Mixtures of Experts: As an Attempt To Integrate the Dual Route Cascaded and the Triangle Models for Reading English Words

Shin-ichi Asakawa

Centre for Information Sciences,
Tokyo Womens' Christian University,
2-6-1 Zempukuji, Suginami, Tokyo 1678585, Japan
asakawa@twcu.ac.jp [*]

Abstract. An implementation of neural network models for reading English words aloud is proposed. Since 1989, there has been existing a debate in neuropsycholgy and cognitive science about the models of reading. One is the Dual Route Cascaded model, another is the Triangle model. Since there exist arbitrary variables in both models, it was difficult to decide which model would be appropriate to explain the data from psychological experiments and neuropsychological evidence. Therefore, in order to provide a solution of this debate, an attempt to integrate both models was attempted. By introducing the Mixtures of Experts Network model, a solution to overcome the arbitrariness of both models could be given. The Mixtures of Experts Network model could include both models as a special case. From the Mixtures of Experts network's point of view, the difference between the Dual Route Cascaded model and the Triangle model would be considered as a quantitative difference of the dispersion parameters.

keywords: Mixtures of Experts, Dual Route Cascaded Model, Triangle Model, Reading English words aloud,

1 Introduction

We discuss here an implementation of neural network models for reading English words aloud. Neuropsychologists and speech therapists, who have to take care of dyslexic patients, ask for neural network modelers to develop an efficient model to explain the performance of the language abilities of their patients. Among models proposed previously, two models have been considered as important, the Dual Route Cascaded (DRC)[1–3] and the Triangle model[13, 14]. Although these models can describe dyslexic symptoms, some problems remains still unsolved. We can point out several problems; the arbitrariness of the blending parameter, the existence of the lookup table, and the problem of division of labor. Therefore, nobody could judge which model is able to give a better description. The debate between them still continues, no consensus has not hitherto been reached. In this

[*] Special thanks to Eddy

paper, we tried to elucidate the features of the DRC and the Triangle model. This paper will show that these models can be regarded as just a special case of the more general model, the Mixtures of Experts (ME) model originally proposed by Jordan and Jacobs [5, 6]. This paper will also prove that the qualitative differences between the DRC and the Triangle models could be integrated as the quantitative difference in terms of the dispersion parameter in the ME.

This paper is organized as follows: Section 2 will try make terminology clear to prompt understanding the neuropsychological symptoms of reading disorders and related neural network models. Section 3, will introduce the two major models: the Dual Route Cascaded and the Triangle models, and will clarify problems to be solved. Section 4 will introduce the Mixture of Experts model in order to integrate the Dual Route Cascaded and the Triangle model. In section 5, focuses on attempts to confirm the validity of the Mixture of Experts model by numerical experiment. Section 6, will wrap things up with a discussion and some conclusions.

2 Terminology

Here, we will try to make some terms clear: the distinction between regular words and exception words, and between consistent words and inconsistent words. Regular words are the words which is in accordance with the Grapheme–to–Pronunciation–Corresponding (GPC) rule. Irregular words are the ones that the pronunciation of the words is not accordance with the GPC rules, for example "yacht". With regard to consistency, since the words like "hint", "mint", "saint", and "lint" share the same pronunciation /int/, they are consistent words. But the word "pint" is inconsistent, because it does not share the pronunciation /páint/. Consistent words have many neighbor words like "hint" and "mint" and inconsistent words have few neighbors like "yacht". Exception words, as the definition per se., are inconsistent (Glushko, 1979, p.676). Therefore, the concept "regular–irregular" and the concept "consistent–inconsistent" are not independent.

The surface dyslexic patients can read regular words and non words, but they cannot read exception words, especially low frequency non words. On the contrary, the symptom of phonological dyslexia is described as that the phonological dyslexic patients can read real words but they cannot read non words.

3 The DRC and the Triangle models

3.1 The DRC model

The DRC model has one to one corresponding between orthographic and phonological lexicons. All the real words have been registered into the orthographic and the phonological lexicons in advance. And each orthographic lexicon has a connection to the corresponding unit in the phonological lexicon[3]. Coltheart and his colleagues employed 7981 real words, which means there were 7981 entries, which can be regarded as a lookup table) in the orthographic and the phonological lexicons. This path way from orthography to phonology is called a lexical route. On the other hand, non-words and pseudo words can be read via GPC route. The GPC route was consisted of general rules so that it can translate given words to sound. The GPC rule are not always perfect since English as a orthographic language has many exception words, but almost all non–words can be pronounced by the GPC route. Real words might be read through the lexical route, because there are entries in the lookup table. However, since non–words and pseudo words do not have any entries in the lookup table, these words would be pronounced through the GPC route.

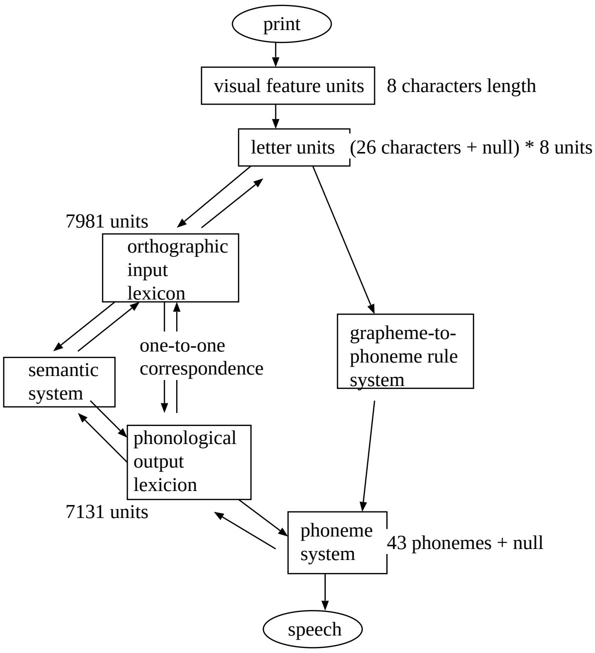
print
visual feature units
8 characters length
letter units
(26 characters + null) * 8 units
7981 units
orthographic input lexicon
one-to-one correspondence
grapheme-to-phoneme rule system
semantic system
phonological output lexicion
7131 units
phoneme system
43 phonemes + null
speech

Fig. 1. The DRC model

In the original DRC model, a discrete switch was postulated to decide which route have to be adopted when we read a word. If there is an entry in the lookup table, then the word is pronounced via the lexical route. However, in the latest version of the DRC (Coltheart et al.,2001), a parameter was introduced in order to merge the outputs from two routes. Here, we can point out the problem how we can adjust the value of this parameter by hand.

3.2 The Triangle model

In the framework of the Triangle model, on the contrary, dyslexic symptoms can be explained as follows. The surface dyslexia might be caused by the lesion in a single route (Plaut et al[11], simulation 4). The letters in the orthography can be pronounced both the direct route and the indirect route via semantics. The pronunciations are affected both routes. In the direct route, regular words and high frequency exception words will be learned, exception words with low frequency need a support of semantics. The degree of dependency on the semantics is called the "division of labor".

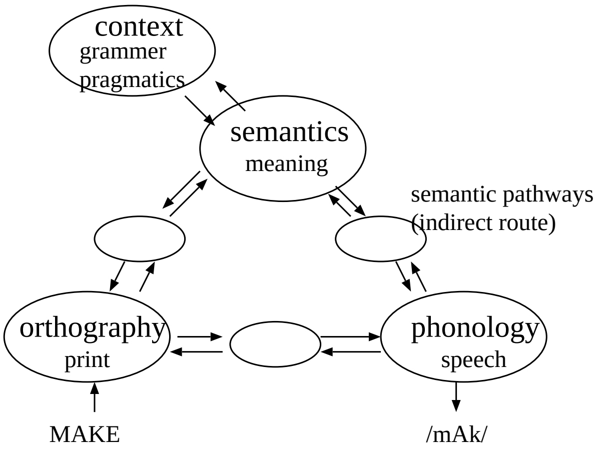

Fig. 2. The Triangle model

Suppose we can extent the concept of the lexical route in the DRC model such that the lexical route can deal with not only the words which it could recognize, but also it can deal with the words which the GPC route could not deal with. Then we can regard that there are no difference between the DRC and the Triangle models, because we cannot point out through which route the word was pronounced. The point is that the DRC model has an arbitrariness to decide the parameter to blend the lexical and the GPC route. Also in the Triangle model, as O'Reilly and Munakata[9, p.322] pointed out as follows: "Note that PMSP (Triangle model) did not actually simulate the full set of pathways. Instead, they simulated the effect of a semantic pathway by providing partial correct input to the appropriate phonological representation during training of their direct pathway model, and then removed these inputs to simulate semantic damage (p.322)"

The Triangle model has an arbitrariness to the degree of contribution of the semantic system. As discussed above, the model should be sufficient to cover all the dyslexic symptoms for reading English words aloud so that how it deals with the problem of blending between the outputs of the lexical and the GPC route in the DRC model. In other words, how it can implement the division of labor problem in the Triangle model.

4 Introduction of the Mixtures of Experts model

In this paper, we propose to introduce the Mixtures of Experts model[5, 6] so that we can let the model learn the GPC rules and classify regular and exception words automatically at the same time. Also, it could become a model which can suggest a solution for the problem of the division of labor if it can learn to classify distinction with regular words and exception words automatically. The ME can learn both the GPC rules and an automatic classification of the lexicons simultaneously. Mixtures of Experts (ME) model has been applied to many problems such as the problem of control of robot arms[6], the problem of character recognition and its location[**?**]. However, no attempts to apply the ME model as a psychological model of reading English words aloud has not been done. The ME is a technique to solve a complicated problem so that it divides the input space into a set of regions and fits simple surfaces to the data that fall in these regions. The division of input space into a set of regions and the rule of the regions were called "divide and conquer" strategy, which would take effectively in many cases. The regions have "soft" boundaries, meaning that data points may lie simultaneously in multiple regions. This "soft" boundaries seems to be roughly "fusion parameter" between the lexical route and the GPC route in the DRC model, or the solution of the "division of labor" problem in the Triangle model, because the boundaries between regions are themselves simple parameterized surfaces that are adjusted by the learning algorithm.

If we trained one large hierarchal neural network by the back propagation algorithm for the data comprising the problems that we can divide into small tasks, then we would observe that learning became slow and we would get only poor generalization because of interference among tasks to be solved. If we know in advance that training data set can be divided into some small regions, then we can apply expert networks to the divided regions by some kinds of gating mechanisms. This kind of strategy would lead us to let each small expert network do effective learning. Learning in the ME stands for letting the gating networks discover ways of the division of input space and let the experts find out the most suitable output for the data belonging to each divided region. The ME model is a kind of supervised learning algorithms. The ME consists of experts networks and gating networks. The gating networks are used to divide problem space, and each expert network is a comparatively simple network producing an output in divided regions. The ME is able to divide problem space automatically and the ME is also able to allocate expert networks for suitable spaces which gating network divided. A two level of the ME architecture is shown in Figure 1. The original Mixtures of experts allows hierarchical multi tree structures more than two layers, but for the sake of our purpose, a two layers' architecture is sufficient here.

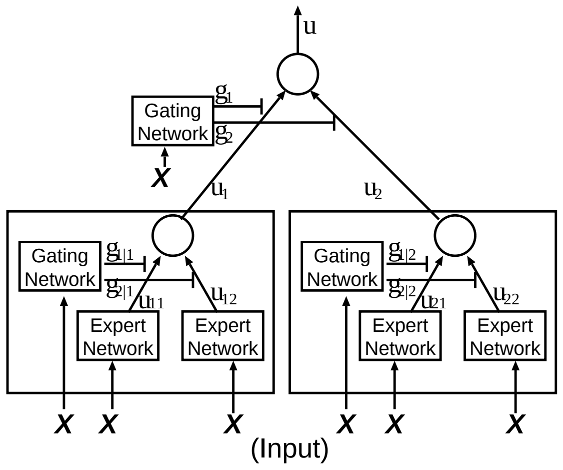

Fig. 3. A two–level mixtures of experts. Each expert network is a simple feed forward network. All the experts are given the same input and have the same number of output units. The gating networks are also feed forward networks and were given the same inputs as the inputs of the experts. The symbol g in the figure is an output (as a probability) of a gating network, and the sum of the values of all the gs is 1.0. The symbol u is the output of an expert. The outputs of experts is the mixtures of weighted sum of variables.

4.1 The dispersion parameter, and the Dirac's delta function

We can formulate the probability of an output $\boldsymbol{y}_i$ of the ith expert network as a conditional probability in which the value is in accordance with a density function with parameter $\boldsymbol{\theta}_i$ as follows:

$$P_i\left(\boldsymbol{y}_i|\boldsymbol{x},\boldsymbol{\theta}_i\right) = \frac{1}{\left(2\pi\sigma_i^2\right)^{n/2}} e^{-\left(1/2\sigma_i^2\right)\left(\boldsymbol{y}-\boldsymbol{\mu}_i\right)^T\left(\boldsymbol{y}_i-\boldsymbol{\mu}_i\right)} \ . \tag{1}$$

where $\boldsymbol{\theta}_i$ is a parameter vector which determine the density function. If P_i is in accordance with a multi dimensional normal distribution, and its covariance matrix is given as $\sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is a n dimensional unitary matrix, then we can get the final probability of the output vector $\boldsymbol{y}$:

$$P\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}\right) = \frac{1}{\left(2\pi\sigma_i^2\right)^{n/2}} \sum_i g_i e^{-\left(1/2\sigma_i^2\right)\left(\boldsymbol{y}-\boldsymbol{\mu}_i\right)^T\left(\boldsymbol{y}-\boldsymbol{\mu}_i\right)} \ . \tag{2}$$

where we postulate that g is known in advance as a producing from a Gaussian density function. We can regard that the dispersion parameter σ_i^2 determines a radius of a hyper sphere. At the limitation $\sigma^2 \to 0$, it tends to the Dirac's delta function. The Dirac's delta function is a function which satisfy:

$$\int_{-\infty}^{\infty} \delta(x)\ dx = 1 \ , \tag{3}$$

and,

$$\begin{cases} \delta(x) = \infty, \text{when } x = 0, \\ \delta(x) = 0, \text{otherwise} \end{cases} \tag{4}$$

The function $\delta(x)$ is 0 everywhere except for the point $x = 0$. The value of the $\delta(x)$ at the point of $x = 0$ is ∞, and the value of the integral which the interval contains $x = 0$ is 1. There are several definitions of the Dirac's delta function. As one of them, there exists a definition in the limitation as we approximate $\sigma^2 \to 0$ in the normal distribution, where σ^2 is a variance of the normal distribution.

4.2 An explanation of reading English words in the ME

The DRC model has two routes, the lexical route and the GPC route. The Triangle model has also two routes, the direct route and the indirect route. The meanings, the purposes, the processes, and the mechanisms of the two routes in both models are different. However, whatever the names these routes are, and whatever these routes' implementations are, we might have to postulate at least two routes in order to explain the data from dyslexic patients (surface and phonological dyslexia). The ME is able to have more than two routes, or expert networks, and is able to have gating networks. We could consider that the gating networks in the ME might be regarded as a solution of the blending parameter in the DRC model, or a solution of the arbitrariness in the Triangle model. In the limitation of $\sigma^2 \to 0$, the output of the gating network is the Dirac's delta function, which means that the expert network controlled the expert network become to respond the only one input vector $\boldsymbol{x}$, or the only one word. This word might be special, an exception word or a low frequency inconsistent words. On the contrary, when we set the value of σ^2 greater, then the expert network controlled by this gating network can deal with many similar words. This network might have to read regular words or consistent words.

One of the main points in this paper is that the ME can learn the dispersion parameter automatically, in which we do not need to look for a high dimensional parameter space. Also, we do not need to prepare an arbitrary input like the equation which was adopted Plaut et al.(p.96, eq. 16). Therefore, introducing the ME, we can implement that the gating networks which would become to respond the word 'pint' only, but it would not respond other neighbor words like 'hint', 'mint', 'print', 'lint' and so on. We would be able to have the model in which the high dimensional space consist of many monosyllabic English words and the model could divide this space according to the regularity and irregularity of the words in this corpus. Also we can regard that when the small value of σ^2, almost 0, it can be identified the same as the lookup table in the DRC model, because these networks could respond the only one word in the corpus. On the other hand, when the dispersion parameter σ^2 is large, the expert network controlled by this gating network can be regarded as an implementation of the GPC rule, because this network could read many words. In addition, the gating networks force their expert networks to learn words shared the same pronunciations, and in other case the gating networks force their expert networks to learn words with specific pronunciations. In this way, the ME model could explain the frequency effect as well. Thus, we can consider that the ME model is possible model to explain both the lookup table and the division of labor simultaneously.

5 Numerical experiment

All the 2998 words which Plaut et al.[11] adopted were used in our experiment[1] . We set the learning coefficient 0.01. All the initial values of connection weights were randomized with uniform random numbers $[-0.1, 0.1]$ The criterion to complete the learning were set the mean squared error as below 0.1. Almost every trial, the iterations were within 20–50 times, and we could get the almost the same results. Plaut et al.[11] checked the generalization ability of their Triangle model by applying the non words list in Glushko[4]. They asserted the validity the Triangle model to compare the result of the model and the data of human subjects. If the ME proposed in this paper showed the same performance as the human subjects, then it might be possible to claim that the ME is one of the candidates models to solve the problems of the way of implementation. This way of implementation is not clear in both the DRC and the Triangle models. Then, we presented the Glushko's non words list to the ME after learning completed, then compared the results with others. The ratios of percent correct are shown in Table1 The results of the human subjects and the Triangle model in the figure are from Plaut et al.(1996) simulation 1, p.69, Table 3.

Table 1. The results of the generalization test of the non words list(Glushko,1979)(%)

	consistent	inconsistent
human	93.8	78.3
Triangle	97.7	72.1
ME	93.0	69.7
bp3(100 hidden units,MSE=0.03)	90.7	53.5
bp3(100 hidden units,MSE=0.05)	95.3	58.1
bp3(30 hidden units,MSE=0.05)	88.4	58.1

For the sake of comparison, the normal back propagation methods were applied with 100 and 30 hidden units and the convergence criteria of the mean square error (MSE) 0.03 and 0.05. All the results of the back propagations are worse in the inconsistent words than other results of human, the Triangle, and the ME. In case of the 100 hidden units and 0.03 MSE, which means the most

[1] All the data we used here was obtained from the URL `http://www.cnbc.cmu.edu/~Plaut/`. Also we obtained the Glushko's non–word list for the generalization experiment from the same URL. Thus, all the data we used in this paper were exactly the same as Plaut et al.(1996).

strict convergence criterion, the performance was the worst of all. This might imply that when we employ a large network to learn the complicated task which can be divided into some regions, it is difficult for the model to extract the statistical characteristics included in the training data. It can be regarded to confirm the findings by Jordan and Jacobs(1994) that we would have poor generalization abilities when we trained large networks to learn complicated problems. It should be considered to employ the "divide and conquer" strategy in such a case.

6 Discussion

As mentioned, the DRC model requires humans to look for the best point of the blending parameter between the lexical and the GPC routes in the high dimensional space. Also, the Triangle has not implemented the division of labor yet. Therefore, these models might not be able to give any substance solutions for simulating dyslexic symptoms even when these models are well mimic human behavior. For the sake of discussion about merits and demerits of the models, we must consider not only the task performances, but also the real nature behind the models. In addition to this point, we should take into consideration about the possibilities of implementations for models as well.

If we could consider that there are expert networks specialized to process exception words, roughly corresponds to the lexical route in the DRC model, and where there exists localized division of regions, roughly corresponds to the division of labor, it might be possible to solve the problems of arbitrariness of both models. In this point of view, when we take into consideration of the limitation the dispersion parameter $\sigma^2 \to 0$, the region divided by this parameter can be identified the lookup table in the DRC model. That is, from the point of the ME model's prospects, we could reinterpret the difference between the DRC model and the Triangle model without discrepancy. Not only the problem of the lookup table and the blending parameter between the lexical route and the GPC route, but also the problem of the division of labor and the arbitrariness of the degree of contribution of semantic pathways, we can provide an unified description.

There is no essential difference between the DRC and the Triangle model in this meaning. In different words, the qualitative difference between the two model can be described as the quantitative difference of the dispersion parameters. It could be considered that the DRC and the Triangle model are particular cases of more general and comprehensive model. When we introduce the ME model as a model of reading English words aloud, it is possible to explain the difficult problem to tune the best point in high dimensional parameter space, and to formulate the arbitrary problem remained unsolved.

Numerous articles have cited the works of Plaut et al.[11] and Coltheart et al.[3]. Thus, it is obvious that both the models are the most valuable model for reading English words and its impairments. On the other hand, in this paper we showed an only one result shown in Table1. Therefore, it is difficult to insist that the ME is superior to previous two models. Not so much as saying so, this model still can be uncompleted. A number of points remain unclear. However, this model might be considered to formalize clearly the problems remained to be unclear in the previous models. Rather than closing the debate between two models, it might be worth attempting to integrate both of them.

References

1. Max Coltheart, B. Curtis, P. Atkins, and M. Haller. Models of reading aloud: Dual-route and parallel-distriputed-processing approaches. *Psychological Review*, 100(4):589–608, 1993.
2. Max Coltheart and K. Rastle. Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, 20:1197–1211, 1994.
3. Max Coltheart, Kathleen Rastle, Corad Perry, Robyn Langdon, and Johannes Ziegler. Drc: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108:204–256, 2001.
4. R. J. Glushko. The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psyhology: Human Perception and Performance*, 5:674–691, 1979.
5. Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffery E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.
6. Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 6:181–214, 1994.
7. J. L. McClelland and D. E. Rumelhart. An interactive activation model of context effects in letter perception: Part 1. an account of basic findings. *Psychological Review*, 88:375–407, 1981.
8. J. Morton. *The logogen model and orthographic structure*. Academic Press, 1980. In U. Firth(Ed.), Cognitive processes in spelling.
9. Randall C. O'Reilly and Yuko Munakata. *Computational Explorations in Cognitive Neuroscience: Understanding in mind by simulating the brain*. MIT Press, 2000.
10. David C. Plaut, James L. McClelland, and Mark S. Seidenberg. Reading exception words adn pseudowords: Are two routes really necessary? In J. P. Levy, D. Bairaktaris, J. A. Bullinaria, and P. Cairns, editors, *Connectionist Models of Memory and Language*, pages 145–159. University College London Press, London, 1995.
11. David C. Plaut, James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103:56–115, 1996.
12. D. E. Rumelhart and J. L. McClelland. An interactive activation model of context effects in letter perception: Part 2. the contextual enhancement effect and some tests and extension of the model. *Psychological Review*, 89:60–94, 1982.
13. Mark S. Seidenberg and James L. McClelland. A distributed, developmetal model of word recognition and naming. *Psychological Review*, 96(4):523–568, 1989.
14. Mark S. Seidenberg, Alan Petersen, David C. Plaut, and Maryellen C. MacDonald. Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(1):48–62, 1996.
15. Mark S. Seidenberg, David C. Plaut, Alan S. Petersen, James L. McClelland, and Ken McRae. Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 20(6):1177–1196, 1994.