Interactive activation model as Bayesian inference

Pramit Mallick[1]
pm2758@nyu.edu

Srinidhi Goud[1]
sgm400@nyu.edu

[1]**Department of Computer Science**
New York University

Abstract

The Interactive Activation and Competition (IAC) model is an connectionist model that uses bidirectional excitatory connections and pools of mutually inhibitory units as features to emulate the behaviors of human memory and generalizations we draw from them. While these connectionist artificial neural network models are largely successful, they draw criticism for not satisfying Bayesian conditions that can explain simple human characteristics. This work attempts to review those claims and reimagines the IAC as a Bayesian inference problem and studies the effects of noise on some of the characteristics of human memory on the model.

1 Introduction

There has been extensive work both in psychology and in cognitive modeling to help better understand how humans deal with visual word recognition and reading. Some of the early works ((Cattell, 1886), (Bagley, 1900)) recognized the influence of context in perception of letters and words when experimenting with visual and auditory stimuli. This observation was reinforced by (Pillsbury, 1897) on finding that context could override sensory input as well. Further work by (Reicher, 1969) also showed that when trying to guess and remember words (on being shown brief visual stimuli, sometimes occluded), subjects were more accurate when the letters formed an actual word. This word superiority effect was later studied by many (cite others here) in varied circumstances that only corroborated the influence of context in tasks such as visual word recognition and reading.

Largely, it can be seen that there has been quite a clear distinction in the type of work that tried to model the phenomenon mentioned above. Some tried to address the issue using the connectionist approach that tried to model the context while others questioned the very non-independence assumption between context and the stimuli. Below, we try to summarize some of the related work.

1.1 Related Work

One of the first successful attempts at trying to model this behavior came in the form of a the Interactive Activation and Competition (IAC) network formulated by (McClelland and Rumelhart, 1981) and (Rumelhart and McClelland, 1982), based on the work by (Grossberg, 1978). The model is essentially an artificial neural network used to model memory and intuitive generalizations by way of bidirectional excitatory connections between related nodes and competitive pools of mutually inhibitory units. One of the best illustrations of the network was by taking the ”West Side Story example” as shown by (McClelland, 1981)) to demonstrate the flexibility of the network in retrieving specifics and generalizations from a small dataset. While this connectionist type of model seemed largely successful, work by (Massaro, 1989) pointed out that the original IAC model failed to produce the pattern of logistic additivity that could easily be captured by earlier models and Bayesian considerations and critiqued the bidirectional flow of activation signals. Another work TRACE by (McClelland and Elman, 1986) tried a similar approach to model speech perception.

Other models like the fuzzy logical model of perception (FLMP) by (Oden and Massaro, 1978) also tried to model the effect of context in letter perception and contrasted against other methods as in the work of (Massaro, 1979) and (Massaro, 1989) and some (Massaro and Cohen, 1991) even went on to question the non-independence influences between stimulus and context information on perceptual recognition. Efforts ((McClelland et al., 2006)) to model the context were challenged ((Norris et al., 2000), (Norris et al., 2008)) by arguing that bi-directional propagation of information would lead to violations of correct probabilistic Bayesian inference.

Some other work by (Pearl, 1982) (which presented a generalized method of using Bayesian likelihood-ratio updating for propagation of beliefs in hierarchically organized structures with multi-hypotheses variables) provided an interactive framework to perform inference which could be used to help substantiate the IA model. A newer version of the IA model, called the MIA (Multinomial Interactive Activation model) by (Khaitan and McClelland, 2010) addressed the concern of producing the correct posterior probabilities by replacing the inhibition between units within pools with the selection of a single unit using the softmax function to assign probabilities to candidate alternatives. The detailed explanation of integrating probabilistic models and interactive networks was explored in (McClelland, 2013) as well.

Our work picks up from the same thread and attempts to model the "Jets and Sharks" dataset using a Bayesian network and shows how it converges to the outputs of the interactive model. We also explore how the Bayesian network can form a memory system like the IA, exhibiting content adressability, graceful degradation, retrieval by name and spontaneous generalization.

2 Methods

The canonical view of Bayesian formula expresses the posterior probability of one of two (h_1, h_2) mutually exhaustive and exclusive hypotheses given some evidence e as below:

$$p(h_i/e) = \frac{p(h_i)p(e/h_i)}{p(h_1)p(e/h_1) + p(h_2)p(e/h_2)} \quad (1)$$

where i can be 1 or 2. Here p(h_i) is called a Prior as it is a given or an assumed quantity. Given the priors and probabilities of evidence given hypotheses (p(e/h_i)) we can calculate the posterior probabilities from this equation.

The equation (1) can be extended to multiple alternative hypotheses as follows:

$$p(h_i/e) = \frac{p(h_i)p(e/h_i)}{\sum_j p(h_j)p(e/h_j)} \quad (2)$$

where j,i $\epsilon\{1, N\}$ for N number of hypotheses.

Now consider the case where there are multiple conditionally independent evidences. The equation (2) will be further extended to:

$$p(h_i/e_1\&e_2\&...\&e_n) = \frac{p(h_i)\Pi_{i'}p(e_{i'}/h_i)}{\sum_j p(h_j)\Pi_{i'}p(e_{i'}/h_j)} \quad (3)$$

where $i'\epsilon\{1, n\}$ for n number of evidences.

Upon replacing the variables in the equation (3) by a variable called Support(S_i) as used by (McClelland, 2013), we arrive at the following equation:

$$p(h_i/\{e\}) = \frac{S_i}{\sum_{i'} S_{i'}} \quad (4)$$

where $S_i = p(h_i)\Pi_{i'}p(e_{i'}/h_i)^{v_j}$ and v_j is just a encoding of the evidences in a d dimension vector. Taking logs on S_i we see that:

$$log(S_i) = log(p(h_i)) + \sum_j v_j log(p(e_j/h_i)) \quad (5)$$

This is similar to the representation of an output of a unit in a neural network:

$$net_i = b_i + \sum_j w_{ij}a_j \quad (6)$$

Comparing equations (5) and (6), we can see that both are interchangeable when we replace the bias b_i with $log(p(h_i))$, the weights w_{ij} with $log(p(e_j/h_i))$ and the activation a_j with v_j. Moreover, taking softmax ($p_i = \frac{e^{net_i}}{\sum_{i'} e^{net_{i'}}}$) on the equation (5) and replacing the variables we can show that we arrive back at the equation (4).

Moreover, taking log on (3) we arrive at the equation:

$$\begin{aligned} log(p(h_i/e_1\&e_2\&...\&e_n)) = log(p(h_i)) + \\ \sum_{i'} v_{i'} log(p(e_{i'}/h_i)) - \\ log(\sum_j p(h_j)\Pi_{i'}p(e_{i'}/h_j)) \end{aligned} \quad (7)$$

which is very similar to the equation used to calculate the net input in IAC by (McClelland and Rumelhart, 1981):

$$input_i(t) = p_i(t) + E\sum_j e_{ij}(t) - I\sum_j i_{ij}(t) \quad (8)$$

Therefore, we can conclude that the neural network used in the IAC model by (McClelland and Rumelhart, 1981) is relatable to a Bayesian network that calculates posteriors at its units.

2.1 Inference using Gibbs sampling/Metropolis Hastings

The goal of the Bayesian inference is to maintain a full posterior probability distribution over a set of random variables. However, maintaining and using the distribution often involves computing integrals which, for most non-trivial models, is intractable. Sampling algorithms based on Monte Carlo Markov Chain (MCMC) techniques are one possible way to go about inference in such models.

The underlying logic of MCMC sampling is that we can estimate any desired expectation by ergodic averages. That is, we can compute any statistic of a posterior distribution as long as we have N simulated samples from that distribution:

$$E[f(s)]_P \approx \frac{1}{N}\sum_{i=1}^{N} f(s^{(i)}) \tag{9}$$

where P is the posterior distribution of interest, f(s) is the desired expectation, and $f(s^{(i)})$ is the i^{th} simulated sample from P. Gibbs sampling is one MCMC technique suitable for the task. The idea in Gibbs sampling is to generate posterior samples by sweeping through each variable (or block of variables) to sample from its conditional distribution with the remaining variables fixed to their current values. For instance, consider the random variables X_1, X_2, and X_3. We start by setting these variables to their initial values $x_1^{(0)}, x_2^{(0)}$, and $x_3^{(0)}$ (often values sampled from a prior distribution q). At iteration i, we sample $x_1^{(i)} \sim p(X_1 = x_1 | X_2 = x_2^{(i-1)}, X_3 = x_3^{(i-1)}$, sample $x_2^{(i)} \sim p(X_2 = x_2 | X_1 = x_1^{(i)}, X_3 = x_3^{(i-1)}$, and sample $x_3^{(i)} \sim p(X_3 = x_3 | X_1 = x_1^{(i)}, X_2 = x_2^{(i)}$. This process continues until convergence (the sample values have the same distribution as if they were sampled from the true posterior joint distribution). However, Gibbs sampling has several limitations. The alternative to handle these limitations as explained by Scott(2007) (p.108) is to use Metropolis-Hashtings (MH) algorithm.

2.2 Bayesian Networks and Inference using Loopy Belief Propagation

One of the ways to represent conditional dependencies in a joint probability distribution using a compact factorized representation is to use Bayesian networks which are a probabilistic graphical model that takes advantage of conditional independence (edges) via a directed acyclic graph (DAG). We can model the features of the data as the parent nodes to the latent hidden class (in our case, the instance value), thus giving us the ability to use inference methods to help evaluate various probabilistic queries. While it is possible to use variable elimination as the inference algorithm (because the network is small), we choose the approximate method of loopy belief propagation in favor of a more generalized approach (other approximate methods like MCMC and importance sampling can also be used here).

Belief propagation ((Pearl, 1982)) is an iterative algorithm to calculate the estimated marginal distribution (called beliefs) for each unobserved node, conditional on any observed nodes. It is achieved by first converting the Bayesian network into a Factor graph (which is a bipartite graph representing the factorization of a function) by adding factor nodes on the edge in between two nodes. Then an series of messages are passed between the nodes that express the beliefs of the marginal probabilities till they converge. This sum-product algorithm is an exact inference method and guarantees to converge to the true marginals, and while there is no guarantee of convergence in general graphs (with loops), it acts as a good approximate algorithm to calculate the marginal probabilities. Specifically,

- A message from a variable node v to a factor node a is the product of the messages from all other neighboring factor nodes (except the recipient)

$$\forall x_v \in Dom(v),\ \mu_{v\to a}(x_v) = \prod_{a^* \in N(v)\setminus\{a\}} \mu_{a^*\to v}(x_v). \tag{10}$$

- A message from a factor node a to a variable node v is the product of the factor with messages from all other nodes, marginalized over all variables except the one associated with v

$$\forall x_v \in Dom(v),\ \mu_{a\to v}(x_v) = \sum_{\mathbf{x}'_a : x'_v = x_v} f_a(\mathbf{x}'_a) \prod_{v^* \in N(a)\setminus\{v\}} \mu_{v^*\to a}(x'_{v^*}) \tag{11}$$

where $N(a)$ is the set of neighboring (variable) nodes to a. If $N(a) \setminus \{v\}$ is empty then $\mu_{a \to v}(x_v) = f_a(x_v)$, since in this case $x_v = x_a$ (in the case of edge nodes).

These messages are iteratively updated till convergence using an optimal scheduling algorithm which can reach convergence for the cases where the factor graph is a tree. For general graphs (with loops) there is no optimal scheduling algorithm and thus all the messages are updated simultaneously.
In our case, the network is extremely simple and has no cycles, and thus we see, demonstrably, that the algorithm actually calculates the true marginal probabilities.

3 Data and Model

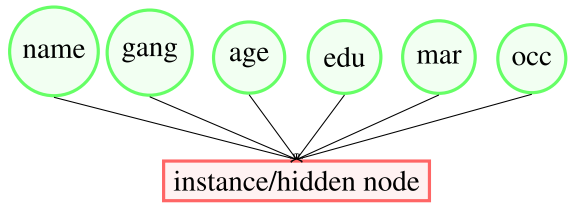
We try two approaches. One as outlined in (McClelland, 2013), to relate the weights and activations of the IAC network to the actual probabilities. Another approach explored is to define the model as a Beyesian network (a probabilistic graphical model) and use approximate inference algorithms to calculate the posterior probabilities used to answer queries.

3.1 IAC as probabilistic model

The graph is constructed similar to the original IAC graph. However, the goal is to achieve the posterior probabilities by sampling from the joint posterior of the generative model. The approached used here is a Bayesian procedure known as Gibbs sampling 2.1. At hidden node level, each hidden instance has a prior value $p(h_i)$. Since, hidden nodes are one-to-one mapping to the names, the priors are calculated from the feature counts of 'name' feature. The connection weights linking each feature to the hidden state is equal to $p(e_i/h_j)$. The information is passed as logarithmic additions from features to hidden states and from hidden states to features, iteratively. The weights are considered only from features to hidden states. At each step, the active node from each pool is selected based on Softmax. The state of activation in the network after settling for many iterations will be a sample from the joint posterior of the generative model (McClelland, 2013). While the model showed some progress in its updates, our model failed to replicate the results of IAC.

3.2 Using Bayesian Network

For the purposes of modeling the "Sharks and Jets" dataset as a Bayesian network, we can take the observed features/evidences such as 'name', 'gang', 'age', 'edu', 'mar', 'occ' and use them as parent nodes (conditional variables) to the hidden node which would take on the values of the instance variable as shown below.



For each of the evidences (e), we can easily calculate the $p(e) = \frac{\#\text{count of e}}{\#\text{total count of data}}$. This leaves us with task of filing up the conditional probability table (CPT) of the hidden node. Studying the dataset, we see that there is a since each sample belongs to an individual, it can be classified as its own class (leading to 27 classes/instances in total). Thus the probability of the instance given the evidences is 1. Thus we can create the CPT by creating a cross product of all the possible combinations of the evidences and assigning the conditional probability to 1 if it exists in the data and 0 otherwise. Thus we are now ready to perform inference on network and compare it to the IAC model.

3.3 Adding noise

It is also interesting to see how various models perform under the influence of some noise (whether they can still correctly determine the posterior probabilities). To inject noise in our simple Bayesian network, we would have to alter the CPT table. In our previous simplified assumption, we set the value of $p(f_i|hidden_k) = 1 \forall f_i \in F$ (feature) if there existed a data point which reflected the condition. We alter this probability to reflect the noise in our data. We sample from a gaussian distribution ($\mu = 0, \sigma = 0.005$) and set it as the value of $p(f_j|hidden_k) \vee i = j$ for the features that do not occur with the given hidden class and set the actual feature occurance to be $p(f_i|hidden_k) = 1 - \sum_{j \in F \setminus \{f_i\}} p(f_j|h_k)$. Thus embedding our noise in our probability distributions, i.e we do not say with absolute certainty (of 1) that given the hidden class, it has a feature. We then use these noisy probabilities to construct our CPT using the equation 3. Giving rise to a noisy

CPT as shown in the figure below -

gang	age	edu	mar	occ	name	noisy prob	actual
Jets	20s	HS	single	bookie	Pete	0.0213138	1
Jets	20s	HS	single	bookie	Ned	4.0607e-07	0
Jets	20s	HS	single	bookie	Jim	3.65463e-06	0
Jets	20s	HS	single	bookie	Ralph	3.65463e-06	0
Jets	20s	HS	single	bookie	Nick	7.30927e-06	0
Jets	20s	HS	single	bookie	Clyde	6.57834e-05	0
Jets	20s	HS	single	bookie	Rick	4.0607e-07	0
Jets	20s	HS	single	bookie	John	3.65463e-06	0
Jets	20s	HS	single	bookie	Lance	3.65463e-06	0
Jets	20s	HS	single	bookie	Ken	0.000131567	0
Jets	20s	HS	single	bookie	Mike	6.57834e-05	0
Jets	20s	HS	single	bookie	Fred	0.0011841	0
Jets	20s	HS	single	bookie	Dave	4.0607e-07	0
Jets	20s	HS	single	bookie	Ike	7.30927e-06	0
Jets	20s	HS	single	bookie	Earl	4.0607e-07	0
Jets	20s	HS	single	bookie	Ol	2.25595e-08	0

Figure 1: Noisy CPT

We then use this CPT to perform inference ans study the variance in the behavior of the model.

4 Observations

From our experiments[1] we can clearly see that the inference results (posterior distributions) of the described setup converges to the softmax values of the activations in IAC model. One of the concrete examples is when we activate the node of "bookie" from the "occ" pool and the "single" node from the "mar" pool and try to infer the posterior likelihood of belonging to the "gang" of "Jets" or "Sharks".

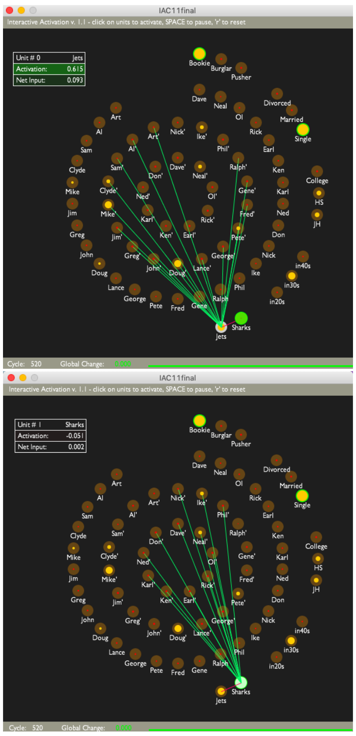


Figure 2: Activations according to the IAC model

Here we can see that the activation for "Jets" is 0.615 and for "Sharks" is -0.051. Taking the softmax according to $p_i = \frac{e^{net_i}}{\sum_{i'} e^{net_{i'}}}$, we get the probabilities for "Jets" to be 0.66 and 0.33 for "Sharks". Comparing it to the probabilities assigned with belief propagation using our Bayesian network, we get a probability of 0.691 "Jets" and 0.308 for "Sharks". Thus we can demonstrably see the convergence of the two models towards the true posterior probabilities. Some other examples include We now study some the characteristics of the human memory that can be emulated using the IAC and compare it our model with and without the noise.

4.1 Content Addressability

Humans have shown the property of being reminded of facts and specific instances from obscure and sometimes even incomplete descriptions of the episode. This characteristic is known as content addressability. The original IAC model can be shown to display the property by activating some of the feature nodes and let it retrieve the specific instance. For example, by activating the "Sharks" and the "20s" feature nodes, we can see that the IAC correctly retrieves the instance of "Ken" (by activating it the most). This can be studied in the Bayesian network by calculating the marginal posterior distribution given the evidence. So we would want to calculate $p(name|gang, age)$ which would provide us with a distribution over the names, and in our case we can see that "Ken" is given a probability of 1 here. In the case of noise injection, we see that expectedly the probability is distributed over the names, but the model is able to remember the context and assign the highest probability of 0.2 to "Ken" which proves that the model successfully emulates the content addressability. Some of the other probabilities are shown in the Table below

Thus we can say that the model is able to

Names	Noisy Prob.	Naive Prob
Pete	0.0481	0
Fred	0.0488	0
Ken	0.2043	1
Neal	0.0240	0

demonstrates the inherent property of content addressability of the human memory system.

[1]Our code can be found here

4.2 Graceful Degradation

In humans memories do not vanish immediately but fade gradually our memory systems have a property known as graceful degradation. For example, in a computer if a part of a memory or a hard rive becomes corrupted, the entire fact or data is lost. Whereas brain works differently by splitting up information. So as memory fades there is a graduate degradation in the ability to recall. So even if one some part of the memory is corrupt or lost, human brains manage to retrieve the information almost correctly.

This 'Graceful Degradation' can be noticed in IAC model too, when we probe all the properties of Ike ("30s", "JH", "Sharks", "bookie") except choosing "divorced" instead of "single", the model still activates the Ike node the highest. Similarly, in our Bayesian framework we can query the posterior distribution of $p(name|features)$ and see the results of the inference. Some of results are

Names	Noisy Prob.	Naive Prob
Ned	0.0400	0.0370
Ike	0.7205	0.0370
Ken	0.0001	0.0370
Karl	0.0022	0.0370

Here we see that the naive version of assigning CPT values to 1s will give us incorrect results (all the names are assigned 0.0370 prob.). This is because since we enforced the conditional distribution to a sort of delta dirac distribution over the data sample, (i.e $p(f_i|hidden_k) = 1$), the model cannot infer anything outside of the given data. Thus when presented with the evidence of "divorced", the model is equally unsure about every hidden class and assigns them equal probabilities (i.e 1/27).
However, with noise, we see that the model is able to identify "Ike" with very high probability. Thus suggesting that injection of noise in Bayesian systems make them more robust and closer to the characteristics of human memory.

4.3 Retrieval by name

Humans can retrieve the properties of a person with almost certainty given the person's name (provided it's a unique name). This is a property that can be expected from the models. IAC model manages to retrieve all the correct properties upon probing a name. So is the case in our Bayesian Network model (through querying $p(features|name)$). We see this as a consistent property across both with and without noise.

4.4 Spontaneous Generalization

Another property of human memory and cognition is its ability to establish a general trend over a set of instances and answer questions regarding the same. It essentially forms representations of the features that are shared by the exemplars of a category and uses it as a the basis for generalization. In concrete terms, it can make inferences about a typical member of a class, even though no specific instance might actually follow the pattern. The IAC network can accomplish spontaneous generalisation of this kind by activating a property and cycling. One example of this kind tries to find the common attributes of "Jets". By activating the node, we see that in general, members of "Jets" can be known to be in their 20s, single and have only had junior high education. The property can also be seen in the case of Bayesian networks by calculation the joint posterior distribution over the features. For our example we can ask the question "What are single gang members like?". Thus, we can simply query $p(features|mar = "single")$ and see the following distributions

"gang"	Noisy Prob.	Naive Prob
Jets	0.5555	0.6616
Sharks	0.4444	0.3383

"age"	Noisy Prob.	Naive Prob
20s	0.3703	0.3441
30s	0.4814	0.5915
40s	0.1481	0.0643

"edu"	Noisy Prob.	Naive Prob
HS	0.4074	0.5465
JH	0.3703	0.3569
COL	0.2222	0.0964

Here we can clearly see that the model is correctly and consistently giving the probabilities of the generalized notion of what single gang members must be like. We observe that most of the single gang members are members of the "Jets" gang, are in their "30s" and most have an education of "HS". This can be verified using the IAC model as well as we can see the corresponding activations of the nodes to be the maximum in their respective pools. While there is only a slight difference in the probabilities between the naive and the noisy method, we can see that the noisy approach would generalize better with larger datasets. Thus both with and without noise, the Bayesian network model is also able to display the spontaneous generalization property of human cognition and memory.

5 Conclusion

In this work we revisit some of the key aspects of the IAC model and how they relate to a probabilistic framework. We draw parallels between the update rules and the activations of the nodes in the connectionist network to log probabilities of the posterior and joint distributions. We also frame the memory network model as a Bayesian network (PGM) and observe how the reformulated model performs with respect to the original IAC model. We also introduce noise in the Bayesian network and observe how the noise makes the model more robust and generalizable when compared to the naive approach. Thus we can see that the IAC model can indeed be formalized as a Bayesian inference problem of a probabilistic graphical model.

References

William Chandler Bagley. 1900. *The apperception of the spoken sentence: A study in the psychology of language*. Cornell university.

James McKeen Cattell. 1886. The time it takes to see and name objects. *Mind* 11(41):63–65.

Stephen Grossberg. 1978. A theory of visual coding, memory, and development. *Formal theories of visual perception* pages 7–26.

Pranav Khaitan and James L McClelland. 2010. Matching exact posterior probabilities in the multinomial interactive activation model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 32.

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

Dominic W Massaro. 1989. Testing between the trace model and the fuzzy logical model of speech perception. *Cognitive psychology* 21(3):398–421.

Dominic W Massaro and Michael M Cohen. 1991. Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology* 23(4):558–614.

James L McClelland. 1981. Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the third annual meeting of the cognitive science society*. Citeseer.

James L McClelland. 2013. Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Frontiers in psychology* 4:503.

James L McClelland and Jeffrey L Elman. 1986. The trace model of speech perception. *Cognitive psychology* 18(1):1–86.

James L McClelland, Daniel Mirman, and Lori L Holt. 2006. Are there interactive processes in speech perception? *Trends in cognitive sciences* 10(8):363–369.

James L McClelland and David E Rumelhart. 1981. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review* 88(5):375.

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

Dennis Norris, James M McQueen, and Anne Cutler. 2000. Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences* 23(3):299–325.

Gregg C Oden and Dominic W Massaro. 1978. Integration of featural information in speech perception. *Psychological review* 85(3):172.

Judea Pearl. 1982. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science .

Walter Bowers Pillsbury. 1897. A study in apperception. *American Journal of Psychology* 8:315–393.

Gerald M Reicher. 1969. Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of experimental psychology* 81(2):275.

David E Rumelhart and James L McClelland. 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.