_R(l) \prod_{i=1}^{m} \eta_{X_i}(x_i, l)}, \qquad (60)$$

for $i = 1, \cdots, m$; $x_i \in \mathcal{X}_i$ and $k \in \mathcal{R}$. Thus the model with parameters η is indistinguishable from the original model. Moreover the η parameters have the following constraints

$$\eta_R(1) = 1, \qquad (61)$$

$$\eta_{X_i}(x_i, 1) = 1, \text{ for } i = 1, \cdots, m;\ x_i \in \mathcal{X}_i, \qquad (62)$$

$$\prod_{x_i \in \mathcal{X}_i} \eta_{X_i}(x_i, k) = 1, \text{ for } i = 1, \cdots, m;\ k \in \mathcal{R}. \qquad (63)$$

Thus in order to reproduce an arbitrary factorized model with parameters $\tilde{\eta}$ we just need the following parameters: $r - 1$ parameters η_R, and $(r - 1)(|\mathcal{X}_i| - 1|)$ parameters η_{X_i} per treatment factor. This gives us a total of $(r - 1)(|\mathcal{X}_1| + \cdots + |\mathcal{X}_m| - m + 1)$ free parameters.

9.2 Optimality Theorem.

Let $(\Omega, \mathcal{F})$, be the measurable space in which our random variables are defined. For simplicity we work with discrete random variables since generalization to continuous random variables is transparent. The random variables $X_1, \cdots, X_m$ represent external sources of information. The random variable A represents a property of these sources which the perceptual systems needs to infer, e.g., the correct response alternative. The random variable R represents the subject's response to $X_1, \cdots, X_m$. The set of values taken by X_i with non-zero probability is symbolized $\mathcal{X}_i$. The set of values taken by R or A with non-zero probability is symbolized $\mathcal{R}$. The task of the subject is to guess A based on $X_1, \cdots, X_m$. We let the joint probability mass function $p_{X_1 \cdots X_m AR}$ factorize as follows:

$$\begin{aligned} &p_{X_1 \cdots X_m AR}(x_1, x_2, a, r) = \\ &p_{X_1 \cdots X_m}(x_1, \cdots, x_m)\, p_{A|X_1 \cdots X_m}(a \mid x_1, \cdots x_m)\, p_{R|X_1 \cdots X_m}(r \mid x_1, \cdots, x_m)\ , \end{aligned} \tag{64}$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ and $a, r \in \mathcal{R}$. This factorization corresponds to the assumption that the subject has no direct access to A and must guess it from the information contained in $(X_1, \cdots, X_m)$. The subject's knowledge is captured by the conditional probability mass function $\tilde{p}_{A|X_1 \cdots X_m}$. From this we define the joint probability mass function $\tilde{p}_{X_1 \cdots X_m AR}$ as follows

$$\begin{aligned} &\tilde{p}_{X_1 \cdots X_m AR}(x_1, \cdots, x_m, a, r) = \\ &p_{X_1 \cdots X_m}(x_1, \cdots, x_m)\, \tilde{p}_{A|X_1 \cdots X_m}(a \mid x_1, \cdots x_m)\, p_{R|X_1 \cdots X_m}(r \mid x_1, \cdots, x_m)\ , \end{aligned} \tag{65}$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ and $a, r \in \mathcal{R}$. Let P and $\tilde{P}$ be measures on $(\Omega, \mathcal{F})$ consistent with $p_{X_1 \cdots X_m AR}$, and $\tilde{p}_{X_1 \cdots X_m AR}$. We refer to them as the objective (P) and subjective $(\tilde{P})$ probability measures[10]. We shall assume that $\tilde{p}_{X_1 \cdots X_m A}(x_1, \cdots, x_m, k) \neq 0$, for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, k \in \mathcal{R}$.

OPTIMALITY THEOREM: *A Bayesian system that satisfies the two following conditions conforms to the Morton-Massaro law:*

1. *It maximizes a weighted sum of subjective discriminant value and entropy*

$$E^P(\log \tilde{p}_{X_1 \cdots X_m A}(X_1, \cdots, X_m, R)) + \beta\, H^P(X_1, \cdots, X_m, R), \tag{66}$$

where E^P, H^P respectively symbolize expected values and entropy with respect to the probability measure P, and $\beta > 0$ is a fixed parameter.

[10]Note that the only subjective term in $\tilde{P}$ is due to $\tilde{p}_{A|X_1 \cdots X_m}$.

2. *It assumes a distribution of information sources which factorizes as follows:*

$$\tilde{p}_{X_1\cdots X_m|A}(x_1,\cdots,x_m \mid a) = \phi_A(a)\phi_X(x_1,\cdots,x_m)\phi_1(x_1,a)\cdots\phi_m(x_m,a), \tag{67}$$

for $(x_1,\cdots,x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$. *Here* ϕ_R *is a term selectively influenced by* A, *not* X, ϕ_X *is selectively influenced by* X, *not* A, *and the terms* ϕ_i *are selectively influenced by the* i^{th} *element of* X *and by* A.

COROLLARY: $p_{R|X_1\cdots X_m}(a \mid x_1,\cdots,x_m)$ *satisfies the Morton–Massaro law for* $(x_1,\cdots,x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$, *if and only if there is a Bayesian system that satisfies conditions (1) and (2) of the Optimality Theorem and whose response probabilities equal* $p_{R|X_1\cdots X_m}(a|x_1,\cdots,x_m)$ *for* $(x_1,\cdots,x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$.

Proof: We seek a probability mass function $p_{R|X_1\cdots X_m}$ that maximizes

$$E^P(\log \tilde{p}_{X_1\cdots X_m A}(X_1,\cdots X_m,R)) + \beta\ H^P(X_1,\cdots,X_m,R)\ . \tag{68}$$

First we decompose (68) into terms controlled by the environment and terms controlled by the perceptual system

$$\begin{aligned}&\beta\ H^P(X_1,\cdots X_m) + \sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_m\in\mathcal{X}_m} p_{X_1\cdots X_m}(x_1,\cdots,x_m) \\ &\Big(E^P(\log \tilde{p}_{X_1\cdots X_m,A}(X_1,\cdots,X_m,R \mid X_1=x_1,\cdots,X_m=x_m)) \\ &+ \beta\ H(R \mid X_1=x_1,\cdots X_m=x_m)\Big)\ ,\end{aligned} \tag{69}$$

where $H^P(X_1,\cdots,X_m)$, the joint entropy of $X_1,\cdots,X_m$, and $p_{X_1\cdots X_m}$ are controlled by the environment. The terms affected by $p_{R|X_1\cdots X_m}$ are: (1) $H^P(R \mid X_1=x_1,\cdots,X_m=x_m)$, the response entropy for fixed values of $X_1\cdots X_m$; and (2) $E^P(\log \tilde{p}_{X_1\cdots X_m A}(X_1,\cdots,X_m,R) \mid X_1=x_1,\cdots,X_m=x_m)$, the subjective discriminant value for fixed values of $X_1,\cdots,X_m$. Thus for each $(x_1,\cdots,x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ an optimal policy should maximize

$$\begin{aligned}&E^P(\log \tilde{p}_{X_1\cdots X_m A}(X_1,\cdots,X_m,R \mid X_1=x_1,\cdots,X_m=x_m)) + \\ &\beta\ H^P(R \mid X_1=x_1,\cdots,X_m=x_m)\ .\end{aligned} \tag{70}$$

Equation (70) is a Helmholtz energy function, with β playing the role of thermal temperature and thus it is uniquely optimized by a Boltzmann distribution (Reif, 1967)

$$p_{R|X_1\cdots X_m}(r \mid x_1, \cdots x_m) = \frac{1}{Z(x_1, \cdots, x_m)} \exp\left(\frac{1}{\beta} \log \tilde{p}_{X_1\cdots X_m A}(x_1, \cdots, x_m, r)\right), \tag{71}$$

where

$$Z(x_1, \cdots, x_m) = \sum_{k\in\mathcal{R}} \exp\left(\frac{1}{\beta} \log \tilde{p}_{X_1\cdots X_m A}(x_1, \cdots, x_m, k)\right), \tag{72}$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, $r \in \mathcal{R}$. The above expression can be written as

$$p_{R|X_1\cdots X_m}(r \mid x_1, \cdots, x_m) = \left(\frac{\tilde{p}_{X_1\cdots X_m A}(x_1, \cdots x_m, r)}{\sum_{k\in\mathcal{R}} \tilde{p}_{X_1\cdots X_m A}(x_1, \cdots, x_m, k)}\right)^{1/\beta}. \tag{73}$$

Using Condition 2 we find that

$$p_{R|X_1\cdots X_m}(r \mid x_1, \cdots, x_m) = \left(\frac{\phi_A(r)\ \phi_1(x_1, r) \cdots\ \phi_m(x_m, r)}{\sum_{k\in\mathcal{R}} \phi_A(k)\ \phi_1(x_1, k) \cdots\ \phi_m(x_m, k)}\right)^{1/\beta}. \tag{74}$$

We now take probability ratios between two response alternatives, and find that they factorize

$$\frac{p_{R|X_1\cdots X_m}(r \mid x_1, \cdots, x_m)}{p_{R|X_1\cdots X_m}(k \mid x_1, \cdots, x_m)} = \left(\frac{\phi_A(r)\ \phi_1(x_1, r)}{\phi_A(k)\ \phi_1(x_1, k)}\right)^{1/\beta} \cdots \left(\frac{\phi_m(x_m, r)}{\phi_m(x_m, k)}\right)^{1/\beta}. \tag{75}$$

The "only if" part of the Corollary is obvious. The "if" part can be proven by construction. Since $p_{R|X_1\cdots X_m}$ follows the Morton-Massaro law, there are $\eta_{X_i}(\cdot,\cdot) > 0$ terms such that

$$p_{R|X_1\cdots X_m}(k \mid x_1, \cdots, x_m) = \frac{\eta_{X_1}(x_1, k) \cdots \eta_{X_m}(x_m, k)}{\sum_{l=1}^{r} \eta_{X_1}(x_1, l) \cdots \eta_{X_m}(x_m, l)}, \tag{76}$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$. Consider now a system with subjective likelihood functions such that

$$\tilde{p}_{A|X_1 \cdots X_m}(x_1 \cdots, x_m \mid k) = \frac{1}{Z(k,\beta)} \Big(\eta_{X_1}(x_1, k) \cdots \eta_{X_m}(x_1, k) \Big)^{\beta}, \qquad (77)$$

for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m, a \in \mathcal{R}$. Here β is a positive constant and Z is a proportionality function needed to make sure probabilities add up to one. Such a system assumes conditional independence, thus satisfying Condition 2 of the Optimality Theorem. Moreover, if we use uniform subjective priors, i.e., $\tilde{p}_A(k) = 1/r$ for $k \in \mathcal{R}$, then from (74) it follows that a system maximizing

$$E^P(\log \tilde{p}_{X_1 \cdots X_m A}(X_1, \cdots, X_m, R)) + \beta \; H^P(X_1, \cdots, X_m, R), \qquad (78)$$

exhibits the desired distribution of external responses.

□

9.3 Condition 2 of the Optimality Theorem

In this section we analyze useful cases under which adherence to Condition 2 of the Optimality Theorem can be proven: *Case 1, conditionally independent sources:* In this case

$$p_{X_1, \cdots, X_n|A}(x_1, \cdots, x_n \mid r) = p_{X_1|A}(x_1 \mid r) \cdots p_{X_n|A}(x_n \mid r), \qquad (79)$$

for all for $(x_1, \cdots, x_m) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, $r \in \mathcal{R}$. Thus, in this case likelihood ratios factorize, satisfying Condition 2 of the Optimality Theorem. An example of this case is shown in Figure 7 A.
Case 2, Gaussian categories with equal covariance within classes: In this case

$$p_{X|A}(x \mid k) = \frac{1}{(2\pi)^{n/2}|\Sigma|} \exp(-\frac{1}{2}(x - \mu_k)' \Sigma^{-1}(x - \mu_k)), \qquad (80)$$

for $x \in \mathbb{R}^n$, $k \in \mathcal{R}$. Here Σ is the within class covariance matrix and μ_k the within class expected value. After a few steps, likelihood ratios can be expressed as follows

$$\frac{p_{X|A}(x \mid k)}{p_{X|A}(x \mid l)} = \exp(x' \Sigma^{-1}(\mu_l - \mu_k)) \, \frac{\exp(-\frac{1}{2}\mu_k' \Sigma^{-1} \mu_k)}{\exp(-\frac{1}{2}\mu_l' \Sigma^{-1} \mu_l)}, \qquad (81)$$

factorizing as prescribed by Condition 2 of the Optimality Theorem. An example of this case is shown in Figure 7 B.
Case 3, monotonic transformations of sources with factorized likelihood ratios:

Let $Z_1, \cdots, Z_n$ be discrete random variables with factorized likelihood ratios. Let $X_i = h_i(Z_i)$ for $i = 1, \cdots, n$, where h_i is a strictly monotonic function. It follows that

$$p_{X|A}(x|k) = p_{Z|A}(h_i^{-1}(x)|k), \tag{82}$$

for $x \in \mathbb{R}^n$, $k \in \mathcal{R}$. Since $p_{Z|A}$ factorizes, so does $p_{X|A}$. A similar argument can be made for continuous random variables. An example of this case is shown in Figure 7 C.

9.4 Implementation theorem

PROPOSITION: *Consider a diffusion network as defined in* (38) *through* (43) *with symmetric weights and external input* $X = (X_1, \cdots, X_m)'$. *Let* $\mathcal{C}_j$ *be the channel for* X_j, $j = 1, \cdots, m$ *as defined on page 59. If* $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ *for* $j = 1, \cdots m$, $j \neq i$ *then the equilibrium density of responses factorizes into a term controlled by* X_i *and a term jointly controlled by all the other inputs.*
Proof: Some of the symbols used in this proof are defined in (38) through (43). Without loss of generality we let $\rho_i = 1$ in (39) for $i = 1, \cdots, n$. Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ be a vector of gain parameters. Let $y \in \mathbb{R}^n$ represent the internal state of a diffusion network and z the activation vector corresponding to state y, i.e., $z_i = \varphi(\alpha_i y_i)$ for $i = 1, \cdots, n$. We assume the activation function is bounded monotonic and differentiable (e.g., the logistic function). This satisfies the conditions for existance of stochastic equilibrium specified in Gidas (1986). Call X_i the stimulus and the rest of the external inputs the context. Let z^s, z^c and z^r represent the components of z for the units in the stimulus channel, context channel and response specification module. Let $x \in \mathbb{R}^m$ be a vector representing an external input and let x^s, x^c be the components of x for the external stimulus and context. Let $Z_\alpha(t)$ a random vector representing the activation at time t of a diffusion network with gain vector α. Let $Z_\alpha = \lim_{t\to\infty} Z_\alpha(t)$, represent the activations at stochastic equilibrium. In Movellan (1998) it is shown that if the weights are symmetric i.e., $w = w'$ then the probability density of Y is as follows

$$p_{Z_\alpha|X}(z \mid x^s, x^c) = \frac{1}{K_\alpha(x_s, x_c)} \exp((2/\sigma^2)\, G_\alpha(z \mid x_s, x_c)) \,, \tag{83}$$

where

$$K_\alpha(x_s, x_c) = \int \exp((2/\sigma^2)\, G_\alpha(z \mid x_s, x_c))\, dz, \tag{84}$$

$$G_\alpha(z \mid x) = H(z \mid x) - \sum_{i=1}^{n} S_{\alpha_i}(z_i), \tag{85}$$

$$H(z \mid x) = z'\ w\ z/2 + z'\ v\ x, \tag{86}$$

$$S_{\alpha_i}(z_i) = \frac{1}{\alpha_i} \int_{\varphi(0)}^{z_i} \varphi^{-1}(u)\, du - \frac{\sigma^2}{2} \log \frac{dz_i}{dy_i} - \frac{\gamma}{2} y_i^2, \tag{87}$$

and φ^{-1} is the inverse function of φ. The H function is sometimes known as the "harmony" and S_i as the "stress". The equilibrium density is derived from the fact that the drift is the gradient of G with respect to the internal states and that the activations are a monotonic function of the internal states (Movellan, 1998).

Without loss of generality hereafter we set $\sigma^2 = 2$. When there are no direct connections between the stimulus and context units there are no terms in the goodness function in which x^s or z^s occur jointly with x^c or z^c. Because of this, the goodness can be separated into three additive terms, that depend on x^s, x^c and a third term which depends on the response units:

$$G_\alpha(z^s, z^c, z^r \mid x^s, x^c) = G^s_\alpha(z^s, z^r \mid x^s) + G^c_\alpha(z^r, z^c \mid x^c) + G^r_\alpha(z^r)\ , \tag{88}$$

where

$$G^s_\alpha(z^s, z^r \mid x^s) = (z^s)' w_{s,s} z^s/2 + (z^s)' w_{s,r} z^r + (z^s)' v_{s,s} x^s + (z^r)' v_{r,s} x^s - \sum_i S_{\alpha_i}(z^s_i)\ , \tag{89}$$

$$G^c_\alpha(z^c, z^r \mid x^s) = (z^c)' w_{c,c} z^c/2 + (z^c)' w_{c,r} z^r + (z^c)' v_{c,c} x^c + (z^r)' v_{r,c} x^c - \sum_i S_{\alpha_i}(z^c_i)\ , \tag{90}$$

$$G^r_\alpha(z^r) = (z^r_i)' w_{r,r} z^r/2 - \sum_i S_{\alpha_i}(z^r_i)\ . \tag{91}$$

Here $w_{s,r}$ is a sub-matrix of w connecting the stimulus and response units. Similar notation is used for the other sub-matrices of w and v. It follows that we can write the ratio of the joint probability density of any two activation

states $z, \tilde{z}$ as follows:

$$\frac{p_{Z_\alpha|X}(z^s, z^c, z^r \mid x^s, x^c)}{p_{Z_\alpha|X}(\tilde{z}^s, \tilde{z}^c, \tilde{z}^r \mid x^s, x^c)} = \frac{\exp(G^s_\alpha(z^s, z^r \mid x_s) + G^c_\alpha(z^c, z^r \mid x^c) + G^r_\alpha(z^r))}{\exp(G^s_\alpha(\tilde{z}^s, \tilde{z}^r \mid x_s) + G^c_\alpha(\tilde{z}^c, \tilde{z}^r \mid x^c) + G^r_\alpha(\tilde{z}^r))}, \tag{92}$$

which factorizes as desired. To get probability densities for the response units, we integrate over the states of all the other units

$$p_{Z^r_\alpha|X}(z^r \mid x^s, x^c) = \int\int p_{Z_\alpha|X}(z^s, z^c, z^r \mid x^s, x^c)\, dz^s\, dz^c, \tag{93}$$

and after rearranging terms

$$p_{Z^r_\alpha|X}(z^r \mid x^s, x^c) = \frac{1}{K_\alpha(x^s, x^c)} \left(\int \exp(\, G^s_\alpha(z^s, z^r \mid x^s) + \, G^r_\alpha(z^r))\, dz^s \right) \left(\int \exp(\, G^c_\alpha(z^c, z^r \mid x^c))\, dz^c \right), \tag{94}$$

which also factorizes.

□

9.5 Application of the Implementation Theorem to Discrete Responses

Here we show that an "active-wins" policy results in external response probability ratios which factorize as we increase the gain parameter of the response units. Besides the previous assumptions on the activation function φ, in this section we further assume that it takes values on $[0, 1]$ and its derivative is an even function (e.g., φ could be the logistic function). Let $z^{(1)} = (1, 0, 0, \cdots, 0)'$, $z^{(2)} = (0, 1, 0, \cdots, 0)'$ be two r-dimensional vectors representing two activation states of the response specification units. For $i \in \{1, 2\}$ and $\Delta \in (0, 0.5)$ let

$$z^{(i)}_\Delta = (1 - z^{(i)})\Delta + (z^{(i)})(1 - \Delta), \tag{95}$$

$$R^{(i)}_\Delta = \{x \in \mathbb{R}^r : x_j \in ((1 - \Delta)z^{(i)}_j, \Delta + (1 - \Delta)z^{(i)}_j), \text{ for } j = 1, \cdots, r\}. \tag{96}$$

Consider an active–wins policy in which we choose external response i when the activation of response unit i is larger than $1-\Delta$ and the activation of all the other response units is smaller than Δ. In this case the sets $R_\Delta^{(1)}$ and $R_\Delta^{(2)}$ are regions of the $[0,1]^r$ space which map into the external responses 1 and 2. We now investigate the convergence of the probability ratio of these two external responses as we let $\Delta \to 0$, i.e., as the response regions collapse into corners of $[0,1]^r$,

$$\lim_{\Delta\to 0} \frac{P(Z_\alpha^r \in R_\Delta^{(2)} \mid X = x)}{P(Z_\alpha^r \in R_\Delta^{(1)} \mid X = x)} = \lim_{\Delta\to 0} \frac{\int_{R_\Delta^{(2)}} p_{Z_\alpha^r|X}(u \mid x)du}{\int_{R_\Delta^{(1)}} p_{Z_\alpha^r|X}(u \mid x)du} = \tag{97}$$

$$\lim_{\Delta\to 0} \frac{\Delta^r p_{Z_\alpha^r|X}(z_\Delta^{(2)} \mid x)}{\Delta^r p_{Z_\alpha^r|X}(z_\Delta^{(1)} \mid x)} = \lim_{\Delta\to 0} \frac{\int\int e^{G_\alpha^r(z_\Delta^{(2)}, z^s, z^c \mid x)} dz^s\, dz^c}{\int\int e^{G_\alpha(z_\Delta^{(1)}, z^s, z^c \mid x)} dz^s\, dz^c}. \tag{98}$$

Now note that

$$G_\alpha(z_\Delta^{(1)}, z^s, z^c \mid x) = H(z_\Delta^{(1)}, z^s, z^c \mid x) - \sum_{i=1}^r S_{\alpha_i}(z_{\Delta,i}^{(1)}) - \sum_i S_{\alpha_i}(z_i^s) - \sum_j S_{\alpha_j}(z_j^c), \tag{99}$$

where $Z_{\Delta,i}^{(1)}$ represents the i^{th} component of the vector $Z_\Delta^{(1)}$. Since φ' is even it can be shown that $S_{\alpha_i}(x) = S_{\alpha_i}(1-x)$ for $x \in (0,1), i = 1, \cdots, r$. Thus

$$\sum_{i=1}^r S_{\alpha_i}(z_{\Delta,i}^{(1)}) = \sum_{i=1}^r S_{\alpha_i}(z_{\Delta,i}^{(2)}), \tag{100}$$

and therefore

$$\lim_{\Delta\to 0} \frac{P(Z_\alpha^r \in R_\Delta^{(2)} \mid X = x)}{P(Z_\alpha^r \in R_\Delta^{(1)} \mid X = x)} = \frac{\int\int e^{H(z_\Delta^{(2)}, z^s, z^c \mid x) - \sum_i S_{\alpha_i}(z_i^s) - \sum_j S_{\alpha_j}(z_j^c)} dz^s\, dz^c}{\int\int e^{H(z_\Delta^{(1)}, z^s, z^c \mid x) - \sum_i S_{\alpha_i}(z_i^s) - \sum_j S_{\alpha_j}(z_j^c)} dz^s\, dz^c}. \tag{101}$$

It is easy to show that this ratio factorizes. Moreover, for $\Delta > 0$ if we let $\alpha_1 = \cdots = \alpha_r = \alpha$, where $\alpha > 0$ then $\lim_{\alpha\to\infty} P(Z_\alpha^r \in [\Delta, 1-\Delta]^r) = 0$, since as the gain of the response units increases (see Figure 10) the distribution of Z_α^r converges to a distribution with probability mass at the corner of the $[0,1]^r$ hypercube and with factorized probability ratios as expressed on (101). This tells us that in the limit, as we increase the gain, the response probabilities of the external responses associated with units 1 and 2 factorize. Since the indexing of the response units was arbitrary this argument holds for all the external responses.

9.6 Simulating Diffusion Networks

We simulated diffusion networks using a forward Euler approach, i.e., stochastic differentials were replaced by discrete time differences of the form

$$Y_i(t + \Delta t) = Y_i(t) + \mu_i(Y_i(t), X_i)\Delta t + \sigma\sqrt{\Delta t}\, N_i(t), \tag{102}$$

for $i = 1, \cdots, n$, where the $N_1(t), \cdots, N_n(t)$ are independent standard Gaussian random variables (i.e., with mean zero and variance 1). As $\Delta t \to 0$, the process defined in (102) converges in distribution to the continuous time process defined by the stochastic differential equation

$$dY_i(t) = \mu_i(Y_i(t), X_i)dt + \sigma dB_i(t)\, N_i(t). \tag{103}$$

The results presented in this paper hold for the underlying continuous time process and are expected to hold in discrete time simulations provided Δt is small. How small Δt should be is an empirical question whose answer may depend on the problem. In our simulations we found that for $\Delta t < 0.01$ the simulations closely matched the expected theoretical results. The network used for the simulation displayed in Figure 13 had a single unit with no external input. The core of the program, looked as follows:

```
sigma=1; alpha=1; gamma= 0.08; dt=0.1; /* Network parameters */
/* Obtain 50 million sample paths */
for(trial=0;trial < 50000000;trial++){
  y=0.0; /* Initial internal state */
  /* We cycle 600 times */
for(cycle=0;cycle<600; cycle++){
  z= 1.0/(1.0+exp(-alpha*y)); /* Activation */
  ybar=0; /* In this simple network the net input is always zero */
  drift= alpha*z*(1.0-z)*(ybar - y) - gamma*y;
  y += drift*dt + sigma*gaussrav()*sqrt(dt);
  }
}
```

The theoretical probability density displayed in Figure 13 was obtained as follows:

```
for(z=0.0025;z<=1.0;z+=0.005){
  y= log(z/(1.0 -z))/alpha;
```

```
    /* The goodness function  */
    g = z*x -  (z* log(z) + (1.0-z)*log(1.0-z))/alpha - gamma*y*y/2.0;
    pdensity = (exp(2.0*g/sigma/sigma))/z/(1.0-z)/0.005;
    pdensity=pdensity/15.488505;  /* The number 15.488505 was computed
                                     using a Riemann sum approximation.
                                     It makes the density integrate to 1.
                                  */
}
```

For the simulation displayed in Figure 19, the network consisted of 10 units: 6 letter units and 4 word units (see Figure 15). The four word units were fully interconnected with inhibitory connections. Letter units had negative connection to and from other letter units within the same position (e.g., "M" and "R" had bidirectional negative connections). Letter units had bidirectional positive connections with consistent word units and bidirectional negative connections with inconsistent word units (e.g., The letter unit "E" had positive connections to and from word unit "RED" and negative connections to and from "MUD"). In the symmetric case the magnitude of all the weights in the network was 1.5. We also tried a case in which the magnitude of the feed–forward letter–to–word weights was 4.5, and a case in which the magnitude of the feed–back word–to–letter weights was 4.5, three times larger than all the other weights. There were 4 context conditions representing: M_N, R_D, M_D, R_N. On each context condition the external input to the appropriate context letters was set to 0.5 and the external input to the other context letters was set to 0. There were 6 stimulus conditions in which the external inputs to the "E" and "U" units were as follows: $\{(0, 0.5), (0.1, 0.4), (0.2, 0, 3), (0.3, 0.2), (0.4, 0.1), (0.5, 0)\}$. Thus there was a total of 24 different input patterns. For each pattern we run 40000 trials. On each trial the internal states Y were initialized to 0 and equation (102) was sequentially applied 1000 times. At that point the activation of the two response units was collected and the response alternative with largest activation value was chosen. The core of the program looked as follows:

```
dt = .02,sigma = 0.2, alpha=1.0, beta=  0.00001, bias= 1.0, nunits=10;
/* Set the 24 input patterns */
for(i=0.0;i<1.1;i+=1.0){
  for(j=0.0;j<1.1;j+=1.0){
    for(k=0.0;k<1.1;k+=0.2){
      x[pat][0] = i*0.5; x[pat][1] = (1.0- i)*0.5;
```

```
      x[pat][2] = j*0.5; x[pat][3] = (1.0- j)*0.5;
      x[pat][4] = k*0.5; x[pat][5] = (1.0- k)*0.5;
      pat++;
    }
  }
}
/* Start the experiment */
for(pat=0;pat<npats;pat++){
  probE[pat]=0.0;
  for(trial=1;trial<=40000;trial++){
    for(i=0;i<nunits;i++) y[i]=0.0; /* Initial states */
    for(cycle=1;cycle<1000;cycle++){
      for(i=0;i<nunits;i++){
      /* The units are ordered as follows: M,R,E,U,D,N,MEN,MUD,RED,RUN */
      /* M is unit 0, E and U are units 2 and 3, the response units    */
        ybar[i] = x[pat][i]-bias/alpha;    /* Add external input and bias
                                              to net input of each unit
                                           */
        z[i] = 1.0/(1.0+exp(-alpha*y[i])); /* Compute activation values */
      }
    }
    /* Add the incoming activations to the net input of each unit */
    for(i=0;i<nunits;i++){
      for(j=0;j<nunits;j++){
        ybar[i]+= z[j]*w[i][j];
      }
    }
    for(i=0;i<nunits;i++){
      adel = alpha*(z[i])*(1-z[i])*dt*( ybar[i] - y[i])
        - beta*y[i] + sigma*sqrt(dt)*gnoise();
                            /* gnoise generates standard
                                       Gaussian samples with
                                       zero mean and unit variance.
                                    */
      y[i]+= adel; /* y is the internal state */
    }
    if(z[2]>z[3]) probE[pat] +=1.0; /* Response determined
                                       by the unit with
```

```
                                        largest activation
                                      */
    }
  }
```

10 References

Allman, J., Miezin, F., & McGuiness, E. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review Neuroscience*, *8*, 407–430.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Earlbaum.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Bülthoff, H. H., & Yuille, A. L. (1996). A Bayesian framework for the integration of visual modules. In T. Inui, & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 49–70). Cambridge, MA: MIT Press.

Clark, J. J., & Yuille, A. L. (1990a). *Data fusion for sensory information processing systems.* Boston: Kluwer Academic Publishers.

Clark, J. J., & Yuille, A. L. (1990b). *Data fusion for sensory processing.* Kluwer Academic.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley & Sons.

Dayan, P., & Sejnowski, T. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*, 5–22.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Journal of Machine Learning*, *29*, 103–130.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: John Wiley.

Efron, A. (1982). *The jacknife, the bootstrap and other resampling plans.* Philadelphia, Pennsylvania: SIAM.

Estes, W. K. (1975). The locus of inferential and perceptual processes in letter identification. *Journal of Experimental Psychology: General*, *104*, 122–145.

Felleman, D. J., & Essen, D. C. V. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.

Garner, W. R., & Morton, J. (1969). Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, *72*, 233–259.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *4*, 1–58.

Ghazanfar, A. A., & Nicolelis, A. L. (1997). Nonlinear processing of tactile information in the thalamocortical loop. *Journal of Neurophysiology*, *78*, 506–510.

Gidas, B. (1986). Metropolis-type monte carlo simulation algorithms and simulated annealing. In J. L. Snell (Ed.), *Topics in contermporary probability and its applications* (pp. 159–232). Boca Raton: CRC Press.

Golden, R. M. (1986). A developmental neural model of visual word recognition. *Cognitive Science*, *10*, 241–276.

Gray, M. S., Movellan, J. R., & Sejnowski, T. (1997). Dynamic features for visual speechreading: A systematic comparison. In Mozer, Jordan, & Petsche (Eds.), *Advances in neural information processing systems*, Vol. 9. MIT Press.

Green, D. M., Weber, D. L., & Duncan, J. E. (1977). Detection and recognition of pure tones in noise. *Journal of the Acoustical Society of America*, *62*(4), 948–954.

Hennecke, M. E., Stork, D. G., & Ventakesh Prasad, K. (1996). Visionary speech: looking ahead to practical speech reading systems. In D. G. Stork, & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 331–349). New York: NATO/Springer-Verlag.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74–95.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science*, *81*, 3088–3092.

Hornik, K., Stinchombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 259–366.

Hubel, D. H., & Weisel, T. N. (1962). Receptive fields, binocular orientation and functional architecture in the cat's visual cortex. *Journal of Physiology*, *166*, 106–154.

Hubel, D. H., & Weisel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Karatzas, I., & Shreve, S. E. (1988). *Brownian motion and stochastic calculus*. New York: Sprienger-Verlag.

Keppel, G. (1993). *Design and analysis: A researcher's handbook*. New Jersey: Prentice Hall.

Kloeden, P. E., & Platen, E. (1992). *Numerical solutions to stochastic differential equations*. Berlin: Springer.

Knill, D. C., Kersten, D., & Yuille, A. L. (1996). Introduction: A bayesian formulation of visual perception. In D. C. Knill, & W. Richards (Eds.), *Perception as Bayesian inference*. Cambridge University Press.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Kuhl, P. K., & Meltzoff, A. M. (1982). The bimodal perception of speech in infancy. *Science*, *218*, 1138–1141.

Lamme, V. A. (1995). The neurophysiology of figure-ground segregation in primary visual cortex. *Journal of Neuroscience*, *15*(2), 1605–1615.

Lee, T., Mumford, D., Romero, R., & Lamme, V. (1998). The role of primary visual cortex in higher level vision. *Vision Research*, 2429–2454.

Lorente de No, R. (1922). La corteza cerebral del raton. (primeral contribucion. la corteza acustica. *Trab. Lab. Invest. Biol. Univ. Madrid*, *20*, 41–78.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D., & Suppes, P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology,vol iii*. New York: Willey.

Luettin, J., Thacker, N., & Beet, S. (1996). Statistical lip modelling for visual speech recognition. *Proceeding of the VIII European Signal Processing Conference*.

Marr, D. (1982). *Vision*. New York: Freeman.

Massaro, D. W. (1987a). Categorical partition: A fuzzy logical model of categorization behavior.

Massaro, D. W. (1987b). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Massaro, D. W. (1987c). *Speech perception by ear and eye: A paradigm for psychological research*. Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1989a). *Perceiving talking faces*. Cambridge, Massachusetts: MIT Press.

Massaro, D. W. (1989b). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*, 398–421.

Massaro, D. W. (1996). Integration of multiiple sources of information in language processing. In T. Inui, & J. L. McClelland (Eds.), *Attention and performance XVI: Information integration in perception and communication* (pp. 397–432). Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1983a). Phonological constraints in speech perception. *Perception and Psychophysics*, *34*, 338–348.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology*, *23*, 558–614.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170–172). Berkeley, CA.

McClelland, J. L. (1991). Stochastic interactive activation and the effect of context on perception. *Cognitive Psychology*, *23*, 1–44.

McClelland, J. L. (1993). Toward a theory of information processing in graded, random, and interactive networks. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience* (pp. 655–688). Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. volume 2: Psychological and biological models.* Cambridge, Massachusetts: MIT Press.

McFadden, D. (1978). Modeling the choice of residential location. In L. Lundqvist, & J. Weibull (Eds.), *Spatial interaction and residential location* (pp. 75–96). Amsterdan: North Holland.

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

Mineiro, P. (1998). *Personal communication.*

Mineiro, P., Movellan, J. R., & Williams, R. J. (1998). Learning path distributions using nonequilibrium diffusion networks. In M. Kearns (Ed.), *Advances in neural information processing systems* (pp. 597–599). Cambridge, Massachusetts: MIT Press.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Movellan, J. R. (1994). A reinforcement algorithm to learn trajectories with stochastic neural networks. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems.* San Mateo: Morgan Kaufman.

Movellan, J. R. (1998). A learning theorem for networks at detailed stochastic equilibrium. *Neural Computation*, *10*(5), 1157–1178.

Movellan, J. R., & Chadderdon, G. (1996). Channel separability in the audio visual integration of speech: A Bayesian approach. In D. G. Stork, & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 473–487). New York: NATO/Springer-Verlag.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, *17*, 463–496.

Movellan, J. R., & McClelland, J. L. (1995). *Stochastic interactive processing, channel separability and optimal perceptual inference: an examination of morton's law* (Technical Report PDP.CNS.95.4, Available at http://hydra.psy.cmu.edu/pub/pdp.cns/). Carnegie Mellon University.

Myers, R. H. (1990). *Classical and modern regression with applications.* Belmont: Duxbury Press.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Oksendal, B. (1992). *Stochastic differential equations*. Berlin: Springer Verlag.

Oliveira, J. T. (1961). La representation des distributions extremales bivariees. *Bulleting of the international statistical institute*, *33*, 477–480.

Paap, K. R., Newsome, S. L., McDonald, J. E., & Schvaneveldt, R. W. (1982). An activation-verification model of letter and word recognition: The word superiority effect. *Psychological Review*, *89*, 573–594.

Peon, R. H. (1961). Reticular mechanisms of sensory control. In W. Rosenblith (Ed.), *Sensory communication*. Cambridge: MIT Press.

Phaf, R. H., van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, *22*, 273–341.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Prasad, K. V., Stork, D. G., & Wolff, G. J. (1993). *Preprocessing video images for neural learning of lipreading* (Technical Report No. TR93-26). Ricoh California Research Center.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Ramon y Cajal, S. (1892). *Nuevo concepto de la histologia de los centros nerviosos*. Barcelona: Imprenta de Henrich y Ca.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Reif, F. (1967). *Statistical physics*. New York: McGraw-Hill.

Repp, B. H., Healy, A. F., & Crowder, R. G. (1983). Exploring the mcgurk effect. *Bulletin of the Psychonomic Society*, 358.

Rumelhart, D. E., & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review*, *81*(2), 99–117.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2 (Chap. 14, pp. 7–57). Cambridge, MA: MIT Press.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Sperry, R. W. (1969). A modified concept of conciousness. *Psychological Review*, *76*, 532–536.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (1988). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Szentagothai, J., & Arbib, M. A. (1974). *Conceptual models of neural organization.* Cambridge: MIT Press.

Townsend, J. T., & Hu, G. G. And Ashby, F. G. (1981). Perceptual sampling of orthogonal straight line features. *Psychological Research*, *43*, 259–275.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* Springer-Verlag.

Yellot, J. I. (1977). The relationship between luce's choice axiom, thustone theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, *15*, 109–144.

Zadeh, L. A. (1988). *Fuzzy logic* (Technical Report CSLI-88-116). Stanford University. Center for the Study of Language and Information.