COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 38 (2014) 1139–1189

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12146

Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition

James L. McClelland,[a] Daniel Mirman,[b] Donald J. Bolger,[c] Pranav Khaitan[d]

[a]*Department of Psychology, Stanford University*
[b]*Department of Psychology, Drexel University and Moss Rehabilitation Research Institute*
[c]*Department of Psychology, University of Maryland*
[d]*Department of Computer Science, Stanford University*

Received 25 September 2011; received in revised form 1 November 2013; accepted 2 November 2013

Abstract

In a seminal 1977 article, Rumelhart argued that perception required the simultaneous use of multiple sources of information, allowing perceivers to optimally interpret sensory information at many levels of representation in real time as information arrives. Building on Rumelhart's arguments, we present the Interactive Activation hypothesis—the idea that the mechanism used in perception and comprehension to achieve these feats exploits an interactive activation process implemented through the bidirectional propagation of activation among simple processing units. We then examine the interactive activation model of letter and word perception and the TRACE model of speech perception, as early attempts to explore this hypothesis, and review the experimental evidence relevant to their assumptions and predictions. We consider how well these models address the computational challenge posed by the problem of perception, and we consider how consistent they are with evidence from behavioral experiments. We examine empirical and theoretical controversies surrounding the idea of interactive processing, including a controversy that swirls around the relationship between interactive computation and optimal Bayesian inference. Some of the implementation details of early versions of interactive activation models caused deviation from optimality and from aspects of human performance data. More recent versions of these models, however, overcome these deficiencies. Among these is a model called the multinomial interactive activation model, which explicitly links interactive activation and Bayesian computations. We also review evidence from neurophysiological and neuroimaging studies supporting the view that interactive processing is a characteristic of the perceptual processing machinery in the brain. In sum, we argue that a computational analysis, as well as behavioral and neuroscience evidence, all support the Interactive Activation hypothesis. The evidence suggests that contemporary

Correspondence should be sent to James L. McClelland, Department of Psychology, 344 Jordan Hall, Bldg 420, 450 Serra Mall, Stanford University, Stanford, CA 94305. E-mail: mcclelland@stanford.edu (or) Daniel Mirman, Department of Psychology, Stratton Hall, 3141 Chestnut St., Drexel University, Philadelphia, PA 19104. E-mail: daniel.mirman@drexel.edu

versions of models based on the idea of interactive activation continue to provide a basis for efforts to achieve a fuller understanding of the process of perception.

Keywords: Perception; Interactive activation; Parallel distributed processing; Connectionist models; Optimal perceptual inference; Neural networks

1. Introduction

One of the foundational concepts in the parallel distributed processing (PDP) framework is interactive activation. Interactive activation is synonymous with the concept of mutual constraint satisfaction: The idea is that, as a general principle, perceptual, linguistic, and other mental representations arise through the bidirectional propagation of activation among simple, neuron-like processing units. The concept was central to the interactive activation (IA) model of letter and word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1981, 1982) and the TRACE model of speech perception (McClelland & Elman, 1986). In these models, the focus was on bidirectional interactions between units standing for wholes and parts, such as words and letters or phonemes; letters and letter features; and phonemes and their features. In these models, individual neuron-like processing units were assigned to represent explicitly enumerable perceptual units such as words, letters, phonemes, and features. The processing units might be viewed as standing for populations of neurons dedicated to the corresponding cognitive units (Bowers, 2009), but we hold a different view. In line with the proposal of Smolensky (1986), the processing units in the model stand for informational states encoded as alternative patterns of activity over populations of neurons each of which participates in the representation of many different items (Hinton, McClelland, & Rumelhart, 1986; Plaut & McClelland, 2010). IA models track the time evolution and content of such states, a useful projection of the full complexity of the underlying neural activity into what Smolensky called a conceptual representational space, where their relationship with overt behavior such as letter, phoneme, or word identification is easier to track.

As detailed below, the empirical motivation for interactive activation models is the observation that, in experiment after experiment, the identification or interpretation of any element or aspect of a visual, auditory, or other input is influenced by the identity and interpretation of every other element or aspect of the input. Correspondingly, there is a motivation at the level of a theory of optimal perceptual interpretation: In general, direct sensory evidence for the interpretation of an input at any level of perceptual description can be inconclusive when considered in isolation, and the most likely interpretation of each element can only be determined when the interpretation of all elements and many sources of evidence are considered together. Indeed, a single coherent interpretation of all elements may well be strongly determined by the totality of the evidence, even though all of the individual elements of evidence are highly ambiguous (Fig. 1a).

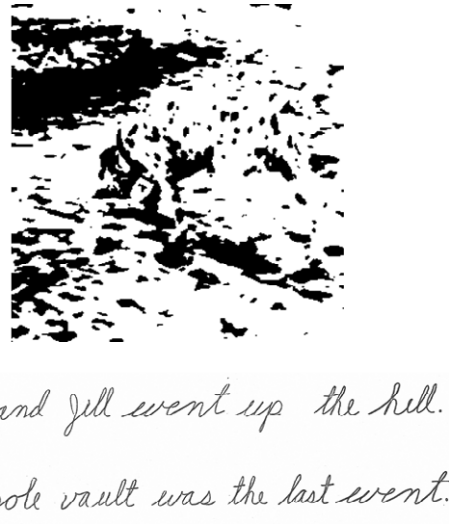


Fig. 1. Top: A Dalmatian dog emerges from an assemblage of individually uninterpretable blotches. From James (1965). Copyright © Ronald C. James, reprinted with permission. Bottom: The hand-written words "went" in the first sentence and "event" in the second are identical, but they are perceived differently in the two different contexts. From fig. 3, p. 579 of Rumelhart (1977). Copyright © Taylor and Francis Group, reprinted with permission.

Beyond the domain of perception and comprehension, multiple simultaneous constraints apply to selection of aspects of contextually appropriate actions and reconstruction of memories, as well as many other aspects of cognition. Likewise, goals and task demands provide additional constraints that are integrated into perception, interpretation, remembering, and action, thus influencing, and often being influenced by, the outcome of processing. Chapter 1 of the PDP volumes (McClelland, Rumelhart, & Hinton, 1986) argued that these same considerations arise in all other areas of cognitive processing, including action selection, problem solving, and memory.

The idea that all aspects of perception and cognition involve parallel distributed processing in this way is an alternative to modular approaches to perception and cognition. Interactive processing allows for the possibility that specific neurons or neural populations in particular brain areas may be specialized to represent one or another type of information, so a certain kind of compartmentalization of information remains. In order, however, for all sources of information to simultaneously constrain all others, any outcome in which a particular ensemble of such neurons is active is thought to be the consequence of processing that is distributed across neural populations in multiple brain areas, including neurons that represent information of many different types. Thus, for example, while there can be brain regions dedicated to the representation of visual, semantic, auditory, and articulatory aspects of a visually presented word, the activations of neurons in all of the participating brain regions are taken to be mutually interdependent within the interactive activation/mutual constraint satisfaction framework.

1.1. Precursors to interactive activation models

The motivation for an interactive approach to perception and comprehension was laid out in a paper by Rumelhart (1977). Rumelhart reviewed existing data going back to the 19th century on the role of context in letter, phoneme, and word perception (Fig. 1b), and on the use of a range of sources of information in resolving ambiguities in syntactic and semantic interpretation of both spoken and written words and sentences. He took the goal of perception and comprehension to be to find a joint interpretation of an input at many different levels of representation, through a mutual constraint satisfaction process guided by knowledge of the prior probabilities of alternative hypotheses and of conditional probabilistic relations between these alternatives. Rumelhart went on to envision how a process of settling on such an interpretation might take place. Drawing inspiration from *Hearsay* (Reddy, Erman, Fennell, & Neely, 1973), an early artificial intelligence model of speech perception, he envisioned a data structure called a "message center" or "blackboard," where estimates of the probabilities of possible elements of the interpretation of an input could be "chalked in" for inspection and adjustment by specialized experts, each working in parallel on the contents of the blackboard. For example, for the case of written input, the estimate of the probability that the letter in a particular position in a word might be the letter A might be increased by a lexical expert that used information about a preceding C and a subsequent T along with lexical information that C, followed by A and T, spells the familiar word CAT. The lexical-level CAT hypothesis might be further strengthened if the participant has just viewed a picture containing a drawing of a cat. At the feature, letter, and word levels, the model drew on an earlier model by Rumelhart and Siple (1974) that relied on knowledge of word and letter probabilities and the conditional probabilities of letters given words to account for data on the identification of letters in displays of three-letter sequences.

2. The computational problem addressed by interactive activation models

The arguments laid out by Rumelhart (1977) support the following statement of the computational challenge faced in perception and language comprehension:

> *Search for the most probable interpretation*. Perception and language understanding are the process of seeking the most probable interpretation of a written or spoken input at many different levels of representation. An interpretation, for example, of a written or spoken linguistic expression represents the visual or auditory features present; the letters or speech sounds; the words, phrases, and sentences; and the meaning and syntactic structure of these items. The goal of the process is to find the interpretation that has the highest probability overall.

Exploitation of prior knowledge and context. Because of the ubiquity of ambiguity and noise, maximizing the probability of finding the correct interpretation of any given aspect of the perceptual input depends on exploitation of prior knowledge and information from context, including adjacent elements in the expression itself, prior input, and input from other domains such as accompanying visual information.

Although Rumelhart (1977) did not stress it, we add the following important real-time constraint on a model of perception and comprehension:

Real-time processing constraint. Perception and comprehension must deliver results as quickly as possible, allowing information of all different types to influence interpretation of information of all other types as it becomes available.

Our inclusion of this constraint in the formulation of the problem of perceptual inference differs from typical computational-level formulations (Feldman, Griffiths, & Mrogan, 2009; Marr, 1982), in which only inputs and outcomes are considered, without consideration of the time or processing steps required to compute the outcome. Clearly, though, time is precious, and in a dynamic world, failure to comprehend (and act) quickly can lead to missed opportunity and sometimes, catastrophe. Thus, achieving results as quickly as possible in real time is part of the computational-level challenge facing the perceptual system. Researchers coming from a computational-level starting point have begun to consider the importance of this issue (Norris, 2013; Vul, Goodman, Griffiths, & Tenenbaum, 2014).

2.1. Human perception and comprehension as an approximation to optimal perceptual inference in real time

The above statements characterize the computational problem a system of perception and comprehension must solve. Our next proposition states that human perception and comprehension mechanisms are organized to address these computational considerations:

Humans approximate optimal real-time perceptual inference. Human perceivers approximate the patterns of behavior we would expect from an optimal system of perception and comprehension, exploiting context and prior knowledge to guide perception and comprehension and reflecting the influence of all sources of external input on all aspects of the interpretation as the input becomes available in real time.

There are limits on speed and accuracy that are imposed by the characteristics of neural hardware, affecting the extent to which humans can achieve a close approximation to optimality. We also note that experience is required for optimization, so that speed and accuracy both increase gradually with practice and exposure. The consequences of experience involve learning about the statistical structure of the perceptual world, tuning of perceptual and other cognitive systems to exploit this structure, and allocation of brain

resources (neurons and synapses) to support performance. In the present article, we focus on perception and comprehension by skilled adults perceiving and comprehending spoken and written input from their native language, assuming that experience-dependent optimization has already occurred.

2.2 The interactive activation hypothesis

The statement of the problem and the characterization of human performance given above appear to be widely accepted, but several alternative approaches have been taken to characterizing the mechanisms that allow human perceivers to succeed in exploiting context and prior knowledge effectively. In this article, we consider the following hypothesis:

> *Interactive activation hypothesis.* Implementation of perceptual and other cognitive processes within bidirectionally connected neural networks in the brain provides the mechanism that addresses the key computational challenges facing perceptual systems, and it gives rise to the approximate conformity of human performance to optimal perceptual inference in real time.

In what follows, we discuss the history of research on interactive models in perception. We examine the early IA and TRACE models and the experimental evidence relevant to their fundamental assumptions. We consider how well they address the computational challenges specified above, and we consider how consistent they are with evidence from behavioral experiments. We examine empirical and theoretical controversies surrounding the idea of interactive processing, including a controversy that has swirled around the relationship between interactive computation and optimal Bayesian inference. We also review evidence from neurophysiological and neuroimaging studies of the neural basis of perception. To anticipate our conclusions: Computational analysis as well as behavioral and neuroscience evidence are all consistent with the Interactive Activation hypothesis. Although there have been and will likely remain those who advocate for alternative approaches, the evidence suggests to us that contemporary versions of models based on these ideas have considerable merit. At the end of the article, we revisit this conclusion and consider ways in which interactive approaches may develop in the future.

3. The interactive activation and TRACE models

Testing the IA hypothesis requires the development of explicit models that embody its assumptions, as well as the analysis of these models to understand their properties and to examine the extent of their ability to account for patterns in human behavior. The Interactive Activation model of letter and word perception (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1981, 1982), and its offspring, the TRACE Model of speech perception (McClelland & Elman, 1986), represented initial steps in such a research program, focusing primarily on modeling patterns in data.

The IA model of letter and word perception addresses the perception of letters presented in one of four display locations, either alone or together with neighboring letters in the other locations. Position-specific pools of neuron-like processing units are posited at feature and letter levels, and a word level spans the array of input positions (Fig. 2a). There are bidirectional excitatory connections between mutually consistent units in adjacent levels and bidirectional inhibitory connections among units within each pool. Before presentation of a stimulus, all units' activation values are set to a resting level slightly below 0. External input, once presented, drives feature units, which in turn activate consistent letter units and inhibit inconsistent letter units.[1] Letter units in turn activate

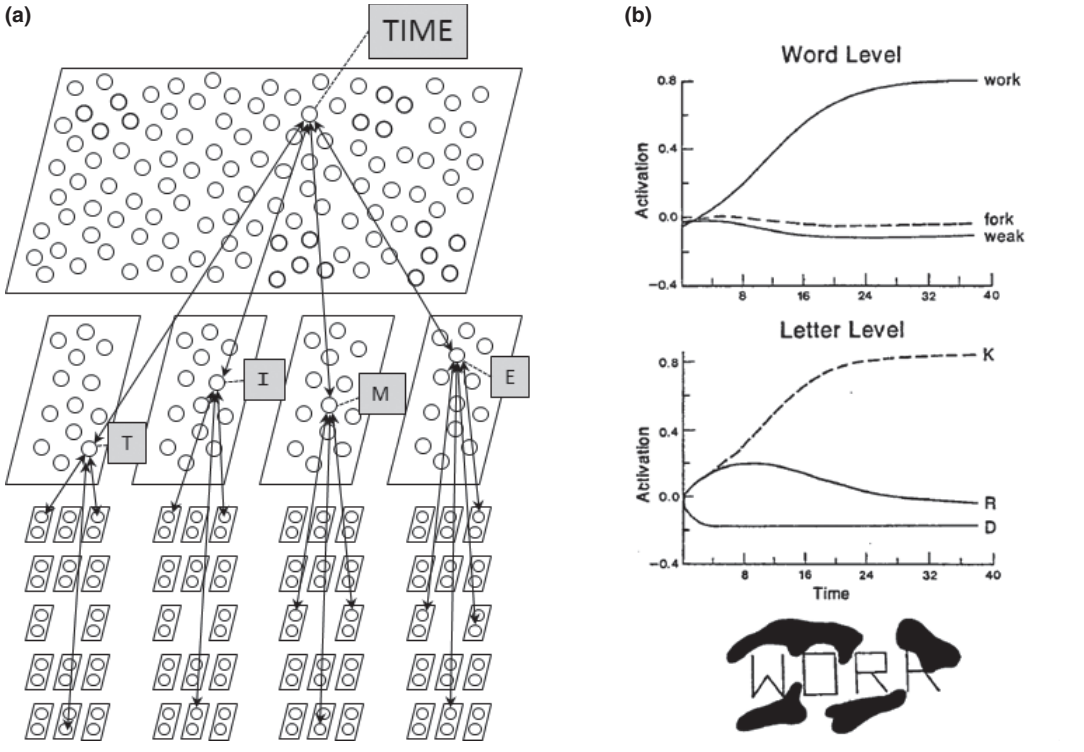


Fig. 2. (a) The interactive activation model, indicating the pools of units corresponding to words, letters in each of four positions, and features in the same positions. Excitatory connections for the word TIME and the letters and features of this word are shown. Units within each pool are mutually inhibitory, though the inhibitory connections are not drawn in. At the feature level, units are organized into pools consisting of two units, one for the presence and one the absence of each possible feature. Reprinted from fig. 6, p. 14 of McClelland (2013). Copyright ©James L. McClelland, reprinted with permission. (b) The time course of activation of letter units in the fourth position and word units in the original version of the interactive activation model, after presentation of the display shown below the figure. The visible segments in the last position are equally consistent with the letters K and R, and inconsistent with other letters. At the word level, only one known word, WORK, is consistent with the active letters in each of the four positions. This word feeds back to support the unit for K, which then dominates R in the fourth position. Reprinted from fig. 8, p. 23 of McClelland et al. (1986). Copyright © MIT Press, reprinted with permission.

consistent word units which compete with each other and also send feedback to support letters consistent with possible words. An illustration of this process, as it applies to the ambiguous input indicated in Fig. 2b, shows the time course of activation, demonstrating how it can find the contextually most likely interpretation within a few processing cycles. Although the featural input in the fourth position is equally consistent with R or K, only K makes a word (WORK) with the context letters. Due to bottom-up support from active letters in all four positions, this word becomes more active than any other word; it suppresses other competing word alternatives and provides top-down support for K, which then suppresses R via competition, leading to a state in which there is a consistent interpretation of the input at both the letter and word levels.

Details of the interactive activation process. We describe the details of the activation process as it was conceived in the IA and TRACE models. These details will be relevant later to our discussion of the relationship between interactive activation and optimal inference. The activation process, as originally formulated (adapting proposals of Grossberg, 1978), assigned continuously varying activation values to units for letters and words. The process is in principle viewed as a completely continuous process, approximated in simulations as a series of fine-grained time steps. During each time step, for each unit, its *net input* is first calculated. This is the sum, over all units projecting to it, of the activation of the sending unit times the value of the incoming connection weight from that unit, plus any direct external input to the unit:

$$net_i = \sum [a_j]^+ w_{ij} + e_i$$

The strengths of excitatory and inhibitory weights were determined by separate parameters for feature-to-letter, letter-to-word, word-to-letter, and within-layer influences. The notation $[a_j]^+$ indicates that a unit's activation value is only propagated if greater than 0.

Once the net input to each unit has been established, activations are adjusted as follows:

$$\begin{aligned} \text{If } (net_i \geq 0) &: \Delta a_i = net_i(1 - a_i) - d(a_i - r) \\ \text{otherwise} &: \Delta a_i = net_i(a_i - m) - d(a_i - r) \end{aligned}$$

These equations implement a process in which a positive net input pushes activation up toward a maximum value of 1, while a negative net input pushes activation down toward a minimum (m), usually set to $-.2$ or $-.3$. The rightmost term in each equation implements a restoring force sometimes thought of as corresponding to a decay or leakage process that tends to pull activation values toward their resting level (r); the parameter d represents the strength of this tendency.

Processing in the model is completely deterministic. To address human performance in perception experiments, where performance is probabilistic, predicted response probabilities are derived by applying the Luce choice rule to a running average of the resulting activation values, so that the probability of choosing alternative i is given by:

$$p(r_i) = e^{g\bar{a}_i} \Big/ \sum_{i'} e^{g\bar{a}_{i'}}$$

For example, the probability of choosing the letter K as the identification response for the letter in the fourth position in the display in Fig. 2a would be calculated by setting i to be the index of the unit corresponding to the letter K in the fourth position. The index i' runs over all the letters in the same position, including the one indexed by i, and g is a scaling parameter. The quantity $\bar{a}_i$ corresponds to the running average activation of the unit for the letter in question at the time when the network is interrogated. For most of the experiments modeled in McClelland and Rumelhart (1981), this time was taken to be the time post-stimulus onset that resulted in highest possible probability of correct responding.

The TRACE model extends the ideas from the IA model to the processing of a stream of speech by postulating a much larger number of position-specific feature and letter unit arrays, as well as corresponding banks of position-aligned word units, so that there is a unit for every feature and phoneme at each position, and a unit for every word starting at every position. As spoken input arrives sequentially in real time, each successive time sample of the spoken input is directed to the next input position. In this way, the same bidirectional activation process as captured in the IA model of letter perception could be applied to the processing of spoken inputs corresponding to one or a few words. The architecture allowed phoneme-level and word-level constraints to be applied to sequences of input samples regardless of where in the input stream these samples occurred. The structure of the TRACE model should not be viewed as a literal claim about the neural mechanism. Instead, it should be seen as a higher-level characterization capturing the relative rather than absolute constraints between phoneme- and word-level information: If there is a /k/ at a particular time, it supports the word "cat" starting in the same time, and the word "ticket" starting two phonemes earlier (among many other possibilities), and these constraints are captured in the connections between units for the corresponding items in the corresponding positions.[2] Activation in this array of units formed a dynamic memory trace of the results of processing a spoken input, hence the name of the model. The architecture was inspired by the earlier concept of the blackboard as discussed by Rumelhart (1977), and a model developed at about the same time (McClelland, 1985, 1986) explored how neural hardware might implement these computations without the reduplication of units and connections.

4. Behavioral evidence

4.1. Empirical foci of the IA and TRACE models

The IA and TRACE models targeted letter and phoneme perception, addressing a large body of relevant data illustrating effects of word context on recognition of letters and speech sounds. Much of the early behavioral evidence can be summarized as explorations

of *word superiority effects*: Letters are recognized more accurately when presented within words than when presented in isolation or in random sequences of letters (e.g., Reicher, 1969). The models also addressed the ubiquitous finding that ambiguous visual and speech inputs are likely to be identified as letters or phonemes consistent with surrounding lexical context (e.g., Ganong, 1980; Massaro, 1979). For example, Ganong (1980) showed that an ambiguous sound between /k/ and /g/ was more likely to be identified as /k/ in an "_iss" context (where it fits to form the word "kiss") and as /g/ in an "_ift" context (where it fits to form the word "gift"). The advantage for letters in words also extended to letters in pronounceable, word-like pseudowords (such as LEAT or TOVE, McClelland & Johnston, 1977). The IA model of word perception provided a novel account of the mechanism by which letters in pseudowords like LEAT were perceived more accurately than letters in unword-like non-words (e.g. LTAE) or single letters presented without context; in the model, the pseudoword advantage occurred through the partial activation of units for words sharing several letters with the presented input. Such items are called *neighbors* of the given input. These word units then fed back support to the units for the constituent letters, many of which are partially supported by activations of several different words (Fig. 3). Newman, Sawusch, and Luce (1997) demonstrated neighborhood effects in identification of ambiguous speech segments, consistent with this account. The IA model predicted that letters in unpronounceable strings that nevertheless had many word "neighbors" (e.g., the "L" in SLNT) would show as much facilitation as letters in comparable pronounceable strings (SLET), and an experiment reported in Rumelhart and McClelland (1982) confirmed this prediction.

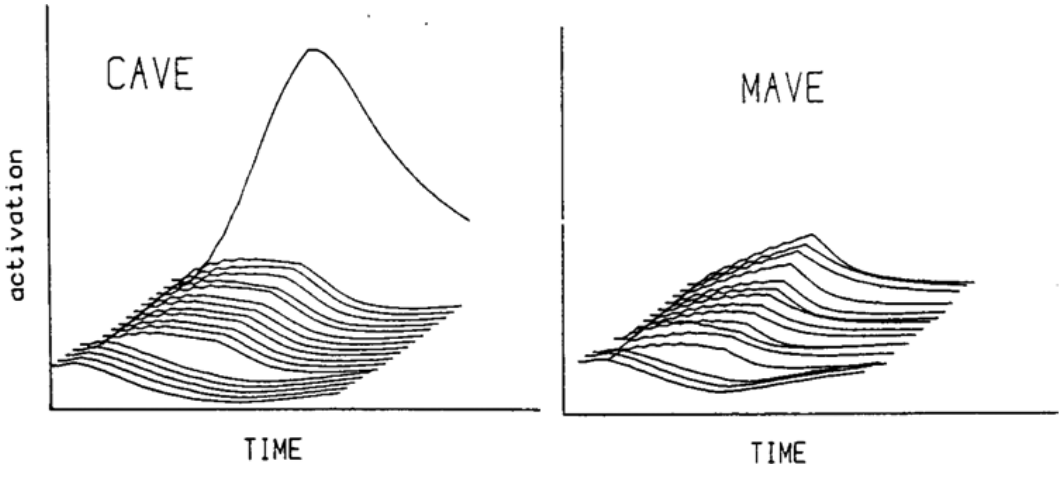


Fig. 3. Activations at the word level produced by CAVE and MAVE in the interactive activation model. Activations of all units whose activation exceeds 0 at any time during processing are shown. Activation traces are offset spatially with those reaching higher maximal activations starting behind and to the right. In the case of MAVE, several words contribute top-down support to the presented letter in each of the four-letter positions. From fig. 13, p. 396 and fig. 9, p. 393 of McClelland and Rumelhart (1981). Copyright © American Psychological Association. Reprinted with permission.

The bidirectional interactive processing in the IA and TRACE models predicts that context effects can occur for contextual elements that come after a target input element, as well as for elements that come before the target. This prediction was tested and confirmed in experiments that separately manipulated the duration of each context letter and examined its effect on the recognition of target letters in each letter position (Rumelhart & McClelland, 1982). In general, *all* context letters influence accuracy of perception of *each* target letter. Similarly, lexical effects on phoneme recognition occur for word-initial as well as embedded or word-final phonemes (Ganong, 1980; Warren, 1970), and the effects extend to contextual information in subsequent words in some studies (Sherman, 1971; Warren & Warren, 1971). Of course, if perceivers in a phoneme identification task are required to respond too soon after an ambiguous segment, subsequent context has little effect (Fox, 1984), and this was captured in simulations using the TRACE model. A wide range of additional phenomena in speech perception, including lexically based segmentation of a stream of spoken sounds into words and the perceptual magnet effect (Kuhl, 1991), were also addressed by the TRACE model.

Evidence of human conformity to the real-time processing constraint. One of the motivating phenomena leading to the development of the TRACE model was evidence supporting the view that word identification occurs in real time during speech perception. Marslen-Wilson and colleagues were the first to focus on this point, showing that identification occurs very shortly after a spoken input becomes uniquely consistent with a single possible word (Marslen-Wilson & Welsh, 1978). A large body of subsequent work examining eye movements during spoken word-to-picture matching tasks further supported the general principle that context and stimulus information mutually constrain processing in real time. Several of these studies include both non-linguistic visual input as well as spoken auditory input, as envisioned in the 1977 paper by Rumelhart. The initial experiments using this method showed that visual context influenced the immediate interpretation of a syntactically ambiguous prepositional phrase (Chambers, Tanenhaus, & Magnuson, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Subsequent studies also showed that syntactic and semantic expectations can also constrain which lexical candidates are considered. For example, Dahan and Tanenhaus (2004) showed that participants rule out possible target objects upon hearing a verb (such as "climb") that rules out some of the objects as potential objects of the action named by the verb (e.g., a watch). Critically, these contextual influences became evident very soon after the constraining information was presented (Dahan & Tanenhaus, 2004; Magnuson, Tanenhaus, & Aslin, 2008) and were continuously updated as new information became available. This was demonstrated particularly clearly by Allopenna, Magnuson, and Tanenhaus (1998) in a study that showed that about 200 ms after word onset—the minimum required to plan and execute an eye movement—listeners were already more likely to fixate objects whose names matched the initial consonant and vowel of the word. Furthermore, their results showed that word candidates that did not match an input's onset could still become activated if supported by enough subsequent phonological input, consistent with the idea of a set of candidates whose activations are continuously updated in light of ongoing input. Many of

these papers simulated their findings using the TRACE model or simplified models based on similar assumptions (Spivey & Tanenhaus, 1998).

4.2. Evidence of the generality of context effects

Word context effects on recognition of letters and phonemes have served as a major focus for research on interactive processing, but the principle is very general and recurs across many different domains of perception and cognition. For example, just as in word recognition, there is a tendency for phonological errors in speech production to result in existing words rather than non-words, and such effects are well explained by interactive models of speech production (Dell, 1986; see also Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Rapp & Goldrick, 2000).

Interactive processing also plays an important role in visual object perception. Just as in the word advantage effects, perception of an ambiguous color can be biased by object context (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Kubat, Mirman, & Roy, 2009). For example, an ambiguous color halfway between yellow and orange is perceived as more yellow in the context of a school bus and as more orange in the context of a carrot. Furthermore, paralleling a result from Elman and McClelland (1988) discussed below, Mitterer and de Ruiter (2008) showed that object-context feedback can recalibrate color categories. The well-known illusory contours phenomenon in Kanizsa figures (Kanizsa, 1979; Fig. 4) demonstrates that a simple figure context can even induce the perception of contours that are completely absent from the input, as expected from interactive activation.

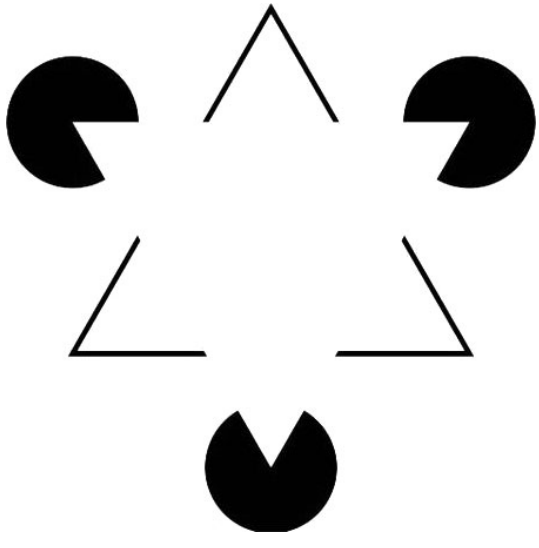


Fig. 4. Illusory contours in the Kanizsa triangle. Image source: Kanizsa triangle, Wikimedia Commons, http://en.wikipedia.org/wiki/File:Kanizsa_triangle.svg. Copyright © Wikipedia Commons. Reprinted under the GNU Free Documentation License.

Moving to higher level phenomena, it has been clear for many years that context affects the resolution of lexical ambiguity, as Rumelhart (1977) predicted. There are models of such effects that restrict context effects to a post-access selection process (Swinney, 1979), but interactive models predict that if the context is sufficiently constraining, then it could constrain which meanings of an ambiguous word are initially activated (e.g., Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982) and even cause pre-activation before the input is presented (McClelland, 1987). Eye-tracking studies have revealed such anticipatory effects in language processing in adults (e.g., Altmann & Kamide, 1999; Magnuson et al., 2008; see also the Dahan & Tanenhaus, 2004 study mentioned above) as well as infants (e.g., McMurray & Aslin, 2004). Electrophysiological scalp recordings (ERP) also suggest that words can be pre-activated by sentence contexts (van den Brink, Brown, & Hagoort, 2001; DeLong, Urbach, & Kutas, 2005).

The study of context effects has focused on how perception of elements such as letters or phonemes (or edges or colors) is affected by their immediate context (e.g., words or objects). However, processing is also affected by other contextual factors, including task instructions and relative probability of different types of stimuli. For example, lexical context effects are reduced when the proportion of non-words in a block of trials in a perceptual experiment is relatively high. Specifically, if the non-word proportion is high, the speed advantage for recognition of phonemes in words compared to non-words is reduced (Mirman, McClelland, Holt, & Magnuson, 2008), the word bias in speech errors is reduced (Hartsuiker, Corley, & Martensen, 2005), the short-term memory advantage for words over non-words is reduced (Jefferies, Frankish, & Ralph, 2006), and there is an increase in regularization errors in reading words that have inconsistent letter-sound mappings (e.g., reading "pint" to rhyme with "mint"; Monsell, Patterson, Graham, Hughes, & Milroy, 1992). These results can be interpreted as reflecting reduced activation of lexical (or possibly semantic) representations so that representations of words are less active and consequently have a smaller feedback effect (for implementations of these effects within TRACE, see Mirman et al., 2008).

The modulation of processing through attention can be implemented in networks of bidirectionally connected processing units—that is, interactive activation networks. One example of such a model is the model of attentional modulation of processing in the Ericksen flanker task (Cohen, Servan-Schreiber, & McClelland, 1992). In this model, units standing for different spatial locations are bidirectionally connected with units for features in these locations, and these units are, in turn, bidirectionally connected with position-independent units of the alternative possible target letter identities. Directing attention to a location is thought to arise from top-down activation of the unit standing for that spatial location; this enhances the activation of units for features in the corresponding position, giving them an eventual upper hand in subsequent processing, but allowing activations from inconsistent flankers nevertheless to retard identification of the item in the target location (as is observed in experiments, for example, Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988). Although some implementations of the Cohen et al. model have simplified the architecture such that not all connections are bidirectional, we take it as a given that attention to locations and stimulus features involves a bidirectional

propagation of activation such that salient inputs, as well as goals and task demands, can participate in determining the focus of attention (Phaf, Van der Heijden, & Hudson, 1990). Furthermore, a recent model of interactive engagement between dorsal (action) and ventral (object) processing systems illustrates how interactive processing can facilitate the simultaneous identification of two or more objects present in a display at the same time (Henderson & McClelland, 2011).

Finally, we note that interactive activation processes may also play an important role in memory (Kumaran & McClelland, 2012; McClelland, 1981). A cue (such as an individual's name) can activate a representation of the item in memory, and this in turn can activate known features of the item, which then, through recurrence, activate other similar items. These items then in their turn can fill in additional features that are then attributed to the cued item. This use of interactivity extends similarity-based generalization models to cases in which relevant items in memory do not overlap with the cue (the individual's name may be unique) but do overlap on other dimensions that are brought into the computation via recurrent, interactive computation.

5. Interactive processing and optimal perceptual inference

While the above indicates some of the empirical support for the IA and TRACE models and demonstrates that the applicability of the principle of interactive activation extends beyond the domain of perception, it does not explicitly address the question of the relationship between the IA model and optimal perceptual inference. The topic has been the source of a heated critique in the literature on visual and auditory context effects (Massaro, 1989; Massaro & Cohen, 1991; Norris & McQueen, 2008; Norris, McQueen, & Cutler, 2000). The papers just cited argue that interactive processing will distort perception away from the pattern that is both seen in behavioral data and expected if information integration is consistent with principles of Bayesian inference, and that interactive activation causes undue contextual influence, producing, for example, inappropriate "hallucination" of lexically consistent phonemes.

It is ironic that the IA hypothesis would face such critiques, given that Rumelhart's early ideas about context effects on perception (Rumelhart, 1977; Rumelhart & Siple, 1974) were explicitly formulated in terms of probabilistic, Bayesian inference. Furthermore, the "hallucinatory" perception of contextually consistent phonemes observed in the models is, for us, exactly what the model *should* produce, both from the point of view of optimal performance in natural contexts and from the point of view of accounting for the findings in human perception. Consider what happens when a brief noise burst occurs coincident with the production of a phoneme in a spoken sentence. Listeners are likely to perceive (perceptually restore) the correct speech sound in such cases, even when the noise replaces the sound rather than being played over it (Samuel, 1981; Warren, 1970). The perception of the phoneme is in some sense a hallucination, but in a natural context, the inference that the speaker has produced the contextually appropriate sound is far more likely to be correct that the inference that he suspended his speech for the exact duration

of the noise burst. In general, exploiting context to determine what we hear is more likely to lead us to hear what was really said, except in experiments where natural probabilistic contingencies can be broken.

It is true, however, that in developing the interactive activation model, Rumelhart and McClelland (1981, 1982) and McClelland and Rumelhart (1981) gave no explicit consideration to a probabilistic formulation of the problem of perception per se, drawing instead on the non-probabilistic, neurally inspired processing models proposed by Grossberg (1978, 1980) without considering whether this formulation corresponded exactly to optimal probabilistic inference. In retrospect, this appears to have led to unfortunate misunderstandings and needless controversies that we hope to put to rest in the present article. Specifically, subsequent research on interactive activation models supports two key points:

1. The IA and TRACE models, in their original formulation, did not provide an exact implementation of a principled Bayesian computation; indeed, the initial formulation of these models did distort these computations, in ways that deviate both from optimality and from human data.
2. However, variants of the models that retain their essential interactive character are consistent with Bayesian principles and can capture data that were problematic for the original formulation.

Regarding point (1), flaws in the original IA and TRACE models are discussed in McClelland (1991). There, it was observed that the activation assumptions of the model together with the assumptions about the translation of these activations into response probabilities produced patterns of choice responses that deviated from Bayesian probabilistic models and from human choice responding. These deviations occurred even in the absence of any interactivity in processing: That is, they occurred even when two sources of bottom-up information were combined to determine the activation of units standing for possible choice alternatives. Thus, the shortcomings of the original model may not have been a consequence of interactivity per se.

Here, we consider point (2) above in more detail. Specifically, we describe how a variant of the IA model called *the multinomial interactive activation* (MIA) model (Khaitan & McClelland, 2010) operates according to Bayesian principles of perceptual inference, considering the case of a display containing a sequence of four letters, as in most of the experiments modeled by the original IA model. A fuller treatment of the probabilistic principles and their relationship with computations in artificial networks is provided in McClelland (2013), and that article should be consulted by those interested in the details behind the briefer presentation here.

The MIA model draws heavily on insights brought into research on artificial neural networks by Hinton and Sejnowski in the form of the *Boltzman Machine*, first presented in a conference proceedings paper describing how such a machine could perform optimal perceptual inference (Hinton & Sejnowski, 1983), and subsequently described in the PDP volumes (Hinton & Sejnowski, 1986). We begin by describing the relevant ideas from the original Boltzmann machine.

5.1. States, their goodness, and their probability in the Boltzmann machine

In Boltzmann machines, units take on binary activation values (0 or 1). Units (which we index with the subscripts i and j) are thought of as corresponding to perceptual predicates about a sensory input (e.g., the input contains a particular line segment at a particular location, or it signals the presence of a particular object at some location). A consequence of using binary activation values is that it makes it relatively easy to consider, not only unit-by-unit probabilities but also the probability of different overall states of the network. Each state S_π corresponds to a specific pattern of [0, 1] values over all of the units, and each state has a *Goodness* G_π, corresponding to how well the state satisfies the graded constraints encoded in the connection weights (w_{ij}) among active units (a_i and a_j) and the bias terms associated with the units (b_i). Weights can be thought of as encoding probabilistic constraints between pairs of predicates, and biases can be thought of as encoding prior probabilities of individual predicates, in ways we will make precise for the MIA model below. The goodness of a state is defined as

$$G_\pi = \sum\nolimits_{i>j} w_{ij} a_i a_j + \sum_i a_i b_i$$

The subscripts i and j run over all units in the network, and the notation $i > j$ simply indicates that the connection between a pair of units, which is assumed to be symmetric ($w_{ij} = w_{ji}$) is only counted once in measuring goodness. The goodness is greater when the bias terms on active units are more positive and when the weights between active units are more positive.

When performing perceptual inference in a Boltzmann machine, some of the units may be forced or *clamped* into specified 0 or 1 values, corresponding to a sensory input pattern, while the activation values of the remaining units are set by a probabilistic updating process. The resulting states of these unclamped units are thought of as a possible interpretation of the sensory input. In the original Boltzmann machine, this updating process consisted of a sequence of updates, each of which involved selecting an unclamped unit at random. Indexing this unit as unit i, we then set its activation depending on its net input, $net_i = \sum_j a_j w_{ij} + b_i$, where j runs over units with connections to unit i. Once the net input is computed, the units' activation is set to 1 with probability

$$p_i = \frac{1}{1 + e^{-net_i/T}}$$

or to 0 with probability 1-p_i. T is a parameter called *temperature*, determining how strongly the activation is constrained by the unit's net input.

If this process is allowed to iterate for a sufficient number of updates, the probability that the network will be in any particular state S_π is equal to the exponential function of the goodness of the state scaled by the temperature, divided by the sum of corresponding quantities for all possible states (indexed by π'), including state π:

$$p(S_\pi) = \frac{e^{G_\pi/T}}{\sum_{\pi'} e^{G_{\pi'}/T}}$$

Here, a possible state is any state in which all the clamped units have their clamped values; each such state is one of the possible patterns of binary activation values over all of the remaining, unclamped units in the network. Since the sum over all the states in the denominator is the same no matter which state we are considering, we can express this relationship by saying that the probability of a state is proportional to the exponential function of the goodness of the state scaled by the temperature:

$$p(S_\pi) \propto e^{G_\pi/T}.$$

5.2. Generative model of the knowledge embodied in the IA model

The multinominal interactive activation model encodes specific probabilistic constraints in the biases and connections among units in a slight variant of the Boltzmann machine. Our next step is to define the probabilistic knowledge that we will be encoding in the network. We adopt a specific hypothetical formulation of the probabilistic knowledge that might underlie a perceiver's (implicit) beliefs about the process that might produce the arrays of visual input features in a letter perception experiment. This knowledge has the form of a *probabilistic generative model*. The concept of a generative model is a useful tool for characterizing the probabilistic structure of an environment and of the information reaching the sensory surface from the environment, and also as a hypothetical abstract characterization of the knowledge a perceiver uses in performing perceptual inference. Although the phrase was not used to describe it, a simple generative model lies at the heart of signal detection theory (Green & Swets, 1966): According to this theory, perceivers are thought to receive signals selected from either a signal plus noise distribution or a noise alone distribution. The parameters of the model are the probabilities of signal plus noise versus noise alone, and the means and standard deviations of each of the two distributions. Signal detection theory provides a theory of optimal perceptual inference in this situation. The generative model we offer here for letter displays is a bit more elaborate, but similar in spirit. It is very similar to the formulation of the beliefs about the probabilistic structure of letter displays used in the model of Rumelhart and Siple (1974), although these authors did not use this terminology.

According to our generative model, the feature array that reaches a perceiver's eye is generated by first selecting a word w_i at random from the possible words in a target lexicon (here, a set of English words that are all four letters long), with a probability $p(w_i)$ monotonically related to the word's language frequency. Once a word is selected, a sequence of letters is generated probabilistically based on the word. The probability of

generating letter j in position k given that word i was selected is represented $p(l_{jk}|w_i)$. With high probability (assumed to be .9 in our simulations), the letter in each position is the correct letter for the given word, but there is a small probability that one of the other letters of the alphabet may be generated instead (given that the correct letter's probability is .9, the probability of each of the other letters is .1/25, or .004). Letters, in turn, give rise to a specification of presence or absence values for each of a set of possible letter features treated (following Rumelhart & Siple, 1974) as line segments (Fig. 5). For example, the letter T specifies that line segments should be *present* across the top of the corresponding feature array and down the middle of the array, and that other possible line segments that could occur in a feature array should be *absent*. Generation of feature values from letters and/or their registration by the perceptual system is also treated as probabilistic. Specifically, for a given letter position k, the probability of generating value v (which can be *present* or *absent*) for feature dimension f given letter j is represented $p(v_{fk}|l_{jk})$. The probability of generating the correct value of a given feature is relatively high (.9 in our simulations), and the probability of generating the incorrect value is equal to one minus this high value (.1).

Given the generative model above, it is possible to calculate the probability of every possible *path* through the generative model, where a path consists of a choice of one word, a choice of one letter in each position in the word, and a choice of one value (*present* or *absent*) for each feature in each letter position. We use the notation P_π to represent a particular path, using the same subscript π that we used previously for the states of a Boltzmann machine. This usage is appropriate, since patterns of activation in the MIA model will correspond to paths through the generative model.

The probability of a particular path P_π, represented $p(P_\pi)$, is simply the product of the probabilities of each of the individual probabilistic events assumed to underlie the creation of the path according to the generative model:

Fig. 5. The letters A–Z as they are represented in the Rumelhart & Siple font, with the full set of features shown in a single block below the letters. From fig. 2, p. 101 in Rumelhart and Siple (1974). Copyright © American Psychological Association. Reprinted with permission.

$$p(P_\pi) = p(w_i) \prod_k \left(p(l_{jk}|w_i) \prod_f (v_{fk}|l_{jk}) \right).$$

5.3. Perceptual inference under the generative model

The problem of perceptual inference (for our case) is to take a set of specified feature values $\{V\}$ and infer which of the possible paths consistent with this set of feature values gave rise to it. The possible paths are all the paths that have the given set of specified feature values. There is one such path for each combination of one word and one letter in each position (in the model, there are 1,179 possible words, and 26 possible letters per position, for $1{,}179 \times 26^4$, or approximately 540 million such paths). The probability of path π given the specified feature values, represented as $p(P_\pi|\{V\})$, is called the posterior probability of the path. The posterior probability of path P_π is given by

$$p(P_\pi|\{V\}) = p(P_\pi) / \sum_{\pi'} p(P_{\pi'}),$$

where the summation in the denominator runs over all possible paths consistent with the specified feature values $\{V\}$.

In principle, we could calculate the probability of each such path, given the set of observed features, and choose the one that is most likely to have generated the observed features under the generative model. The multinomial IA model does not carry out this explicit calculation. Instead, the model *samples* from the set of possible activation states S_π, corresponding to possible paths through the generative model. While the model does not always sample the most probable state, it has the following property: The more probable a state is under the generative model, the more likely the state is to be sampled. We shall make this statement more precise below.

5.4. The MIA model: Using interactive activation to sample from the posterior distribution of the generative model

We now describe the MIA model and explain how it can sample from the correct posterior probability distribution over alternative possible interpretations of the set of specified feature values produced by the generative process above, where an interpretation corresponds to a path, specifying one word and one letter in each position.

As in the original IA model, the model (shown in Fig. 2) contains a unit for each possible word; a unit for each possible letter in each of four positions; and a unit for each possible value (*present* or *absent*) of each feature (e.g., horizontal across the top) of each of the four input feature arrays.[3] Units are organized into pools corresponding to sets of mutually exclusive alternatives. One pool consists of the set of units corresponding to the possible words and four other pools correspond to the sets of units for each of the possi-

ble letters in each of the four-letter positions. There are also four sets of 14 pools of units at the feature level: Each of these pools contains a "present" and an "absent" unit for a specific feature in a specific letter position.

The MIA model replaces the original model's pair-wise inhibitory connections between units in the same pool with the constraint that only one unit in a pool can be active at one time. Under this constraint, each pool now corresponds to a multinomial random variable—a variable that can take one of n alternative values, where n corresponds to the number of units in the pool. This is the feature of the model that gives rise to the word "multinomial" in its name. (Dean, 2005 proposed such a scheme in his computational model of neocortex; see also Lee & Mumford, 2003). Like the mutual inhibition assumption in the original model, the mutual exclusivity assumption in the MIA model is considered to be an idealized, conceptual-level consequence of the local inhibitory circuitry found throughout the brain; it plays a role similar to the role of the mutual inhibition between units in the same pool in the original model. This way of treating inhibition is similar to the divisive normalization model proposed by many modelers (e.g., Grossberg, 1978) and used by neuroscientists to model neural responses in visual cortex (Heeger, 1992).

In the MIA model, the probabilistic information that characterizes the generative model described above is used explicitly to set the bias terms and connection weights of the network. For reasons discussed below, the biases and weights correspond to logarithms of the relevant probabilistic quantities. Specifically, bias weights are assigned to each word unit. The value of the bias weight b_i on the unit for word i is set equal to $\ln(p(w_i))$, that is, the natural logarithm of the probability that word i would be sampled by the generative process described above (in what follows, the word "logarithm" always refers to the natural logarithm). The connection weight between each word unit w_i and each letter unit l_{jk} for letter j in position k is set to $\ln(p(l_{jk}|w_i))$, the logarithm of the probability that the letter would be generated given that word i was the word selected by the generative process. Similarly, the connection weight between the unit for letter j in position k and the feature unit for each of the two possible values of feature f in that position is set to $\ln(p(v_{fk}|l_{jk}))$, the logarithm of the probability that the feature would be generated under the generative model, given that the letter had been generated.

In summary, the MIA model embodies in its connection weights a logarithmic transformation of the probabilistic information in the generative model described above. If the model's knowledge exactly corresponded to the logs of the probabilities in a generative model that actually produced the displays used in a particular experiment, its outputs could be related to the true probabilities of events in the world that generated these inputs. Alternatively, we can think of the model as representing subjective estimates of these probabilities as they are employed by perceivers. In that case, to the extent that there are differences between the knowledge embedded in perceivers' perceptual systems and the true statistics of the world, perception that would be optimal with respect to the estimates might be non-optimal with respect to the statistics of the real world.

For the sake of our present goal of demonstrating that the multinomial IA model can sample from the posterior of the probability distribution defined by the generative model,

we consider a case in which the *present* or the *absent* values of a subset of the features of a presented letter string are specified by an external input. For the example in Fig. 6, none of the features in the first position were specified, whereas the features in the second, third, and fourth positions were the features of the letters O, O, and D, respectively. According to the generative model (bars labeled calculated probability in the figure), the letters that form words with the context (F, G, H, M, and W) are all fairly likely; differences among them mostly reflect differences in the values of $p(w)$ for the associated words (FOOD, GOOD, HOOD, MOOD and WOOD).[4]

5.5. Processing in the MIA model

As in the Boltzmann machine, feature specifications are presented to the model by turning on the unit corresponding to the value of each specified feature. Processing begins with feature units clamped as specified above, and with no units active in any of the letter pools or in the word pool. Processing takes place over a number of cycles, similar to the random updating process in the Boltzmann machine. However, in our case the cycle is

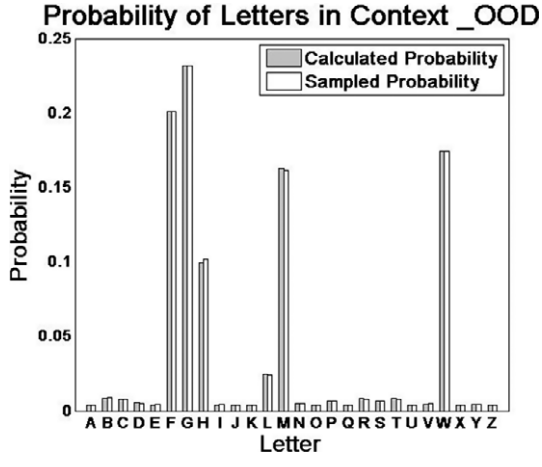


Fig. 6. Comparison of directly computed posterior probabilities and the results of the Gibbs sampling process in the multinomial interactive activation (IA) model, for letters in the first position of a four-letter display. The figure shows the calculated posterior probability of each possible letter in the first position of a four-letter array, following the presentation of a display in which no feature values are specified in the first position followed by full specification of features of the letters O, O, and D in the second, third, and fourth positions, respectively. The gray bars represent the calculated Bayesian posterior probabilities for the first letter position. These probabilities reflect the lexical knowledge embodied in the generative model. For this calculation, $p(l|w)$ was set to .9 for the correct letter in each position of the word, and .1/25 for each of the other possible letters, and $p(f|l)$ was assumed to be .9 for the correct value of each feature of each letter, and .1 for the incorrect value. The white bars represent the sampled probability that each of the letters in the first position was active after 50 iterations of the multinomial IA model. The weights in the model were set to correspond to the logarithms of the probabilities used for the Bayesian calculation, as described in the text. A total of 10,000 simulated trials were run for 100 iterations. Results are mean probabilities averaged over iterations 51–100. Slight differences between sampled and calculated probabilities are within the range of sampling error.

not random (although this detail is not critical for the functioning of the model, it makes discussion of the meaning of the computation somewhat simpler). Within a cycle, activations are first determined for each of the four-letter pools, using the existing activations at the feature and word levels; then activation is determined for the word pool, using the activations in each of the four-letter positions as well as the bias weights associated with each of the word units. Determination of activation in each pool begins by calculating each unit's net input, based on the weights, biases, and activations of other units as usual.

As previously stated, the model differs from the original IA model and indeed the original Boltzmann machine in that, at each time step, only one letter unit in each position and only one word unit may be active; the active unit is chosen probabilistically using the *softmax* function, so that, for each unit within the pool, the probability that a unit is chosen depends on the exponential function of its own net input divided by the corresponding quantities for all the units in the pool (itself included):

$$p(a_i = 1) = \frac{e^{net_i/T}}{\sum_{i'} e^{net_{i'}/T}}.$$

Here, i and i' index the units in the pool being updated and T corresponds to temperature as in the Boltzmann machine. The softmax function can be viewed as an extension of the logistic function used in the Boltzmann machine, where the logistic function sets the activation of a single unit into either the *on* or the *off* state, while the softmax function sets a multinomial random variable into one of its n alternative states, in which exactly one of the units in the pool is active.

Let us now consider the relationship between this computation and sampling from the posterior probabilities of possible letters, given a set of observed features. For specificity, consider the computation of the activation for the pool of units corresponding to the letter in the second position of a four-letter display, on the first cycle of activation, when no units are active at the word level. In this case, the sending units are units corresponding to values of features in the second letter position, the receiving units are the units for possible letters in the second position of the string, and the weights are the connection weights between the letter and feature units, each of which corresponds to the log of the probability of the particular value of the feature (present or absent) given the letter. Noting that the activation of a sending unit is equal to 1 for the unit corresponding to the specified value v of feature f, and that there are no bias terms specified at the letter level in the model, the expression for the net input to letter unit j in position k can be rewritten as

$$net_{jk} = \sum_f \ln(p(v_{fk}|l_{jk})).$$

Now, when we compute $e^{net_{jk}}$ for use in the softmax function to compute the probability of activating the unit, this expression turns into $\prod_f p(v_{fk}|l_{jk})$, the probability that we

would have generated the observed values of the features from the given letter, under the generative model.[5] Plugging these values into the softmax function, we see that it is equivalent to:

$$p(a_{jk} = 1) = \frac{\left(\prod_f p(v_{fk}|l_{jk})\right)^{1/T}}{\sum_{j'}\left(\prod_f p(v_{fk}|l_{j'k})\right)^{1/T}}$$

For the case where $T = 1$, this equation corresponds to Bayes' formula for the posterior probability of letter *j,* given the values of the features (McClelland, 2013).[6] In that case, the softmax function will choose a letter to activate with a probability equal to the posterior probability of the letter given the specified features. If T is unequal to 1, these probabilities will be taken to the $1/T$ power, then renormalized. As stated before, we can express this more compactly as

$$p(a_{jk} = 1) \propto \left(\prod_f p(v_{fk}|l_{jk})\right)^{1/T}$$

5.5.1. The roles of logs and exponentials in linking neural and probabilistic computation

The reader may be tempted to ask at this point why we have bothered with using the logarithms of probabilistic quantities in defining the strengths of the connection weights in the MIA model network, since we then proceed to undo this logarithmic transformation when we exponentiate the net input to a unit for use in the softmax function (see note 5) or the closely related logistic function. Indeed, it would be possible to reformulate the MIA model, directly using the prior probabilities of words and the conditional probabilities of letters given words and of features given letters, and then redefining the activation function accordingly. The reason for using the logs of these probabilistic quantities is based ultimately on the inspiration from neuroscience that continues to lie behind the MIA model and other neural network models, and on the previous history of models linking neurons to computation. The MIA model traces its lineage through a marriage of the original IA model, a descendent of an earlier model of Grossberg (1978), with the Boltzmann machine, a descendent of the earlier model of Hopfield (1982). Ultimately, these models can in turn be traced back through the Perceptron (Rosenblatt, 1958) to the McCullough-Pitts neuron (Pitts & McCullough, 1947), a device that added up weighted signals and compared them to a threshold. The idea that neurons additively combine excitatory and inhibitory signals, and then fire when a threshold is reached, is, or course, the standard intuitive simplification of a model neuron relied on by neuroscientists. In the presence of a source of additive Gaussian noise in the inputs to such a simplified model neuron, the probability of firing will closely match the logistic function of the summed or net input. Thus, the McCullough-Pitts neuron with noise added to its input turns out to be a closely approximate implementation of the logistic neuron used in Boltzmann machines,

which in turn implements Bayes' rule if the weights and bias terms are set to the logs of the appropriate probabilistic quantities, as Hinton and Sejnowski (1983) were the first to point out (see McClelland, 2013, for further discussion of a possible neural basis for the softmax function).

Returning to the main thread of our development, we now consider the net input to each unit at the word level. In this case, the net input consists of the bias term representing the log of the subjective probability of the word, plus the sum of terms corresponding to the product of the activation of each letter level unit, times the weight between the word unit and the letter unit. From the first step in the computation described above, one letter unit in each position has an activation value of 1, while all other letter units' activation values are 0, so the net input to word unit i becomes

$$net_i = \ln(p(w_i)) + \sum_k \ln(p(l_{jk}|w_i))$$

where l_{jk} represents the active letter unit in position k. Now, computing e^{net_i}, we obtain the probability, under the generative model, that the word would be chosen for presentation, times the probability that the active letters would have been generated, given that the word had been chosen. Putting this into the softmax function, we obtain

$$p\,(a_i = 1) \propto \left(p(w_i) \prod\nolimits_k p(l_{jk}|w_i)\right)^{1/T}$$

Expressing this in words, the probability that a given word unit is chosen to be the only one active is proportional to the prior probability of occurrence of the word, times the probability that the word would have generated the set of active letters. Again, this implements the basic logic of Bayes rule for calculating a posterior probability that a particular word was presented, in this case given prior information (represented by $p(w_i)$) and the likelihood of evidence (in this case the active letters) given the word.

Finally, let us consider the activation of a unit j in any one of the letter pools on the next cycle, when there is a single-word unit active at the word level. The net input to each letter level unit is the same as before, but with an extra term corresponding to the log of the probability of the letter, given the active word. Once this expression is exponentiated, it corresponds to the probability of the letter given the active word, times the probability of the set of specified features, given the letters. The expression for the probability that a given letter j will be activated in position k is

$$p(a_{jk} = 1) \propto \left(p(l_{jk}|w_i) \prod\nolimits_f p(v_{fk}|l_{jk})\right)^{1/T}$$

Thus, after the second update of letter level activations, the probability that a given letter unit in each position is chosen to be the active unit in that position is proportional to

the probability of the letter, given the active word, times the probability of the set of features in the given position, given the letter, scaled by 1/*T*.

Note that the weights between word and letter units and between letter and feature units were defined in terms of the top-down, generative process that is treated as underlying the creation of the displays. The letter-to-feature weights are used in computing bottom-up input from feature to letter units and the word-to-letter weights are used in computing the bottom-up input from the letter to the word units. The word-to-letter weights are also used to compute the top-down influences from the word units to the letter units, and, although we do not consider it here, the letter-to-feature weights could be used to fill in missing feature-level activations. Thus, the same connection weight values are used symmetrically, in both directions, even though their values are those specified in the top-down generative model. Because the weights are used symmetrically, the model shares an essential characteristic with the Boltzmann machine: The activation updates tend to move the states of the network in the direction of states of higher overall goodness.

In summary, given the order of processing specified above, and running with $T = 1$, the probability that a given letter unit will be active in a given position will correspond to the probability of the letter given the features under the generative model. When the word level is first updated, a single word will be chosen with a probability proportional to the probability of the word given the chosen letters. Thus, our calculation will produce a sample from the possible states of the underlying generative model that could have produced the observed features. However, our estimates of the probabilities of the letters have not yet taken the word-level information into account. The next update at the letter level does take the word-level information into account, so that, for each letter position, the probability that a letter unit will be active is equal to the probability of the letter, given both the active word and the given array of features.

It might seem that the computation is complete at this point, but the probabilities of letter activations after the second update at the letter level do not exactly match their correct posterior probabilities. However, as the sampling progresses through additional cycles alternating between updates and the word and letter levels, the activation probabilities converge toward the correct posterior probabilities. The sampling procedure is a generalization to the multinomial case of the procedure used in the Boltzmann machine to set activation states. Like the Boltzmann machine sampling procedure, our procedure is an instance of *Gibbs sampling* (Geman & Geman, 1984), a widely used procedure that originated in statistical physics, where it has been shown to provide unbiased samples from the posterior of a probability distribution by making local updates of individual variables consistent with the conditional distribution of these individual variables given the current values of other variables (see McClelland, 2013, for details). This is exactly what we are doing in the MIA: We are sampling states of the letter units, conditional on states of the word and feature units; and we are sampling states of the word units, conditional on states of the letter units and feature units (although the feature units only affect the word units indirectly, via the states of the letter units).

5.6. Probabilities of states of the MIA model and pathways through the generative model

If we sample states of the MIA model at some temperature T, the probability that we will be in a given state after an initial "burn-in" period is equal to $e^{G(s_\pi)/T}$, where the goodness is defined as it was above. For the specific case of the MIA model, the goodness becomes

$$G(S_\pi) = \ln(p(w_i)) + \sum_k \left(\ln p(l_{jk}|w_i) + \sum_f \ln p(v_{fk}|l_{jk}) \right)$$

exponentiating this expression, we obtain:

$$e^{G(s_\pi)} = p(w_i) \prod_k \left(p(l_{jk}|w_i) \prod_f (v_{fk}|l_{jk}) \right)$$

The expression on the right is the probability, under the generative model, that the path through the generative model underlying the observed set of features is the one that correspond to state S_π. Plugging this into the probability-goodness equation, we see that the model visits such states with probability proportional to the temperature-scaled probability that they actually generated the observed features:

$$p(s_\pi) \propto \left(p(w_i) \prod\nolimits_k \left(p(l_{jk}|w_i) \prod\nolimits_f p(v_{fk}|l_{jk}) \right) \right)^{1/T}$$

or more simply

$$p(S_\pi) \propto p(P_\pi|\{V\})^{1/T}$$

The temperature parameter T has both an overt and a covert role in the behavior of the model. Overtly, when T is very high, all states become equiprobable, whereas when T becomes very low, only the states with the highest posterior probability have any appreciable chance of being sampled by the network after the "burn-in" period. However, if the network is run at a very low temperature, the burn-in period becomes exceedingly long. The approach initially suggested for the Boltzmann machine was to use simulated annealing, whereby T starts high and is gradually reduced. Instead of this, in the simulations we have conducted with the MIA model, we have run the model at a fixed temperature $T = 1$. In this case, we have found that the network achieves the correct posterior probability distribution in less than 20 cycles, and the approximation is quite good within about 10 cycles.

5.7. Making overt responses based on the state of the model

The development thus far shows how an interactive neural network can sample from the posterior of the probability distribution over entire states of a neural network. These

states are samples from the joint distribution of assignments of both letter and word identities that could have given rise to the actual features present in the network's input. Should we be interested in determining the identity of a particular item—say, the letter in a given position, as in many visual word recognition studies, or the whole word, as in many other studies—we can observe that the probability of being in a state where the unit in question is active (regardless of the activations of other units) corresponds to the correct posterior probability of the item. In other words, the network's states are simultaneously samples from the marginal distribution of each of the multinomial variables and the joint distribution of all of these variables. This is exactly what Rumelhart (1977) envisioned as the outcome of interactive processing in perception.

To generate a response that is a sample from this distribution, say about the identity of the letter in the first position of a word, a perceiver would only need to report the identity of the letter that had been selected through the iterative settling process. Simulations of the model verify this mathematical fact; one example illustrating this is shown in Fig. 6 (see caption for further explanation).

5.8. Sampling as an approximation to optimality

We have described a model in which perception involves sampling from the posterior of the generative model characterizing the stimuli presented to the perceptual system. It should be noted here that the truly optimal policy would be to choose the alternative with the highest posterior probability, rather than sample alternatives in proportion to their relative probability, the policy we follow in the model by setting the temperature parameter T to 1[7]. Alternatively, however, we can see the temperature parameter as reflecting intrinsic processing noise in the perceptual system. In that case, we can see each trial in a perceptual experiment as an attempt to find the single best interpretation subject to the prevailing level of noise. In either case, the higher the temperature, the more random behavior will be. The advantage of higher temperature is that it allows fuller exploration of the range of possible perceptual interpretations and avoids premature commitment early in a computation.

In Boltzmann machines, optimal perceptual interpretation is made possible by gradually reducing temperature, but this policy is only guaranteed to find a global optimum after an infinite time. In view of the real-time constraint, sampling at a fixed intermediate temperature may be the compromise the brain adopts as its approximation to optimal perceptual inference in real time.

5.9. Perceptual facilitation in non-words in the MIA model

As we noted earlier, an important feature of the original IA model was the fact that it accounted for the facilitation of perception of letters in pseudowords, such as MAVE, as well as for facilitation of perception of letters in words. In the original model, this occurred because a non-word could partially activate several words that shared letters in common with the string presented. At first glance, it might be supposed that the MIA

model would not show the same pattern, since only one word is active at a given time. To explore this, we considered the ambiguous displays in Fig. 7, where the letter in the second position is partially occluded but occurs either in a word, in a pseudoword, or by itself. The available features are equally consistent with the letters A and H in the Rumelhart-Siple font used to represent letters in the simulation. Can the model successfully use context to resolve the ambiguity, selecting A as the more likely alternative, even if the ambiguous segment occurs in a pseudoword context?

To address this question, simulations with each of the three displays shown in the figure were conducted. For the single letter alone case, the word level was switched completely off, as a baseline for assessing the influence of the word level in the other two contexts. The results are shown in Fig. 7. As the figure indicates, in the absence of context (white bars), the alternatives A and H are both chosen about half of the time, since the feature values specified are maximally consistent with both of these alternatives. With either context, the letter A becomes far more likely than the letter H. This occurs to a greater extent when the first position contains a C than when it contains an M, but it occurs to a considerable extent in both cases.

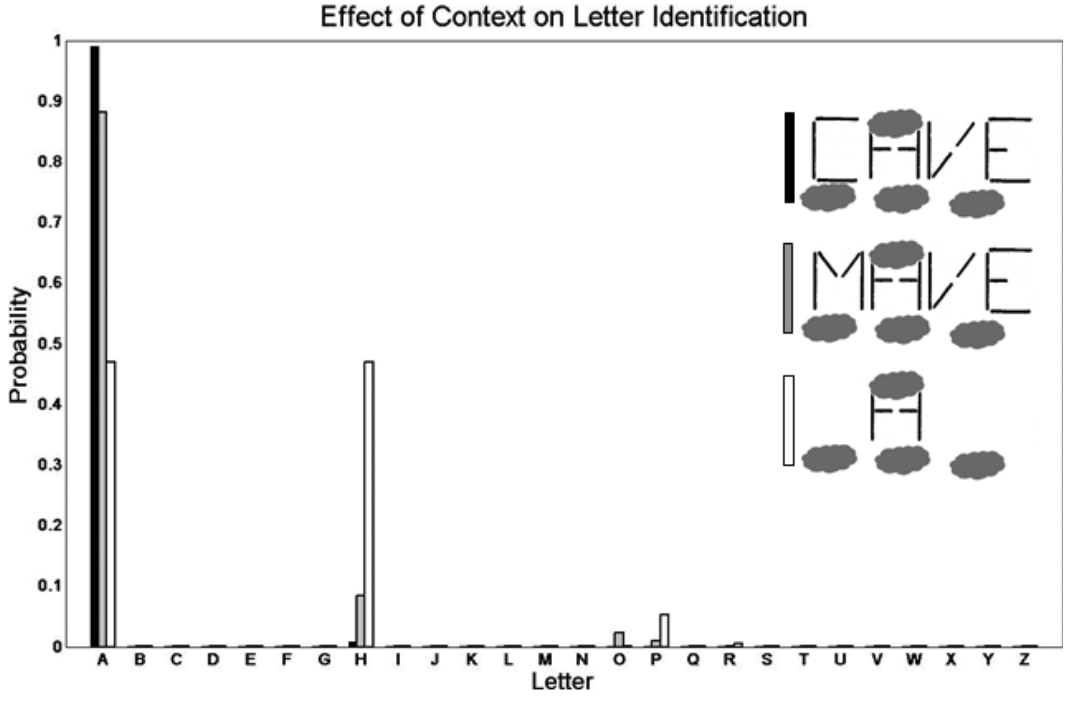


Fig. 7. Probability that different letters are activated in the second letter position when an ambiguous character equally consistent with A or H is presented in different contexts (black bar: C_VE; gray bar: M_VE; white bar: no context). In the generative model based on the Rumelhart-Siple font, both A and H are equally likely to generate the features shown, and the letter P is next most likely. But when the context is C_VE or M_VE, A is far more likely. The M_VE context supports the letter O to some degree, but the feature information is unlikely under the hypothesis that the letter is O, so overall O is much less likely than A.

Why does the model tend to choose the letter A in both contexts? When the rest of the letters form the word CAVE, the entire display is far more likely to have been generated by CAVE than any other word, and thus the letter A is far more likely to have been the letter in the second position than the letter H. When the first letter is M, no single word is highly likely to have generated all of the observed features. In fact, the word MOVE is overall more likely than any other single word (although it is inconsistent with some of the features in position 2, it is consistent will all of the features in all of the other positions). However, many other words, including CAVE, GAVE, HAVE, SAVE, and WAVE, as well as MADE, MAKE, MALE, MARE, and MATE, are all partially consistent with the full set of features. Each of the words listed is sometimes sampled at the word level; when MOVE is sampled, the model can choose O as the letter in the second position, but it can also choose A or H, since these letters can occasionally be generated according to the generative model when the underlying word was MOVE. When one of the words containing A in the second position is sampled, it almost always chooses A as the corresponding letter.[8]

5.10. The MIA model exhibits logistic additivity, addressing a limitation of the original IA model

We have seen that, in the multinomial IA model, if settling occurs at a fixed temperature $T = 1$, exact matching of posterior probabilities according to our generative model can be obtained. Do human perceivers also match these posterior probabilities? Since it is hard to obtain independent evidence of the subjective probabilities involved, the tendency has been to determine whether or not perceivers are combining context and stimulus information according to the functional form we would expect if they were performing optimally. Interestingly, there is a simple functional form that arises in the multinomial IA model and other stochastic variants of the IA model for the way in which a factorial manipulation of stimulus and context information should affect the probability of choosing a particular alternative (McClelland, 2013; Movellan & McClelland, 2001): It is easy to show (for a subset of these models, including the multinomial IA model) that the *logit* of the probability of making a particular response (where $\text{logit}(p)$ is defined as $\ln(p/(1-p))$ a quantity also known as the log-odds) should correspond to a sum of two quantities, one due only to the stimulus (corresponding to the relative probability of the sampled features given the item) and another due only to the context (corresponding to the relative probability of the item given the context).

$$\text{logit}(p_i) = s_i + c_i$$

An additional term b_i can be included to incorporate a bias associated with the alternative's prior probability. This relationship (which Movellan and McClelland called *logistic additivity*) holds at least approximately in the data from many studies investigating the joint effects of context and stimulus information (see Movellan & McClelland, 2001 for a

review; see Pitt, 1995 for one exception). The multinomial IA model exhibits logistic additivity, and its tendency to do so is unaffected by the value of the temperature parameter (T): T can be thought of as scaling the magnitudes of the stimulus and context terms in the model's predictions, but it is not in general separately identifiable from the data.

As Massaro (1989) noted in his early critique of the original IA and TRACE models, these models did not capture the logistic additivity seen in the data from many experiments, and this failure was the basis for his conclusion that interactivity fundamentally distorts perception; similar concerns have contributed to the criticisms offered by Norris et al. (2000) and Norris and McQueen (2008). While the original model's assumptions did distort this relationship, the problem was not in fact due to interactivity: As mentioned above, the influence of multiple sources of input failed to exhibit logistic additivity under the activation functions used in the original models even when propagation of activation was strictly feed forward (McClelland, 1991). In any case, logistic additivity is observed in the MIA model, overcoming this limitation of the original model.

It is important to note that logistic additivity is observed in a number of other variants of the IA model (McClelland, 1991, 1998; Movellan & McClelland, 2001); in particular, it is not necessary to assume the unit activations are binary. Although the result is harder to prove mathematically for such cases, it has been demonstrated to hold in simulations. The variants that exhibit logistic additivity incorporate variability in the input to the model and/or intrinsic to processing within the model.

5.11. Interim summary

It is hoped that the exposition of the MIA model makes clear that interactive activation produces a good approximation to optimal perceptual interpretation in real time, in accordance with the IA hypothesis, and that the MIA model (along with other variants of the IA model) can capture the logistic additivity pattern seen in data. This does not mean, of course, that the MIA model is the best possible model of human perceptual processing or even that interactivity is a part of the process of perception. Indeed, critics have argued that interactivity is not necessary to achieve a good approximation to optimality, leading them to argue for models in which processing is unidirectional. We now turn to consider this issue.

6. Is it advantageous for influences to feed back into the perceptual system?

A number of authors have proposed that context effects on letter or phoneme identification can be adequately explained by relying only on feed-forward processing, with integration of stimulus and contextual information occurring at a subsequent, decision stage (e.g., Massaro, 1989; Norris & McQueen, 2008; Norris et al., 2000; Paap, Newsome, McDonald, & Schvaneveldt, 1982). A post-perceptual decision level that integrates perceptual and contextual information can explain how stimulus and lexical information affect letter or phoneme identification (Fig. 8a). Thus, these authors have argued, interactive activation is of no benefit, and it need not be incorporated into models of perception.

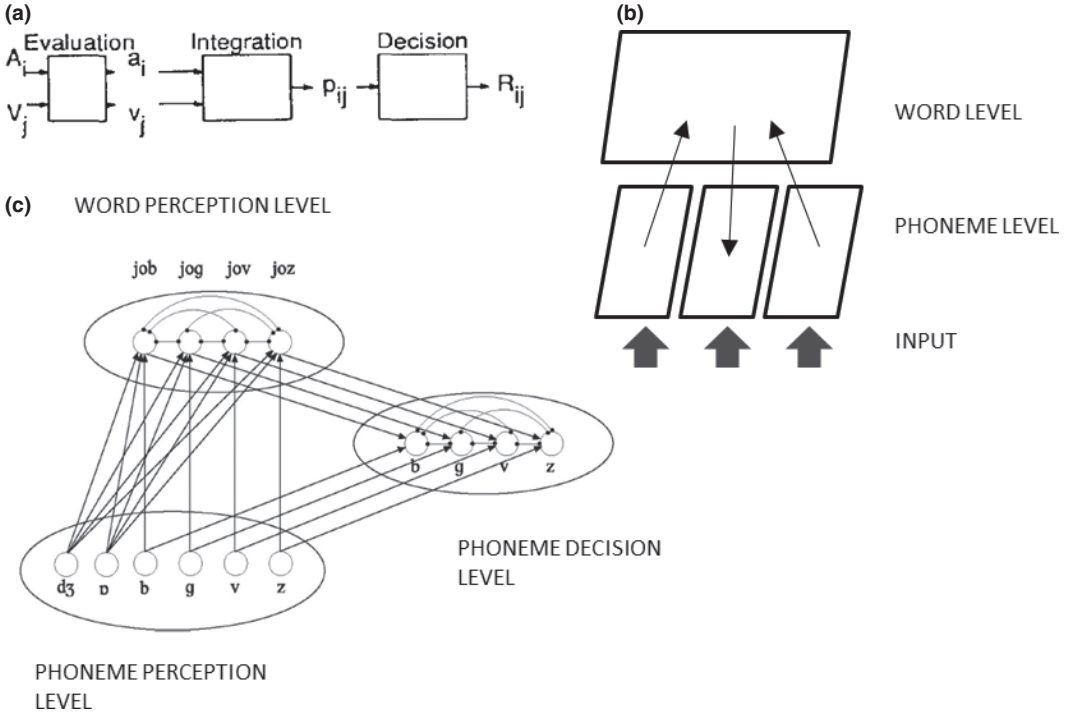


Fig. 8. (a) Massaro's schematic representation of the integration of stimulus and context information according to his Fuzzy Logical Model of Perception, reprinted from fig. 1, p. 401 of Massaro (1989). Copyright © Elsevier Ltd, reprinted with permission. The A and V variables in Massaro's figure correspond to the stimulus and context variables presented in the text. (b) Schematic diagram indicating a unidirectional propagation of information for computing the contextual and stimulus factors used in Massaro's model for the identification of the segment in the middle position of a three-phoneme syllable. (c) The architecture of the MERGE model of speech perception (Norris et al., 2000), reprinted from fig. 11, p. 384 of Norris and McQueen (2008). Copyright © American Psychological Association, reprinted with permission.

We argue that there are two important ways in which interactive activation can be beneficial:

1. It implements optimal perceptual identification over many representational levels and at many positions within a level at the same time.
2. It allows the consequences of these processes to be available inside the perceptual system itself, thereby allowing for the possibility of knock-on consequences for processing of other inputs or for processing the same item on later occasions.

We consider these points in the next two sections.

6.1. Implementing optimal inference over many levels and positions simultaneously

To underscore the first advantage of an interactive approach, we contrast it with the approach proposed by Massaro (1989), who has advocated strictly feed-forward

computation for the integration of context and stimulus information in perception. Similar points apply to the approach favored by Norris and collaborators (Norris & McQueen, 2008; Norris et al., 2000), as we shall discuss below.

Massaro's model focuses on the perceptual identification of a single specified target item. For example, in one experiment (Massaro & Cohen, 1983) of the type considered by Massaro, the target item was the second speech sound in a monosyllable beginning with either /t/, /s/, /p/, or /v/, and ending in the vowel /i/ ('ee'). Seven different sounds were presented in each context, between the initial consonant and the vowel, organized on a continuum from /l/-like to /r/-like, for a total of 28 distinct stimuli in all. Each stimulus was presented many times to each participant, with the task of identifying the second segment as either /l/ or /r/.

From a Bayesian point of view, one could propose that perception depends on calculating an estimate of the posterior probability that a given input is /r/ or /l/, using both stimulus and context as sources of constraining information. This can be done by calculating, for each context *c,* the quantities $p(r|c)$ and $p(l|c)$; and also by calculating for each stimulus *s*, the quantities $p(s|r)$ and $p(s|l)$. The correct posterior for $p(r|s,c)$ is then given by:

$$p(r|s,c) = \frac{p(s|r)p(r|c)}{p(s|r)p(r|c) + p(s|l)p(l|c)}$$

Massaro's model (Fig. 8a) assumes that participants calculate quantities corresponding to normalized estimates of the probabilistic quantities in the above formulation.[9] Notably, the representation of context used in the calculation described above excludes the stimulus information from the second segment; the first and last segments specify the context, while the second provides the stimulus information used in the calculation.

The information encoded in the connection weights in a three-phoneme-slot processing system could be used to calculate the terms needed for Massaro's model, although we would then be using this knowledge in a feed-forward, rather than an interactive fashion (Fig. 8b; although arrows go up and down in this figure, each arrow goes in only one direction, and there are no feedback connections). The featural information in the first and last positions would be used to calculate $p(p|f)$ for each possible phoneme in the first and last positional slots; then, at the word level, one can calculate $p(w|\{p_1\},\{p_3\})$ for each word in the lexicon, relying as before on the assumptions of our generative model (the expression $\{p_1\}$ denotes the vector of $p(p|f)$ values for all possible phonemes in position 1, and similarly for $\{p_3\}$). The quantity $p(r|c)$ can now be calculated as the sum over all words of the probability of the word given the input in the first and last position, times the probability of r in the second position given each word, and similarly for $p(l|c)$. This corresponds to using the connection weights between the phoneme and word units in one direction in the first and last position, and in the opposite direction in the second position, as illustrated in the figure. The desired quantity $p(r|s,c)$ is then calculated by combining the lexical input with the bottom-up stimulus support calculated for the phonemes in the second position and then using the above equation. This calculated probability is then used to generate the *r* response with probability $p(r|c,s)$ or the *l* response with probability

$p(l|c,s) = 1 - p(r|c,s)$. An alternative, sampling-based approach that would produce *r* responses with the same probability would proceed by selecting a single phoneme in positions 1 and 3, based on the feature input to these positions; then selecting a single word based only on these phonemes; then selecting between *r* and *l* for the middle position based on the selected word and the feature input in the second position. In either case, the calculations are unidirectional, and contextual and stimulus support for the target item are calculated separately, as Massaro's model proscribes.

We can now contrast Massaro's feed-forward proposal with the interactive activation approach, in which a bidirectional computation is applied across all positions, as previously described. In Massaro's model, the computations outlined above are only valid for calculating the posterior probability of the phoneme in the second position. This may not seem problematic when considering the experimental paradigm used by Massaro and Cohen (1983), where the target was always the phoneme in the second position (See Fig. 8b). However, in most experiments on the perception of letters in words, including the experiments of Reicher (1969), Massaro and Klitzke (1979), and nearly all of the experiments addressed by the IA model, participants are not cued prior to the trial on which letter will be the target letter. For these cases, the multinomial IA model simultaneously samples from the correct Bayesian posterior in all four positions. Furthermore, the MIA model uses the same representation at the word level both as its sample from the distribution of possible words and as the basis for constraining perception of each possible letter. For Massaro's model, the context representation for each position excludes the bottom-up information from that position, and thus is an incomplete representation of the information relevant to the identification of the word. In short, for an input containing three letters, four different word-level quantities are needed, one for word level, and one for each letter position.[10]

Feed-forward computation in MERGE and related models. An approach very similar to Massaro's is advocated by Norris and colleagues in their models of perceptual processing of words and letters or phonemes (Norris & McQueen, 2008; Norris et al., 2000). Just as in Massaro's model, the correct feed-forward calculation of the necessary top-down constraints for each letter or phoneme is different for each item at a lower level (e.g., for phonemes in each position, the lexical context must be based on the phonemes in all other positions). In particular, when considering the role of context on identification of a target segment (e.g., the effect of the first two segments in *job* on the identification of the final segment, see Fig. 8c), bottom-up information about the target segment is not allowed to affect values at the word perceptual level until after the top-down influence from the first two segments has been combined with the target segment information in the phoneme decision layer (D. Norris, personal communication, July 2011). This would be difficult to implement, since information about speech segments overlaps in the spoken input. The difficulty is compounded when we consider the effects of subsequent context, as in the classic experiment of Ganong (1980), where the target segment is the first segment in a word context – a /g/ or /k/ followed by "iss" or "ift," or in experiments where disambiguating context occurs in a subsequent word (Warren & Warren, 1971). To

explain this effect, segments subsequent to the target segment must be allowed to affect the word level, but the target segment must be prevented from doing so. In interactive models, this complication is unnecessary. Context phonemes in all positions can affect processing of each phoneme in each position simultaneously, with decisions about each being updated as information becomes available, either about prior or subsequent elements of the input.

In summary, non-interactive models in the psychological literature have not addressed the simultaneous use of context and stimulus information at multiple levels and multiple positions within a level. They have tended to focus on joint use of context and stimulus information in identifying a specified target item at one level of processing, without dealing with the fact that in natural perceptual situations, the goal is to simultaneously interpret multiple items at many different levels of processing. In contrast, interactive models allow representations of alternatives at different levels and different positions within a level to mutually constrain each other in an integrated parallel, distributed, and interactive computation.

6.2. *Knock-on consequences of interactive processing*

We now consider the second advantage of interactive models over feed-forward models: Interactivity allows effects of context to affect subsequent processing within the perceptual system. Such effects include effects on processing of neighboring items present in the immediate context of a presented item, and effects on processing of similar inputs on subsequent occasions.

Knock-on consequences for neighboring input items. A case of the first type was considered by Elman and McClelland (1988). They focused on a phenomenon in speech perception known as compensation for coarticulation (Mann & Repp, 1981; Stephens & Holt, 2003): The perceptual system seems to compensate for the effects that articulation of one phoneme has on the acoustic realization of neighboring phonemes. For example, the lip formations associated with /s/ and /ʃ/ (“sh”) persist into the articulation of subsequent stop consonants like /t/ and /k/, shifting the frequency content of the successor. Perceivers compensate for this, allowing more accurate recognition of the successor. Thus, when an ambiguous sound between /t/ and /k/ is preceded by /s/, it will tend to be heard as /k/; when preceded by /ʃ/, it will tend to be heard as /t/. In this situation, the presence of background noise or articulatory variability could obscure the identity of the preceding fricative sound, robbing a strictly feed-forward system of information to allow compensation. But if that fricative sound occurred in a lexically constraining context, and feedback were allowed to influence the activation of the contextually more likely fricative, compensation could nevertheless occur, improving identification of subsequent phonemes. Elman and McClelland (1984) included a mechanism for producing such compensatory effects in one version of the TRACE model, simulating the lexically mediated compensation for coarticulation effect.

Elman and McClelland (1988) subsequently designed an experiment to determine whether lexical context could trigger compensation for coarticulation, as the TRACE model predicted. They presented ambiguous /t/ or /k/ sounds preceded by an ambiguous fricative sound halfway between /s/ and /ʃ/. In turn, the ambiguous fricative was preceded by one of two different lexical contexts, one consistent with /s/ (e.g., "Christma_") and one consistent with /ʃ/ (e.g., "fooli_"). If lexical information feeds back to influence phoneme processing, then the ambiguous fricative in "Christma_" should behave like an acoustic /s/ and cause a shift in the perception of the following phoneme toward /k/. Conversely, the same ambiguous fricative in "fooli_" should behave like an acoustic /ʃ/ and cause a shift in the perception of the following phoneme toward /t/. This is precisely what Elman and McClelland found. Although this result has been questioned (Pitt & McQueen, 1998), it has been replicated in multiple different laboratories, and with different sets of materials (Magnuson, McMurray, Tanenhaus, & Aslin, 2003; Samuel & Pitt, 2003). Those who favor non-interactive approaches have, however, presented recent evidence further contesting the source of the effect (McQueen, Jesse, & Norris, 2009), and research on the topic continues.

Knock-on consequences for processing similar inputs on subsequent occasions. Other researchers have explored other knock-on effects of lexical context on phoneme identification that are also predicted by the interactive account. One such effect has been demonstrated using selective adaptation, a domain-general phenomenon in which repeated presentation of a particular stimulus causes a perceptual shift such that neutral stimuli are perceived as being less like the repeatedly presented stimulus. In the case of speech perception, after repeated presentation of a phoneme (e.g., /s/), perception of an ambiguous phoneme (e.g., halfway between /s/ and /ʃ/) is shifted toward the alternative interpretation (in this case, /ʃ/; e.g., Samuel, 1986; Samuel & Kat, 1996). To demonstrate lexically mediated selective adaptation, a neutral sound (an ambiguous phoneme or a noise burst) was repeatedly presented in lexical contexts that were consistent with only one interpretation. If the neutral sound was presented in /s/-biased contexts such as "bronchiti_", "arthriti_", etc., the selectively adapted representation was /s/; if it was presented in /ʃ/-biased contexts such as "aboli_", "demoli_", etc., the selectively adapted representation was /ʃ/ (Samuel, 1997, 2001). Thus, the lexical information determined which sublexical representation was selectively adapted, influencing subsequent phoneme and word identification.

A third example of knock-on consequences of lexical feedback—one that was predicted in the McClelland and Elman (1986) TRACE model paper—is lexically guided tuning of speech sound categories. Such tuning is essential for listeners to be able to correctly identify different speakers' productions, since phoneme category boundaries vary between individuals. For example, speakers of English and Spanish center their /b/ and /p/ categories at different points along a dimension called voice onset time. Furthermore, regional dialects are often distinguished by differences in details of both consonant and vowel production. Lexical information provides a ready source of information for tuning speech perception in response to such shifts in speech sounds, and several studies

beginning with Norris, McQueen, and Cutler (2003) now indicate that such tuning does in fact occur in speech perception (van Linden & Vroomen, 2007 showed an analogous shift in use of visual cues from the lips; for a review see Samuel & Kraljic, 2009). The pre-lexical locus of this effect is supported by evidence that the tuning effect generalizes to influence perception of words not used in the induction of the effect (McQueen, Cutler, & Norris, 2006). In TRACE, lexical information feeds back to influence pre-lexical phoneme unit activations, and Mirman, McClelland, and Holt (2006) augmented TRACE with a simple Hebbian learning rule to adjust the feature to phoneme connections, allowing it to simulate the relevant experimental findings.

More generally, Friston (2003; see also Spratling & Johnson, 2004) has argued that top-down feedback is necessary to learn the hierarchical representations that are found throughout perceptual and cognitive systems, and indeed some form of feedback is used in many different neural network learning algorithms. Proponents of autonomous/feed-forward accounts of perception acknowledge the necessity of feedback for learning but insist that this feedback is not equivalent to the "online" feedback that influences processing in interactive activation models (e.g., Norris et al., 2003). We argue that a system in which feedback can guide learning as well as perception provides a parsimonious account. Furthermore, if feedback guides learning, then the learned representations will necessarily reflect a combination of bottom-up and top-down information, making the representations themselves both consequences of and intrinsic to their roles in interactive processing.

In sum, feedback not only allows contextual constraints to determine the identity of elements (such as letters and phonemes) of larger units (such as words) but also allows the results of this contextually determined identification process to influence processing of neighboring elements (compensation for coarticulation) and subsequent occurrences of the same elements (adaptation, retuning). Knock-on consequences of feedback provide both motivation for and evidence of direct top-down feedback in perception.

7. Neural basis of interactive processing

7.1. Basic neuroscience findings

Evidence from research on the neural basis of perception supports the presence of interactive processing in the brain. Interactive processing is supported by a basic feature of brain architecture: Wherever in the neocortex there is a "forward" path from area A to area B there tends to be a strong (sometimes much stronger) return pathway (Felleman & van Essen, 1991). Many studies correspondingly show that reversible inactivation of putatively higher level or downstream cortical areas (e.g., higher level visual or auditory cortex) affects stimulus-driven activity in primary areas (e.g., Hupé et al., 1998; Carrasco & Lomber, 2010), implicating reciprocal interactions in cortical processing. Neural recordings in rhesus monkeys indicate that the same "edge detectors" in V1 that respond to physically present edges also respond to illusory edges in Kanizsa figures. The illusory contour response in V1 was found to occur later than the response in V2, suggesting that

the response in V1 was due to feedback from higher level visual processing (Lee & Nguyen, 2001). Similarly, binocular rivalry appears to be a mutual constraint satisfaction/interactive activation process with neurons in many different visual areas, from V1/V2 to inferotemporal cortical areas, showing consistency with the global percept (Leopold & Logothetis, 1999). Evidence of bidirectional propagation of activity between occipito-temporal and pre-frontal brain areas is also seen in human magneto-encephalography (MEG) studies of visual object recognition (e.g., Bar, 2004).

In addition to top-down feedback from higher levels within a processing modality, neurophysiological studies have shown cross-modal interactions between primary regions of perceptual processing (see Ghazanfar & Schroeder, 2006 for review). To us such mutual constraints between modalities are just as much examples of the fundamental principle of mutual constraint satisfaction as the bidirectional interactions between levels in a hierarchical perceptual system. Although several studies have argued that sensory integration occurs in secondary sensory or association cortex (Bavelier & Neville, 2002; Jones & Powell, 1970) or in frontal cortex (Rizzolatti, Riggio, Dascola, & Umlita, 1980), recent evidence has pointed to the presence of top-down inputs from these association regions to primary sensory cortices in audition (Cappe & Barone, 2005; Schroeder et al., 2001) and vision (Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003) as well as direct input from auditory cortex to primary visual cortex (Falchier et al., 2002; Hall & Lomber, 2008) and vice versa (Bizley & King, 2009). Physical projections from auditory cortex terminating in area V1 have also been observed in the monkey (Falchier et al., 2002; Rockland & Ojima, 2003) and in the adult cat (Hall & Lomber, 2008), suggesting that these connections are not limited to early developmental stages. In addition, evidence from multiunit recordings in the ferret has shown that roughly 20% of the neurons in area A1 respond to visual stimulation (Bizley & King, 2009).

Overall, a growing body of evidence is challenging the idea that there is encapsulation of sensory processing at the neural level (see Ghazanfar & Schroeder, 2006). Instead, the evidence suggests that a highly interactive biological system enables the simultaneous use of information across hierarchical levels from multiple modalities for spatial localization, communication, and various social behaviors (Lewkowicz & Ghazanfar, 2009). This interactive neural system implements cognitive processing that relies on the simultaneous, coherent engagement of representations at many levels and within many modalities at the same time—that is, processing that is distributed, parallel, and interactive.

7.2. Interactivity in the brain mechanisms of human language processing

Interactive processing has also been a key theme in research on human language processing and reading. Much of this work has been conducted within the framework of the "Triangle model" of single-word reading (Seidenberg & McClelland, 1989; and subsequent extensions), which can be viewed as a version of the interactive activation model that relies on learned distributed representations rather than localist representations of units at the orthographic, phonological, and semantic level. Here, we highlight the interactive processing aspects of the framework as illustrated in Fig. 9, focusing on the

timing and locus of mutual influences of phonology and orthography and of lexical effects on phonological and orthographic processing. Note that in the triangle model framework, bidirectional connections throughout the model are sensitive to lexical knowledge as well as knowledge of the patterns of covariation between orthographic and phonological representations. Specifically, the presentation of a visual or spoken word form would induce bidirectional interactions among orthographic, phonological, and semantic representations, leading to the prediction that lexical knowledge and spelling-sound consistency would affect orthographic and phonological representations, at least in skilled readers, once the relevant connections had become strengthened through experience.

Discussions of the neural basis of visual word recognition have focused heavily on the role of a region of the left occipito-temporal cortex known as the Visual Word-Form Area (VWFA; McCandliss, Cohen, & Dehaene, 2003; Dehaene, Cohen, Sigman, & Vinckier, 2005). Some have argued that VWFA functions as an orthographic "input" lexicon, a repository for visual forms of words (Kronbichler et al., 2004, 2007), while others have contended that this region is prelexical in nature (Dehaene et al., 2005), with some possible hierarchical organization of orthographic representations in or near the VWFA. In an interactive framework, a representation can be structured orthographically and still be sensitive to lexical constraints and influences from other input modalities. That is, we can consider the VWFA to be the approximate neural analog of the pool of units labeled "orthography" in the triangle model, which primarily represent orthographic structure but are also sensitive to interactive influences from other representations. A considerable body of evidence supports the view that processing in this region is susceptible to influences from other input modalities, including influences arising from tactile (Braille)

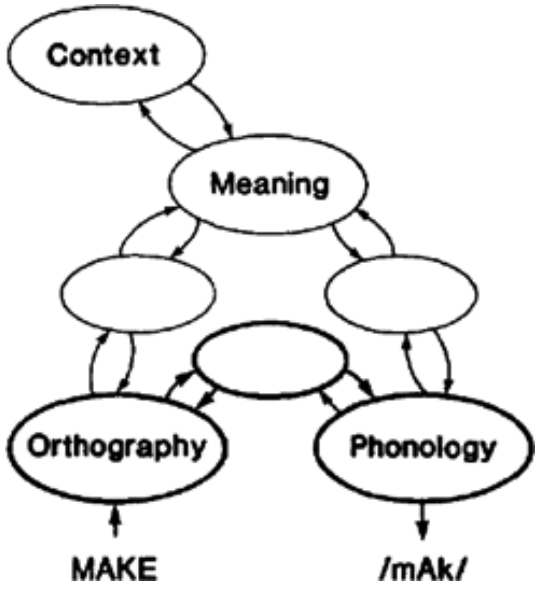


Fig. 9. The triangle model framework for single-word reading. Reprinted from fig. 1, p. 526, of Seidenberg and McClelland (1989). Copyright © American Psychological Association, reprinted with permission.

input for congenitally blind patients (Buchel, Price, Frackowiak, & Friston, 1998; Cohen et al., 1997) for handwriting (Barton, Fox, Sekunova, & Iaria, 2010) and for auditory word processing (Binder, Medler, Westbury, Liebenthal, & Buchanan, 2006; Cone, Burman, Bitan, Bolger, & Booth, 2008; Desroches et al., 2010). As important, studies that have looked at the influence of consistency between a word's spelling and its sound have revealed graded effects of consistency and frequency at the item level mirroring the behavioral findings of consistency effects in naming (Bolger, Hornickel, Cone, Burman, & Booth, 2008; Bolger, Minas, Burman, & Booth, 2008; Graves, Desai, Humphries, Seidenberg, & Binder, 2010). Consistent with predictions from the Triangle model (Harm, McCandliss, & Seidenberg, 2003), Bolger, Hornickel, et al. (2008) and Bolger, Minas, et al. (2008) found that response to grapheme–phoneme consistency in the VWFA increased with reading skill. These findings support the view that interactive processing becomes established as reading skill becomes more and more automatic; this is captured in the triangle model framework in terms of the strengthening of bidirectional connections between the neurons participating in each of the three different types of representations with experience.

Neuroimaging studies of speech perception have also addressed the predictions of interactive models. Whereas accuracy of phonological perception is associated with the superior temporal cortex, decision time is associated with inferior frontal/insula cortex (Binder, Liebenthal, Possing, Medler, & Ward, 2004) and anterior cingulate/medial frontal regions of cortex (Grinband, Hirsch, & Ferrera, 2006; Grinband et al., 2011). Interactive models predict that brain regions involved in phonological processing (e.g., posterior superior temporal gyrus and Heschl's gyrus in the superior temporal sulcus) should show effects of lexical bias. In contrast, autonomous decision-level integration models predict that these lexical bias effects should be limited to brain regions involved in decision-making and response selection (e.g., inferior frontal gyrus and anterior cingulate gyrus). An fMRI study (Myers & Blumstein, 2008; see also Guediche, Salvata, & Blumstein, 2013) found that the lexical bias on categorization of ambiguous phonemes was associated with increased activation in the superior temporal gyrus. This region is also activated during auditory hallucinations of voices in patient populations (Dierks et al., 1999) and imagined speech of others in healthy individuals (McGuire, Silbersweig, & Frith, 1996).

Electrophysiological measures have provided key evidence that lexical and consistency effects occur early, during perceptual and/or lexical processing, rather than during a post-perceptual decision stage, in both visual and auditory modalities. For example, rhyming effects on visual processing of orthographically dissimilar words have been detected around 260 ms after stimulus onset (Kramer & Donchin, 1987), and syllable effects in visual word processing have been shown at around 250–350 ms (Ashby & Martin, 2008; Carreiras, Ferrand, Grainger, & Perea, 2005). Consistency effects in auditory lexical decision tasks (Perre, Midgley, & Ziegler, 2009; Perre & Ziegler, 2008) and semantic categorization tasks (Pattamadilok, Perre, Dufau, & Ziegler, 2008;Pattamadilok, Morais, De Vylder, Ventura, & Kolinsky, 2009) have been shown to occur in ERP roughly 300–350 ms post-stimulus and time locked to the point of inconsistency. Findings from MEG imaging, which provides greater spatial resolution, have localized the early rhyming effects in visual tasks to the left occipito-temporal region (Wilson, Leuthold, Lewis,

Georgopoulos, & Pardo, 2005). In related work, van Linden and colleagues (van Linden, Stekelenburg, Tuomainen, & Vroomen, 2007) found that lexical context induced an early, perceptually based mismatch negativity effect, suggesting that lexical information directly affected perceptual processing stages.

Although neuron-level neuroanatomical precision is difficult to achieve in the domain of human language processing, recent studies combining multiple imaging modalities show promise for increasing both spatial and temporal precision. A study combining MEG and electro-encephalography (EEG) with anatomical MRI (Gow, Segawa, Ahlfors, & Lin, 2008) found reactivation of posterior superior temporal gyrus following activation of a region associated with lexical processing (supramarginal gyrus). An ERP study (Molinaro, Duñabeitia, Marìn-Gutièrrez, & Carreiras, 2010) found that during an early period (180–220 ms after onset) letter-like numbers in word contexts (e.g., M4T3R14L) were processed more like numbers than letters, but only slightly later (250–300 ms after onset) this pattern reversed and letter-like numbers were processed more like letters than numbers. A combined ERP-MEG study (Sohoglu, Peelle, Carlyon, & Davis, 2012) replicated the facilitative effect of prior knowledge (written text) on perceptual clarity of degraded speech and found that this effect was reflected in inferior frontal gyrus activity before superior temporal gyrus activity, consistent with top-down feedback from higher level processing in the inferior frontal gyrus modulating perceptual processing in the superior temporal gyrus.

The exact nature, timing, and location of lexical and consistency effects in visual and auditory word perception remains subject to a range of interpretations, and a considerable body of ongoing work is addressing these issues. One very general open question is whether top-down and between-modality influences should be viewed as an additional sources of constraint on interpretation, as in the interactive activation framework, or whether, instead, top-down signals should be viewed as predictions that are compared with bottom-up signals, generating error signals that then drive learning mechanisms (Friston, 2008; Mumford, 1992; Rao & Ballard, 1999). A further question is the interplay between such influences and synchronization of neural activity within and across brain regions (see Gotts, Chow, & Martin, 2012 and commentaries therein for a recent discussion).

There appears to be little doubt that top-down influences affect relatively early, modality-specific processing areas, both in language processing and in other tasks. Brain regions tend to be connected bidirectionally and there is strong neurophysiological evidence that these bidirectional connections implement interactive activation in perceptual and conceptual processes (Ghuman, Bar, Dobbins, & Schnyer, 2008; Gotts et al., 2012). Specifically within the domain of language processing, the neural evidence indicates that feedback and audio-visual interactions directly influence perceptual processing, consistent with interactive models.

8. Summary and future directions

Over the course of this article, we have laid out the case for interactive activation and mutual constraint satisfaction in perception and cognition. We have focused primarily on

visual and spoken word recognition, the target phenomena first addressed by IA models, but we have also considered other applications of interactive approaches. We have explored computational theory-level considerations and neuroscience evidence as well as evidence on the role of context in perception as revealed by behavioral studies.

We have argued that interactive activation addresses key computational challenges facing perceptual systems and is consistent with a wide range of evidence, including behavioral and neuroscience evidence on the mechanisms of perception and language processing in the brain. Overall, it appears that both computational analyses and the behavioral and neuroscience evidence are consistent with the IA hypothesis.

While the computational and empirical considerations seem strongly to support an interactive perspective, there are several important challenges requiring future investigation within an interactive activation framework.

Dynamics of perception in probabilistically grounded interactive activation models. The IA hypothesis states that processing approaches the ideal of achieving optimal results in real time as information becomes available. A good deal of experimental work has been carried out showing that participants in perceptual and language-processing tasks use all of the available information and start to show sensitivity to it within a third of a second of its arrival at the sensory surface. Simulations of such findings have been provided using the original TRACE model and related, simple Luce-ratio-based models (Spivey & Tanenhaus, 1998). Future work should explore these issues in more detail, relying on probabilistically grounded models like the multinomial IA model.

We have begun to explore a related issue in the multinomial IA model (Khaitan & McClelland, 2010), namely, the build-up of performance—and of contextual influence on performance—as participants are given increasing amounts of exposure to target information (Massaro & Klitzke, 1979). This issue is important because Massaro and Cohen (1991) specifically posed it as a challenge to the interactive activation model that was not fully addressed by the stochastic version of the model presented in McClelland (1991). Specifically, if input feature information builds up over time according to the empirical function proposed by Massaro and Cohen (1991), would the processing machinery provided by the multinomial IA model show the right pattern of enhancement for perception of letters in words compared to letters in random sequences? The simulation reported in Khaitan and McClelland (2010) suggests that the answer to this question may be yes, but the simulation is preliminary, and more work is needed.

Adaptive optimization to task and instructional constraints. An important topic for further research is the adaptive optimization of processing in interactive activation models in response to task and instructional constraints. There are a number of important open issues here. First, as we have noted, participants do adjust the extent to which they show lexical influences on processing as a function of changes in the probability that stimulus items will be words or non-words. Such influences are easily incorporated into Bayesian models (Rumelhart & Siple, 1974 consider this issue extensively) and have also been incorporated into the original IA and TRACE models (Mirman et al., 2008). It appears,

however, that there are limits on the extent to which participants can actually suspend the use of their knowledge of lexical constraints on speech sound identity. For example, in one recent study (Hawthorne, 2011) participants showed lexical influences on perception of speech sounds whether or not they were informed that each sound occurred equally often in each of two possible contexts, as one might expect if the knowledge of lexical constraints were hard wired into connections among the neurons involved in naturalistic language processing, and these same neurons and connections were relied upon independent of the instructional manipulation. There are empirical and theoretical questions here that deserve further consideration.

Incorporating learning and distributed representations in interactive models of perception. Research on interactive activation models of perception pre-dated the development of powerful learning models for parallel distributed processing systems that were developed in the mid-1980s. Models using learned distributed representations have been successful in addressing a wide range of aspects of linguistic and semantic processing, and we look forward to full integration of learning and distributed representation in further explorations of perceptual processing tasks. Recent developments of fast and powerful learning methods for deep belief networks (Hinton, 2014) should facilitate these explorations.

Meeting the computational challenges facing perceptual and cognitive systems in naturalistic perceptual contexts. The IA and TRACE models that have been the focus of our investigations here finesse many challenges facing the development of models that will be robust and efficient enough to succeed in matching human capabilities in naturalistic perceptual situations. These challenges are the focus of intense research among a wide range of researchers in the fields of AI, machine vision, and machine learning. Much of this work builds on neural network ideas with origins in IA models and precursors of such models, and of course a great deal of this work incorporates explicit probabilistic inferencing mechanisms. In turn, much of this work should feed back into the effort to understand human perceptual processing mechanisms, as they are instantiated in the neural mechanisms provided by the brain. The further development of interactive activation models of perception will benefit greatly from these developments.

Fully grounding IA models in the neural mechanisms provided by the brain. The final challenge we will mention is the goal of understanding exactly how the IA process is implemented in the neural machinery in the brain. Neurons and their properties have been a source of inspiration in the development of these models, and evidence from neuroscience supports the view that perceptual processing in the brain is an IA-like process, as we have reviewed. Building an integrated understanding of the way in which neural mechanisms give rise to perception is a goal that many researchers strive for; if the IA hypothesis is correct, such an integrated understanding will rely on principles of interactive activation.

Acknowledgments

This research was supported in part by Air Force Research Laboratory Grant FA9550-07-1-0537. The authors would like to acknowledge the useful comments of the members of the Parallel Distributed Processing laboratory at Stanford University.

Notes

1. Although the original IA model employed between-level inhibition as well as excitation, the TRACE model and other subsequent models used excitatory-only connections between levels with inhibition restricted to within-level interactions. The primary reason for eliminating between-level inhibition was to allow even a poorly fitting interpretation to become active when there is no better interpretation. We will return to this issue in discussing the multinomial interactive activation model below.
2. For simplicity, the IA and TRACE models assumed discrete slots for letters and phonemes, although TRACE assumed some spread of phonological features producing overlap between adjacent slots. Recent evidence reviewed in Norris (2013) suggests that both models should allow for positional uncertainty, so that letters near the appropriate position can still activate the corresponding word-level unit (e.g., TRCK should activate the word TRUCK much more than TRXY does).
3. Presentations of the original IA model did not stress that it contained separate units for the presence and for the absence of each possible feature. Fig. 2 makes this feature of the model more explicit than in earlier diagrams of the model.
4. Note that the $p(w)$ values used in the model are not raw word frequencies; instead, as in the original IA model, these probabilities are compressed (McClelland & Rumelhart, 1981). Without this compression, there would be a much larger range of variation in the posterior probabilities shown in Fig. 6. The compression of the $p(w)$ values amounts to an (implicit) "assumption" about stimulus frequency incorporated in the model. The bias terms on the word units are the natural logarithms of these already-compressed $p(w)$ values.
5. This result follows from the fact that the sum of the logarithms of a set of quantities is equal to the logarithm of the product of the quantities, for example, $\ln(a)+\ln(b) = \ln(ab)$, and the fact that e to the log of a quantity is simply the quantity itself, that is, $e^{\ln x} = x$. We also rely on the fact that $e^{x/T} = (e^x)^{1/T}$.
6. The complete Bayes' formula would contain factors for the prior probabilities of letters. However, in the generative model, letters do not have independent prior probabilities; instead, letter probabilities depend on the word level, whose influence on the letter level is incorporated on the second and subsequent updates of the units at the letter level. On the first update, letters are treated as equally probable. Such a constant factor would cancel out and is therefore not expressed in the equation.

7. It should be noted here that changing the temperature parameter is equivalent to scaling the weights and biases in the model, and these in turn represent relative probabilities and relative conditional probabilities in the generative model. Thus, a lower temperature corresponds to assuming less randomness in the generative model.
8. If the model was required to read out from the word level, it would always produce a word response, but the same would have been true of the original IA model. When asked to report all four letters, human observers do not always report words when pseudowords are presented (McClelland & Johnston, 1977). Further research is needed to determine if the pattern of whole report responses obtained with pseudowords can be explained by the MIA model, assuming readout from the four-letter positions.
9. In Massaro's model (Massaro, 1989), the relative stimulus support for r, called s_r, corresponds to $p(s|r)/(p(s|r)+p(s|l))$; and the relative context support c_r corresponds to $p(r|c)/(p(r|c)+p(l|c))$. The stimulus and context support for l are defined similarly. Since $s_r + s_l = 1$, s_l can be replaced by $1 - s_r$; similarly, c_l can be replaced by $1 - c_r$. Thus, for the two alternative case his model then becomes $p(r|s,c) = s_r\, c_r \,/\, (s_r\, c_r + (1 - s_r)(1 - c_r))$. Participants then choose the r response with a probability equal to the resulting estimate of $p(r|s,c)$.
10. As Pearl (1982) showed, it is possible to keep a record of the information passed up from each position to a higher level, and then cancel this back out of the top-down signal broadcast down to all lower levels from above, and a precursor of this idea was described in Rumelhart (1977). We view Pearl's proposal as an alternative implementation of an interactive model of perception; a comparison of this approach to the MIA model is provided in McClelland (2013).

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language*, *38*(4), 419–439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Ashby, J., & Martin, A. E. (2008). Prosodic phonological representations early in visual word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(1), 224–236. doi:10.1037/0096-1523.34.1.224.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629.

Barton, J. J. S., Fox, C. J., Sekunova, A., & Iaria, G. (2010). Encoding in the visual word form area: An fMRI adaptation study of words versus handwriting. *Journal of Cognitive Neuroscience*, *22*(8), 1649–1661. doi:10.1162/jocn.2009.21286.

Bavelier, D., & Neville, H. J. (2002). Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience*, *3*, 443–452.

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, *7*(3), 295–301.

Binder, J. R., Medler, D. A., Westbury, C. F., Liebenthal, E., & Buchanan, L. (2006). Tuning of the human left fusiform gyrus to sublexical orthographic structure. *Neuroimage*, *33*(2), 739–748.

Bizley, J., & King, A. (2009). Visual influences on ferret auditory cortex. *Hearing Research*, *258*, 55–63.

Bolger, D. J., Hornickel, J., Cone, N. E., Burman, D. D., & Booth, J. R. (2008). Neural correlates of orthographic and phonological consistency effects in children. *Human Brain Mapping*, *29*(12), 1416–1429.

Bolger, D. J., Minas, J., Burman, D. D., & Booth, J. R. (2008). Differential effects of orthographic and phonological consistency in cortex for children with and without reading impairment. *Neuropsychologia*, *46*(14), 3210–3224.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*, 220–251.

van den Brink, D. l., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience*, *13*(7), 967–985.

Buchel, C., Price, C., Frackowiak, R. S. J., & Friston, K. (1998). Different activation patterns in the visual cortex of late and congenitally blind subjects. *Brain*, *121*, 409–419.

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, *22*, 2886–2902.

Carrasco, A., & Lomber, S. G. (2010). Reciprocal modulatory influences between tonotopic and nontonotopic cortical fields in the cat. *The Journal of Neuroscience*, *30*(4), 1476–1487.

Carreiras, M., Ferrand, L., Grainger, J., & Perea, M. (2005). Sequential effects of phonological priming in visual word recognition. *Psychological Science*, *16*(8), 585–589. doi:10.1111/j.1467-9280.2005.01579.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 687–696.

Cohen, L. G., Celnik, P., Pascual-Leone, A., Corwell, B., Falz, L., Dambrosia, J., Honda, M., Sadato, N., Gerloff, C., Catala, M. D., & Hallett, M. (1997). Functional relevance of cross-modal plasticity in blind humans. *Nature*, *389*(6647), 180–183. doi:10.1038/38278

Cohen, J. D., Servan-Schreiber, D., & McClelland, J. L. (1992). A parallel distributed processing approach to automaticity. *American Journal of Psychology*, *105*, 239–269.

Cone, N. E., Burman, D. D., Bitan, T., Bolger, D. J., & Booth, J. R. (2008). Developmental changes in brain regions involved in phonological and orthographic processing during spoken language processing. *Neuroimage*, *41*(2), 623–635.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 498.

Dean, T. (2005). A computational model of the cerebral cortex. In *Proceedings of AAAI-05* (pp. 938–943). Cambridge, MA: MIT Press.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341.

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

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117.

Desroches, A. S., Cone, N. E., Bolger, D. J., Bitan, T., Burman, D. D., & Booth, J. R. (2010). Children with reading difficulties show differences in brain regions associated with orthographic processing during spoken language processing. *Brain Research*, *1356*, 73–84.

Dierks, T., Linden, D. E. J., Jandl, M., Formisano, E., Goebel, R., Lanfermann, H., & Singer, W., et al. (1999). Activation of Heschl's gyrus during auditory hallucinations. *Psychiatry: Interpersonal and Biological Processes*, *22*, 615–621.

Elman, J. L., & McClelland, J. L. (1984). Speech perception as a cognitive process: The interactive activation model. In Norman Lass (Ed.), *Speech and Language*. Vol. *10*. New York: Academic Press.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory & Language*, *27*(2), 143–165.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, G. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience*, *22*, 5749–5759.

Feldman, N. H., Griffiths, T. L., & Mrogan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

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

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 526–540.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*, 1325–1352.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology*, *4*(11), e1000211.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, *6*(1), 110–125.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721741.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Science*, *10*, 278–285.

Ghuman, A. S., Bar, M., Dobbins, I. G., & Schnyer, D. M. (2008). The effects of priming on frontal-temporal communication. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(24), 8405–8409.

Gotts, S. J., Chow, C. C., & Martin, A. (2012). Repetition priming and repetition suppression: A case for enhanced efficiency through neural synchronization. *Cognitive Neuroscience*, *3*(3–4), 227–237.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: A Granger causality analysis of meg and eeg source estimates. *NeuroImage*, *43*, 614–623.

Gratton, G., Coles, M. H., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre- and post stimulus activation of response channels: A psychophysiological analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 331–344.

Graves, W. W., Desai, R., Humphries, C., Seidenberg, Mark S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex*, *20*(8), 1799–1815. doi:10.1093/cercor/bhp245.

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

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*(5), 757–763.

Grinband, J., Savitskaya, J., Wager, T. D., Teichert, T., Ferrera, V. P., & Hirsch, J. (2011). The dorsal medial frontal cortex is sensitive to time on task, not response conflict or error likelihood. *Neuroimage*, *57*(2), 303–311.

Grossberg, S. A. (1978). A theory of coding, memory, and development. In E. L. J. Leeuwenberg & H. F. J. M. Buffart (Eds.), *Formal theories of visual perception* (p. 1978). New York: Wiley.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review*, *87*, 1–51.

Guediche, S., Salvata, C., & Blumstein, S. E. (2013). Temporal cortex reflects effects of sentence context on phonetic processing. *Journal of Cognitive Neuroscience*, *25*(5), 706–718.

Hall, A. J., & Lomber, S. G. (2008). Auditory cortex projections target the peripheral field representation of primary visual cortex. *Experimental Brain Research*, *190*(4), 413–430.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368.

Harm, M. W., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading*, *7*(2), 155–182.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al (1975). *Journal of Memory & Language*, *52*(1), 58–70.

Hawthorne, D. J. (2011). Can instructions diminish the influence of phonetic categories on the perception of speech sounds? Unpublished research paper, Department of Psychology, Stanford University.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197. doi:10.1017/S0952523800009640.

Henderson, C. M., & McClelland, J. L. (2011). A PDP model of the simultaneous perception of multiple objects. *Connection Science*, *23*, 161–172.

Hinton, G. E. (2014). Where do features come from? *Cognitive Science*, *38*, 1078–1101.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume I. (Ch. 7, pp 282–317). Cambridge, MA: MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558.

Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature*, *394*, 784–787.

James, R. C. (1965). Photo of a dalmation dog. *LIFE Magazine*, *58*(7), 120.

Jefferies, E., Frankish, C. R., & Ralph, M. A. L. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language*, *54*(1), 81–98.

Jones, E. G., & Powell, T. P. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, *93*, 793–820.

Kanizsa, G. (1979). *Organization in vision*. New York: Praeger.

Khaitan, P., & McClelland, J. L. (2010). Matching exact posterior probabilities in the Multinomial Interactive Activation Model. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (p. 623). Austin, TX: Cognitive Science Society.

Kramer, A. F., & Donchin, E. (1987). Brain potentials as indices of orthographic and phonological interaction during word matching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 76.

Kronbichler, M., Bergmann, J., Hutzler, F., Staffen, W., Mair, A., Ladurner, G., & Wimmer, H. (2007). Taxi vs. Taksi: On orthographic word recognition in the left visual occipitotemporal cortex. *Journal of Cognitive Neuroscience*, *19*(10), 1584–1594.

Kronbichler, M., Hutz, F., Wimmer, H., Mair, A., Staffen, W., & Ladurner, G. (2004). The visual word form area and the frequency with which words are encountered: Evidence from a parametric fMRI study. *Neuroimage*, *21*(3), 946–953.

Kubat, R., Mirman, D., & Roy, D. K. (2009). Semantic context effects on color categorization. In N. A. Taatgen & H. V. Rijn (Eds.), *Proceedings of the 31st Annual Cognitive Science Society Meeting* (pp. 491–495). Austin, TX: Cognitive Science.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*, 93–107.

Kumaran, D., & McClelland, J. L. (2011). Beyond Episodic memory: A complementary learning systems account of the hippocampal contribution to generalization. *Psychological Review*, *119*, 573–616.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *20*(7), 1434–1448.

Lee, T. S., & Nguyen, M. (2001). Dynamics of subjective contour formation in early visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1907–1977.

Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: Changing views in perception. *Trends in Cognitive Sciences*, *3*(7), 254–264.

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, *13*(11), 470–478.

van Linden, S., Stekelenburg, J. J., Tuomainen, J., & Vroomen, J. (2007). Lexical effects on auditory speech perception: An electrophysiological study. *Neuroscience letters*, *420*(1), 49–52. doi:10.1016/j.neulet.2007.04.006.

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1483–1494.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science*, *27*(2), 285–298.

Magnuson, J. S., Tanenhaus, M. K., & Aslin, R. N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition*, *108*(3), 866–873.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, *69*, 546–558.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63.

Massaro, D. W. (1979). Letter information and orthographic context in word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 595–609.

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

Massaro, D. W., Cohen, M. M. (1983). Phonological context in speech perception. *Perception & 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.

Massaro, D. W., & Klitzke, D. (1979). The role of lateral masking and orthographic structure in letter and word perception. *Acta Psychologica*, *43*, 413–426.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*(7), 293–299.

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

McClelland, J. L. (1985). Putting knowledge in its place: A scheme for programming parallel processing structures on the fly. *Cognitive Science*, *9*, 113–146.

McClelland, J. L. (1986). The programmable blackboard model of reading. In J. L. McClelland, D. E. Rumelhart, & the PDP research group. *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume II (pp. 122–169). Cambridge, MA: MIT Press.

McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention & performance XII: The psychology of reading* (pp. 1–36). London: Erlbaum.

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

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford, England: Oxford University Press.

McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, *4*, 503. doi:10.3389/fpsyg.2013.00503.

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

McClelland, J. L., & Johnston, J. C. (1977). The role of familiar units in perception of words and nonwords. *Perception & Psychophysics*, *22*, 249–261.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception, I: An account of basic findings. *Psychological Review*, *88*(5), 375–407.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The Appeal of Parallel Distributed Processing. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume I. (pp. 3–44). Cambridge, MA: MIT Press.

McGuire, P. K., Silbersweig, D. A., & Frith, C. D. (1996). Functional neuroanatomy of verbal self-monitoring. *Brain*, *119*, 907–917.

McMurray, B., & Aslin, R. N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, *6*(2), 203–229.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language*, *61*(1), 1–18.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, *13*(6), 958–965.

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science*, *32*(2), 398–417.

Mitterer, H., & de Ruiter, J. P. (2008). Recalibrating color categories using world knowledge. *Psychological Science*, *19*(7), 629–634.

Molinaro, N., Duñabeitia, J. A., Marìn-Gutièrrez, A., & Carreiras, M. (2010). From numbers to letters: Feedback regularization in visual word recognition. *Neuropsychologia*, *48*(5), 1343–1355.

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., & Milroy, R. (1992). Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 452–467.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics*, *66*, 241–251.

Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: An fMRI investigation. *Cerebral Cortex*, *18*(2), 278–288.

Newman, R. S., Sawusch, J. R., & Luce, R. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 873–889.

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299–370.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

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

Pattamadilok, C., Morais, J., De Vylder, O., Ventura, P., & Kolinsky, R. (2009). The orthographic consistency effect in the recognition of French spoken words: An early developmental shift from sublexical to lexical orthographic activation. *Applied Psycholinguistics*, *30*, 441–462.

Pattamadilok, C., Perre, L., Dufau, S., & Ziegler, J. (2008). On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience*, *21*(1), 169–179.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of AAAI-82*. (pp. 133–136).

Perre, L., Midgley, K., & Ziegler, J. C. (2009). When beef primes reef more than leaf: Orthographic information affects phonological priming in spoken word recognition. *Psychophysiology*, *46*(4), 739–746. doi:10.1111/j.1469-8986.2009.00813.

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain research*, *1188*, 132–138. doi:10.1016/j.brainres.2007.10.084.

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.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *21*(4), 1037–1052.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory & Language*, *39*(3), 347–370.

Pitts, W., & McCullough, W. S. (1947). How we know universals: The perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, *9*, 127–147.

Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: A critique of Bowers' (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review*, *117*, 284–288.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extraclassicalreceptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–499.

Reddy, D. R., Erman, L. D., Fennell, R. O., & Neely, R. B. (1973). The hearsay speech understanding system: An example of the recognition process. In *Proceedings of the 3rd international joint conference on Artificial Intelligence* (pp. 185–194). San Francisco, CA: Morgan Kaufmann.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness ofg stimulus material. *Journal of Experimental Psychology*, *81*, 274–280.

Rizzolatti, G., Riggio, L., Dascola, I., & Umlita, C. (1980). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, *25*(1a), 31–40.

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *Journal of Psychophysiology*, *50*, 19–26.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and Performance VI*. (pp. 573–603). Hillsdale, NJ: LEA.

Rumelhart, D. E., & McClelland, J. L. (1981). Interactive processing through spreading activation. In C. Perfetti & A. Lesgold (Eds.), *Interactive processes in reading* (pp 37–60). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94.

Rumelhart, D. E., & Siple, P. (1974). The process of recognizing tachistoscopically presented words. *Psychological Review*, *81*, 99–118.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General*, *110*(4), 474.

Samuel, A. G. (1986). Red herring detectors and speech perception: In defense of selective adaptation. *Cognitive Psychology*, *18*(4), 452–499.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *32*(2), 97–127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–351.

Samuel, A. G., & Kat, D. (1996). Early levels of analysis of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 676–694.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language*, *48*(2), 416–434.

Schroeder, C. E., Lindsley, R. W., Specht, C., Marcovici, A., Smiley, J. F., & Javitt, D. C. (2001). Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology*, *85*, 1322–1327.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*(4), 489–537.

Sherman, G. (1971). The phonemic restoration effect: An insight into the mechanisms of speech perception. Unpublished master's thesis, University of Wisconsin, Milwaukee.

Smolensky, P. (1986). Neural and conceptual interpretation of PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP research group. *Parallel distributed processing: Explorations in the microstructure of cognition*. Volume II, (pp. 390–431). Cambridge, MA: MIT Press.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *The Journal of Neuroscience*, *32*(25), 8443–8453.

Spivey, M., & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1521–1543.

Spratling, M. W., & Johnson, M. H. (2004). A feedback model of visual attention. *Journal of Cognitive Neuroscience*, *16*(2), 219–237.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech (L). *Journal of the Acoustical Society of America*, *114*(6,Pt1), 3036–3039.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 632–634.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.

Warren, R. M., & Warren, R. P. (1971). Some age differences in auditory perception. *Bulletin of the New York Academy of Medicine*, *47*(11), 1365.

Wilson, T. W., Leuthold, A. C., Lewis, S. M., Georgopoulos, A. P., & Pardo, P. J. (2005). Cognitive dimensions of orthographic stimuli affect occipitotemporal dynamics. *Experimental Brain Research.*, *167* (2), 141–147. doi:10.1007/s00221-005-0011-4.

Interactive Activation and Mutual Constraint Satisfaction in Perception and Cognition

James L. McClelland and Daniel Mirman and Donald J. Bolger and Pranav Khaitan

知覚と認知における相互作用的な活性化と相互制約充足

要旨

ルメルハートは 1977 年に発表した重要な論文で，知覚には複数の情報源を同時に利用することが必要であり，知覚者は情報が到着するとリアルタイムに様々な表現レベルで感覚情報を最適に解釈することができると主張した。これは，知覚や理解のメカニズムが，単純な処理ユニット間の活性化の双方向の伝搬を通じて実行される，相互活性化過程を利用しているという考えであり，ルメルハートの議論に基づいている。次に，この仮説を探る初期の試みとして，文字・単語知覚の相互活性化モデルと音声知覚の TRACE モデルを取り上げ，それらの仮定と予測に関連する実験的証拠を検討する。これらのモデルが，知覚の問題がもたらす計算上の課題にどの程度対応しているかを検討し，行動実験から得られた証拠とどの程度一致しているかを検討する。相互的な計算と最適なベイズ推論との関係をめぐる論争を含め，相互処理という考えをめぐる経験的・理論的な論争を検証する。初期のバージョンの相互活性化モデルの実装の詳細は，最適性や人間の成績データの側面からの逸脱を引き起こした。しかし，最近のモデルでは，これらの欠点を克服している。これらのモデルの中には，多項相互活性化モデルと呼ばれるものがあり，相互活性化とベイズ計算を明示的に結びつけている。また，神経生理学や神経画像学の研究から，相互的処理が脳内の知覚処理機構の特徴であることを裏付ける証拠を紹介する。以上のことから，我々は，計算論的分析，行動学的および神経科学的証拠のすべてが，「相互活性化仮説」を支持していると主張する。これらの証拠は，相互活性化という考えに基づいた現代版のモデルが，知覚過程の完全な理解を達成するための努力の基礎となり続けることを示唆している。

キーワード: 知覚，相互活性化，並列分散処理，コネクショニストモデル，最適知覚推論，ニューラルネットワーク

1. 導入 Introduction

並列分散処理 (PDP) 枠組みの基礎となる概念の一つに，相互活性化がある。相互活性化は，相互制約充足の概念と同義である。一般原理として，知覚や言語などの心的表現は，単純なニューロンのような処理単位間で活性化が双方向に伝搬することで生じるという考え方である。この概念は，文字や単語の知覚の相互活性化 (IA) モデル (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1981, 1982) や，音声知覚の TRACE モデル (McClelland & Elman, 1986) の中心的なものであった。これらのモデルでは，単語と文字や音素，文字と文字特徴，音素とその特徴など，全体と部分を表すユニット間の双方向的相互作用に焦点が当てられていた。これらのモデルでは，単語，文字，音素，特徴などの明示的に列挙可能な知覚単位を表すために，個々のニューロンのような処理単位が割り当てられていた。この処理ユニットは，対応する認知ユニットに特化したニューロンの集団を表していると考えられる (Bowers, 2009)，我々は違う見方をしている。Smolensky (1986) の提案に沿って，モデル内の処理ユニットは，それぞれが多くの異なるアイテムの表現に関与するニューロンの集団の活動の代替パターンとして符号化された情報状態を表している (Hinton, McClelland, & Rumelhart, 1986; Plaut & McClelland, 2010)。IA モデルは，このような状態の時間変化と内容を追跡する。これは，基礎となる神経活動の完全な複雑さを，Smolensky が概念的な表現空間と呼んだものに投影するのに役立つ。

以下に詳述するように，相互活性化モデルの経験的な動機は，実験に次ぐ実験で，視覚，聴覚，その他の入力のいずれかの要素や側面の識別や解釈が，入力の他のすべての要素や側面の識別や解釈に影響されるという観察である。これに対応して，最適な知覚的解釈の理論のレベルでの動機付けがある。一般的に，知覚記述のどのレベルにおいても，入力を解釈するための直接的な感覚的証拠は，単独で考えると結論が出ないことがあり，各要素の最も可能性の高い解釈は，すべての要素の解釈と多くの証拠源を一緒に考えたときにのみ決定される。実際，個々の要素の証拠が非常に曖昧であっても，すべての要素の 1 つの一貫した解釈が，証拠の全体性によって強く決定されることがよくある (図1a)。

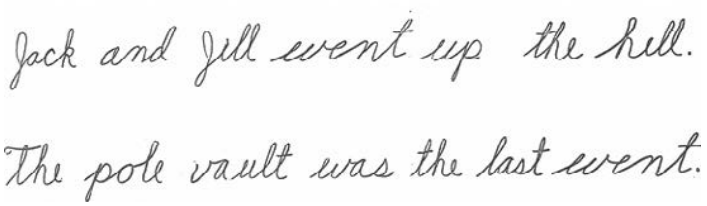


図1. 上: 個別には解釈できない滲みの集合体からダルメシアン犬が出てくる。 James (1965)より。Copyright © Ronald C. James, reprinted with permission.
下: 手書きで書かれた1文目の “went”と2文目の “event “は同じだが, 2 つの異なる文脈の中では異なって認識される。 Rumelhart (1977) p. 579, fig.3 より。Copyright © Taylor and Francis Group, reprinted with permission.

知覚と理解の領域を超えて, 複数の同時制約が, 文脈的に適切な行動の側面の選択や記憶の再構成など, 認知の他の多くの側面に適用される。 同様に, 目標や課題要求は, 知覚, 解釈, 記憶, 行動に統合される追加的な制約となり, 処理の結果に影響を与え, しばしば影響を受ける。
PDP 編の第 1 章 (McClelland, Rumelhart, & Hinton, 1986) では, 行動選択, 問題解決, 記憶など, 認知処理の他のすべての領域で, これらの同じ考慮事項が生じることを論じている。

このように知覚や認知のすべての側面に並列分散処理が関わっているという考え方は, 知覚や認知のモジュール化アプローチに代わるものである。 並列分散処理では, 特定の脳領域にある特定のニューロンや神経集団が, ある種の情報を表現するために特化されている可能性があるため, 情報のある種の区分けが残っている。 しかし, すべての情報源が同時に他の情報源を制約するためには, そのようなニューロンの特定のアンサンブルが活性化する結果は, 複数の脳領域の神経集団に分散された処理の結果であると考えられる。 例えば, 視覚提示された単語の視覚的, 意味的, 聴覚的, および口音的な側面を表現するための脳領域がある一方で, 関与する全脳領域のニューロンの活性化は次のように考えられる。 関与する全脳領域のニューロンの活性化は, 相互活性化／相互制約充足の枠組みの中で, 相互に依存していると考えられる。

1.1. 相互活性化モデルの前身 Precursors to interactive activation models

知覚と理解の相互的アプローチの動機は, Rumelhart (1977) の論文にある。 Rumelhart は, 文字, 音素, 単語の知覚における文脈の役割 (Fig.1b) や, 話し言葉や書き言葉, 文章の統語的・意味的解釈の曖昧さを解消するための様々な情報源の利用について, 19 世紀にさかのぼって既存のデータを検討した。 ルメルハートは, 知覚と理解の目的を, 代替仮説の事前確率と代替仮説間の条件付き確率関係の知識に導かれた相互制約充足過程を通じて, 多くの異なる表現レベルで入力の共同解釈を見つけることだと考えた。 さらに, ルメルハートは, このような解釈を決定する処理がどのように行われるかを想定している。 初期の音声知覚の人工知能モデルである Hearsay (Reddy, Erman, Fennell, & Neely, 1973) からヒントを得て 「メッセージセンター」または「黒板」と呼ばれるデータ構造を想定した。 これは, 入力の解釈に含まれる可能性のある要素の確率の推定値が「チョークで書き込まれ」専門のエキスパートがそれぞれ黒板の内容を並行して作業し, 検査や調整を行うことができる。
例えば, 文字入力の場合, ある単語の特定の位置にある文字が A である可能性の推定値は, 先行する C と後続する T の情報と, Cに続いて A と T が続くとおなじみの単語 CAT になるという語彙情報を用いた語彙専門家によって高められるかもしれない。 また, 参加者が猫の絵が描かれた写真を見たばかりであれば, 語彙レベルの CAT 仮説はさらに強化されるかもしれない。 このモデルは, Rumelhart and Siple (1974) による初期のモデルを参考にしている。 このモデルは, 単語と文字の確率, および単語に文字が与えられた場合の条件付き確率の知識に基づいて, 3 文字の配列を表示したときの文字の識別に関するデータを説明している。

2. 相互活性化モデルが取り組む計算論上の課題

ラメルハート (1977) の議論は, 知覚や言語理解において直面する計算上の課題について, 次のように述べている。

- **最も確率の高い解釈の探索**. 知覚や言語理解とは, 書かれたり話されたりした入力に対して, さまざまな表現レベルで最も確率の高い解釈を模索する処理である。 例えば, 書かれた言語表現や話された言語表現の解釈は, 存在する視覚的または聴覚的な特徴, 文字や音声, 単語, 句, 文章, そしてこれらの項目の意味や構文構造を表す。 この処理の目的は, 全体として最も確率の高い解釈を見つけることである。

- **事前知識と文脈の活用**. 曖昧さとノイズが普遍的に存在するため, 知覚入力の任意の側面の正しい解釈を見つける確率を最大化するには, 事前知識と, 表現自体の隣接要素, 事前入力, 付随する視覚情報などの他の領域からの入力を含む文脈からの情報を活用することが必要である。

Rumelhart (1977) は強調しなかったが, 我々は知覚と理解のモデルに以下の重要な実時間制約を加える。

- **実時間処理の制約**. 知覚と理解は可能な限り迅速に結果を出さなければならず, すべての異なるタイプの情報が利用可能になったときに, 他のすべてのタイプの情報の解釈に影響を与えることができるようにしなければならない。

この制約を知覚的推論の問題の定式化に取り入れることは, 典型的な計算レベルの定式化 (Feldman, Griffiths, & Mrogan, 2009; Marr, 1982) とは異なり, 入力と結果のみを考慮し, 結果を計算するために必要な時間や処理手順を考慮しない。 しかし, 時間が貴重であることは明らかであり, ダイナミックな世界では, 迅速に理解 (および行動) できなければ, 機会損失や時には大惨事につながる可能性がある。
このように, 実時間でできるだけ早く結果を出すことは, 知覚システムが直面している計算論レベルの課題の一部である。 計算機レベルの出発点に立つ研究者たちは, この問題の重要性を考え始めている (Norris, 2013; Vul, Goodman, Griffiths, & Tenenbaum, 2014)。

2.1. 実時間での最適な知覚的推論の近似値としての人間の知覚と理解

以上の記述は, 知覚と理解のシステムが解決しなければならない計算上の問題を特徴づけている。 次の命題は, 人間の知覚と理解の機構は, これらの計算上の問題に対処するために組織されているということである。

- **人間は最適な実時間知覚推論を近似する** 人間の知覚者は, 知覚と理解の最適なシステムから期待される行動パターンに近似する。 文脈と事前の知識を利用して知覚と理解を導き, 入力が実時間で利用可能になると, 外部入力のすべてのソースの影響を解釈のすべての側面に反映させる。

速度と精度には, 神経ハードウェアの特性による限界があり, 人間が最適性に近い状態を実現できる範囲に影響を与える。 また, 最適化のためには経験が必要であり, 練習や経験によって速度と精度が徐々に向上していくことがわかっている。 経験の結果として, 知覚世界の統計的構造の学習, この構造を利用するための知覚およびその他の認知システムの調整, 成績 (到達, 達成) をサポートするための脳資源 (ニューロンおよびシナプス) の割り当てが行われる。 本稿では, 経験に依存した最適化がすでに行われていると仮定して, 熟練した大人が母国語の話し言葉や書き言葉を知覚・理解することに焦点を当てる。

2.2 相互活性化仮説 The interactive activation hypothesis

上述の問題提起と人間の成績の特徴付けは, 広く受け入れられているように思われるが, 人間の知覚者が文脈や事前の知識を効果的に利用することに成功する機構を特徴付けるため, いくつかの別のアプローチが取られている。 本稿では, 次のような仮説を検討する。

- **相互活性化仮説** 脳内で双方向に接続された神経ネットワークに知覚やその他の認知処理が実装されていることは，知覚系が直面している重要な計算上の課題を解決する機構であり，人間の成績が実時間で最適な知覚的推論に近似していることの原因となっている。

以下では，知覚における対話型モデルの研究の歴史について説明する。初期の IA モデルと TRACE モデル，そしてそれらの基本的な仮定に関連する実験的な証拠を調べる。上記の計算上の課題にどの程度対応しているか，また，行動実験から得られた証拠とどの程度一致しているかを検討する。相互的計算と最適ベイズ推論との関係を巡って巻き起こった論争を含め，相互的処理という考えを取り巻く経験的・理論的な論争を検証する。また，知覚の神経基盤に関する神経生理学的および神経画像学的研究から得られた証拠を概観する。我々の結論を予想すると計算論的分析，行動学的・神経科学的証拠は，すべて相互活性化仮説と一致している。これまでも，そしてこれからも，別のアプローチを主張する人たちはいるだろうが，証拠からは，これらの考えに基づいた現代版のモデルにはかなりの利点があると考えられる。本稿の最後に，この結論を再検討し，相互的なアプローチが今後どのように発展していくのかを考えてみたい。

3. 相互活性化モデルと TRACE モデル

IA 仮説を検証するためには，その仮定を具現化した明示的なモデルを開発するとともに，これらのモデルを分析してその特性を理解し，人間の行動パターンをどの程度説明できるかを検討する必要がある。文字や単語の知覚の相互活性化モデル (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1981, 1982) と，その子孫である音声知覚の TRACE モデル (McClelland & Elman, 1986) は，このような研究プログラムの最初のステップであり，主にデータのパターンをモデル化することに焦点を当てていた。

文字と単語の知覚の I Aモデルは，4 つの表示位置のうちの1つに表示された文字を，単独で，または他の位置にある隣接する文字とともに知覚することを扱っている。ニューロン様処理ユニットの位置固有のプールは，特徴レベルと文字レベルに配置されており，単語レベルは入力位置の配列に広がっている (図2a)。隣接するレベルの互いに矛盾しないユニット間には双方向の興奮性接続があり，各プール内のユニット間には双方向の抑制性接続がある。刺激が提示される前に，すべてのユニットの活性化値は 0 よりわずかに低い休止レベルに設定されている。外部からの入力が提示されると，特徴ユニットが駆動し，一貫性のある文字ユニットが活性化され，一貫性のない文字ユニットが抑制される (脚注 1)。文字ユニットは一貫性のある単語ユニットを活性化し，互いに競合するとともに，可能性のある単語に一致する文字をサポートするフィードバックを送る。この処理を 図2b の曖昧な入力に適用した場合，活性化の時間経過を示し，数回の処理サイクルで文脈的に最も可能性の高い解釈を見つけることができることを示している。4 番目の位置にある特徴的な入力は，R でも K でも同じように一致するが，K だけが文脈上の文字で単語 (WORK) を作る。この単語は，他の競合する単語候補を抑制し，K にトップダウンのサポートを提供し，K は競合により R を抑制し，文字レベルと単語レベルの両方で入力の一貫した解釈がある状態になる。

脚注1 オリジナルの IA モデルでは，レベル間の抑制と興奮が採用されていたが，TRACE モデルをはじめとするその後のモデルでは，レベル間の接続は興奮のみとし，抑制はレベル内の相互作用に限定した。レベル間の抑制を排除した最大の理由は，より良い解釈がない場合に，適合性の低い解釈であっても有効とするためである。この問題については，後述の多項相互活性化モデルの説明で触れる。

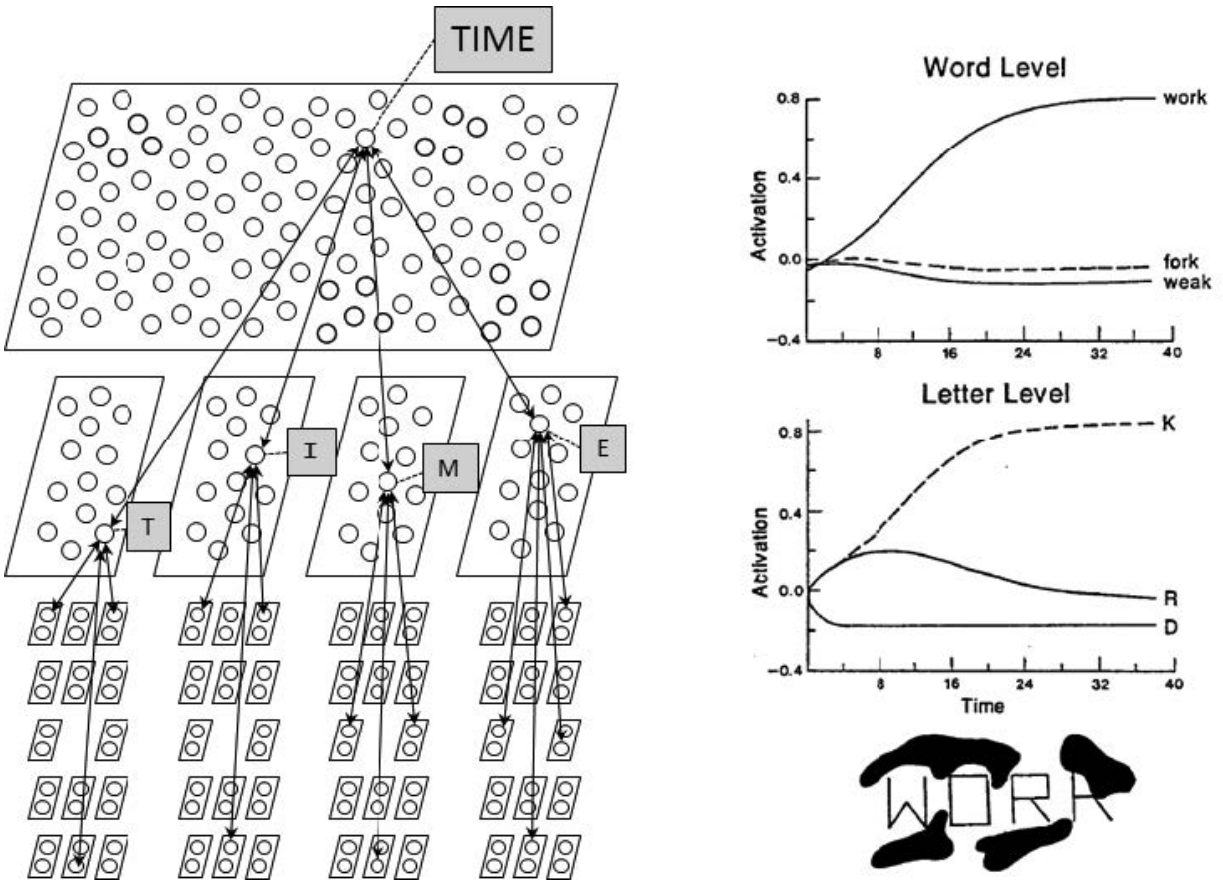


図2. (a) 対話型活性化モデル。単語，4 つの位置にある文字，同じ位置にある特徴に対応するユニットのプールを示す。TIME という単語と，その単語に含まれる文字や特徴に対応する興奮性の接続が示されている。各プール内のユニットは相互に抑制的であるが，抑制的な接続は描かれていない。特徴レベルでは，ユニットは 2 つのユニットからなるプールに編成されており，1 つは可能な各特徴の有無に対応し，もう 1 つは非対応である。McClelland (2013)のp.14、fig.6 より転載。Copyright ©James L. McClelland, reprinted with permission.
(b) 対話型活性化モデルのオリジナル版で，図の下に示したディスプレイを提示した後の，4 番目の位置にある文字ユニットと単語ユニットの活性化の時間経過。最後の位置にある見えているセグメントは，K と R という文字と同じように一致しており，他の文字とは一致していない。単語レベルでは，WORK という 1 つの既知の単語だけが，4 つの位置のそれぞれの活性な文字と一致している。この単語

は，Kのユニットをサポートするためにフィードバックされ，4番目の位置でRを支配している。McClelland ら (1986) p.23, 図 8より転載。

3.1 相互活性化モデルの詳細

IA と TRACE モデルについて詳細を述べる。両モデルは，相互活性化と最適推論についての後述する議論に関連する。活性化過程により，Grossberg(1978) の提案を修正して原著で定式化されたように，文字ユニット，単語ユニットへ絶えず変動する活性値を割り当てる。原理的に，この処理過程は完全に連続過程であり，シミュレーションでは，細分化された時間ステップによって近似される。書く時間ステップでは，ユニットとそのネット入力がまず計算される。これは，そのユニットへ照射する全ユニットについて総和であり，送信ユニット掛ける結合重みの値，プラス，直接の外部入力である。

$$\text{net}_i = \sum \left[a_j\right]^+ w_{ij} + e_i. \tag{1}$$

興奮性と抑制性の重みの強さは，特徴から文字，文字から単語，単語から文字，そして層内の影響についてそれぞれ別のパラメータで決定した。$\left[a_j\right]^+$ という表記は，あるユニットの活性化値が 0 より大きい場合にのみ伝搬されることを示している。

各ユニットへネット入力が到達すると，活性値が次式に従って調整される:

$$\begin{aligned} \Delta a_i &= \text{net}_i(1 - a_i) - d(a_i - r) && \text{if net}_i \geq 0 \\ \Delta a_i &= \text{net}_i(a - m) - d(a_i - r) && \text{otherwise} \end{aligned}$$

これらの式は，正の純入力が活性化を最大値の 1 に向かって押し上げる一方で，負の純入力が活性化を最小値m に向かって押し下げる過程を表している。各式の最右項は，活性値を休止レベル r に引き寄せる傾向のある，減衰または漏出過程に対応すると考えられる復元力を実装しており，パラメータ d はこの傾向の強さを表している。

このモデルでの処理は完全に決定論的である。成績が確率的である知覚実験における人間の成績に対応するために，予測された応答確率は，結果として得られた活性化値の連続的な平均値に Luce 選択則を適用することによって導出され，選択肢 i を選択する確率は次式で与えられる:

$$p(r_i) = \frac{e^{g\bar{a}_i}}{\sum_j e^{g\bar{a}_j}}, \tag{2}$$

例えば，図 2a のように 4 番目の位置で文字 k を選択する確率は，4 番目の位置で文字 k に対応するインデックス i をセットすることで計算される。インデックス j は，同位置の i を含む全文字について実行される。g はスケールパラメータである。$\bar{a}_i$ は移動平均に対応する。McClelland and Rumelhart 1981 のモデル実験では，刺激提示後の最高応答確率である。

TRACE モデルは，IA モデルのアイデアを音声ストリームの処理に拡張したもので，より多くの位置固有の特徴および文字ユニット配列と，位置に合わせた単語ユニットの対応するバンクを仮定しており，各位置のすべての特徴および音素に対応するユニットと，各位置から始まるすべての単語に対応するユニットが存在する。音声入力がリアルタイムで順次到着すると，音声入力の各連続した時間サンプルが次の入力位置に向けられる。このようにして，文字認識の IA モデルと同じ双方向の活性化処理過程を，1 つまたは数個の単語に対応する音声入力の処理に適用することができた。このアーキテクチャにより，音素レベルおよび単語レベルの制約を，入力ストリームのどの位置にあるかに関わらず，入力サンプル系列に適用することが可能となった。TRACE モデルの構造は，神経メカニズムに関する文字通りの主張としてではなく，音素レベルと単語レベルの情報の間の絶対的ではなく相対的な制約を捉えた高レベルの特徴と見なすべきである。ある時刻に /k/ があれば，同じ時刻に cat という単語が始まることも，2 音前に ticket という単語が始まることも (他の多くの可能性も含めて) 支持され，これらの制約は，対応する位置にある対応するアイテムのユニット間の接続に取り込まれる (脚注2) このようなユニットの配列の活性化は，音声入力の処理結果の動的な記憶トレースを形成し，それがモデルの名前の由来となった。また，ほぼ同時期に開発されたモデル (McClelland, 1985, 1986) では，ユニットや接続を再複製することなく，神経ハードウェアがこれらの計算をどのように実装するかが検討された。

脚注 2 簡略化のため，IA と TRACEのモデルでは，文字と音素のスロットを個別に想定しているが，TRACE では，隣接するスロットの間で音素が重なるような広がりを想定している。Norris (2013) によれば，どちらのモデルも位置の不確実性を考慮すべきであり，適切な位置に近い文字でも対応する単語レベルのユニットを活性化することができる (例えば，TRCK は TRXY よりもはるかに多く単語TRUCK を活性化するはずである)。

4. 行動学的証拠

4.1. IAモデルとTRACEモデルの経験上の焦点

IA モデルとTRACE モデルは，文字と音素の認識を対象としており，文字や音声の認識における単語の文脈の影響を示す多くの関連データに対応している。初期の行動学的証拠の多くは，単語の優位性効果の研究としてまとめられる。文字は，単語の中で提示された方が，単独で提示された場合や文字のランダムな配列で提示された場合よりも，より正確に認識される (例えば Reicher, 1969)。また，これらのモデルは，曖昧な視覚入力や音声入力は，周囲の語彙的な文脈に合致した文字や音素として認識される可能性が高いという一般的な知見にも対応している (例えば Ganong, 1980; Massaro, 1979)。例えば Ganong (1980) は /k/ と /g/ の間の曖昧な音が “_iss “ の文脈では /k/ と識別されやすく、”_ift” の文脈では /g/ と識別されやすいことを示した。単語の中の文字に対する優位性は，発音可能な単語のような擬似単語 (LEAT や TOVE など McClelland & Johnston, 1977) の中の文字にも及ぶ。単語知覚の IA モデルは，LEAT のような擬似単語の文字が，単語らしくない非単語 (LTAE など) の文字や，文脈なしに提示された単一の文字よりも正確に知覚されるという機構を説明する新しいモデルであった。このような単語は，与えられた入力の近隣語と呼ばれる。これらの単語ユニットは，構成する文字のユニットに支持をフィードバックし，その多くは複数の異なる単語の活性化によって部分的に支持される (図3)。Newman, Sawusch, and Luce (1997) は，この説明と一致するように，曖昧な音声セグメントの識別における近隣効果を実証した。IA モデルでは，発音できない文字列に含まれる文字であっても，多くの単語の「近接語」がある場合 (例えば SLNT の L) は，同等の発音可能な文字列 (SLET) に含まれる文字と同じくらいの促進効果を示すと予測されており，Rumelhart and McClelland (1982) で報告された実験では，この予測が確認された。

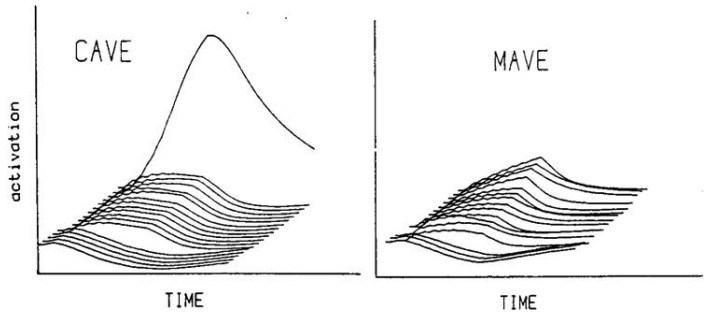


図3. CAVE と MAVE による対話型活性化モデルでの単語レベルの活性化。 処理中の任意の時点で活性化が 0 を超えたすべてのユニットの活性化を示す。 活性化の痕跡は空間的にオフセットされており， 最大活性化が高いものほど後ろから右に向かっている。 MAVE の場合，4 つの文字位置のそれぞれにおいて， 複数の単語が提示された文字をトップダウンで支持する。McClelland and Rumelhart (1981)の fig.13, p.396 および fig.9, p.393 より。

IA モデルとTRACE モデルの相互活性化処理から， 文脈効果は標的となる入力要素の後に来る文脈要素でも、標的の前に来る要素でも起こりうることが予測される。 この予測は， 各文脈文字の持続時間を個別に操作し， 各文字位置での標的文字の認識に対する影響を調べた実験で検証され， 確認された (Rumelhart & McClelland, 1982)。 一般に， すべての文脈文字が各標的文字の認識精度に影響を与える。 同様に， 音素認識における語彙効果は， 単語の初頭音素だけでなく， 埋め込み音素や単語の最終音素に対しても生じ(Ganong, 1980; Warren, 1970)， その効果は後続単語の文脈情報にまで及ぶ研究もある (Sherman, 1971; Warren & Warren, 1971)。 もちろん， 音素識別課題で知覚者が曖昧なセグメントの直後に応答することを要求された場合， 後続文脈はほとんど影響しない (Fox, 1984)。 TRACE モデルでは， 音声認識における様々な現象 (音声の流れを単語に分割する語彙ベースの分割， 知覚的磁石効果(Kuhl, 1991) など) にも対応している。

4.1.1 実時間処理制約への人間の適合性の証拠

TRACE モデルを開発する動機となった現象の一つに， 音声知覚中に単語の識別が実時間で行われるという見解を裏付ける証拠があった。 Marslen-Wilson らはこの点に最初に着目し， 音声入力が 1 つの可能性のある単語と一意に一致するようになった直後に識別が行われることを示した (Marslen-Wilson & Welsh, 1978)。 その後， 単語と絵のマッチング課題での眼球運動を調べた多くの研究により， 文脈と刺激の情報が実時間での処理を相互に制約するという一般原則が支持された。 これらの研究の中には， 1977 年に Rumelhart が発表した論文で想定されていたように， 非言語的な視覚入力と音声的な聴覚入力の両方を含むものがある。 この方法を用いた最初の実験では， 視覚的な文脈が， 統語的に曖昧な前置詞句の即時解釈に影響を与えることが示された (Chambers, Tanenhaus, & Magnuson, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995)。 その後の研究では，統語的・意味的な期待が， どの語彙候補を考慮するかを制約することも示された。 例えば， Dahan and Tanenhaus (2004) は， climb という動詞を聞いたときに， その動詞で指定された動作の対象となりうるものを除外することを示した。 重要なことは， これらの文脈的な影響は， 制約となる情報が提示された後すぐに明らかになり （Dahan & Tanenhaus, 2004; Magnuson, Tanenhaus, & Aslin, 2008)， 新しい情報が利用可能になると継続的に更新されることである。 このことは， Allopenna, Magnuson, and Tanenhaus (1998) の研究で特に明確に示されている。 さらに， 入力のオンセットと一致しない単語候補であっても， その後に十分な音韻入力があれば活性化されることが示された。 これらの論文の多くは TRACE モデルや， 同様の仮定に基づいた簡略化されたモデルを用いて， その結果をシミュレーションしている (Spivey & Tanenhaus, 1998)。

4.2. 文脈効果の一般性を示す証拠

文字や音素の認識における単語の文脈効果は， 対話型処理に関する研究の主要な焦点となってきたが， この原理は非常に一般的で， 知覚や認知のさまざまな領域で繰り返されている。 例えば， 単語認識と同様に， 音声生成における音韻誤りは， 非単語ではなく既存の単語になる傾向があり， このような効果は音声生成の相互化モデルでよく説明される (Dell, 1986; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997; Rapp & Goldrick, 2000 も参照)。

相互的処理は， 視覚的な物体知覚においても重要な役割を果たす。 言葉優位性効果と同様， 曖昧な色の知覚は対象物の文脈によって偏ることがある (Hansen, Olkkonen, Walter, & Gegenfurtner, 2006; Kubat, Mirman, & Roy, 2009)。 例えば， 黄色とオレンジの中間色の曖昧な色は， スクールバスの文脈ではより黄色に， ニンジンの文脈ではよりオレンジに知覚される。 さらに， 後述する Elman and McClelland (1988) の結果と同様に， Mitterer and de Ruiter (2008) は，物体-文脈フィードバックが色カテゴリを再調整することを示した。 また， Kanizsa図形における有名な錯視的輪郭現象 (Kanizsa, 1979; 図4) は， 相互活性化から予想されるように， 単純な図形の文脈でも， 入力からは全く見えない輪郭を知覚させることができることを示している。

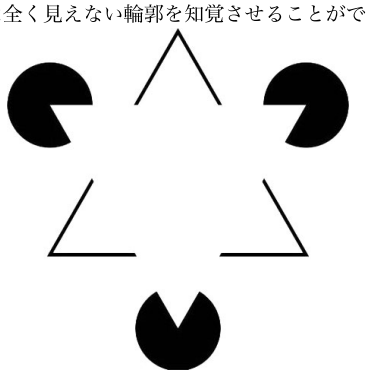


図 4. カニッツァ三角形の錯視的輪郭。 画像の出典: Kanizsa triangle, Wikimedia Commons, http://en.wikipedia.org/wiki/File:Kanizsa_triangle.svg. Copyright © Wikipedia Commons. Reprinted under the GNU Free Documentation License

より高度な現象に目を向けると， Rumelhart (1977) が予測したように， 文脈が語彙の曖昧さの解決に影響を与えることは長年にわたって明らかにされてきた。 文脈効果をアクセス後の選択過程に限定したモデルもあるが (Swinney, 1979)， 相互型モデルでは， 文脈が十分に制約されていれば， 曖昧な単語のどの意味が最初に活性化されるかが制約され (Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982)， さらには入力が提示される前に事前活性化が引き起こされると予測されている (McClelland, 1987)。 アイトラッキング研究では， 成人 (Altmann & Kamide, 1999; Magnuson et al., 2008; 前述の Dahan & Tanenhaus, 2004 の研究

も参照) や乳児 (McMurray & Aslin, 2004など) の言語処理において，このような予期効果が明らかにされている。 また，頭皮の電気生理学的記録 (ERP) は，文脈によって単語が事前に活性化されることを示唆している (van den Brink, Brown, & Hagoort, 2001; DeLong, Urbach, & Kutas, 2005）。

文脈効果の研究では， 文字や音素などの要素 (あるいはエッジや色) の知覚が， その直前の文脈 (単語や物体など) によってどのように影響されるかに注目してきた。 しかし， 処理は， 課題の指示や異なるタイプの刺激の相対的な確率など， 他の文脈的要因によっても影響を受ける。 例えば， 知覚実験において， あるブロックの試行における非単語の割合が相対的に高い場合， 語彙的文脈効果は減少する。 具体的には， 非単語の割合が高いと， 非単語に比べて単語の音素を認識する際の速度の優位性が低下し(Mirman, McClelland, Holt, & Magnuson, 2008)， 発話エラーの単語バイアスが低下し (Hartsuiker, Corley, & Martensen, 2005)， 非単語に比べて単語の短期記憶の優位性が低下し (Jefferies, Frankish, & Ralph, 2006)， 文字と音の対応関係が一致しない単語を読む際の規則化エラーが増加する (例 Monsell, Patterson, Graham, Hughes, & Milroy, 1992)。 これらの結果は， 語彙表現 (あるいは意味表現) の活性化が減少したために， 単語の表現があまり活性化されず， その結果， フィードバック効果が小さくなったと解釈できる (TRACEにおけるこれらの効果の実装については Mirman et al. 2008)。

注意による処理の調節は， 双方向に接続された処理ユニットのネットワーク， すなわち相互活性化ネットワークで実現することができる。 このようなモデルの一例として， Ericksen の脇添え課題 (flanker task) における注意による処理の変調モデルがある(Cohen, Servan-Schreiber, & McClelland, 1992)。 このモデルでは， 異なる空間的位置を表すユニットは， その位置の特徴を表すユニットと双方向に接続され， さらにこれらのユニットは， 代替可能な標的文字の同一実体の位置に依存しないユニットと双方向に接続されている。 ある場所に注意を向けることは， その空間的位置を表すユニットがトップダウンで活性化されることで生じると考えられている。 これにより， 対応する位置の特徴を表すユニットの活性化が促進され， その後の処理で最終的に優位に立つことができるが， それにもかかわらず， 一貫性のない脇添え刺激 (フランカー) からの活性化が標的の位置にある項目の識別を遅らせることができる (例えば，Gratton, Coles, Sirevaag, Eriksen, & Donchin, 1988 などの実験で観察されている)。 Cohen らのモデルの一部の実装では， すべての接続が双方向ではないようにアーキテクチャを単純化しているが， 我々は， 場所や刺激の特徴への注意は， 顕著な入力， 目標や課題要求が注意の焦点を決定するのに関与するような活性化の双方向の伝播を伴うことを当然のことと考えている (Phaf, Van der Heijden, & Hudson, 1990)。 さらに， 背側 (行動) と 腹側 (物体) の処理系が相互に作用するという最近のモデルでは， 相互的処理によって， ディスプレイに表示されている 2 つ以上の物体を同時に識別することができることが示されている (Henderson & McClelland, 2011)。

最後に， 記憶においても， 相互活性化過程が重要な役割を果たしている可能性があることを指摘しておく(Kumaran & McClelland, 2012; McClelland, 1981)。 個人名などの手掛かりによって， 記憶の中のその項目の表現が活性化され， その結果， その項目の既知の特徴が活性化され， それが再帰的に他の類似した項目を活性化する。 その結果， 再帰的に他の類似した項目が活性化され， これらの項目がさらに特徴を補い， その結果， 手がかりとなった項目に帰属することになる。 このような相互作用を利用することで， 類似性に基づく一般化モデルを， 記憶内の関連項目が手がかりとは重ならないが (個人の名前はユニークかもしれない)， 他の次元では重なっていて， それが再帰的で相互作用的な計算によってもたらされる場合にも拡張することができる。

5. 相互処理と最適知覚推論

以上， IA モデルと TRACE モデルの実証的な裏付けの一部を 示し， 相互活性化の原理の適用範囲が知覚の領域 を超えていることを示したが， IA モデルと最適知覚推論との関係についての疑問には明確に触れていない。 このテーマは， 視覚および聴覚の文脈効果に関する文献において， 激しい批判の的となっている (Massaro, 1989; Massaro & Cohen, 1991; Norris & McQueen, 2008; Norris, McQueen, & Cutler, 2000)。 今引用した論文では， 相互的処理は， 行動データに見られるパターンや， 情報統合がベイズ推論の原理に合致している場合に予想されるパターンから知覚を歪めてしまうこと， また， 相互活性化は， 過度な文脈の影響を引き起こし， 例えば， 語彙的に一貫した音素の不適切な「幻覚」を引き起こすことを主張している。

IA 仮説がこのような批判を受けるのは皮肉なことだが， Rumelhart の知覚における文脈効果に関する初期のアイデア (Rumelhart, 1977; Rumelhart & Siple, 1974) は， 確率的なベイズ推論の観点から明確に定式化されていた。 さらに， モデルで観察された文脈的に一貫した音素の「幻覚」的な知覚は， 自然な文脈での最適な成績という観点からも， 人間の知覚における発見を説明するという観点からも， 我々にとってはまさにモデルが生み出すべきものなのである。 話し言葉の中の音素の生成と同時に短いノイズバーストが発生した場合を考えてみよう。 このようなケースでは， ノイズが音の上に再生されるのではなく， 音を置き換える場合でも， 聞き手は正しい音声を知覚する (知覚的に復元する) 可能性がある (Samuel, 1981; Warren, 1970)。 音素知覚はある意味で幻覚のようなものであるが， 自然な文脈の中では， 話し手が文脈的に適切な音を出したという推論は， ノイズバーストの正確な持続時間の間， 話し手が発話を中断したという推論よりもはるかに正しい可能性が高い。 一般的には， 自然な確率的偶然性が破られるような実験を除いては， 文脈を利用して聞き手を決定する方が， 実際に言われたことを聞き取ることができる可能性が高いのである。

しかし， Rumelhart と McClelland (1981, 1982), McClelland と Rumelhart (1981) は，相互活性化モデルを開発する際に， 知覚の問題を確率論的に定式化することを明確に考慮せず， 代わりに Grossberg (1978, 1980) が提案した非確率論的で神経学的に触発された処理モデルを用い， この定式化が最適な確率論的推論に正確に対応するかどうかを考慮しなかったことは事実である。 振り返ってみると， このことが不幸な誤解と無用な論争を招いたと思われる。 具体的には，相互活性化モデルに関するその後の研究では， 2 つの重要なポイントが支持されている。

1. IA モデル と TRACE モデルは， 当初の定式化では， 原理的なベイズ計算を正確に実装することはできなかった。 実際， これらのモデルの最初の定式化は， 最適性と人間のデータの両方から逸脱した方法で， これらの計算を歪めた。
2. しかし， その本質的な相互的性格を維持したモデルの変形は， ベイズ則に合致し，当初の定式化では問題となったデータを捉えることができる。

1）については， IA モデルと TRACE モデルの欠陥が McClelland (1991) で議論されている。 そこでは， モデルの活性化の仮定と， これらの活性化の反応確率への変換についての仮定が， ベイズ確率モデルや人間の選択反応から逸脱した選択反応のパターンを生み出すことが観察された。 このような逸脱は， 処理に相互作用がない場合にも生じた。 つまり， 2 つのボトムアップ情報を組み合わせて， 可能性のある選択肢を表すユニットの活性化を決定した場合にも生じたのである。
このように， オリジナルモデルの欠点は， 相互性そのものに起因するものではないのかもしれない。

ここでは， 上記 (2) の点について， より詳細に検討する。 具体的には， 多項式相互活性化 (MIA) モデル (Khaitan & McClelland, 2010) と呼ばれる IA モデルの変形が， ベイズの知覚推論原理に従ってどのように動作するかを説明する。 このモデルでは， オリジナルの IA モデルでモデル化されたほとんどの実験と同様に， 4 つの文字系列を含む表示の場合を考慮している。 確率論的原理と AI ネットワークにおける計算との関係については， McClelland (2013) に詳しく書かれているので， ここでの簡単な説明の裏にある詳細に興味のある方は， そちらの論文を参照のこと。

MIA モデルは， Hinton と Sejnowski が 人工ニューラルネットワークの研究に持ち込んだ， ボルツマンマシンという形の洞察を大きく利用してる。 ボルツマンマシンは， そのようなマシンがどのようにして最適な知覚的推論を行うことができるかを説明した会議録論文 (Hinton & Sejnowski, 1983) で初めて発表され， その後， PDP の巻物 (Hinton & Sejnowski, 1986) で説明されている。
まず， オリジナルのボルツマン・マシンの関連するアイデアを説明する。

5.1. ボルツマンマシンでの状態， その良さ， その確率

5.2. IA モデルに具現化された知識の生成モデル

多項式相互活性化モデルは， ボルツマンマシンのわずかな変形で， ユニット間のバイアスと接続における特定の確率的制約を符号化する。 次のステップは， ネットワークに符号化される確率的知識を定義することである。 ここでは， 文字の知覚実験において， 視覚入力の特徴の配列を生成するプロセスに関する知覚者の (暗黙の) 信念の根底にある可能性のある確率的知識を，特定の仮説的な形で採用する。 この知識は， 確率的な生成モデルの形をしている。 生成モデルの概念は， 環境と環境から感覚面に到達する情報の確率的な構造を特徴づけるための有用な道具であり， また， 知覚者が知覚の推論を行う際に使用する知識の仮想的な抽象的特徴づけとしても用いられる。 この言葉は使われていないが， 信号検出理論 (Green & Swets, 1966) の中核には， 単純な生成モデルがある。 この理論によると， 知覚者は， 信号＋雑音の分布または雑音のみの分布のいずれかから選択された信号を受け取ると考えられる。 このモデルのパラメータは， 「信号＋雑音」と「雑音のみ」の確率， および 2 つの分布の平均と標準偏差である。 信号検出理論は， このような状況における最適な知覚的推論の理論を提供する。 ここで提案する文字表示の生成モデルは， もう少し精巧なものだが， 精神的には類似する。 これは Rumelhart and Siple (1974) のモデルで使われた文字表示の確率構造に関する信念の定式化に極く類似しているが， 彼らはこの用語を使っていない。

我々の生成モデルによれば， 知覚者の目に到達する特徴配列は， まず， ターゲット語彙 (ここでは， すべて 4 文字の英単語の集合) に含まれる可能性のある単語から， その単語の言語頻度に単調に関連する確率 $p(w_i)$ で， 単語 w_i をランダムに選択することによって生成される。 単語が選択されると，その単語に基づいて確率的に文字列が生成される。 単語iが選択された場合， k の位置に文字 j が生成される確率は $p(l_{jk}|w_i)$ と表される。 高い確率 (シミュレーションでは 0.9 と仮定) で， 各位置の文字は与えられた単語の正しい文字であるが， 代わりにアルファベットの他の文字が生成される可能性もわずかにある (正しい文字の確率が 0.9 とすると， 他の各文字の確率は0.1/25，つまり 0.004)。 また、文字は、（Rumelhart & Siple, 1974 に倣って）線分として扱われる可能性のある文字の特徴のセットのそれぞれについて， 存在または不在の値の指定を生じさせる (図 5)。 例えば， T という文字は， 対応する特徴の配列の上部と中央部に線分が存在し， 特徴の配列に存在しうる他の線分は存在しないことを指定している。 知覚システムによる文字からの特徴値の生成とその登録も， 確率的なものとして扱われる。 具体的には， 与えられた文字位置 k に対して， 文字jが与えられた特徴次元 f の値 v (存在してもしなくてもよい) を生成する確率は，$p(v_{fk}|l_{jk})$ と表される。 与えられた特徴の正しい値が生成される確率は比較的高く (シミュレーションでは 0.9)， 正しくない値が生成される確率はこの高い値から 1 を引いた値 (0.1) になる。

Fig. 5. The letters A–Z as they are represented in the Rumelhart & Siple font, with the full set of features shown in a single block below the letters. From fig. 2, p. 101 in Rumelhart and Siple (1974). Copyright American Psychological Association. Reprinted with permission.

上記の生成モデルを考えると， 生成モデルを通過するすべての可能な経路の確率を計算することができる。 経路は， 1 つの単語の選択， 単語の各位置にある 1 つの文字の選択， 各文字の位置にある各特徴に対する 1 つの値 (存在するか存在しないか) の選択からなる。 特定のパスを表すのに $p(P_\pi)$ という表記を使う， これは以前ボルツマンマシンの状態を表すのに使ったのと同じ添え字 p を使っている。 これは，MIA モデルにおける活性化のパターンが， 生成モデルのパスに対応するからである。

特定のパス P_π の確率 ($p(P_\pi)$ と表記) は， 生成モデルに基づいてパスの生成を前提とした個々の確率的事象の確率の積に過ぎない。

$$p(P_n) = p(w_i) \prod_k \left(p(l_{jk}|w_i) \prod_f (v_{fk}|l_{jk}) \right). \tag{3}$$

5.3. 生成モデルによる知覚の推論

知覚的推論の問題 (我々の場合) は，指定された特徴値のセット {V} を取り， この特徴値のセットと一致する可能な経路のうち， どの経路が生じたかを推論することである。 可能な経路とは， 指定された特徴値のセットを持つすべての経路のことである。 このような経路は， 各位置にある 1 つの単語と 1 つの文字の組み合わせごとに 1 つずつ存在する (モデルでは， 1 つの位置に1,179 個の単語と 26 個の文字が存在するため， $1,179X26^4$，つまり 約 5 億 4 千万個の経路が存在する)。 指定された特徴量が与えられたときの経路 p の確率は， $p(P_\pi|V)$と表され，経路の事後確率と呼ばれる。 パス P_π の事後確率は次式で与えられる：

$$p(P_\pi|\{V\}) = \frac{p(P_\pi)}{\sum_{\pi'} p(P_{\pi'})} \tag{4}$$

ここで， 分母の合計は， 指定された特徴値 {V} に一致するすべての可能なパスに対して実行される。

原理的には， 観測された特徴集合が与えられたとき， そのような経路のそれぞれの確率を計算し， 生成モデルのもとで観測された特徴を生成した可能性が最も高いものを選択することができる。 多項式 IA モデルでは， このような明示的な計算は行わない。 代わりに， モデルは生成モデルの可能な経路に対応する， 可能な活性化状態のセット S_π からサンプルする。 多項 IA モデルは， 常に最も確率の高い状態をサンプリングするわけではないが， 次のような特性を持つ。 生成モデルの下で状態の可能性が高ければ高いほど， その状態がサンプルされる可能性は高くなる。 以下では， この性質をより正確に説明する。

5.4. MIAモデル: 生成モデルの事後分布からサンプリングするための相互活性化の使用

ここで，MIA モデルについて説明し， MIA モデルが， 上記の生成処理によって生成された指定された特徴値のセットの代替可能な解釈について， 正しい事後確率分布からどのようにサンプリングできるかを説明する。 ここで，解釈は， 各位置に 1 つの単語と 1 つの文字を指定するパスに対応する。

オリジナルの IA モデルと同様， このモデル (図2) には， 可能な単語ごとのユニットが含まれている。 また， 4 つの入力特徴配列の各特徴 (例えば， 横長の上部) の可能な値 (存在するか， 存在しないか) に対応するユニットも含まれている。 3 ユニットは， 互いに排他的な選択肢のセットに対応するプールに編成される。
1 つのプールは， 可能性のある単語に対応するユニットのセットからなり， 他の 4 つのプールは， 4文字の位置にある可能性のある文字のそれぞれに対応するユニットのセットに対応している。 また， 特徴レベルのユニットプールは，4 セット 14 個ある。 それぞれのプールには， 特定の文字位置にある特定の特徴に対する「存在」ユニットと「不在」ユニットが含まれている。

MIAモデルでは， オリジナルモデルの同じプール内のユニット間の一対の抑制性接続に代わって， プール内の 1 つのユニットのみが一度に活性化するという制約が設けられている。 この制約下で， 各プールは多項確率変数， つまり n 個の代替値のいずれかを取ることができる変数に対応する。 これが， このモデルの特徴であり， 名前に「多項式」という言葉がついている (Dean, 2005 は， 新皮質の計算モデルでこのような方式を提案している Lee & Mumford, 2003 も参照)。 MIA モデルにおける相互抑制の仮定は， オリジナルモデルの相互抑制の仮定と同様に， 脳全体に存在する局所的な抑制回路の理想化された概念レベルの帰結と考えられている。 これは， オリジナルモデルにおける同一プール内のユニット間の相互抑制の役割と同様の役割を果たしている。 このような阻害の扱い方は， 多くのモデラーによって提案され (Grossberg, 1978など)， 神経科学者が視覚野の神経応答のモデル化に用いた分割正規化モデルに類似する (Heeger, 1992)。

MIA モデルでは， 上述の生成モデルを特徴づける確率的な情報を， ネットワークのバイアス項と接続の重みを設定するために明示的に使用する。 後述の理由により， バイアスと重みは， 関連する確率的な量の対数に対応している。 具体的には， 各単語ユニットにバイアス重みが割り当てられる。 単語 i のユニットに対するバイアス重み b_i の値は， $\ln(p(w_i))$， すなわち， 上述の生成プロセスによって単語 i がサンプリングされる確率の自然対数に等しく設定される (以下では「対数」という言葉は常に自然対数を意味する)。 位置 k にある文字 j についての各単語ユニット w_i と各文字ユニット l_{jk} との間の接続重みは，単語iが生成過程で選択された単語であることを前提に， その文字が生成されるであろう確率の対数である $\ln(p(l_{jk}|w_i))$ に設定される。 同様に，位置 k の文字 j のユニットと，その位置にある特徴 f の 2 つの可能な値のそれぞれに対する特徴ユニットとの間の接続重みは， 文字が生成された場合にその特徴が生成モデルの下で生成される確率の対数である $\ln(p(v_{fk}|l_{jk}))$ に設定される。

要約すると，MIA モデルは， 上述した生成モデルの確率情報を対数変換したものを接続重みに組み込んでいる。 もし， このモデルの知識が， 特定の実験で使用されたディスプレイを実際に生成した生成モデルの対数確率に正確に対応していれば， その出力は， これらの入力を生成した世界の事象の真の確率に関連していることになる。 また， このモデルは， 知覚者が主観的に推定した確率を表していると考えることもできる。 この場合， 知覚者の知覚システムに組み込まれた知識と世界の真の統計との間に違いがある限り， 推定値に関しては最適な知覚が， 真の統計に関しては非最適な知覚となる可能性がある。 その場合， 知覚者の知覚システムに埋め込まれた知識と世界の真の統計との間に差がある限り， 推定値に関しては最適な知覚が， 現実の世界の統計に関しては最適ではないかもしれない。

多項式 IA モデルが， 生成モデルで定義された確率分布の事後からサンプリングできることを実証するという現在の目的のため， 提示された文字列の特徴のサブセットの値の有無が外部入力で指定された場合を考える。
図 6 の例では， 1 位の特徴は何も指定されていないが， 2 位，3 位，4 位の特徴は，それぞれ O，O，D の文字の特徴であることがわかった。 生成モデル (図の計算確率と書かれた棒) によると， 文脈に沿った単語を構成する文字 (F, G H, M, W) は， いずれもかなりの確率であり， それらの違いは， 関連する単語 (FOOD, GOOD, HOOD, MOOD, WOOD) の $p(w)$ の値の違いを反映していることがほとんどである。(footnote 4)

脚注 4: このモデルで使われている $p(w)$ の値は，生の単語頻度ではなく， オリジナルの IA モデルと同様， これらの確率が圧縮されていることに注意 (McClelland & Rumelhart, 1981)。 この圧縮がなければ， 図 6 に示した事後確率の変動幅はもっと大きくなっていただろう。 この $p(w)$ 値の圧縮は， モデルに組み込まれた刺激頻度に関する (暗黙の) 「仮定」に相当する。 単語単位のバイアス項は， これらの圧縮された $p(w)$ 値の自然対数である。

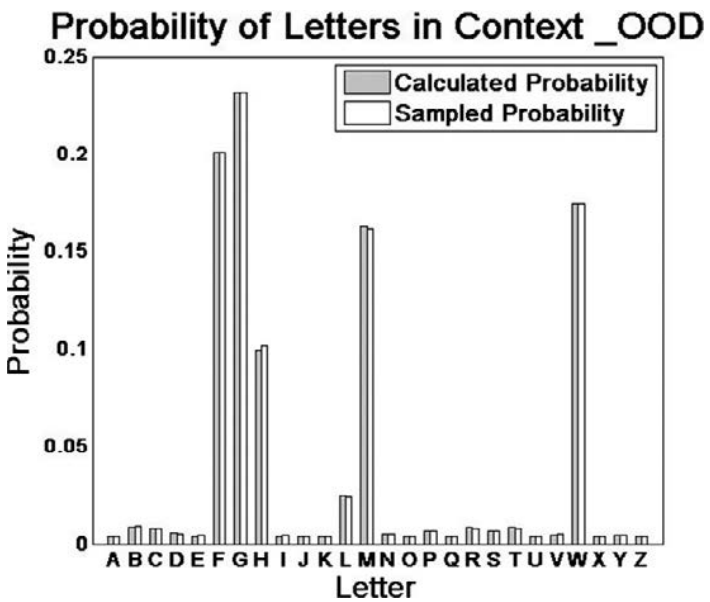


図 6. 4 文字のディスプレイの最初の位置にある文字について， 直接計算された事後確率と， 多項式相互活性化 (IA) モデルのギブスサンプリングプロセスの結果との比較。 図は， 4 文字の配列の 1 番目の位置にある文字の事後確率を計算したもので， 1 番目の位置に特徴量が指定されていないディスプレイが提示された後， 2 番目， 3 番目， 4 番目の位置にそれぞれ O, O, D の特徴量が完全に指定された場合のものである。 灰色のバーは， 1 番目の文字の位置について計算されたベイズ事後確率を示している。 これらの確率は， 生成モデルに組み込まれた語彙知識を反映している． また，$p(f|l)$ は， 各文字の特徴が正しい場合は 0.9，正しくない場合は 0.1 であると仮定した。 白い棒は，多項式 IA モデルを 50 回反復した後に， 1 位の文字のそれぞれが活性化確率をサンプリングしたものである。 モデルの重みは， 本文で述べたよう

に，ベイズ計算に用いた確率の対数に対応するように設定した。合計 10,000 回の模擬試験を 100 回繰り返し行った。結果は，反復回数 51〜100 回を平均した平均確率である。サンプリングされた確率と計算された確率のわずかな違いは，サンプリングエラーの範囲内である。

5.5. MIA モデルでの処理

ボルツマンマシンと同様，特徴の指定は，指定された各特徴の値に対応するユニットをオンにすることで，モデルに提示される。処理は，特徴のあるユニットが上記のようにクランプされ，文字プールと単語プールのいずれのユニットもアクティブになっていない状態で開始される。処理は，ボルツマンマシンのランダムな更新処理と同様，いくつかのサイクルにわたって行われる。しかし，この場合，サイクルはランダムではない (この詳細はモデルの機能には重要ではないが，計算の意味についての議論を多少簡単にする)。1 つのサイクルの中では，まず 4 文字プールのそれぞれについて，特徴レベルと単語レベルの既存の活性化を用いて活性化が決定される。次に 4 文字の各位置の活性化と，各単語ユニットに関連するバイアス重みを用いて，単語プールの活性化が決定される。各プールの活性化の決定は，通常通り，他のユニットの重み，バイアス，活性化に基づいて，各ユニットの純入力を計算することから始まる。

前述したように，このモデルは，オリジナルの IA モデルやオリジナルのボルツマンマシンとは異なり，各時間ステップにおいて，各位置にある 1 つの文字ユニットと 1 つの単語ユニットのみが活性化することができる。活性化したユニットは，ソフトマックス関数を使用して確率的に選択されるため，プール内の各ユニットにおいて，あるユニットが選択される確率は，自身のネット入力をプール内のすべてのユニットの対応する量で割った指数関数に依存する。

$$p(a_i = 1) = \frac{e^{\text{net}_i/T}}{\sum_j e^{\text{net}_j/T}}. \tag{5}$$

ここで，i と j は更新されるプールのユニットを示し，T はボルツマンマシンと同様に温度に対応する。ソフトマックス関数は，ボルツマンマシンで使用されているロジスティック関数を拡張したものとみなすことができる。ロジスティック関数では，1 つのユニットの活性化をオンまたはオフのいずれかの状態に設定するが，ソフトマックス関数では，多項確率変数を n 個の代替状態の 1 つに設定し，プール内のユニットのうち正確に 1 つが活性化される。

ここでは，この計算と，観測された特徴のセットが与えられたときに可能な文字の事後確率からのサンプリングとの関係を考えてみよう。具体的には 4 文字のディスプレイの 2 番目の位置にある文字に対応するユニットのプールの活性化を，活性化の第 1 サイクルで計算することを考える。活性化の第 1 サイクルにおいて，4 文字ディスプレイの 2 番目の位置にある文字に対応するユニットプールの活性化を計算してみよう。この場合，送信ユニットは 2 番目の文字の位置にある特徴の値に対応するユニットであり，受信ユニットは文字列の 2 番目の位置にある可能性のある文字のユニットであり，重みは文字と特徴のユニットの間の接続重みであり，それぞれが文字が与えられたときの特徴の特定の値 (存在するか存在しないか) の確率の対数に対応している。送出ユニットの活性化は，特徴 f の指定された値 v に対応するユニットの場合は 1 に等しく，モデルには文字レベルで指定されたバイアス項がないことに注意すると，位置 k の文字ユニット j へのネット入力の式は次のように書き換えられる:

$$\text{net}_{jk} = \sum_f ln(p(v_{jk}|l_{jk})) \tag{6}$$

ここで，ユニットを起動する確率を計算するソフトマックス関数で使用するために $e^{\text{net}_{jk}}$ を計算すると，この式は $\prod_f p(v_{jk}|l_{jk}$，つまり生成モデルの下で，与えられた文字から特徴の観測値を生成したであろう確率となる(脚注 5)。これらの値をソフトマックス関数に代入すると，次式を得る:

$$p(a_{jk} = 1) = \frac{\left(\prod_f p(v_{fk}|l_{jk}\right)^{1/T}}{\sum_g \left(\prod_f p(v_{fk}|l_{gk}\right)^{1/T}} \tag{7}$$

脚注 5. この結果は，例えば $\ln(a) + \ln(b) = \ln(ab)$ のように，量の集合の対数の和が量の積の対数に等しいという事実と，量の対数に対する e が単に量そのものであるという事実，すなわち $e^{\ln x} = x$ から導かれる。

$T = 1$ の場合，この式は，特徴量の値が与えられたときの文字 j の事後確率を表すベイズの式に対応する(McClelland, 2013)(脚注 6)。この場合，ソフトマックス関数は，指定された特徴量が与えられたときの文字の事後確率に等しい確率で，活性化する文字を選択する。T が 1 に等しくない場合，これらの確率は 1/ T乗された後，再正規化される。前述のように，これをよりコンパクトに表現すると

$$p(a_{jk} = 1) \propto \left(\prod_f p(v_{jk}|l_{jk})\right)^{1/T} \tag{8}$$

脚注 6. 完全なベイズの式には，文字の事前確率を表す係数が含まれている。しかし，生成モデルでは，文字は独立した事前確率を持っていない。代わりに，文字の確率は単語レベルに依存し，単語レベルが文字レベルに与える影響は，文字レベルのユニットの2 回目以降の更新に組み込まれる。1 回目の更新では，文字は同じ確率として扱われる。このような定数要素は相殺されてしまうため，式では表されていない。

5.5.1. 神経計算と確率的計算を結びつけるための対数と指数の役割

読者はこの時点で，なぜ MIA モデルのネットワークの接続重みの強さを定義する際に，わざわざ確率的な量の対数を使ったのかと聞きたくなるかもしれない。なぜなら，ソフトマックス関数 (注5) や密接に関連するロジスティック関数で使用するために，ネットの入力を単位に指数化する際に，この対数変換を元に戻すからである。実際，単語の事前確率と，単語に与えられた文字の条件付き確率，および文字に与えられた特徴の条件付き確率を直接使用し，それに応じて活性化関数を再定義することで，MIA モデルを再構築することができる。これらの確率量の対数を用いる理由は，MIA モデルをはじめとするニューラルネットワークモデルの背景にある，神経科学からのインスピレーションと，ニューロンと計算を結びつけるモデルの歴史に基づく。MIA モデルは，Grossberg (1978) の初期モデルの流れを汲むオリジナルの IA モデルと，Hopfield (1982) の初期モデルの流れを汲むボルツマンマシンを融合させたもので，その系譜を辿る。これらのモデルは最終的に，パーセプトロン (Rosenblatt, 1958) を経て，重み付けされた信号を加算して閾値と比較する装置である マッカロック・ピッツ・ニューロン (Pitts & McCullough, 1947) まで遡ることができる。ニューロンは興奮性と抑制性の信号を加算し，閾値に達すると発火するという考えは，もちろん，神経科学者が信頼するモデルニューロンの標準的な直観的単純化である。このように単純化されたモデルニューロンの入力に加法性ガウスノイズが存在する場合，発火確率は，合計または正味の入力のロジスティック関数と密接に一致する。これは，Hinton and Sejnowski (1983) が最初に指摘したように，重みとバイアス項が適切な確率量のログに設定されていれば，ベイズ則を実装することになる (ソフトマックス関数の神経基盤の可能性については McClelland, 2013 参照)。

話を本筋に戻すと，今度は単語レベルの各ユニットへの純入力を考える。この場合，純入力は，単語の主観的確率の対数を表すバイアス項と，文字レベルの各ユニットの活性化の積に，単語ユニットと文字ユニットの間の重みをかけた項の合計で構成される。前述の計算の最初のステップから，各位置の 1 つの文字ユニットの活性化値は 1 で，他のすべての文字ユニットの活性化値は 0 なので，単語ユニット i への純入力は以下のようになる:

$$\text{net}_i = \ln p(w_i) + \sum_k \ln p(l_{jk}|w_i), \tag{9}$$

ここで，l_{jk} は k の位置にあるアクティブな文字ユニットを表す。ここで，e^{net_i} を計算すると，生成モデル下で，その単語が提示のために選択される確率と，その単語が選択された場合に能動文字が生成される確率が求められる。これをソフトマックス関数に入れると以下のようになる:

$$p(a_i = 1) \propto \left(p(w_i) \prod_k p(l_{jk}|w_i \right)^{1/T}. \tag{10}$$

これを言葉で表現すると，ある単語ユニットが唯一のアクティブなものとして選択される確率は，その単語の事前の出現確率に，その単語がアクティブな文字のセットを生成したであろう確率を掛けたものに比例する。ここでも，これは，特定の単語が提示された事後確率を計算するベイズ則の基本的な論理を実装している。この場合，事前情報 ($p(w_i)$ で表される) と，単語が与えられた場合の証拠の可能性 （この場合はアクティブな文字) が与えられる。

最後に，単語レベルで 1 つの単語ユニットがアクティブになっている場合，次のサイクルで文字プールのいずれかのユニット j がアクティブになることを考えてみよう。各文字レベルのユニットへの純入力は先ほどと同じだが，アクティブな単語が与えられたときの文字の確率の対数に対応する項が追加されている。この式を指数化すると，アクティブな単語が与えられたときの文字の確率に，文字が与えられたときの指定された特徴のセットの確率を掛けたものになる。与えられた文字 j が k の位置で活性化される確率を表す式は以下のようになる:

$$p(a_{jk}) \propto \left(p(l_{jk}|w_i) \prod_f p(v_{jk}|l_{jk}) \right)^{1/T}. \tag{11}$$

このように，文字レベルの活性化の 2 回目の更新の後，各位置にある所定の文字ユニットがその位置の活性化ユニットとして選択される確率は，活性化された単語が与えられた場合の文字の確率に，文字が与えられた場合の所定の位置にある特徴量のセットの確率を $1/T$ でスケーリングしたものに比例する。

なお，単語と文字の間の重み，および文字と特徴の間の重みは，ディスプレイ作成の基礎となるトップダウンの生成プロセスの観点から定義されている。文字と特徴の間の重みは，特徴から文字ユニットへのボトムアップ入力を計算する際に使用され，単語と文字の間の重みは，文字から単語ユニットへのボトムアップ入力を計算する際に使用される。単語から文字への重みは，単語ユニットから文字ユニットへのトップダウンの影響を計算するためにも使用される。また，ここでは考慮していないが，文字から特徴への重みは，不足している特徴レベルの活性化を埋めるために使用することができる。重みが対称的に使用されているので，このモデルはボルツマンマシンと本質的な特徴を共有している。活性化の更新は，ネットワークの状態を，全体的に良い状態の方向に移動させる傾向がある。

要約すると，上述の処理順序が与えられ，$T = 1$ で実行された場合，ある文字ユニットがある位置で活性化する確率は，生成モデル下での特徴が与えられた文字の確率に対応する。単語レベルが最初に更新されたとき，選択された文字が与えられたときの単語の確率に比例した確率で，1 つの単語が選択される。このように，私たちの計算は，観測された特徴を生み出す可能性のある，基礎的な生成モデルの状態からのサンプルを生成する。しかし，文字の確率の推定では，まだ単語レベルの情報が考慮されていない。文字レベルでの次の更新では，単語レベルの情報が考慮されるので，各文字の位置について，文字ユニットが活性化する確率は，活性化した単語と与えられた特徴配列の両方が与えられた場合の文字の確率と等しくなる。

この時点で計算が完了したように見えるが，文字レベルでの 2 回目の更新後の文字の活性化確率は，正しい事後確率とは一致しない。しかし，単語レベルと文字レベルの更新を交互に繰り返しながらサンプリングを進めていくと，活性化確率は正しい事後確率に収束していく。このサンプリング手順は，ボルツマンマシンで活性化状態を設定する際に用いられる手順を多項式の場合に一般化したものである。ボルツマンマシンのサンプリング手順と同様に，我々の手順は，Gibbs サンプリング (Geman & Geman, 1984) のイ実体である。Gibbs サンプリングは，統計物理学を起源として広く用いられている手順であり，他の変数の現在の値が与えられたときに，個々の変数の条件付き分布と一致するように個々の変数を局所的に更新することで，確率分布の事後から不偏のサンプルを提供することが示されている (詳細は McClelland, 2013)。これはまさに MIA で行っていることである。文字ユニットの状態を，単語ユニットと特徴ユニットの状態を条件にしてサンプリングし，単語ユニットの状態を，文字ユニットと特徴ユニットの状態を条件にしてサンプリングしている (ただし，特徴ユニットは，文字ユニットの状態を介して間接的に単語ユニットに影響を与えるだけである)。

5.6. MIA モデルの状態と生成モデルのパスウェイ確率

ある温度 T で MIA モデルの状態をサンプリングした場合，最初の「バーンイン」期間の後に特定の状態になる確率は $e^{G(s_\pi)/T}$ に等しく，ここで「良さ」は上記のように定義される。MIA モデルの具体的なケースでは，良さは次のようになる。

$$G(S_\pi) = \ln p(w_i) + \sum_k \left(\ln p(l_{jk}|w_i) + \sum_f \ln p(v_{fk}|l_{jk}) \right), \tag{12}$$

上式を指数化することにより次式を得る:

$$e^{G(S_\pi)} = p(w_i) \prod_k \left(p(l_{jk}|w_i) \prod_f p(v_{jk}|l_{jk}) \right). \tag{13}$$

右辺は，生成モデルの下で，観測された特徴のセットの基礎となる生成モデルの経路が，状態 S_π に対応するものである確率を表している。これを確率善悪の式に差し込むと、モデルは、観測された特徴を実際に生成した温度スケールの確率に比例した確率で、そのような状態を訪れることがわかります。

5.7. モデルの状態に基づいて顕在反応の形成

これまでの開発では，相互活性化ニューラルネットワークが，ニューラルネットワーク全体の状態に関する確率分布の事後分布から，どのようにサンプリングできるかを示してきた。これらの状態は，ネットワークの入力に存在する実際の特徴を生じさせる可能性のある，文字と単語の両方の同一性の割り当ての同時分布からのサンプルである。例えば，多くの視覚的単語認識研究のように，与えられた位置にある文字や，他の多くの研究のように，単語全体など，特定の項目の同一性を決定することに興味がある場合，問題のユニットが活性化している状態にある確率が (他のユニットの活性化に関わらず)，アイ

テムの正しい事後確率に対応していることを観察することができる。 言い換えれば， ネットワークの状態は， 多項変数のそれぞれの限界分布と、これらの変数のすべての同時分布からのサンプルであると同時に， 多項変数のそれぞれの周辺分布と， これらの変数のすべての同時分布からのサンプルである。 これは， Rumelhart (1977) が， 知覚における相互的処理の結果として想定したものと全く同じである。

例えば， 単語の最初の位置にある文字の識別について， この分布からのサンプルである応答を生成するためには， 知覚者は， 反復的な決済プロセスによって選択された文字の識別を報告するだけでよい。 このモデルのシミュレーションにより， この数学的事実が証明された。 その一例を図 6 に示す (詳細は図注参照)。

5.8. 最適性への近似としてのサンプリング

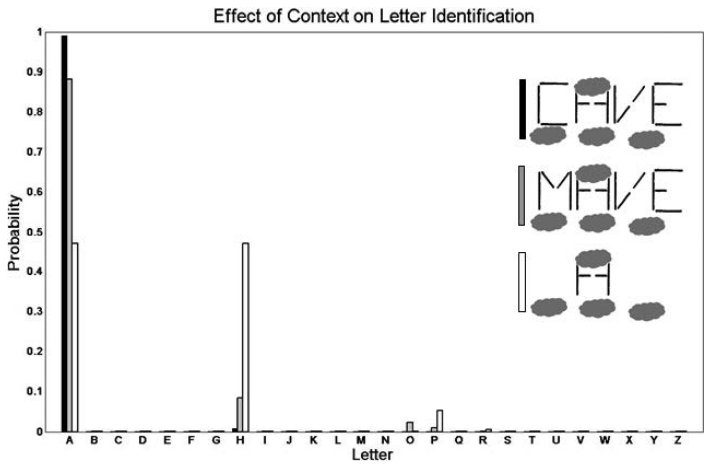
ここまでは， 知覚系に提示された刺激を特徴づける生成モデルの事後確率から， 知覚がサンプリングされるモデルを説明してきた。 ここで注意しなければならないのは， 本当に最適な方策は， 相対的な確率に比例して選択肢をサンプリングするのではなく， 事後確率が最も高い選択肢を選択することである (モデルでは温度パラメータ T を 1 に設定している) (脚注 7)。 しかし， この温度パラメータは， 知覚システムに内在する処理ノイズを反映していると考えることもできる。 その場合， 知覚実験の各試行は， 一般的なレベルのノイズの下で， 単一の最良の解釈を見つけようとする試みであると考えることができる。 いずれにしても， 温度が高ければ高いほど， ランダムな挙動が見られる。 温度が高いことの利点は， 可能な知覚的解釈の範囲を十分に探索することができ， 計算の初期段階での早すぎるコミットメントを避けることができることである。

> 脚注 7 ここで注意しなければならないのは， 温度パラメータを変更することは， モデルの重みと偏りをスケーリングすることと同じであり， これらは生成モデルにおける相対的な確率と相対的な条件付き確率を表しているということである。 したがって， 温度を低くすることは， 生成モデルにおいてより少ないランダム性を仮定することに相当する。

ボルツマンマシンでは， 温度を徐々に下げていくことで最適な知覚解釈が可能になるが， この方針では無限の時間が経過した後にしか大域的な最適値を見つけることができない。 実時間での制約を考えると， 中間温度でのサンプリングは， 脳が実時間で最適な知覚解釈に近づけるための妥協点なのかもしれない。

5.9. MIA モデルにおける非単語での知覚の円滑化

先述のとおり，オリジナルの IA モデルの重要な特徴は， MAVE のような疑似語の中の文字の知覚の促進を， 単語の中の文字の知覚の促進と同様に説明するという事実であった。 オリジナルのモデルでは， 単語ではないものでも， 提示された文字列と共通の文字を持つ複数の単語を部分的に活性化することができるため， このような現象が起こる。 MIA モデルでは， 一度に 1 つの単語しか活性化されないため， 一見すると同じパターンを示さないと思われるかもしれません。 そこで， 図 7 のように 2 番目の位置にある文字が部分的に隠されているが， 単語の中にあるか，擬似単語の中にあるか， あるいは単独であるか， という曖昧な表示を考えてみた。 シミュレーションで文字を表現するのに使用した Rumelhart-Siple フォントの A と H の文字と， 利用可能な特徴が同じように一致している。 このモデルは， 曖昧な部分が仮の単語の中にあったとしても， 文脈を利用して曖昧さを解消し， A をより可能性の高い選択肢として選択することができるだろうか？



> 図 7. A または H に等しく一致する曖昧な文字が異なるコンテクストで提示された場合に， 2 文字目の位置で異なる文字が活性化される確率 (黒棒: C_VE， 灰色棒: M_VE, 白棒: 無コンテクスト)。 Rumelhart-Siple フォントに基づいた生成モデルでは， A も H も同じように示された特徴を生成する可能性が高く， 文字 P はその次に可能性が高いとされている。 しかし， 文脈が C_VE や M_VE の場合は， A の可能性がはるかに高い。 M_VE の場合は O の可能性がある程度高いが， 文字が O であるという仮説の下では特徴情報はあり得ないので， 全体としては O の可能性は A よりもずっと低い。

この疑問を解決するために， 図に示す 3 つの表示方法でシミュレーションを行った。 一文字だけの場合には， 他の 2 つの文脈における単語レベルの影響を評価するためのベースラインとして， 単語レベルを完全にオフにした。 その結果を図 7 に示す。 図が示すように，文脈がない場合 (白棒)，A と H の 2 つの選択肢が約半分の割合で選択されている。 どちらの文脈でも、Aという文字はHという文字よりもはるかに可能性が高くなる。 これは， 最初の位置に C がある場合の方が Mが ある場合よりも大きくなるが， どちらの場合でもかなりの確率で起こる。

なぜモデルは両方の文脈で A の文字を選択する傾向があるのだろうか？ 残りの文字が CAVE という単語を形成している場合， ディスプレイ全体は他のどの単語よりも CAVE によって生成された可能性が高く， したがって， 文字 A は文字 H よりも 2 番目の位置にあった可能性がはるかに高い。 最初の文字がMの場合、観察された特徴のすべてを生み出した可能性の高い単一の単語はありません。 実際， MOVE という単語は， 他のどの単語よりも高い可能性を持っている (2 番目の位置にある特徴のいくつかとは矛盾しているが， 他のすべての位置にある特徴のすべてとは矛盾していない)。 しかし， CAVE, GAVE, HAVE, SAVE, WAVE, MADE, MAKE, MALE, MARE, MATE などの他の多くの単語は， すべての特徴のセットと部分的に一致している。 MOVE がサンプリングされた場合，モデルは 2 番目の位置の文字として O を選択することができるが，A や H を選択することもできる。 これは， 基礎となる単語が MOVE であった場合， 生成モデルに従ってこれらの文字が生成されることがあるからである。 2 番目の位置に A を含む単語の 1 つがサンプリングされると， ほとんどの場合， 対応する文字として A が選ばれる。MOVEがサンプリングされた場合、モデルは2番目の位置の文字としてOを選択することができますが、AやHを選択

することもできます。これは、基礎となる単語がMOVEであった場合、生成モデルに従ってこれらの文字が生成されることがあるからです。 2 番目の位置に A を含む単語の 1 つがサンプリングされると， ほとんどの場合， 対応する文字として A が選ばれる。(脚注8)

脚注 8 このモデルに単語レベルからの読み上げを要求した場合， 常に単語の応答が得られるが， 元の IA モデルでも同じであっただろう。 人間の観察者は，4 文字すべてを報告するように求められた場合， 疑似単語が提示されても常に単語を報告するわけではない (McClelland & Johnston, 1977)。 疑似単語で得られた全報告応答のパターンが， 4 文字の位置からの読み出しを想定した MIA モデルで説明できるかどうかは、さらなる研究が必要である。

5.10. MIA モデルはロジスティック加法性を示しオリジナルの IA モデルの限界を解決

多項 IA モデルにおいて， 沈降が一定の温度 $T = 1$ で起こる場合， 我々の生成モデルに従った事後確率の正確なマッチングが得られることを見てきた。 人間の知覚者もこの事後確率と一致するのだろうか？ 主観的な確率の独立した証拠を得るのは難しいので， これまでは， 知覚者が最適な成績をしている場合に期待される関数形に従って， 文脈と刺激の情報を組み合わせているかどうかを判断する傾向があった。 興味深いことに， 刺激と文脈の情報の要因操作が特定の選択肢を選択する確率に影響を与える方法として， 多項 IA モデルや IA モデルの他の確率的変形で生じる単純な関数形式がある (McClelland, 2013; Movellan & McClelland, 2001)。 多項 IA モデルを含むこれらのモデルの下位集合では， 特定の反応をする確率のロジット (ここで logit(p) は $\ln(p/(1-p))$ として定義され，対数オッズ比としても知られている量) は， 刺激のみに起因する量 (項目が与えられたときのサンプルされた特徴の相対確率に対応する) と文脈のみに起因する量 (文脈が与えられたときの項目の相対確率に対応する) の 2 つの量の合計に対応することを簡単に示すことができた。

$$\text{logit}(p_i) = s_i + c_i. \tag{14}$$

さらに， 代替案の事前確率に関連するバイアスを組み込むために， 項 b_i を含めることができる。 この関係 (Movellan & McClelland はロジスティック加法性と呼んだ) は， 文脈と刺激情報の共同効果を調べた多くの研究のデータにおいて， 少なくとも近似的に成立する (レビューは Movellan & McClelland, 2001 参照。一つの例外として Pitt, 1995 を参照)。 多項 IA モデルはロジスティック加法性を示し， その傾向は温度パラメータ (T) の値に影響されないとされている。 T は，モデルの予測における刺激項と文脈項の大きさをスケーリングしていると考えることができるが， 一般的にはデータから個別に特定することはできない。

Massaro (1989) は，IA モデルと TRACE モデルの初期の批判で， これらのモデルが多くの実験データに見られるロジスティック加法性を捉えていないことを指摘し， この失敗が， 相互性が知覚を根本的に歪めるという彼の結論の根拠となった。 当初のモデルの仮定はこの関係を歪めていたが， 実はこの問題は相互性に起因するものではなかった。 前述のように， 複数の入力源の影響は， 活性化の伝播が厳密にフィードフォワードであっても， オリジナルモデルで使用されている活性化関数の下ではロジスティック加法性を示さない (McClelland, 1991)。 いずれにしても， MIA モデルではロジスティックな加法性が見られ， オリジナルモデルのこの限界を克服している。

ロジスティック加法性は，IA モデルの他の多くのバリエーションで観察されることに注意する必要がある(McClelland, 1991, 1998; Movellan & McClelland, 2001)。 このようなケースでは数学的に証明することは困難だが， シミュレーションではこの結果が成り立つことが実証されている。 ロジスティック加算性を示す変種は， モデルへの入力やモデル内の処理に内在する変動性を取り入れたものである。

5.11. 中間まとめ

MIA モデルの説明により， IA 仮説に従って， 相互活性化が実時間で最適な知覚解釈の良い近似値を生み出すこと， そして MIA モデルが (IA モデルの他の変形とともに) データに見られるロジスティック加法性パターンを捉えることができることが明らかになったことを期待している。 もちろん， MIA モデルが人間の知覚処理の最良のモデルであるとか， 相互性が知覚過程の一部であるということを意味するものではない。 実際， 批判者は， 最適性に近づくためには相互作用は必要ないと主張し， 処理が一方向であるモデルを主張している。 ここで， この問題について考えてみよう。

6. 影響が知覚システムにフィードバックされることは有利なのか？

多くの研究者が， 文字や音素の識別における文脈効果は， フィードフォワード処理のみに依存し， 刺激と文脈情報の統合はその後の決定段階で行われることで適切に説明できると提案している (例えば Massaro, 1989; Norris & McQueen, 2008; Norris et al. 2000; Paap, Newsome, McDonald, & Schvaneveldt, 1982)。 知覚情報と文脈情報を統合した知覚後の決定レベルであれば， 刺激や語彙情報が文字や音素の識別にどのように影響するかを説明することができる (図 8a)。 したがって， これらの著者は， 相互活性化な活性化は何の利益にもならないので， 知覚のモデルに組み込む必要はないと主張している。

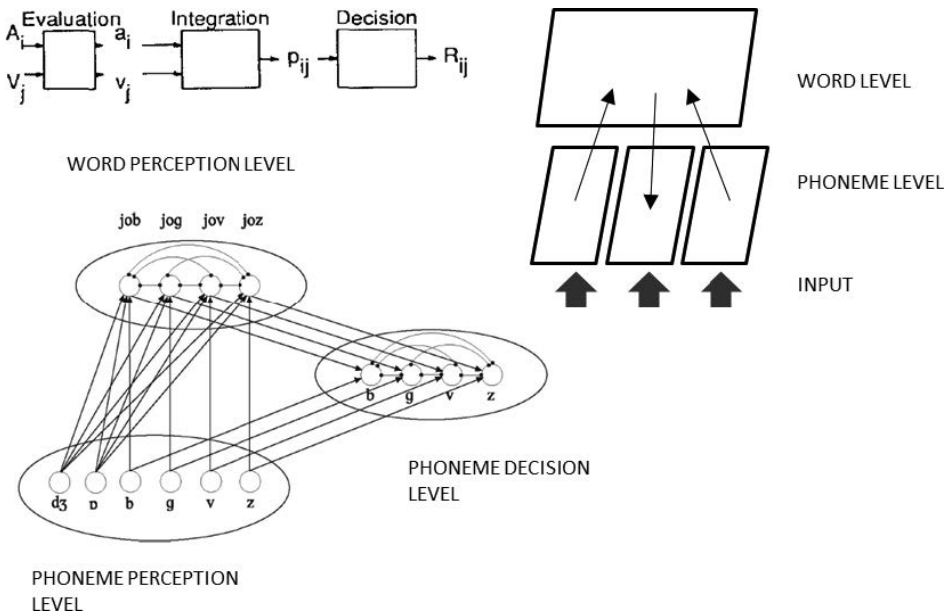


図 8. * (a) Massaro の ファジー論理知覚モデル による刺激と文脈の情報の統合の模式図 (Massaro, 1989 p. 401 図 1 から転載 Copyright ©Elsevier Ltd, reprinted with permission.) Massaro の 図 の A と V の変数は， 本文で紹介した刺激と文脈の変数に対応している。 * (b) 3 音節の中間位置にあるセグメントを識別するための Massaro モデルで使用される文脈因子と刺激因子を計算するための情報の一方向の伝搬を示す模式図 * (c) MERGE model of speech perception (Norris et al., 2000) のアーキテクチャ (Norris and McQueen, 2008 fig. 11, p.384 より転載 Copyright © American Psychological Association, reprinted with permission.

我々は，相互活性化が有益であるためには，2 つの重要な方法があると主張する。

1. 多くの表現レベルにわたって， またレベル内の多くの位置で，最適な知覚的識別を同時に実行する
2. これらの処理過程の結果を知覚系自体の内部で利用できるようにすることで， 他の入力の処理や， 同じ項目を後から処理する際に， ノックオンの結果をもたらす可能性を可能にする

これらの点について，続く 2つの節で検討する。

6.1. 多段階と多位置で最適な推論の同時実装

相互的なアプローチの利点を強調するために， Massaro (1989) が提案したアプローチと対比してみよう。Massaro は， 知覚における文脈と刺激の情報を統合するために， 厳密な前向き計算を提唱した。 Norris とその共同研究者 (Norris & McQueen, 2008; Norris et al., 2000) が提案したアプローチにも， 同様の点が当てはまる。

Massaro のモデルは， 指定された 1 つの目標項目を知覚的に識別することに焦点を当てている。 例えば， Massaro が考えたタイプの実験 (Massaro & Cohen, 1983) では， 目標項目は，/t/, /s/, /p/, /v/ のいずれかで始まり，/i/ の母音で終わる単音節中の 2 番目の音声であった(ee)。 それぞれの文脈で， 頭の子音と母音の間に 7 種類の音が提示され，/l/ のような音から /r/ のような音へと連続して配置され， 全部で 28 種類の刺激が提示された。 それぞれの刺激は各参加者に何度も提示され， 2 番目のセグメントが /l/ または /r/ のいずれかであることを識別するという課題が与えられた。

ベイズ流の観点からは， 知覚は， 刺激と文脈の両方を制約情報源として用いて， 与えられた入力が /r/ または /l/ である事後確率の推定値を計算することに依存すると提案できる。 これは，各文脈 c について， 量 $p(r|c)$ と $p(l|c)$ を計算し， また， 各刺激 s について，量 $p(s|r)$ と $p(s|l)$ を計算することで行うことができる。 そして，$p(r|s,c)$ の正しい事後分布は次のように与えられる:

$$p(r|s,c) = \frac{p(s|r)\ p(r|c)}{p(s|r)\ p(r|c) + p(s|l)\ p(l|c)}. \tag{15}$$

Massaro モデル (図8a) では， 被験者は上記定式化における確率的な量の正規化された推定値に対応する量を計算すると仮定している (脚注9)。 注目すべきは， 上述の計算に用いられる文脈の表現は， 2 番目のセグメントから刺激情報を除いたものであり， 最初と最後のセグメントは文脈を指定し， 2 番目のセグメントは計算に用いられる刺激情報を提供している。

脚注 9. Massaro モデル (Massaro, 1989) では，s_r と呼ばれる r の相対的な刺激支持は，$p(s|r)/(p(s|r)+p(s|l))$ に相当し，相対的な文脈支持 c_r は，$p(r|c)/(p(r|c)+p(l|c))$ に相当する。 l の刺激と文脈サポートも同様に定義される。 $s_r + s_l = 1$ なので，s_l は $1 - s_r$ に置き換えることができ， 同様に c_l は $1 - c_r$ に置き換えることができる。 したがって， 2 つの選択肢がある場合のモデルは，
$p(r|s,c) = s_r c_r/(s_r c_r + (1 - s_r)(1 - c_r))$ となる。 被験者は，結果として得られた $p(r|s,c)$ の推定値に等しい確率で r の回答を選択する。

3 音節処理系の結合重みに符号化された情報は， Massaro モデルに必要な項を計算するために使用することができる。 図8b 矢印は上下にあるが， それぞれの矢印は一方向にしか動いておらず， フィードバック接続はない。 最初と最後の位置にある特徴的な情報は， 最初と最後の位置のスロットにある可能性のある各音素について $p(p|f)$ を計算するために使用される。 次に， 単語レベルでは， 以前のように生成モデルの仮定に基づいて， 語彙の各単語について $p(w|\{p1\},\{p3\})$ を計算することがでる ($\{p1\}$ という表記は， 1 の位置にある可能性のあるすべての音素についての $p(p|f)$ の値のベクトルを表し，$\{p3\}$ についても同様)。 量 $p(r|c)$ は， 最初と最後の位置に入力された単語の確率のすべての単語に対する合計に， 各単語が与えられた 2番目の位置のrの確率を掛けたものとして計算できる。 これは， 図のように， 音素ユニットと単語ユニットの間の接続重みを， 1 番目と最後の位置では一方向に， 2 番目の位置では逆方向に使うことに相当する。 次に， 所望の量 $p(r|s,c)$ は， 語彙入力と， 2 番目の位置の音素に対して計算されたボトムアップ刺激支持とを組み合わせて，上式を用いて計算される。 この算出された確率を用いて，確率 $p(r|c,s)$ の r 反応，または確率 $p(l|c,s) = 1 - p(r|c,s)$ の l 反応が生成される。 同じ確率で r 応答を生成する別のサンプリングベースのアプローチでは，1 位と 3 位の特徴入力に基づいて 1 つの音素を選択し， 次にこれらの音素のみに基づいて 1 つの単語を選択し， 選択された単語と 2 位の特徴入力に基づいて真ん中の位置で r と l のどちらかを選択する。 いずれの場合も計算は一方向であり， Massaro モデルが規定しているように， 目標項目に対する文脈上の支持と刺激上の支持は別々に計算される。

ここで， Massaro のフィードフォワード提案と， 前述のようにすべての位置に双方向の計算を適用する相互活性化アプローチを対比してみる。 Massaro モデルでは， 上述の計算は， 2 番目の位置にある音素の事後確率を計算するためにのみ有効である。 Massaro and Cohen (1983) の実験パラダイムでは， 目標は常に 2 番目の位置にある音素であったので， 問題ないと思われるかもしれない (図 8b 参照)。 しかし， Reicher (1969) や Massaro and Klitzke (1979) の実験をはじめ，IA モデルで扱っているほとんどの実験では， どの文字がターゲット文字になるか， 試行の前に参加者は手がかりを得られない。 これらのケースでは，多項 IA モデルは， 4 つの位置すべてにおいて， 正しいベイズの事後確率から同時にサンプリングする。 さらに，MIA モデルは， 可能な単語の分布からのサンプルとして， また可能な各文字の知覚を制約するための基礎として， 単語レベルで同じ表現を使用する。 Massaro モデルでは， 各位置の文脈表現は，その位置からのボトムアップ情報を除外しているため， 単語の識別に関連する情報の不完全な表現となっている。 つまり， 3 つの文字を含む入力に対して， 4 つの異なる単語レベルの量が必要となり， 1 つは単語レベル， 1 つは各文字の位置に必要となる (脚注10)。

脚注 10. Pearl (1982) が示したように， 各位置から上のレベルに渡された情報を記録しておき， 上から下に戻されるトップダウン信号からこれをキャンセルすることが可能である。 このアイデアの先駆者は，Rumelhart (1977) である。 Pearl の提案は， 知覚の相互作用モデルの代替的な実装と考えられる。 このアプローチと MIA モデルとの比較は McClelland (2013) に記載されている。

6.1.1 MERGE および関連モデルにおけるフィードフォワード計算

Norris らは， 単語と文字または音素の知覚処理のモデルにおいて， Massaro のアプローチと非常によく似たアプローチを提唱している (Norris & McQueen, 2008; Norris et al., 2000)。 Massaro モデルと同様， 各文字や音素に必要なトップダウン制約の正しいフィードフォワード計算は， 下位レベルの項目ごとに異なる (例えば， 各位置の音素については， 他のすべての位置の音素に基づいて語彙的文脈を計算する必要がある)。 特に， ターゲットセグメントの識別における文脈の役割を考慮する場合 (例えば， ジョブの最初の 2 つのセグメントが最後のセグメントの識別に与える影響，図8c参照)， 最初の 2 つのセグメントか

らのトップダウンの影響が音素決定層で目標セグメント情報と組み合わされるまでは，目標セグメントに関するボトムアップの情報が単語知覚レベルの値に影響を与えることは許されない (D. Norris, 私信, July 2011)。しかし，音声入力では音声セグメントの情報が重複しているため，これを実現するのは困難である。また，Ganong (1980) の古典的な実験のように，目標セグメントが /g/ または /k/ の後に “iss” または “ift” が続く単語の文脈の最初のセグメントである場合や，曖昧さをなくす文脈が後続の単語に生じる実験 (Warren & Warren, 1971) のように，後続の文脈の効果を考慮すると，困難さはさらに増す。この効果を説明するためには，ターゲットセグメントに続くセグメントが単語レベルに影響を与えることができなければならないが，目標セグメントはそうすることができないようにしなければならない。相互モデルでは，このような複雑さは必要ない。すべての位置にある文脈上の音素は，それぞれの位置にある各音素の処理に同時に影響を与えることができ，各音素に関する決定は，入力の先行または後続の要素に関する情報が得られると更新される。

要約すると，心理学における非相互型モデルは，複数のレベルやレベル内の複数の位置で文脈や刺激情報を同時に利用することを扱っていない。これらのモデルは，特定の目標項目を識別するために，文脈と刺激の情報をある処理レベルで共同利用することに焦点を当てており，自然な知覚状況では，複数項目を多くの異なる処理レベルで同時に解釈することが目標であるという事実を扱っていない。一方，相互モデルでは，異なるレベルやレベル内の異なる位置にある選択肢の表現が，統合された並列・分散・相互的計算の中で，相互に制約し合うことができる。

6.2. 相互的処理のノックオン効果

次に，フィードフォワードモデルに対する相互的モデルの 2 つ目の利点について考える。相互作用により，文脈の影響が知覚系内の後続処理に影響を与えることができる。このような効果には，提示された項目の直近の文脈に存在する隣接項目の処理への影響や，その後の類似した入力の処理への影響などがある。

6.2.1 隣接する入力項目へのノックオンの結果

第一のタイプのケースを Elman & McClelland (1988) が考えた。彼らは，共起語の補償として知られる音声知覚現象に着目した (Mann & Repp, 1981; Stephens & Holt, 2003)。知覚系は，ある音素の発声が隣接音素の音響的実現に与える影響を補償しているようだ。例えば，/s/ や /S/ （“sh”）に関連した唇の形は，/t/ や /k/ のような後続の停止子音の発声にも影響を与え，後続の音の周波数内容を変化させる。知覚者はこれを補うことで，より正確に後続音を認識することができる。したがって，/t/ と /k/ の間の曖昧な音が，/s/ が先行する場合は /k/ として /S/が先行する場合は /t/ として聞こえる傾向があると考えらる。このような状況では，背景音や調音性の変動があると，先行するフリカティブ音の正体が不明瞭になり，厳密なフィードフォワードシステムから補正のための情報が奪われてしまう。しかし，その摩擦音が語彙的に制約のある文脈の中で発生し，フィードバックが文脈上より可能性の高い摩擦音の活性化に影響を与えることができれば，それでも補償が起こり，後続音素の識別が改善される。Elman & McClelland (1984) は，このような補償効果を生み出す機構を TRACE モデルの 1 つのバージョンに組み込み，語彙を媒介とした coarticulation 効果の補償をシミュレートした。

その後，Elman & McClelland (1988) は，TRACE モデルが予測するように，語彙的な文脈が共起語の補償を引き起こすかどうかを調べる実験を行った。実験では /t/ または /k/ の曖昧な音の前に /s/ と /S/ の中間の曖昧な摩擦音を提示した。そして，その曖昧な摩擦音の前には /s/ に一致するもの (例: Christma_) と /S/ に一致するもの (例: “fooli_”) の 2 つの異なる語彙文脈のいずれかが提示された。もし，語彙情報が音素処理に影響を与えるようであれば “Christma_” の曖昧な摩擦音は，音響的には /s/ のように振る舞い，後続の音素の知覚を /k/ にシフトさせるはずである。逆に，“fooli_” の同じ曖昧な摩擦音は，音響的に /S/ のように振る舞い，後続の音素の知覚を /t/ の方向にシフトさせるはずである。これはまさに Elman と McClelland が発見した結果である。この結果は疑問視されているが (Pitt & McQueen, 1998)，複数の異なる研究室で，異なる教材を使って再現されている (Magnuson, McMurray, Tanenhaus, & Aslin, 2003; Samuel & Pitt, 2003)。しかし，非対話型のアプローチを支持する人たちは，最近，効果の原因をさらに論じる証拠を提示しており(McQueen, Jesse, & Norris, 2009)，このテーマに関する研究は続いている。

6.2.2 同様の入力を次の機会に処理する際のノックオン効果

他の研究者たちは，音素識別における語彙的文脈の影響についても研究しているが，これも相互的説明によって予測される。選択的適応とは，特定の刺激を繰り返し提示することで，中立的な刺激が繰り返し提示された刺激に似ていないように知覚されるという，領域一般的な現象である。繰り返し提示された刺激は，中立的な刺激よりも似ていないと知覚されるようになる。音声知覚の場合，ある音素 (例: /s/) を繰り返し提示すると，曖昧な音素 (例: /s/ と /S/ の中間) の知覚が別の解釈 (この場合は /S/) にシフトする。(Samuel, 1986; Samuel & Kat, 1996など)。語彙を介した選択的適応を実証するために，中立音 (曖昧な音素またはノイズバースト)を，1 つの解釈のみに一致する語彙的な文脈で繰り返し提示した。中性音が bronchiti_ や arthriti_ などの /s/ に偏った文脈で提示された場合，選択的に適応された表現は /s/ であり、aboli_ や demoli_ などの /S/ に偏った文脈で提示された場合，選択的に適応された表現は /S/ であった (Samuel, 1997, 2001)。このように，語彙情報がどの部分語彙表現を選択的に適応させるかを決定し，その後の音素・単語識別に影響を与えていた。

McClelland and Elman (1986) の TRACE モデル論文で予測されていた，語彙フィードバックのノックオン結果の 3 つ目の例は，音声音カテゴリーの語彙によるチューニングである。音素カテゴリーの境界は個人によって異なるため，リスナーが異なるスピーカーの音声を正しく識別するためには，このようなチューニングが不可欠である。例えば，英語とスペイン語の話者は，/b/ と /p/ のカテゴリーの中心を，音声開始時間と呼ばれる次元に沿って異なる場所に設定する。さらに，地方の方言は，子音と母音の両方の生成の詳細の違いによって区別されることが多い。そして，Norris, McQueen, Cutler (2003) をはじめとするいくつかの研究により，音声知覚において実際にこのような調整が行われていることが示されている (van Linden & Vroomen (2007) では，唇からの視覚的な手がかりの使用に類似した変化が見られた。この効果の語彙以前の位置は，同調効果が一般化して，効果の誘発に使用されなかった単語の知覚に影響を与えるという証拠によって裏付けられている (McQueen, Cutler, & Norris, 2006)。Mirman, McClelland, and Holt (2006) は，TRACE に単純なヘブの学習則を追加し，特徴と音素の接続を調整することで，関連する実験結果をシミュレートした。

より一般的には，Friston (2003; Spratling & Johnson, 2004) は，トップダウンのフィードバックが，知覚および認知システム全体に見られる階層的な表現を学習するために必要であると主張しており，実際，多くの異なるニューラルネットワーク学習アルゴリズムにおいて，何らかの形でフィードバックが使用されている。知覚の自律的フィードフォワード理論の支持者は，学習のためのフィードバックの必要性を認めているが，このフィードバックは，相互活性化モデルの処理に影響を与える「オンライン」のフィードバックとは同等ではないと主張している (例えば Norris et al, 2003)。我々は，フィードバックが知覚と同様に学習を導くことができる系が，解析的な説明を提供することを主張する。さらに，もしフィードバックが学習を導くのであれば，学習された表現は必然的にボトムアップとトップダウンの情報の組み合わせを反映することになり，表現自体が相互的処理における役割の結果であると同時に本質的なものであると言える。

つまり，フィードバックは，文脈上の制約によって，より大きな単位 (単語など) の要素 (文字や音素など) の同一性を決定するだけでなく，この文脈上の決定された識別処理の結果が，隣接要素の処理 (共起語の補償) や，同じ要素のその後の出現に影響を与えることができる (適応，再チューニング)。フィードバックのノックオン効果は，知覚におけるトップダウンの直接的なフィードバックの動機付けと証拠となる。

7. 相互処理の神経基盤

7.1. 基礎的神経科学の知見

知覚の神経基盤に関する研究から， 脳内での相互的処理の存在が示唆されている。 相互的な処理は， 脳の構造の基本的な特徴によって支えられている。 新皮質では， 領域 A から領域 B への「前向き」経路があるところには， 強い (時にははるかに強い) 「帰還」経路が存在する傾向がある (Felleman & van Essen, 1991)。 これに対応する多くの研究では、上位または下流にあると考えられる皮質領域（例えば、上位の視覚野や聴覚野）を可逆的に不活性化すると、一次領域の刺激駆動活動が影響を受けることが示されており（例えば、Hupe et al, 1998; Carrasco & Lomber, 2010)， 皮質の処理における相互作用が示唆されている。 アカゲザルの神経記録から， 物理的に存在するエッジに反応する V1 の「エッジ検出器」と同じものが， カニッツァ図形の錯視的な輪郭にも反応することがわかった。 V1 での錯視的輪郭の反応は V2 での反応よりも遅れて起こることがわかり， V1 での反応は高次の視覚処理からのフィードバックによるものであることが示唆された (Lee & Nguyen, 2001)。 同様に， 両眼視差は， V1/V2 から視床下部の皮質領域まで， 多くの異なる視覚領域のニューロンが， 大域的な知覚との整合性を示しながら， 相互に制約を満足させ， 相互に活性化する処理過程であると考えられる (Leopold & Logothetis, 1999)。 後頭側頭領域と前頭前野の間で活動が双方向に伝播している証拠は， ヒトの視覚物体認識の脳磁図 (MEG) 研究でも見られる (例えば， Bar, 2004)。

処理モダリティ内のより高いレベルからのトップダウンのフィードバックに加えて， 神経生理学的研究では， 知覚処理の主要領域間のクロスモダリティの相互作用が示されている (レビューは Ghazanfar & Schroeder, 2006 参照)。 我々にとって， このようなモダリティ間の相互制約は， 階層的な知覚系におけるレベル間の双方向の相互作用と同様に， 相互制約充足の基本原理の例である。 感覚統合は， 二次感覚皮質または連合皮質 (Bavelier & Neville, 2002; Jones & Powell, 1970) または前頭皮質 (Rizzolatti, Riggio, Dascola, & Umlita, 1980) で行われると主張する研究がいくつかあるが， 最近の証拠では， 聴覚において， これらの連合領域から一次感覚皮質へのトップダウン入力の存在が指摘されている （Cappe & Barone, 2005; Schroeder et al, 2001) や視覚 (Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003)， また聴覚野から一次視覚野への直接入力 (Falchier et al, 2002; Hall & Lomber, 2008) やその逆の入力 (Bizley & King, 2009) もある。 聴覚皮質からの物理的な投射は， サル (Falchier et al., 2002; Rockland & Ojima, 2003) や成猫 (Hall & Lomber, 2008) でも観察されており，これらの接続は初期の発達段階に限らないことが示唆されている。 さらに， フェレットのマルチユニット記録から， A1 領野のニューロンの約 20% が視覚刺激に反応していることがわかっている (Bizley & King, 2009)。

全体的に， 感覚処理が神経レベルでカプセル化されているという考えに疑問を投げかける証拠が増えている(Ghazanfar & Schroeder, 2006 参照)。 むしろ， 高度に相互作用的な生物学的系が， 空間定位のために複数のモダリティからの情報を階層的に同時に利用することを可能にしていることを示唆している。 その結果， 高度に相互作用的な生物学的系により， 複数のモダリティからの階層的な情報を同時に利用して， 空間定位や通信， さまざまな社会的行動を行うことができると考えられている (Lewkowicz & Ghazanfar, 2009)。 この相互的神経系は， 多くのレベル， 多くのモダリティの表現を同時に， かつ首尾一貫して使用することに依存した認知処理を実行する。

7.2. 人間の言語処理の脳内機構における相互性

人間の言語処理や読解に関する研究では， 相互的処理も重要なテーマとなっている。 この研究の多くは， 単一単語読解の「三角モデル」(Seidenberg & McClelland, 1989; およびその後の拡張) の枠組みの中で行われてきた。 このモデルは， 正書法， 音韻， および意味レベルのユニットの局所的な表現ではなく， 学習された分散表現に依存する相互活性化モデルのバージョンと見なすことができる。 ここでは，図 9 に示した枠組みの相互的処理の側面に焦点を当て， 音韻と正書法の相互影響のタイミングと場所， および音韻と正書法の処理に対する語彙の影響に焦点を当てる。 三角モデルでは， モデル全体の双方向の接続は， 語彙知識と， 正書法と音韻表現の間の共分散のパターンの知識に影響されやすいことに注意。 具体的には， 視覚的または音声的な単語の形が提示されると， 表音文字， 音韻， 意味の各表現間の双方向の相互作用が誘発されることから， 少なくとも熟練した読者においては， 経験によって関連する接続が強化されると， 語彙的知識や綴りと音の一貫性が表音文字と音韻の表現に影響を与えると予測される。

視覚的単語認識の神経基盤については， 左後頭側頭皮質の Visual Word-Form Area (VWFA; McCandliss, Cohen, & Dehaene, 2003; Dehaene, Cohen, Sigman, & Vinckier, 2005) と呼ばれる領域の役割が大きくクローズアップされている。 VWFA は， 正書法の「入力」レキシコンとして機能し， 単語の視覚的な形を保存する場所であるという説 (Kronbichler et al. 2004, 2007) と， この領域は前語彙的な性質を持ち (Dehaene et al. 2005)， VWFA 内またはその付近で正書法の表現が階層的に構成されている可能性があるという説がある。 相互的枠組みでは， 表現は正書法的に構成されていても， 語彙的な制約や他の入力モダリティからの影響を受けやすくなる。 つまり， VWFA は， 三角モデルで「正書法」とラベル付けされたユニットのプールのおおよその神経的比肩物であると考えることができる。 このユニットは， 主に正書法の構造を表すが， 他の表現からの相互作用的な影響にも敏感である。 この領域での処理は， 他の入力モダリティからの影響を受けやすいという見解を裏付ける証拠が数多くある。 例えば， 先天的に盲目の患者の触覚 (点字) 入力から生じる影響 (Buchel, Price, Frackowiak, & Friston, 1998; Cohen et al, 1997)， 手書きの場合 (Barton, Fox, Sekunova, & Iaria, 2010)， 聴覚による単語処理の場合 (Binder, Medler, Westbury, Liebenthal, & Buchanan, 2006; Cone, Burman, Bitan, Bolger, & Booth, 2008; Desroches et al. , 2010)。 また， 単語のつづりと音の一貫性の影響を調べた研究では，一貫性と頻度が項目レベルで段階的に影響することが明らかになっており， 命名における一貫性効果の行動学的知見を反映している (Bolger, Hornickel, Cone, Burman, & Booth, 2008; Bolger, Minas, Burman, & Booth, 2008; Graves, Desai, Humphries, Seidenberg, & Binder, 2010)。 三角モデル (Harm, McCandliss, & Seidenberg, 2003) の予測と一致するように，Bolger, Hornickel, et al.（2008）と Bolger, Minas, et al. (2008) は， VWFA の書記素-音素一貫性への反応が読解力に応じて増加することを見出した。 これらの知見は， 読解力が自動化されていくにつれて， 相互的処理が確立されていくという見解を支持するものである。 このことは， 三角モデルの枠組みにおいて， 3 つの異なるタイプの表現にそれぞれ関与するニューロン間の双方向接続が， 経験によって強化されるという形で捉えられている。

音声知覚の神経画像研究でも， 相互的モデルの予測が取り上げられている。 音韻認識の正確さは上側頭葉と関連しているのに対し， 決定時間は下前頭葉/半島皮質 (Binder, Liebenthal, Possing, Medler, & Ward, 2004) と前帯状皮質/内側前頭葉領域と関連している (Grinband, Hirsch, & Ferrera, 2006; Grinband et al., 2011)。相互的モデルでは， 音韻処理に関与する脳領域 (上側頭溝の後部側頭回やヘッシェル回など) が語彙バイアスの効果を示すはずだと予測される。 一方， 自律的意思決定レベル統合モデルでは， これらの語彙バイアス効果は， 意思決定や反応選択に関与する脳領域 (下前頭回や前帯状回など) に限定されるはずだと予測される。 fMRI 研究 (Myers & Blumstein, 2008; Guediche, Salvata, & Blumstein, 2013 も参照) によると， 曖昧な音素の分類に関する語彙バイアスは， 上側頭回の活性化の増加と関連した。 この領域は， 患者集団における声の幻聴 (Dierks et al., 1999) や， 健常者における他者の発話を想像する際にも活性化される (McGuire, Silbersweig, & Frith, 1996)。

電気生理学的測定により， 語彙効果と一貫性効果は， 視覚と聴覚の両方のモダリティにおいて， 知覚後の決定段階ではなく， 知覚および/または語彙処理の初期段階で生じるという重要な証拠が得られた。 例えば， 正字体の異なる単語の視覚処理における韻律効果は， 刺激開始後 260 ms 程度で検出され (Kramer & Donchin, 1987)， また， 視覚的な単語処理における音節効果は， 250-350 ms 程度で示されている (Ashby & Martin, 2008; Carreiras, Ferrand, Grainger, & Perea, 2005)。 聴覚的語彙決定課題 (Perre, Midgley, & Ziegler, 2009; Perre & Ziegler, 2008) や意味カテゴリー化課題(Pattamadilok, Perre, Dufau, & Ziegler, 2008; Pattamadilok, Morais, De Vylder, Ventura, & Kolinsky, 2009) における一貫性効果は， 刺激後およそ 300-350 ms に ERP で発生し， 不整合点にタイムロックされる

ことが示されている。 また， より高い空間分解能を持つ MEG イメージングの結果から， 視覚課題における初期の韻律効果は， 左後頭側頭部に局在することが明らかになっている (Wilson, Leuthold, Lewis, Georgopoulos, & Pardo, 2005)。 これに関連して，van Linden ら (van Linden, Stekelenburg, Tuomainen, & Vroomen, 2007) は， 語彙的な文脈が知覚に基づく早期のミスマッチネガティブ効果を誘発することを見出し， 語彙情報が知覚処理段階に直接影響を与えていることを示唆した。

ヒトの言語処理の領域では， ニューロンレベルの神経解剖学的な精度を達成することは困難であるが， 複数のイメージングモダリティを組み合わせた最近の研究では， 空間的および時間的な精度を高めることが期待されている。 MEG および脳波と解剖学的 MRI を組み合わせた研究 (Gow, Segawa, Ahlfors, & Lin, 2008) では， 語彙処理に関連する領域 (縁状回) の活性化に伴って， 上側頭回後部が再活性化することが明らかになった。 ERP 研究 (Molinaro, Dunabeitia, Marın-Gutierrez, & Carreiras, 2010) では， 単語文脈の中の文字のような数字 (M4T3R14L など) は， 初期の段階 (発症後 180-220 ms) では文字よりも数字のように処理されるが， わずかに後の段階 (発症後 250-300 ms) では， このパターンが逆転し， 文字のような数字は数字よりも文字のように処理されることがわかった。 ERP-MEG を併用した研究 (Sohoglu, Peelle, Carlyon, & Davis, 2012) では， 劣化した音声の知覚的明瞭度に対する予備知識 (書かれたテキスト) の促進効果を再現し， この効果が上側頭回の活動よりも下前頭回の活動に反映されることを見出した。 これは， 下前頭回の高次処理からのトップダウン・フィードバックが上側頭回の知覚処理を調節していることと一致する。

視覚および聴覚の単語知覚における語彙効果と一貫性効果の正確な性質， タイミング， および位置については， 依然としてさまざまな解釈がなされており， これらの問題を解決するための研究が数多く行われている。 非常に一般的な未解決問題としては， トップダウンやモダリティ間の影響を， 相互活性化の枠組みのように， 解釈を制約する新たな要因とみなすべきなのか， それとも代わりに， トップダウンの信号は， ボトムアップの信号と比較され， エラー信号を生成して， 学習メカニズムを駆動する予測とみなすべきなのか， ということである (Friston, 2008; Mumford, 1992; Rao & Ballard, 1999)。 さらに， このような影響と， 脳領域内および脳領域間の神経活動の同期との間の相互作用についても疑問がある (最近の議論として， Gotts, Chow, & Martin, 2012 およびその解説参照)。

言語処理やその他の課題において， トップダウンの影響が比較的早い段階で， モダリティに特化した処理領域に影響を与えていることは疑う余地がないように思われる。 脳の領域は双方向に接続される傾向があり， これらの双方向の接続が知覚や概念処理において相互に活性化されるという強い神経生理学的証拠がある (Ghuman, Bar, Dobbins, & Schnyer, 2008; Gotts et al., 2012)。 特に， 言語処理領域では， フィードバックや聴覚・視覚の相互作用が知覚処理に直接影響を与えることが神経学的に証明されており， 相互的モデルと一致している。

8. まとめと今後の方向性

本論文では， 知覚と認知における相互活性化と相互制約充足の場合を説明してきた。 IA モデルが最初に取り組んだ対象現象である， 視覚と話し言葉の認識に主に焦点を当ててきたが， 相互的アプローチの他の応用例についても検討してきた。 計算論レベルの考察と神経科学的な証拠， そして行動研究によって明らかにされた知覚における文脈の役割に関する証拠を検討した。

我々は， 相互活性化が， 知覚系が直面している重要な計算上の課題を解決し， 脳における知覚や言語処理機構に関する行動科学や神経科学の証拠を含む幅広い証拠と一致することを主張した。 全体として， 計算論上の解析と行動・神経科学的な証拠の両方が， IA 仮説と一致しているように見える。

計算論と実証に基づく考察は， 相互的視点を強く支持しているように見えるが， 相互活性化の枠組みの中で， 今後の研究を必要とするいくつかの重要な課題がある。

8.1 確率的に根拠づけられた相互活性化モデルにおける知覚のダイナミクス

IA 仮説では， 処理は情報が利用可能になったときに実時間で最適な結果を得るという理想に近づくとしている。 多くの実験が行われ， 知覚や言語処理課題の参加者は， 利用可能なすべての情報を使用し， その情報が感覚表面に到着してから 3 分の 1 秒以内にその情報に対する感度を示し始めることが示されている。 このような発見のシミュレーションは， オリジナルの TRACE モデルや， 関連する単純な Luce 比ベースのモデルを用いて行われている (Spivey & Tanenhaus, 1998)。 今後は， 多項 IA モデルのような確率的に根拠のあるモデルを用いて， これらの問題をより詳細に検討していく必要がある。

我々は、多項 IAモデル (Khaitan & McClelland, 2010) に関連する問題を探究し始めている。 すなわち， 参加者が標的情報への暴露量を増やしていくことで，成績と成績に対する文脈的影響が蓄積されていくという問題である (Massaro & Klitzke, 1979)。 この問題は Massaro and Cohen (1991) が，McClelland (1991) で提示された確率バージョンのモデルでは十分に対応できなかった， 相互活性化モデルへの挑戦として特に提起したため，重要である。 具体的には， Massaro and Cohen (1991) が提案した経験的関数に従って入力特徴情報が時間をかけて蓄積される場合， 多項 IA モデルが提供する処理装置は， ランダムな配列の文字と比較して， 単語の中の文字知覚に対して正しいパターンの増強を示すのだろうか？ Khaitan and McClelland (2010) で報告されたシミュレーションは，この質問に対する答えがイエスであることを示唆しているが， このシミュレーションは予備的なものであり， さらなる研究が必要である。

8.2 課題および教育上の制約に対する適応的な最適化

さらなる研究のための重要な話題は， 課題や命令の制約に応じて， 相互活性化モデルにおける処理の適応的な最適化である。 ここにはいくつかの重要な未解決問題がある。 まず， これまで述べてきたように， 参加者は， 刺激項目が単語または非単語である確率の変化に応じて， 処理に語彙的な影響を与える程度を調整する。 このような影響は， ベイズモデルに容易に組み込むことができ (Rumelhart & Siple, 1974 はこの問題を広範囲に検討している)，オリジナルの IA および TRACE モデルにも組み込まれている (Mirman et al. 2008)。 しかし， 参加者が音声の同一性に関する語彙的制約の知識の使用を実際に中断できる範囲には限界がある。 例えば， 最近の研究 (Hawthorne, 2011) では，2 つの文脈でそれぞれの音が同じ頻度で発生することを知らされていてもいなくても， 被験者は発話音の知覚に語彙的な影響を示した。 これは， 自然主義的な言語処理に関わるニューロンの接続に語彙的制約の知識がハードワイヤードされていて， 指示操作とは無関係に同じニューロンと接続に依存していた場合に予想されることである。 ここには、さらなる検討に値する経験的および理論的な問題がある。

8.3 相互的知覚モデルに学習と分散表現を組み込む

知覚の相互活性化モデルに関する研究は， 1980 年代半ばに開発された並列分散処理システムのための強力な学習モデルの開発に先立って行われた。 学習された分散表現を用いたモデルは， 言語処理や意味処理の幅広い局面で成功を収めており， 今後の知覚処理課題の探求において， 学習と分散表現が完全に統合されることを期待している。 最近開発されたディープ・ビリーフ・ネットワークのための高速で強力な学習方法 (Hinton, 2014) は，このような研究を促進するはずである。

8.4 自然な知覚状況で知覚・認知システムが直面する計算上の課題への対応

IA モデルと TRACE モデルは，自然な知覚状況で人間の能力に匹敵するだけのロバスト性と効率性を備えたモデルを開発するために必要な多くの課題を克服している。 これらの課題は，AI，マシンビジョン， 機械学習の分野の幅広い研究者の間で熱心に研究されている。 これらの研究の多くは，IA モデルやその前身に由来するニューラルネットワークのアイデアをベースにしており， もちろん， 多くの研究では， 明示的な確率的推論機構が組み込まれている。
そして， これら研究の多くは， 人間の知覚処理機構を理解するための努力にフィードバックされるべきであり， それらは脳が提供する神経機構に具現化されるからである。 知覚の相互活性化モデルのさらなる発展は， これらの開発から大きな恩恵を受けるだろう。

8.5 IAモデルを脳が提供する神経メカニズムに完全に基づかせる

最後に挙げる課題は， IA 処理が脳内の神経機構にどのように実装されているかを正確に理解するという目標である。 ニューロンとその特性は， これらのモデルの開発においてインスピレーションの源となってきた。 また， 神経科学からの証拠は， これまでにレビューしてきたように， 脳における知覚処理が IA に似た処理であるという見解を支持している。 知覚を生み出す神経機構を統合的に理解することは， 多くの研究者が目指している目標である。 もし IA 仮説が正しければ， そのような統合的な理解は， 相互活性化の原理に依存することになる。