

A PROPOSAL FOR A COMPUTATIONAL
MODEL OF PHONEMIC ACQUISITION

by

L. Jaime Parchment

Copyright © L. Jaime Parchment 2018

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF LINGUISTICS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

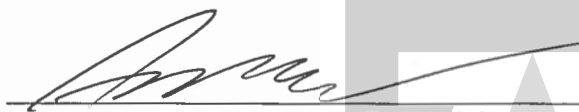
In the Graduate College

THE UNIVERSITY OF ARIZONA

2018

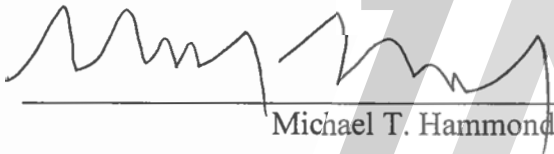
THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by **L. Jaime Parchment**, titled ***A Proposal for a Computational Model of Phonemic Acquisition*** and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.



Andrew B. Wedel

Date: 09 May 2018



Michael T. Hammond

Date: 09 May 2018

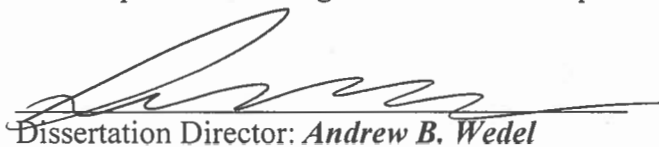


Diane K. Ohala

Date: 09 May 2018

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.



Dissertation Director: **Andrew B. Wedel**

Date: 09 May 2018

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: L. Jaime Parchment

ACKNOWLEDGEMENTS

This is where I'm supposed to wax poetic about the many people who have figured prominently in my journey to this point. I can't do it. There are too many people who have been too important in too many ways over too long a period of time. For the sake of what remains of my sanity, I will limit my acknowledgements to a listing of the people who have been most present and meaningful in my life over the last many months, and hope that the uncounted omissions will understand.

My parents and my sisters

My committee

Listed in the order I met them, back before they had fancy titles...

Prof. Amy Fatzinger

Prof. Lauren Hall-Lew

Dr. Emily Kidder

Dr. Dane Bell

Dr. Emily Bell

From outside the walls...

The RCs, especially my notphews, Cameron, Diego, and Ryan

The Knappys, especially Ryan, Ian, and Claire

All of my students, especially Parviz, Arezou, and Dara

TABLE OF CONTENTS

LIST OF FIGURES	8
ABSTRACT.....	9
CHAPTER 1 Introduction.....	11
1.1 Language Acquisition	11
1.1.1 Statistical learning.....	14
1.1.2 Interaction between levels.....	14
1.1.3 Interaction between measures	16
1.1.4 Summary of acquisition and motivation of focus	17
1.2 The Value of Modeling	20
1.3 Marr’s Three Levels of Analysis	20
1.4 Independence of Levels	22
1.5 Reductionism	22
1.6 Evaluation	22
1.7 The Mirror Heuristic.....	23
1.8 Dissertation Map.....	24
1.8.1 Chapter 2.....	24
1.8.2 Chapter 3	25
1.8.3 Chapter 4.....	26
1.8.4 Chapter 5.....	26
CHAPTER 2 Linguistically Motivated Computational Models.....	27
2.1 Marr’s Three Levels of Analysis	29
2.2 Sound Preprocessing.....	29
2.3 Some Model Families	34
2.3.1 Clustering Models.....	35
2.3.1.1 A Particular Model of Acquisition	44
2.3.2 Information Theory Models Based on Neurons.....	47

2.3.2.1	The Perceptron	47
2.3.2.2	Parallel Distributed Processing	56
2.4	Summary	60
2.4.1	The Mirror Heuristic Revisited	62
CHAPTER 3 Computational Cognitive Neuroscience Models		66
3.1	Neurons	66
3.1.1	Simplifying models	71
3.2	Neuroscience models	74
3.2.1	Compartment models	76
3.2.2	Spiking models	77
3.3	Cognitive neuroscience modeling	81
CHAPTER 4 Proposals for a Model of Phonemic Acquisition		90
4.1	Model Complexity	90
4.2	Quick Review	95
4.3	Modeling Goals	96
4.3.1	Supervision	97
4.3.2	Revising the Goal	98
4.3.3	Secondary Goal	99
4.3.4	Overfitting	101
4.4	Empirical Evidence	103
4.4.1	Categorical Perception	103
4.4.2	Other Animal Experiments	107
4.4.3	Neural Correlates	109
4.5	Designing the Model	111
4.6	Marr Revisited	112
4.7	Summary	114

CHAPTER 5 Conclusion	115
5.1 Overview	115
5.2 Future Directions	115
REFERENCES	118

LIST OF FIGURES

Fig. 1.1 – Timing of acquisition sequence.....	14
Fig. 1.2 – Lexical disambiguation from non-linguistic information.....	15
Fig. 1.3 – Vowel quality depends on the relationship between F1 and F2	16
Fig. 2.1 – A spectrum showing three sinusoids of different frequencies.....	29
Fig. 2.2 – Shape of the basilar membrane in response to different input frequencies.	30
Fig. 2.3 – A trivial clustering problem.....	35
Fig. 2.4 – Clustering can benefit from added dimensions.	36
Fig. 2.5 – An added dimension can change clusters.	36
Fig. 2.6 – Hidden Markov Model (with missing emission probabilities).....	40
Fig. 2.7 – The graph of $y = 2x + 4$	47
Fig. 2.8 – The graph of $y = 4x + 1.5z + 3$	48
Fig. 2.9 – A multiple regression in diagram form.	49
Fig. 2.10 – A logistic, or bounded exponential, curve.....	50
Fig. 2.11 – A logistic curve in three dimensions.	51
Fig. 2.12 – Moving a point across orthogonal axes.	52
Fig. 2.13 – Moving a point across non-orthogonal axes.	53
Fig. 2.14 – A multilayer perceptron.	54
Fig. 3.1 – Parts of a neuron	66
Fig. 3.2 – Ion channel closed (left) and open (right).	67
Fig. 3.3 – Action potential followed by refractory period.	68
Fig. 3.4 – Spike train.....	68
Fig. 3.5 – Logistic curve	70
Fig. 3.6 – tanh function	72
Fig. 3.7 – ReLU and Softmax	72
Fig. 3.8 – Cumulative density function.....	73
Fig. 3.9 – Step function.....	73
Fig. 3.10 – Rulkov map spiking behavior for different values of α	78
Fig. 4.1a – Binary branching tree with nodes A, B, and C.	90
Fig. 4.1b – Inserting a node.	90
Fig. 4.1c – Moving C from one node to another.....	90
Fig. 4.1d – Deleting the empty node.....	91
Fig. 4.2a – A tetrahedral permutation-generator in physical and schematic versions.	91
Fig. 4.2b – 60 degree rotation of ABC yields ACB.....	92
Fig. 4.2c – 60 degree rotation of ACB results in CAB.....	92
Fig. 4.2d – Next rotation gives CBA.	92
Fig. 4.2e – Then BCA.....	93
Fig. 4.2f – And, finally, BAC. One more rotation will return to ABC.....	93
Fig. 4.3 – Overfitting (green line) vs more conservative curve (black line).....	101
Fig. 4.4 – From Steinschneider (1982, p361).	109

ABSTRACT

In the field of computational linguistics, computational modeling of linguistic behavior has been motivated not only by the creation of practical language-related tools such as machine translation, automatic speech recognition, speaker identification, and natural language search, but also by a desire to deepen our understanding of how language works, either in an abstract mathematical sense or in the more literal sense of describing human behavior at various levels of analysis. These models take various forms, some derived from mathematical models of electronic transmission of information (Shannon, 1948), others from abstract models of neural behavior (McCulloch and Pitts, 1943; Rosenblatt, 1958).

In computational neuroscience, computer models are developed to mimic the behavior of brains, with a greater degree of biological realism. These models focus on neural behavior ranging from single neurons to large-scale networks of neurons. Typically, the behavior of interest is the relative activation of groups of neurons, the emergence of synchronized or otherwise patterned activation, and the propagation of signals across networks. The elaboration of relatively high-level cognitive behavior is, at best, secondary to the exploration of low-level physical and electrical interaction (Zednik, 2018).

The growing field of computational cognitive neuroscience has as a goal the development of computational models that are biologically plausible and that exhibit cognitive behavior of interest (Ashby, 2011). Linguistic models of this type are intended

to exhibit the kind of linguistic behavior that is observed in humans, but with an underlying structure and behavior that closely parallels the human brain.

Marr (1982) offers a three-level framework for analyzing models of the brain that has become a standard in neuroscience (Bechtel, 2014). Applying this method of analysis to broad classes of computational models yields insights into the strengths and weaknesses of each. While the ideas in this dissertation may ultimately find broader relevance, it is presently concerned primarily with the modeling of phonemic acquisition in infants. Application of Marr's analysis to the actual system being modeled – the human infant – suggests an approach to the development of acquisition models that departs significantly from traditional computational linguistics models and computational cognitive neuroscience models.

CHAPTER 1

Introduction

1.1 Language Acquisition

Language has been described as a method of organizing thoughts that is unique to humans and only incidentally provides any communicative function (Hauser, Chomsky, and Fitch, 2002). It has also been described as a system of communication of the same general type that some non-human animals have, but with some features that are unique to humans. Hockett (1968) defines a set of nine features common to all primate communication, with four additional features that distinguish human language. What is less controversial is the position that language is important to humans, that it is unique in degree, if not in provenance, and that it can provide a window into the workings of the brain. For all these reasons, it is desirable to develop a scientific understanding of all aspects of language.

The scientific study of language can be approached in many ways. While a complete exploration of these approaches is beyond the scope of this dissertation, it may be fruitful to mention two broad classes. Some researchers explore aspects of the workings of language as possessed by fluent adults. This object of study is a complex system with enough similarity between language users to make communication possible, but with tremendous variation between individuals. The range and details of this variation and possible methods of extracting information from the speech stream and its

higher-level organization have occupied legions of linguists. Other researchers focus on the process by which this sort of system comes into existence in an individual born without the practical ability to use language. The latter approach is of interest here.

The acquisition of language by infants is a puzzle, because children learn language fluently without overt instruction. Of course, children also learn to walk without overt instruction, but the majority of skills that children pick up without help, like walking, tend to be physical in nature. Language is an abstract, complex, combinatorial system of noises that convey compositional meaning, and children learn it effortlessly. Skinner (1957) proposed a behavioral account of language acquisition, in which the learner mimics the speech sounds to which she is exposed and adjusts according to the success of her efforts, eventually converging on grammatical speech. Chomsky (1959) countered that children, under a behaviorist paradigm, do not receive enough input to learn language as quickly as they do. He proposed what was later termed a Language Acquisition Device – an innate and specialized cognitive system that provides the underpinnings of language, leaving the infant to simply identify the grammatical details that distinguish her target language from other possible human languages (Chomsky, 1965). Still other researchers (Stemmer, 1973, inter al.) adopt an empiricist position and posit that infants acquire language through evidence and experience. Since this dissertation is concerned with the modeling of language acquisition, the nativist assumption of an innate language module that reduces acquisition to the recognition of typological differences holds little interest. The empiricist position that acquisition proceeds through the application of statistical reasoning to language input is adopted as a more reasonable starting point.

Language seems to be hierarchical, both in structure and in rough sequence of acquisition. The following list shows the approximate sequence in which various aspects of language are acquired (adapted from Vihman, 2013) and offers definitions adapted from Brown and Miller (2013).

Prosody	Stress, pitch, tempo, loudness, and rhythm
Phonetics	Speech sounds (converging on the target language in time)
Phonology	Organization of speech sounds of the target language
Lexicon	Words of the target language
Morphology	Smallest units of meaning
Syntax	Arrangement of words into phrases, clauses, and sentences

Semantics, concerned with meaning, has a special place in this hierarchy. Under the perspective of language as a system of communication, the ultimate purpose of every part of language is to convey meaning, i.e., to deliver semantics. In some cases, semantic content is purely linguistic in nature, including such examples as comparative constructions (Syrett, 2016) and grammatical aspect (van Hout, 2016). In other cases, semantics serves as the bridge between language and the external world, as exemplified by content words with real-world referents (e.g., “house” or “run” or “purple”). The relationship between semantics and other aspects of language allows a convenient division of the objects of language acquisition into two broad classes – those that carry meaning on their own (meaning units), and those that do not (sub-meaning units). Morphemes are defined as the smallest units that carry meaning, so they are clearly meaning units, as are lexical items and syntactic structures. Sub-meaning units include

the sounds of language. If these are the objects of acquisition, mention must be made of the methods of acquisition. Three points are of particular interest to the present work. First, linguistic acquisition seems to depend on statistical learning mechanisms. Second, separately defined modules of language are not acquired independently. There is interaction between them. Third, particularly with regard to sub-meaning units, acquisition depends not on simple acoustic measures, but on their complex interaction.

1.1.1 Statistical learning

Saffran (2003) proposes that language learning is dependent on statistical learning mechanisms that may not be unique to language, and that the process is guided by constraints on perception and production. Infants' use of statistical learning is evident in a range of language tasks, including learning to identify phrase boundaries (Saffran, 2001), identifying word boundaries (Saffran, et al., 1996), and learning consonants (Maye, Werker, and Gerken, 2002). In each of these cases, the set of possibilities sits on a multi-dimensional continuum, but observed tokens are not uniformly distributed. Instead, inputs tend to pattern in ways that diverge from a random or uniform distribution. These distributional irregularities allow the infant to attend to the details that are meaningful in the target language. Measures that are not meaningful do not pattern in a way that allows the learner to mistakenly identify them as meaningful.

1.1.2 Interaction between levels

Fig. 1.1 shows the overlapping sequence of acquisition of elements of the sound system. There is no element of sound acquisition that begins and ends while the others remain static. There is always overlap in mastery of the various stages.

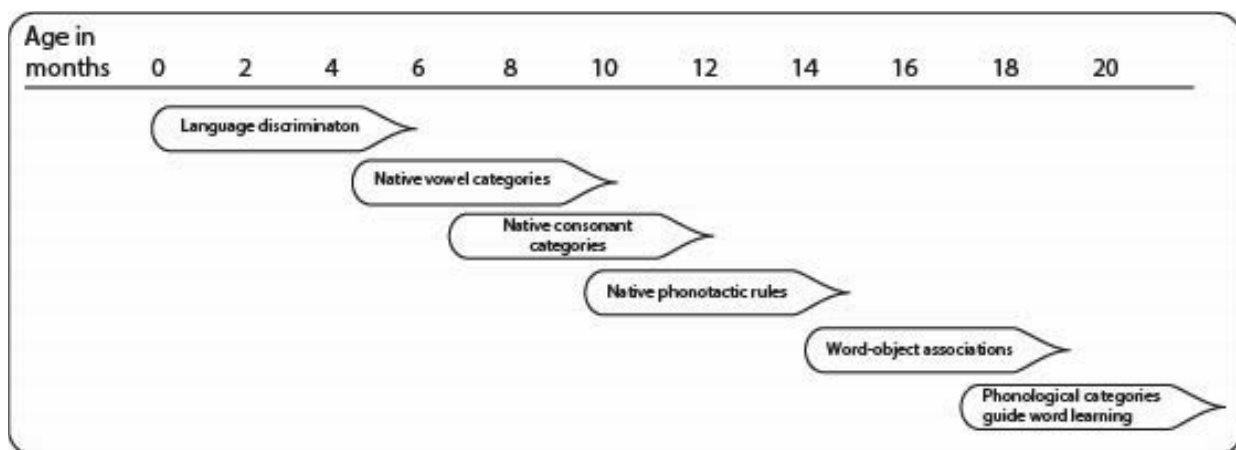


Fig. 1.1 – Timing of acquisition sequence (adapted from Curtin and Zamuner, 2014)

The same can be said for other levels of language, including meaning units.

Acquisition of morphology begins before acquisition of syntax, but is not completed until after the acquisition of syntax has begun.

Disambiguation at one level frequently relies on information from another level. Consider the noun “duck”, referring to the waterfowl, and the verb “duck”, referring to a rapid lowering of the head. The potential ambiguity between these two words can be resolved through syntactic cues. In the sentence, “I fed a ____ at the park”, the blank is likely to be filled with a noun, both because it follows the determiner “a” and because it serves as the direct object of the verb “fed”. In the sentence, “Be sure to ____ through this low doorway”, the blank is likely to be filled with a verb, both because of the preceding infinitive marker “to” and the following adverbial phrase “through this low doorway”. The blanks in both sentences can accept “duck”, and the syntactic context resolves the lexical ambiguity between the noun “duck” and the verb “duck”. In other cases, lexical ambiguity can be resolved by extra-linguistic information (see Fig. 1.2).



Fig. 1.2 – Lexical disambiguation from non-linguistic information

1.1.3 Interaction between measures

Acquiring a sound system requires the learner to distinguish acoustic differences. This statement is deceptively simple, because the differences of interest are seldom signaled by simple acoustic measures. Instead, each contrast is determined by the interaction of a number of acoustic measures. Vowel quality, for example, is determined by a minimum of two acoustic measures – the first and second formants. In Fig. 1.3, a first formant (F1) value of 0.5kHz can indicate as many as five distinct vowels. Only when F1 is combined with the second formant (F2) is a single vowel specified. For example, where $F1 = 0.5\text{kHz}$ and $F2 = 1.5\text{kHz}$, the resulting vowel is [e]. Note that the chart in Fig. 1.3 represents the vowel productions of 76 speakers (Peterson and Barney, 1952) and indicates the variability in vowel production between speakers. Although the first and second formants are often sufficient to determine a vowel, this is not always the case, and additional information is sometimes required. The learner must attend to the full range of acoustic information and the interactions between various measures.

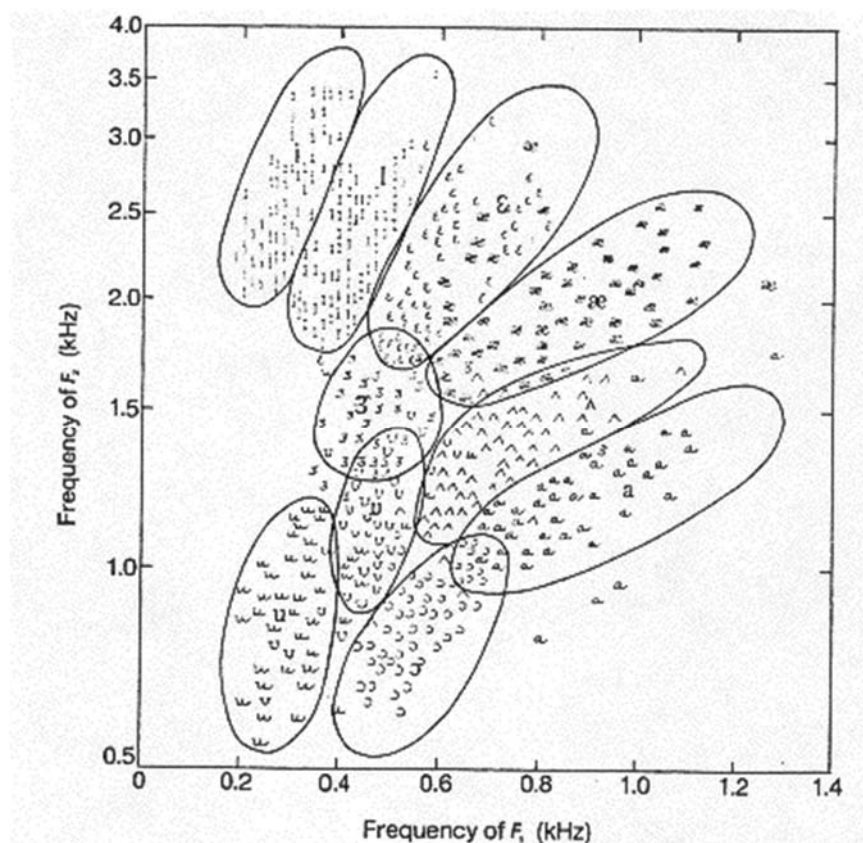


Fig. 1.3 – Vowel quality depends on the relationship between F1 and F2 (Peterson and Barney, 1952)

The interaction between acoustic measures is also evident in the difference between voiced and voiceless plosives ([b] and [p], e.g.), which share an interruption of oral airflow, but differ in the timing between the resumption of oral airflow and the vibration of the vocal folds. (This distinction will figure prominently in Chapter 4.) Other examples support the same fact. Linguistic sound differences are often the result of complex interactions between a number of acoustic measures.

1.1.4 Summary of acquisition and motivation of focus

Linguists commonly divide language into a set of modules – including phonetics, phonology, morphology, and syntax – that can be studied with some degree of independence. These divisions can be of value in studying acquisition since infant

learners do not develop mastery of all aspects of language in parallel, instead following a sequence starting with recognition of basic elements of speech sounds and proceeding through higher levels of organization. However, as outlined above, acquisition research demonstrates that these modules are not entirely independent and that the acquisition process involves the interaction of different aspects of language. The development of syntax requires a lexicon, but the process of developing syntax also helps to develop the lexicon. Lexical acquisition is dependent on knowledge of phonological patterning, but a growing lexicon also reinforces and refines phonological discrimination. In short, earlier stages of acquisition not only inform later stages, but are also reinforced and further developed by their use in later stages. This bidirectionality of influence between the stages of acquisition suggests that the development of an adult-like grammar is a more complex process than can be captured in any linear or sequential representation. As with any scientific research, the feasibility of an acquisition study depends on an initial restriction of the domain of inquiry. One can examine the distribution of babbling sounds in infants or the over-regularization of past tense morphology in older children without regard for the way in which these process might be affected by knowledge of phonotactics. However, it is important to recognize that a theory of acquisition of one restricted element of language will necessarily be incomplete. This is a particularly important point in computational modeling, as will be discussed in Chapter 2.

For the purposes of the present discussion, language acquisition is far too broad a topic to be practical. A restriction of the domain of inquiry is in order. Meaning units, including morphemes, lexical items, and syntactic phrases, all intersect strongly with semantics, which renders them more complex than sub-meaning units. For this reason

alone, they are rejected as topics of further discussion in this work. Among sub-meaning aspects of language, ideas of potential focus include prosody, phonemic contrasts, and phonotactics. Prosody refers to pitch, stress, timing, and rhythm of language. Although prosodic differences do not carry semantic information of the type that distinguishes morphemes or lexical items, it can have an influence on the interpretation of language at all levels. This renders it too complex for the present purpose, so prosody is also rejected as an object of focus. Phonotactics refers to the language-specific rules governing which sounds can be produced in sequence in different parts of a word. For example, in English, the word-initial sequence [bl] is acceptable (as in “blue”, “blog”, and “blasphemy”), but the word-initial sequence [bn] is not. Thus, “blick” is an unattested but possible word of English, while “bnick” is neither attested nor possible. (Example adapted from Chomsky and Halle, 1968.) These rules vary cross-linguistically. Swahili, among other languages, allows word-initial [mb] and [nd] (Polomé, 1967), while Spanish does not (Saporta, 1962). The question of which sounds can appear adjacent to each other obviously requires recognition of those different sounds. For this reason, phonotactics is rejected as the object of focus, in favor of phonemic contrasts.

A phoneme is a speech sound that can make a difference in meaning between two words. The [b] in [blu] “blue” and the [g] in [glu] “glue” are different phonemes, because they represent the sole sound difference between two words, “blue” and “glue”, that have different meanings. By contrast, the unreleased [t̚] at the end of [taɪt̚] “tight” and the aspirated [tʰ] at the end of [taɪtʰ] “tight” represent slightly different ways of producing the final sound of the very same word, so they are not considered different phonemes; they are considered allophones of the same phoneme. Even though they are

acoustically distinct, they are considered members of the same sound category. This identification of acoustically distinct sounds as being of the same type in language use is at the heart of phonemic acquisition. The infant learner is exposed to a multi-dimensional continuum of acoustic input, but that continuum is not smooth. Some regions have much more attested input than others. These regions roughly correspond to sound categories that are meaningful in the target language, despite any acoustic variation. Learning to divide this acoustic space into linguistically meaningful categories is the task that will occupy the remainder of this discussion.

1.2 The Value of Modeling

Modeling the process of phonological acquisition is a worthwhile goal, for at least three reasons. A theory of acquisition that is computationally intractable cannot be an accurate representation of what infant learners do. Creating a computational model to instantiate a theory provides a basic check of the viability of the theory. The shortcomings of the model can also serve to highlight areas where the theory should be refined or reworked. Lastly, with a model that adheres closely to the functioning of the actual system, it is, in principle, possible to conduct experiments on the models that cannot be performed on infants because of ethics, cost, logistic, or countless other reasons.

The question of whether a model adheres closely to the functioning of the actual system is a difficult one that will be addressed beginning in section 1.2.1 and continuing in Chapter 2.

1.3 Marr's Three Levels of Analysis

In his seminal work, *Vision*, David Marr (1982) proposes three levels of analysis for evaluating neuroscience models. This style of analysis has been applied not only to models based on low-level neuronal behavior, but also to higher-level cognitive models and more abstract models that purport to mimic brain behavior. This three-level analysis is widely accepted in neuroscience and related disciplines, although there has been much discussion about the independence of the three levels and the implications for the behavior of the entire model that arise from changing a single level. The three levels are Computation, Algorithm, and Implementation.

The computation level describes the problem the model is intended to solve. It represents the “what” and the “why” of the model. In terms of models of phonemic acquisition, this level might specify that the model is to accept acoustic input and produce phonemic labels.

The algorithm level describes the steps the model will put into effect to accomplish the task laid out in the computation level. This might include a specification of how the raw acoustics will be pre-processed for delivery to the model, any calculations or transformations to be performed by the model, and the style or format of the model output.

The implementation level describes the actual mechanisms the model will use to execute the algorithms from the previous level. Examples of possible implementations include a binary hierarchy of hidden Markov models, a multi-layered perceptron, and a deep-belief network.

1.4 Independence of Levels

Some have argued that the three levels should be treated as independent entities. Others, including Marr, posit a necessary interdependence between levels. For our purposes, the three levels will be discussed as if they are independent, simply for the sake of clarity in comparing different model structures. In reality, a moment's reflection strongly suggests that true independence is a practical impossibility. The details of the implementation level of a model will necessarily place boundaries on possible algorithms, while only certain algorithms will be suitable to the task specified at the computation level (Bechtel, 1993).

1.5 Reductionism

Marr (1982) strongly opposed reductionism in neural modeling, arguing that “both high-level information processing constraints and low-level implementational constraints play mutually reinforcing and constraining roles” (Eliasmith, 2015). This perspective, positing influence between the most abstract and most concrete levels, argues against the reductionist notion that high-level behavior can be completely explained by low-level interactions, and against the opposing extreme that completion of high-level goals renders low-level structures and process irrelevant. On the other hand, as Bickle (2015) points out, the reductionism in current neuroscience is different from the norms of Marr's time, so the concern about reductionism may no longer apply. Here, the adopted position is that interaction between the levels is necessary and appropriate.

1.6 Evaluation

Evaluation of models under discussion will be approached in two ways. The first asks whether the three levels are compatible. The second focuses on the computation level and evaluates the goal of the model. The specific issue of interest is whether the goal of the model is consistent with the purpose and abilities of the system being modeled. The ventral theme of this dissertation is that existing models of phonemic acquisition fail at the computation level. A closer analysis of this failure in Chapter 4 will lead to a proposal for a different approach. Another approach to the evaluation of models is The Mirror Heuristic.

1.7 The Mirror Heuristic

The utility of computational models for elucidating the linguistic behavior of the brain depends on a philosophical assumption that Bechtel (2018) terms “the mirror heuristic”. This assumption is that, if two systems perform the same task, i.e., if they accept the same input and produce the same output, then they are necessarily performing mathematically equivalent functions. If the modeling goal is simply to develop a computational model that performs a language task, the production of the desired output is the only necessary metric of success, and the process that develops the output from the input is of only utilitarian interest. However, if the goal is to improve our understanding of the behavior of the brain during language tasks, the process from input to output is of critical importance, and a trivial satisfaction of the mirror heuristic is insufficient. This heuristic rests on deeper assumptions about the nature of computation in a natural system, the meaning of “mathematically equivalent functions”, and the parts of a model in which this equivalence is meaningful. This issue is discussed in more detail in Chapter 2.

1.8 Dissertation Map

1.8.1 Chapter 2

Chapter 2 describes several types of computational models that develop the ability to discriminate phonemic categories. These models are of a type that Beer (2015) refers to as Information Theory Models, as contrasted with Dynamic Models, which will be discussed in the next section. Information Theory Models are ultimately derived from Shannon's (1948) work on electronic communication and depend on language use being analogous to the transmission of a known signal across a noisy channel. Shannon (1948) considered a known signal traveling through a known channel, and developed metrics for determining the ideal width of a channel and the likelihood of signal loss during transmission. These fundamental concepts have been applied to linguistic modeling because, to an extent, language use can be seen as the transmission of a signal across a noise channel. Applications of these ideas to linguistic acquisition seem a little more questionable, because the learner does not know what the intended signal is and has no real way to evaluate the received signal, if we assume an unsupervised learning paradigm. In reality, during the process of language acquisition, the learner has access to a significant amount of both linguistic and extra-linguistic feedback to provide some supervision to the learning process. Careful consideration of these details is essential to the ecological validity of a model of acquisition.

A second type of Information Theory Model is derived from Rosenblatt's (1958) perceptron, which is based on earlier work by McCulloch and Pitts (1943). These models are inspired by the physical properties of neurons and the potential for emergent behavior in networks of neurons. Their goal is to model the behavior of networks of

neurons, rather than their physical characteristics (Ashby, 2011). The original perceptron is essentially a geometric representation of a logistic regression, as will be demonstrated in Chapter 2. It was argued against by Minsky and Papert (1969) because of its limitations in handling certain logical statements. The perceptron was resurrected by Rumelhart and McClelland (1986), who added a hidden layer, effectively creating the opportunity for much greater complexity in the behavior of the model by allowing the regression variables to interact in limited ways.

Marr's three-level analysis is applied to these models as a way of exploring their adequacy as models of phonemic acquisition. Information Theory models fail as models of acquisition at the Algorithm and Implementation levels. An argument is made that they also fail at the Computation level. The fact that they trivially satisfy the Mirror Heuristic highlights the reason for their inadequacy as models of acquisition.

1.8.2 Chapter 3

Chapter 3 introduces Computational Cognitive Neuroscience models, which fall into a class of models that Beer (2015) calls Dynamic Models. The field of computational cognitive neuroscience has its roots in the development of a mathematical description of the generation of action potentials in the giant squid axon by Hodgkin and Huxley (1953). While earlier models focused on single neurons, some subsequent models have dealt with networks comprising many neurons (Ashby, 2011). While computational neuroscience is largely concerned with modeling the spiking behavior of single neurons or networks of neurons (Trappenberg, 2002), computational cognitive neuroscience attempts to model various cognitive behaviors, by always on a substrate of

biologically plausible models of neurons (Ashby, 2011). In terms of Marr's three levels, computation cognitive neuroscience strives to maintain the Implementation and Algorithm levels as they are in computational neuroscience models. But with a change to the Computation level to include emergent cognitive behavior.

1.8.3 Chapter 4

Chapter 4 proposes a new approach to the modeling of phonemic acquisition that is motivated by the shortcomings of existing models of various types. As will be explained in due course, the two types of computational models broadly defined above are unsatisfactory for the purposes of modeling phonemic acquisition. The information theory models fail at the Implementation and Algorithm levels, although the computational cognitive neuroscience models do better in that regard. Both types of models fail at the Computation level, primarily because the goals they aim to achieve are strongly motivated by the researcher's analysis of the acquisition process, rather than by the process itself. A careful consideration of the task that is actually undertaken by the language learning infant, and the ways in which it diverges from a typical computational analysis of the process, suggests a two-step process leading to phonemic discrimination. The general structure of the proposed model approach is elaborated, along with a detailed set of steps that could lead toward a successful implementation of this new type of model.

1.8.4 Chapter 5

In Chapter 5, the claims made in this dissertation are reviewed and directions for future research are laid out.

CHAPTER 2

Linguistically Motivated Computational Models

One strong motivation for developing computational models of language is to create tools that do useful work. The computerized language task that is most closely related to phonemic acquisition is that of Automatic Speech Recognition (ASR). In an ASR system, the computer takes human speech as input and converts it to a formal representation that the computer can act on. The algorithms that bridge the speech input and the formal output are evaluated in terms of accuracy, speed, and computational efficiency. Any tweak to the model that offers an improvement on these dimensions will be accepted, regardless of how distant it is from the way humans process language. The goals of a model of human language acquisition should be somewhat different. Human infants go through a long, slow, and complex process to acquire language. The result is a robust system that performs well in a variety of acoustic environments, adjusts immediately and effortlessly to unfamiliar voices, simultaneously uses all levels of linguistic analysis for real-time error correction, and engages seamlessly with non-linguistic cognitive functions. Mimicking this system in its entirety is, to say the least, a daunting task best left to the realm of science fiction for the moment. Yet, even if one divorces the broad task of language acquisition from the rest of cognitive development, and further reduces that task to the acquisition of phonemic categories, there are reasons to develop models that disregard the mechanistic metrics of speed of training and accuracy of performance and, instead, mimic as closely as possible some of the elements of the human system being modeled. The elements might include the time-trajectory of

acquisition, the sequence of phonemes learned, the learning algorithms used, the conceptual structure of the model, and many other factors derived from the linguistic and cognitive realities of human infants. Naturally, the choices that drive model design are more complicated than “make it fast or make it act like a baby”. Digital computer technology is fundamentally different from living neurons. A universal Turing machine should, in principle and in the absence of physical and temporal constraints, be able to mimic the behavior of an actual neuron or collection of neurons to an arbitrary degree of precision, but the differences between the two processes are significant. Neurons change and “compute” in real time at the molecular level, in complex ways that do not easily yield to mathematical description. Given that an average neuron, with a mass on the order of 10^{-6} g, comprises in the neighborhood of 10^{16} molecules, has from thousands to tens of thousands of ion channels potentially of at least three-hundred different types, passes up to 10^8 ions per second, and interacts directly with around a thousand other neurons (Jessell, 2000), it is difficult to exaggerate the challenge of a complete mathematical description. Using a digital computer to finely approximate this complex behavior would require a tremendous number of digital calculations at each discrete time step. The process will inevitably run afoul of limits imposed by the scarcity of computational resources. Of course, it is probably not necessary to accurately model each molecule of each neuron, just as it may be unnecessarily simplistic to model a neuron as a simple object that either fires or doesn’t fire at any given moment. Determining some aspects of the system that can be simplified without excessive detrimental effects on the performance of the model will be the subject of the next two chapters. The remainder of this chapter looks at some of the types of models that are

commonly used in computational linguistics, in support of the argument that a step in the direction of more biologically realistic models is desirable for the purposes of modeling phonemic acquisition.

2.1 Marr's Three Levels of Analysis

As noted in Chapter 1, Marr's three levels of analysis are Computation, Algorithm, and Implementation. The Computation level defines the problem to be solved, the Algorithm level specifies the steps to be taken, and the Implementation level provides the structures that will be used. Speech-related learning models typically used in computational linguistics vary primarily at the Implementation and Algorithm levels. Historically, there seems to be a tendency to start with the simplest possible structures and algorithms, and to respond to unsatisfactory performance by incrementally increasing the complexity at these two levels. This is an issue that will be explored more thoroughly in section 2.3. At the Computation level, these models all have the same task of converting acoustic input into a set of labeled tokens, although that process might start with a transformation of the input.

2.2 Sound Preprocessing

Waveforms, as simple graphs of air pressure across time, can show remarkable variation, even when they are captured from successive utterances of the same word from a single speaker. Moreover, they fail to show much of the systematic structure that is known to exist in speech. In mathematical signal processing, a Fourier analysis of a waveform converts the waveform into a spectrum – a set of sinusoids of specified

frequency and amplitude that, when combined additively, can reproduce the waveform. For speech research, since the hearing apparatus does not respond to phase differences, relative phase of component sinusoids is routinely discarded. In a living system, when a sound wave is presented to the hearing apparatus, it causes vibrations in the tympanic membrane which are transferred by the ossicles to the oval window in the cochlea. Those vibrations cause standing waves in the basilar membrane, which is lined with cilia – hair-like nerve cells with different resonant frequencies. The cilia generate nerve impulses in proportion to their degree of activation, and send those impulses to the auditory nerve and on to the auditory cortex (Moore, 2004). While the cochlea performs a type of frequency discrimination, it is notably different from a Fourier transform. Fig. 2.1 shows a spectrum of a sound with component frequencies of 100Hz, 200Hz, and 300Hz. Note that each spectral line shows sound energy at an exact frequency, without spillover into adjacent frequencies. There is no apparent sound energy at 105Hz.



Fig. 2.1 – A spectrum showing three sinusoids of different frequencies.

Contrast this with Fig. 2.2, which shows the shape of the basilar membrane in response to inputs of different frequencies. Each curve represents the shape the basilar membrane adopts in response to a single sinusoid. Note that, while the peak activation is at the input frequency, there is considerable activation at adjacent frequencies. The hearing mechanism simply cannot respond to input frequencies with the mathematical precision of the Fourier transform.

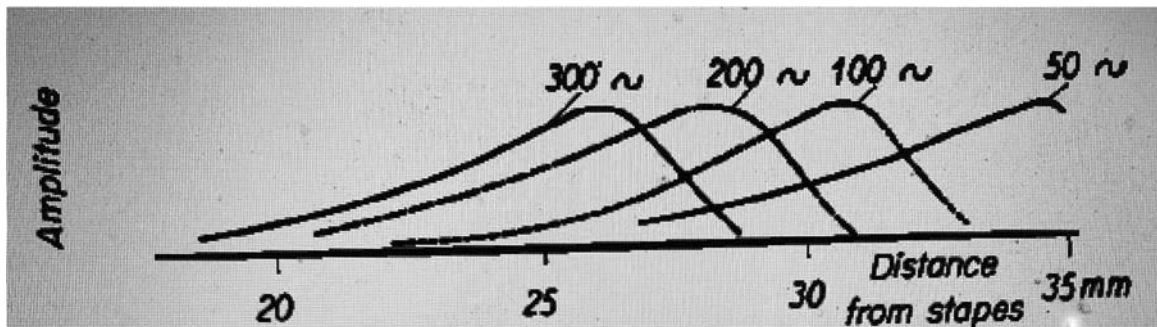


Fig. 2.2 – Shape of the basilar membrane in response to different input frequencies. (von Békésy, 1947).

While the Fourier transform satisfies the mirror heuristic in that, like the cochlea, it takes a waveform as input and separates it into component frequencies, there is no question that, as a model of the cochlea, the Fourier transform falls short in its details. Given a complex waveform consisting of many frequencies, the Fourier transform will produce a spectrum marking exactly those component frequencies. That same waveform presented to the cochlea will result in a shape on the basilar membrane that bears little resemblance to the spectrum. Since an input sinusoid will cause a distortion of the basilar membrane that is centered on the characteristic frequency but spreads a significant distance to either side, a complex input will result in a basilar pattern that represents, not the component frequencies of the input, but the interaction of those component frequencies. While the difference between the Fourier transform and the cochlear response may or may not have a meaningful effect on the performance of a computational

model, it hints at the inadequacy of the mirror heuristic in evaluating models of the acquisition process, as opposed to models of performance. A computational model of the acquisition process should model the natural system in all its details, to the extent possible. The alternative is to adopt a simpler model that might lose meaningful information.

The sound pre-processing typical of computational models departs from the function of the human hearing apparatus in other ways. While the cochlea responds to sound input in real time, the Fourier transform requires a sufficiently long sound input to act on. For this reason, the speech signal is divided into 10ms slices, each of which is treated as a repeating signal to reach an appropriate temporal length for the Fourier algorithm to act on. Clearly, this slicing, copying, and pasting procedure is not carried out by the cochlea. When the spectra from successive time slices are arranged along a temporal continuum, the result is a spectrogram. This is the form of analysis that allows a researcher to visually identify vowel formants, fricatives, and other acoustic features relevant to segment definition. In the human hearing system, any analysis of this type takes place in the auditory cortex (Moore, 2004) and is necessarily different in its details from the transformations performed during computational signal processing.

Since spectra and spectrograms can vary considerably between speakers, another transformation is typically applied in an effort to duplicate the speaker normalization that a human language user apparently accomplishes. The source-filter model of speech production has the glottal source produce a spectrum that is then altered by the vocal tract through constructive and destructive reflections at the glottis and the oral opening. In a spectrum of this resulting sound, the internal harmonic structure is largely due to the

output of the glottis, while the varying spectral envelope is due to the configuration of the vocal tract. Since vocal tracts are minimally variable between speakers when adjusted for length, most interspeaker variation comes from the glottal source. Separation of these two components offers a method of speaker normalization. This is done by passing the spectrum through a mel filter, which adjusts for the uneven response of the human auditory system to different frequencies, and applying a second Fourier-type transform (actually a discrete cosine transform). The result is a set of mel-frequency cepstral coefficients (MFCCs). The higher MFCCs represent the action of the glottis, which is the information we wish to discard during speaker normalization. The lower MFCCs represent the shaping of the vocal tract, which is more consistent across speakers and is clearly related to the production of phonemes. Typically, the first twelve MFCCs are used. Because sonorants and fricatives have more energy than stops, it is useful to calculate the power of the first twelve MFCCs. This produces a thirteenth number. It is also common to calculate the change in each of these thirteen values from one time slice to the next. This change is known as the delta and produces another set of thirteen numbers. The change in the delta from one time slice to the next is known as the delta-delta, and is also commonly calculated, resulting in thirteen more numbers. Ultimately, each time slice is represented by a vector of thirty-nine numbers.

If the goal is simply to produce a system that can, to some extent, normalize between speakers and learn to categorize phonemes, this preprocessing can be effective. However, if the goal is to mimic the behavior of a naturally occurring language-using system (like a human), this method raises a number of questions. (If the goal is to mimic the learning behavior of a human infant, the questions are significantly compounded.

That issue will be addressed in Chapter 4.) While the preprocessing steps outlined above may be roughly analogous to the function of the human auditory system, it is important to consider whether the differences between these two processes are significant. From the perspective of the mirror heuristic, the question is about the definition of the function of the system. Recall that the mirror heuristic states that, if two systems accept the same input and produce the same output, they are performing mathematically identical functions. But is it sufficient to say that the computational pre-processing and the human auditory system both take sound as input and produce something more or less like a spectrum as output, or should we strive for a computational system that produces the very same output as the natural system, in all its details, rather than in broad strokes? This question will be taken up again in Chapter 4.

2.3 Some Model Families

The Algorithm and Implementation levels of analysis highlight the most obvious differences between various models. As mentioned in Section 2.1, there seems to be a historical tendency to create models that are as simple as possible at these two levels, i.e., in their processes and their structures, and then to incrementally make them more complex, in an apparent effort to converge on the minimally complex model that is capable of accomplishing the desired task or representing the relevant natural system. This tendency will be illustrated in two different classes of models. This is not intended as a complete history of computational models, but rather as evidence of a particular philosophical approach to model design that will be argued against in Chapter 4. Section 2.3.1 covers some clustering algorithms which fail as models of acquisition at both the

Implementation and Algorithm levels. Section 2.3.2 discusses models loosely inspired by the workings of the brain that still fail to satisfy Marr’s second and third levels of analysis.

2.3.1 Clustering Models

Jain (2009) defines the goal of a clustering algorithm as “discover[ing] the *natural* grouping(s) of a set of patterns, points, or objects”, and offers the following informal definition of the process.

Given a *representation* of n objects, find K groups based on a measure of *similarity* such that the similarities between objects in the same group are high while the similarities between objects in different groups are low.

Clustering is an important technique in a wide range of disciplines, including computer vision (Shi and Malik, 2000), marketing (Arabie and Hubert, 1994), genomics (Baladi and Hatfield, 2002), and others that rely on multivariate data. A style of clustering known as partitional clustering (Jain, 2009) divides a set of data points into a number of clusters, or regions of high density separated by regions of low density, without positing any structure in the relationships between clusters. This stands in contrast to hierarchical clustering, in which each cluster can be divided into some number of subordinate clusters. This type of clustering will be addressed later in this section. Partitional clustering poses several challenges to the modeler, including defining metrics of similarity between data points, determining the optimal number of dimensions to use, and deciding how many clusters are desired. Although many clustering algorithms have been developed, one of the earliest, simplest, and most common is k-means, introduced

by Steinhaus (1957). Many improvements have been made to the original algorithm in an effort to address the challenges above. For a thorough discussion of these improvements, see Bock (2008). For present purposes, a brief exploration of the general idea of k-means and some of the difficulties it faces as a model of phonemic acquisition will suffice.

Fig. 2.3 illustrates a trivial example of clustering. Tokens with two measured dimensions (X and Y) are plotted on a two-dimensional graph. The clustering algorithm has the task of determining which tokens belong together. In an actual clustering problem, the tokens would be unlabeled, i.e., they would appear identical except for their position. In this example and the ones that follow, tokens are marked with shape and color for expository convenience. At a glance, it should be obvious that the blue stars form one cluster and the red squares another.

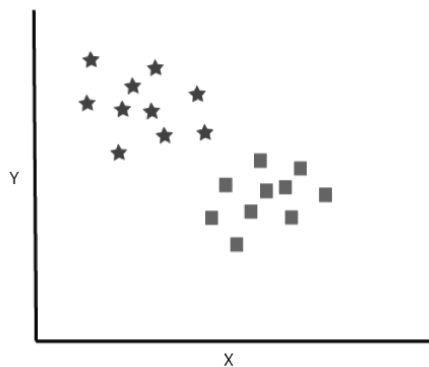


Fig. 2.3 – A trivial clustering problem

In some cases, an added dimension can clarify appropriate clusters. In Fig. 2.4, the left panel shows a number of tokens that overlap in position, seemingly belonging to the same cluster. However, if a third dimension is measured and included in the graph, as in the right panel, we see that the blue stars have low values on the z-axis, while the red

squares have higher values on that dimension. Where clustering was impossible in two dimensions, it becomes possible in higher-dimensional space.

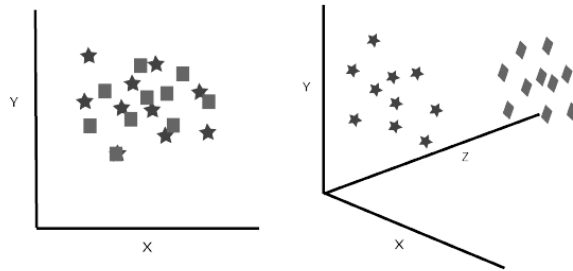


Fig. 2.4 – Clustering can benefit from added dimensions.

In other cases, an additional dimension can change lower-dimensional clustering results. Fig. 2.5 shows such an example. In the two-dimensional left panel, where only the X and Y measures are used, it is clear that the stars form one cluster and the rectangles another, regardless of color. But, when a third measure of each token is taken and plotted on the z-axis, it looks like the blue tokens form one cluster and the red another, regardless of shape.

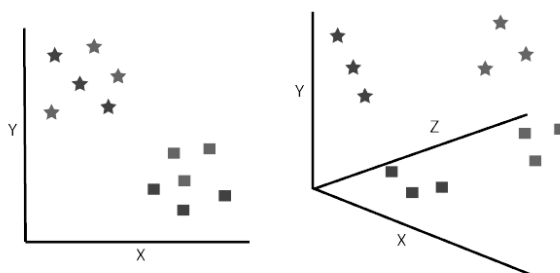


Fig. 2.5 – An added dimension can change clusters.

Because humans generally have difficulty navigating patterns in more than three dimensions, and multivariate datasets often have a far greater number of dimensions, it is

unusual to graph a dataset and determine clusters visually. As an example, in the sound pre-processing scheme laid out in Section 2.2, each time slice of each phoneme is represented as a point in 39-dimensional space. Since a graphical approach is infeasible, the dataset is subjected to mathematical manipulation to separate it into clusters.

A typical algorithm might involve the random selection of k data points to serve as the centroids of k clusters. Each point in the dataset is then assigned to the centroid it is closest to. A cluster consists of a centroid and all the data points that are closer to it than to other centroids. For each cluster, the average position of its data points is calculated and marked as the new centroid. All points in the dataset are again assigned to the nearest centroid, which may have changed, since the centroids have been recalculated. This process of refinement is repeated until the change in centroids is arbitrarily small, until no tokens are moved, or until some other measure of completion is achieved (adapted from Bock, 2008). The improvements alluded to above include variations of this general algorithm. There are also variations in distance metrics, in scaling of various dimensions, in the conditions that stop the refinement process, in whether the centroid is based on the mean or the median value of a cluster, and numerous other details (Jain, 2009).

In terms of modeling phonemic acquisition, the process of developing a centroid for each phoneme is appealingly reminiscent of the perceptual magnet effect (Iverson and Kuhl, 1995). In this theory of phonemic categorization, the infant develops a magnet – a prototypical example of each phoneme – that, much like a k -means centroid, defines the tokens that are perceptually close enough. The idea is that the infant, hearing a token of a phoneme that is reasonably close to the magnet, will experience the perceptual attraction

of the magnet and hear that token as an ideal example of that phoneme class. As strong as the parallels between k-means and perceptual magnets seem to be, all k-means clustering algorithms suffer from problems that are relevant to their utility as models of phonemic acquisition. It is not always clear which dimensions should be included in the clustering process. As illustrated in Fig. 2.5 above, the results of the clustering process can vary with the inclusion or omission of additional measured values. The relative scaling of different dimensions can also have an impact on clustering. Dimensions measured in incomparable units are frequently normalized to span numbers from 0 to 1, where 0 represents the minimum actual value, and 1 represents the maximum actual value (Dhillon, 2004), but many normalization schemes are available, and all affect the distance measures between tokens. The choice of normalization method can influence the development of clusters. One major consideration is selecting or discovering the appropriate number of clusters, i.e., the value of k . In some cases, this value is supplied by the researcher. In other cases, the model is run with different values of k , and the researcher chooses the best result. In still others there is an attempt to automatically determine the appropriate number of clusters (Jain, 2009). The automatic determination of the number of clusters often starts with a large number of clusters that are then combined until some condition is met. Possible conditions include statistical measures of the gaps between clusters (Tibshirani, 2001) and information theoretic measures such as the Minimum Description Length (Hansen and Yu, 2001). At a maximum, a k-means model can develop as many clusters as there are non-identical input tokens. In this outcome, each token would be a phoneme unto itself. At the other extreme, this type of model could assign all tokens to a single large cluster, which would clearly represent a

failure to distinguish phonemes. Whether the model is explicitly told how many clusters to form or is designed to stop dividing clusters according to one of the methods above, this is information that is not available to the human learner. One may propose that the human learner does eventually have this information, so providing it to the computational model is nothing more than a practical simplifying assumption. But, from the perspective of modeling the acquisition process, the provision of information that the human learner must discover in the course of performing the process that is being modeled seems to regrettably taint the ecological validity of the model.

A final problem that restricts the utility of a k-means clustering approach to modeling phonemic acquisition is that the input tokens have a temporal dimension. Each time slice is represented by a fixed-length vector of thirty-nine values, but each phoneme comprises multiple time slices. This variability cannot be accommodated in a model with a fixed number of dimensions.

Other models depend on a different type of algorithm that recognizes the sequential nature of speech information. A given set of cepstral coefficients and related information from a single time slice is more or less likely to belong to a given phoneme, depending on what cepstral information is provided by the next time slice. Recall that the typical preprocessing of sound (see Section 2.2) includes deltas – the change in cepstral coefficients from one time slice to the next – and delta-deltas – the change in deltas from one time slice to the next. In this way, each vector contains information about some of the values in the next two time slices. This built-in sequential information reinforces the connections between time slices by ensuring that the vector corresponding to each time slice is partially predicted by its predecessor and partially predicts its successor. A type

of model that similarly exploits the connections between successive time slices is the Hidden Markov Model (HMM). (For a thorough description of HMMs, see Rabiner, 1989, from which this discussion is derived.) An HMM is a directed graph (a set of nodes with directional connections between them. See Fig 2.6.)

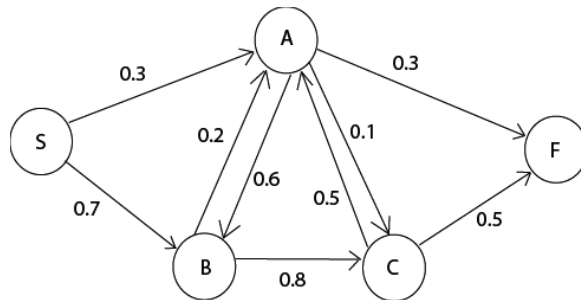


Fig. 2.6 – Hidden Markov Model (with missing emission probabilities)

Movement through the model starts with the start node at the left (labeled “S”) and ends with the finish node at the right (labeled “F”). The connections, or edges, are assigned transition probabilities. For example, from the start node, there is a 0.3 probability of moving to node A and a 0.7 probability of moving to node B (for a total probability of 1 of leaving the start node – staying on a node is not an option). One can calculate the probability of any path by multiplying the transitional probabilities of the individual steps. Starting at S and moving to A, then C, then back to A, and to the final node F has a probability of $(0.3)(0.1)(0.5)(0.3) = 0.0045$. The probabilities of all possible paths sum to 1, since there is a 100% chance of following a possible path. Verification of this fact is left to the reader as an exercise.

In an HMM, each node, in principle, has the ability to emit any value within the set of values that is relevant to the model. (For simplicity’s sake, we will assume that the Start and Finish nodes do not emit anything. This is not the case in all HMM designs.)

In the context of categorizing phonemes, the emission might be the thirty-nine dimensional vector of cepstral coefficients, deltas, delta-deltas, and power values. In this example, possible outputs are the colors red, green, and blue. Each node has emission probabilities for the possible outputs, allowing different nodes to have different likelihoods of emitting any given color. The tables below list the emission probabilities for the three nodes.

	A	B	C
Red	0.5	0.7	0.3
Green	0.3	0.1	0.3
Blue	0.2	0.2	0.4

Emission probabilities for nodes A, B, and C.

In this particular HMM, the output sequence “Blue Red” can only come from a path that has exactly two nodes between start and finish. Returning to Fig. 2.6, there are three paths from start to finish that hit exactly two nodes – Start-A-C-Finish, Start-B-A-Finish, and Start-B-C-Finish. For the first possibility, there is a 0.3 probability of moving from Start to A, then a 0.2 probability that A will emit “Blue”, followed by a 0.1 probability of moving from A to C, and a 0.3 probability that C will emit “Red”. For this path, the probability that the model will emit “Blue Red” is $(0.3)(0.2)(0.1)(0.3)=0.0018$. Probabilities can be calculated for the other paths as well, giving the total probability that the HMM will emit the desired sequence. Note that the possibility of bidirectional movement between A and B and between A and C offers the potential for output

sequences of any length, although each added step reduces the probability of such an output.

In an HMM designed for phonemic categorization, the possible outputs would be the thirty-nine dimensional vectors already discussed. Since MFCC values are not discrete, the emission probability tables take the form of probability density functions, which represent the probabilities of values along a continuum. In a process reminiscent of the way k-means algorithms refine their centroids, learning in an HMM is a matter of adjusting the transition and emission probabilities to maximize the total probability that the HMM will emit the dataset it is training on. If two HMMs are intended to represent two phoneme classes, the dataset is divided in two, with each one assigned to one HMM. The transition and emission probabilities of each HMM are then adjusted to make it the ideal representative of the set of tokens assigned to it. Then each token is reevaluated to determine which of the updated HMMs it fits better, and reassigned, if necessary. The HMMs are updated again, and the process of refinement continues until it reaches a predefined stopping condition.

Phonemes of varying length can be represented by this type of model, because the multiple connections between nodes allow for paths of different lengths from the Start node to the Finish node. While it is not represented in the diagram in Fig. 2.6, in an HMM, a node can connect to itself. A phoneme with a relatively lengthy steady state, like a vowel or a fricative, will likely be best represented by a path through the model that includes several steps from a node to itself. Shorter phonemes, like a flap, will be best represented by a short path through the model. Much of the power of an HMM comes from its flexibility in representing input sequences in several ways. Each input sequence

can, with some degree of probability, be emitted by any path of the right length. The probability of the model emitting that sequence is the sum of the individual probabilities of the various paths producing that sequence. This allows the HMM to represent more than just the most prominent features of an input token, giving it the ability to exploit a broader set of statistical regularities in the data. From the perspective of modeling acquisition, this characteristic of HMMs is appealing, since it has been demonstrated that infants use statistical reasoning in the process of acquisition. However, HMMs only respond to the statistical pattern that are found in the dataset. Saffran (2003) points out that infants not only respond to statistical patterns, but actually constrain the learning process through limitations of their perceptual ability. HMMs respond to whatever statistical information they have access to. Infants do to, but they only have access to the information their sensory apparatus permits. What looks like a similarity between HMMs and infants – their sensitivity to statistical information – is actually an argument for developing ecologically valid and biologically plausible model structures and processes. An analysis of HMMs at the Implementation and Algorithm levels suggests that HMMs are likely to do things infants cannot do and fail to do things that infants can do.

2.3.1.1 A Particular Model of Acquisition

Many models are presented simply as performing categorization of the type that adult humans do. At least one model is explicitly purported to be a model of phonemic acquisition, using HMMs to implement a hierarchical clustering algorithm. Lin (2005) presents a model that features two HMMs that “compete” for a set of inputs. Each input

is assigned to the HMM that is most likely to emit it. The HMMs are adjusted to what amounts to a centroid of their assigned tokens, and all tokens are reevaluated and reassigned. After several rounds of refinement, the model is ready to generate another level in the hierarchy. Each of the initial HMMs spawns two identical daughter HMMs, whose parameters are then randomly adjusted slightly. The daughter HMMs then compete for the input tokens assigned to the mother HMM. This process continues until a binary-branching hierarchy of HMMs is generated with the lowest level representing phonemes defined by acoustic feature bundles that can be read off the tree by following a path to the top node. Much is made of the model's classification of phonemes into a hierarchy of binary acoustic features and the implications this has for acquisition research. The claim that this model mimics an infant's acquisition of phonemic distinctions is problematic. Regardless of whether infants actually learn phonemic distinctions by separating input tokens into binary classes aligned with acoustic features, the model is explicitly designed to do exactly that. An HMM-based hierarchical clustering model in which each cluster spawns two daughter clusters will necessarily produce a binary branching tree of HMMs. This is not a validation of any theory that posits binary acoustic features. The binary nature of the features the model develops are a necessary consequence of the model structure. Similarly, the alignment of the model's developed features with traditional acoustic features says more about the model structure than about the feature system. An input scheme based on specific acoustic transformations would be hard pressed to produce categories that are not correlated with acoustic measures. This model is not so much a validation of feature geometry (Clements, 1985) as it is an expression of feature theory. Even if infants do exactly what

the model does, the model, by virtue of its design, does not tell us anything about what infants do. Thus, it is not a mode of infant phonemic acquisition; it is a model of a specific theory about infant phonemic acquisition. This highlights the ventral difficulty of modeling natural systems as a way of understanding those systems. There can be an overwhelming temptation to model a favored theory of how the natural system works, and the claim that the model's success at performing a task specified by the theory confirms that it is actually a model of the natural system. In the absence of ecological validity or biological plausibility at any level of analysis, this claim of modeling success is overly optimistic.

We have seen that clustering models can vary dramatically in complexity and can learn to categorize phonemes with some degree of success. Some aspects of these models, especially their incremental refinement of categories and their exploitation of statistical regularities in the data, are conceptually similar to what infants do during the acquisition process. Nevertheless, it is clear that the mirror heuristic – the notion that equivalent output indicates equivalent process – is inadequate for models of acquisition. To model the actual acquisition process, not just the end result, the Algorithm and Implementation levels of Marr's analysis must be given more attention. The sound preprocessing that is typical of computational models involving speech diverges from the functioning of the human hearing apparatus in several important ways. An effective k-means clustering algorithm requires information that the infant learner does not have access to. The structure of a putative model of phonemic acquisition makes the results inevitable. In every case, elements of the structure and process of the model diverge so sharply from what infants can do that they eliminate themselves as models of acquisition.

It seems that an effective model of acquisition should begin with an Implementation level that has roots in the actual structure of the natural language learning system. The next section explores some models that are loosely based on the structure and function of the human brain.

2.3.2 Information Theory Models Based on Neurons

Some models make a stronger attempt to align with natural systems on the Implementation level and, to a lesser degree, on the Algorithm level. This section explores some of those models, with an emphasis on the tendency to simplify the structure and algorithms of the model in the beginning, and then make them incrementally more complicated. We begin with the perceptron (Rosenblatt, 1958) and Minsky and Papert's (1966) objections to it, and continue with Rumelhart and McClelland's (1985) development of the modern artificial neural network, and some more recent variants, including convolutional neural networks and LSTM networks.

2.3.2.1 The Perceptron

Building on work by McCulloch and Pitts (1943), Rosenblatt (1958) introduced the perceptron, a rudimentary neural network. This model represents a network of processing elements, each intended to be a simplified version of a neuron. It is worth noting that the perceptron is nothing more than an implementation of a logistic regression – a multiple regression with each input passed through a logistic (or bounded exponential) function to force it into a range between zero and one.

The principle behind the perceptron starts with the idea that an input can influence an output. This is exemplified by the general function in slope-intercept form, $y=mx+b$, where y represents the output or dependent variable, x represents the input or independent variable, m represents the degree to which the value of x influences the value of y , and b represents the value of y in the absence of any influence from x . Fig. 2.7 shows the graph of such a function where $m=2$ and $b=4$.

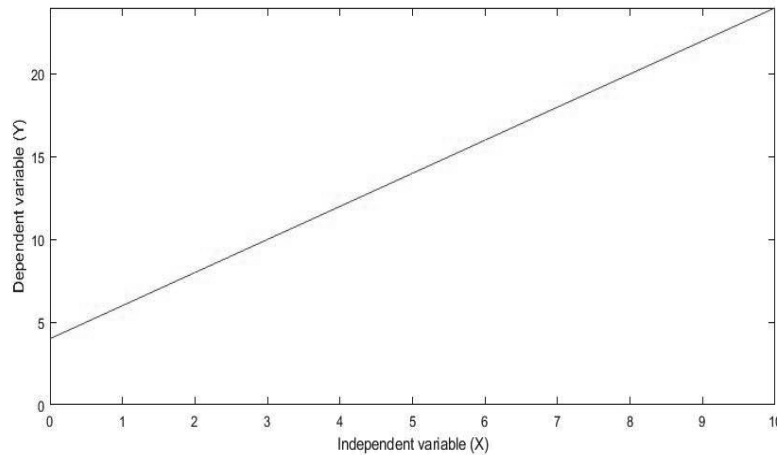


Fig. 2.7 – The graph of $y = 2x + 4$.

Where x is 0, the independent variable does not affect the dependent variable, which has a value of 4. As the independent variable increases in value, the value of the dependent variable increase twice as fast ($m=2$).

In most datasets of interest, including those related to language, the dependent variable tends to be influenced by more than one input or independent variable. This is represented in Fig. 2.8 as a three-dimensional graph of the function $y=4x+1.5z+3$, a specific instance of the general form, $y=mx+nz+b$.

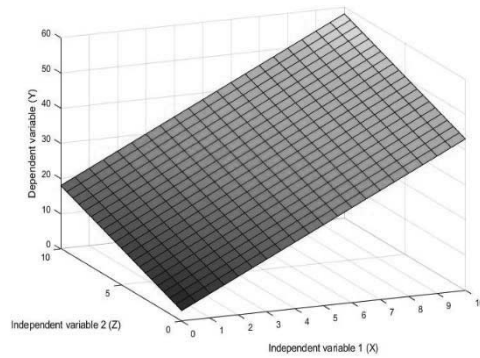


Fig. 2.8 – The graph of $y = 4x + 1.5z + 3$.

The tendency of the dependent variable, absent the influence of the independent variables, is to adopt a value of 3, as indicated by the value of b in the equation. The value of m indicates that a change of 1 in the value of the independent variable labeled x results in a change of 4 in the value of the dependent variable. Similarly, the value of n indicates that a change of 1 in the value of the independent variable labeled z results in a change of 1.5 in the value of the dependent variable. In this way, each independent variable has a weighted effect on the dependent variable. This idea can easily be extended to many more dimensions. To avoid an ever-expanding list of variables, the convention in statistics is to label all the independent variables as x with different subscripts and all the weighting factors as β with subscripts matching the independent variables they are weighting. The value of the uninfluenced dependent variable is labeled β_0 and moved to the front of the equations. This yields the canonical multiple regression equation (Cohen, 2003):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_n x_n$$

Explaining this equation in operational terms, each input variable (x_n) is multiplied by its own weighting factor (β_n) and the results are added together, along with the natural tendency, or bias (β_0), of the dependent variable. The result (y) is the predicted value of the dependent variable. Representing this process in graphical form, with appropriate changes to the variables to align with common usage in computational linguistics, yields the diagram in Fig. 2.9.

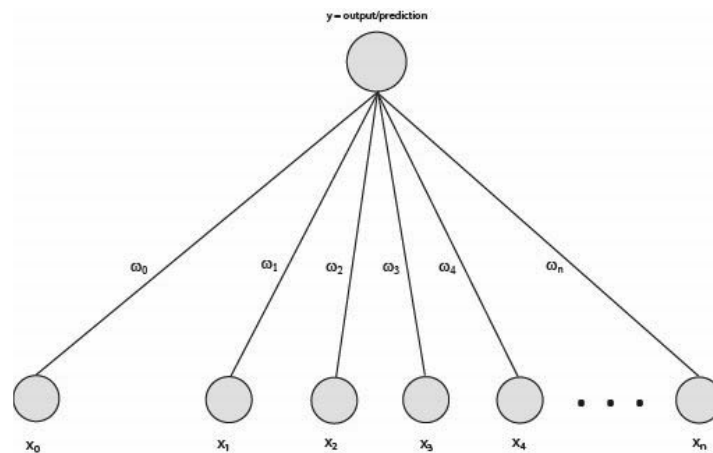


Fig. 2.9 – A multiple regression in diagram form. A perceptron follows this structure, but with a logistic function applied to each input.

Each input variable (x_n), represented by the circles at the bottom, is multiplied by a weighting factor (ω_n), marked next to the connecting lines extending from the circles, and the results are added together to give the predicted output (y), represented by the circle at the top. This diagram represents a perceptron, except for one necessary additional detail.

It has been known since Emil Dubois-Reymond's work in 1849 that nerve cells fire (generate an action potential) at full force or not at all, rather than generating weaker action potentials for weaker inputs (Jessell, 2000). However, neurons do not generate action potentials at precise and predictable moments. Instead, they are increasingly likely

to fire as their membrane potential approaches and then crosses a neuron-specific value. This increase in the probability of firing is not linear, and is best represented as a logistic curve, also known as a bounded exponential curve, illustrated in Fig. 2.10.

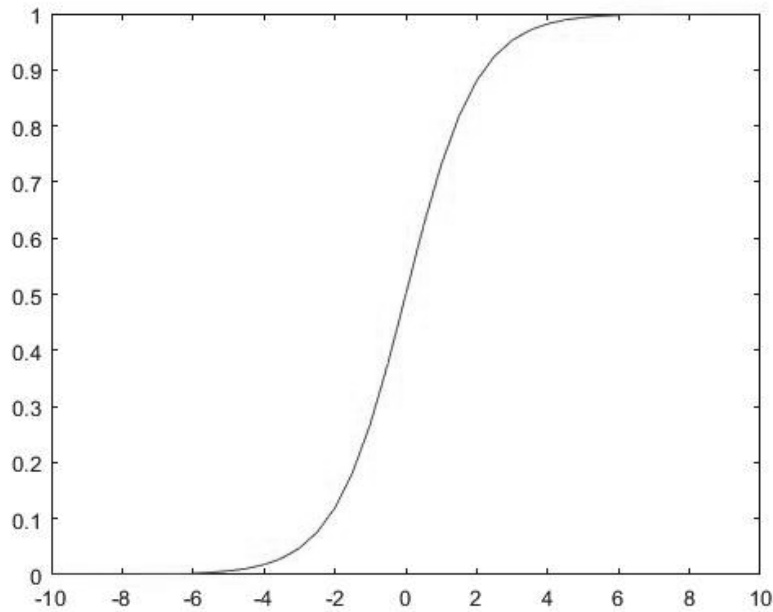


Fig. 2.10 – A logistic, or bounded exponential, curve.

In a perceptron, each input value is taken as the input to a logistic function, with the output of that function then serving as the input to the model illustrated in Fig. 2.9. In effect, each input value is converted to a probability that the analog of a neuron will “fire”, or produce an output that will contribute to the output of the model. This changes the multiple regression to a logistic regression, which is exactly what a perceptron instantiates. Extending the logistic curve to three dimensions (two input variables, one output) gives the graph in Fig. 2.11.

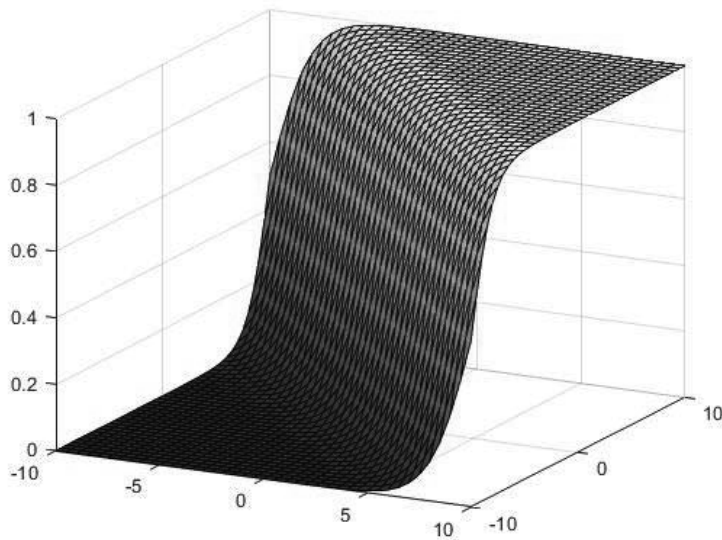


Fig. 2.11 – A logistic curve in three dimensions, showing two independent variables with bounded exponential relationships to the dependent variable.

The perceptron as described above is commonly called a single-layer perceptron, in contrast to the multilayer perceptron that shortly followed (Olazaran, 1996). The multilayer perceptron is intended to address the issue of orthogonality between input variables. The word “orthogonal” refers to statistical independence between two variables. Its origin, from the Greek “ ὀρθο- ” (“straight”) and “ γωνία ” (“angle”) (etymonline.com, 2018), sheds light on its meaning. Consider the graph in Fig. 2.12.

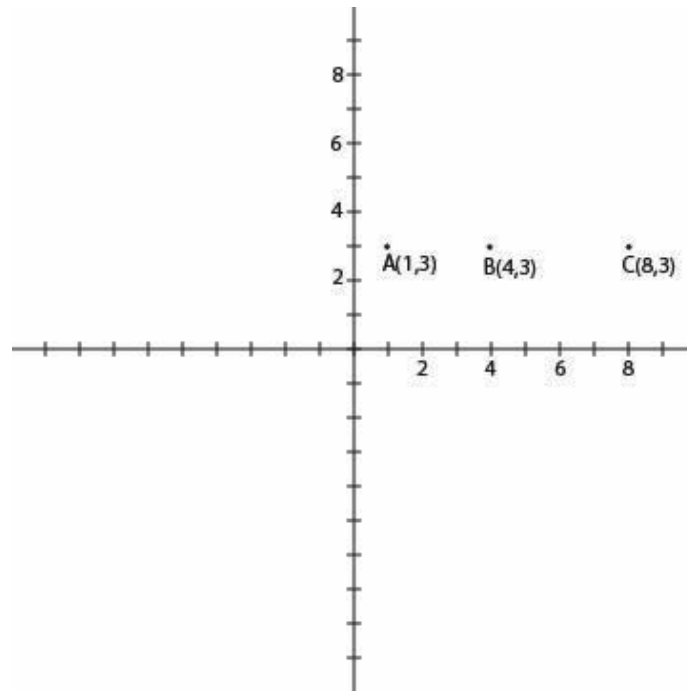


Fig. 2.12 – Moving a point across orthogonal axes.

The horizontal and vertical axes are at right angles to each other – they are orthogonal. (Note that the axes are not intended to represent an independent variable and a dependent variable. Instead, they represent two independent variables, with the dependent variable relegated to a third axis that is omitted for clarity.) This means point A, at 1 on the horizontal axis and 3 on the vertical axis can move horizontally without changing its position relative to the vertical axis. If it moves three steps to the right, it reaches point B, which is still at 3 on the vertical axis. Moving still farther, to 8 on the horizontal axis, it reaches point C, which is still at 3 on the vertical axis. In terms of inputs to a single-layer perceptron, orthogonality between inputs lets each independent variable take on any value, without affecting or being affected by any of the other independent variables. This is evident in the lack of connecting paths between the independent variables in the diagram in Fig. 2.9.

In many datasets, measurable variables are not entirely independent. A change in the value of one might precipitate a change in the value of another. In this case, the variables are non-orthogonal and can be plotted on axes that are not at right angles to each other. This situation is illustrated in Fig. 2.13, where the horizontal axis is rotated 30 degrees counter-clockwise from its original position.

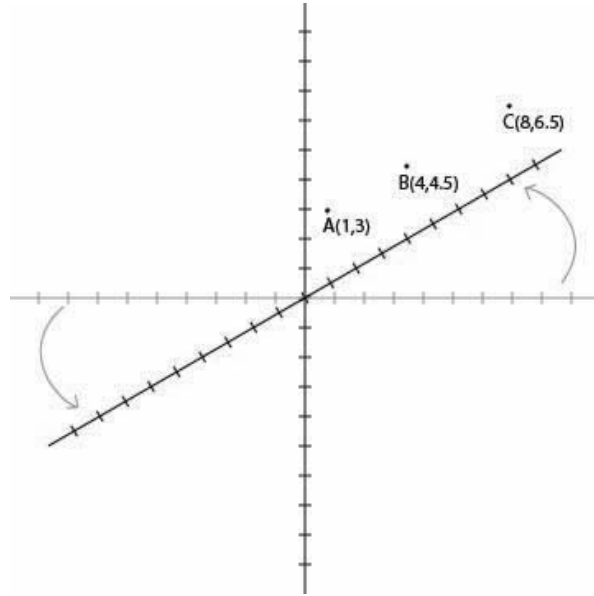


Fig. 2.13 – Moving a point across non-orthogonal axes.

In this situation, the same starting point, A, is at a position of 1 on the formerly horizontal axis and 3 on the vertical axis. Moving point A parallel to the formerly horizontal axis to a position of 4 necessarily changes its position on the vertical axis, from 3 to 4.5. Continuing its movement parallel to the formerly horizontal axis to a position of 8 changes its position on the vertical axis to 6.5. The non-orthogonality of the axes allows interaction between variables. Where two orthogonal variables, x and z , influence the independent variable, y , in the equation $y = mx + nz + b$, the equation in the case of non-orthogonality between x and z would be $y = mx + nx + z(\cos\theta)x + b$, where θ is the angle by which the x axis departs from its original orthogonal position. The

inclusion of the new term, $z(\cos\theta)x$, allows the two independent variables not only to have independent influence on the output variable, but also to moderate that influence depending on the value of the other variable. This can obviously be extended to any number of variables, with each influencing any of the others to whatever degree the model specifies. In a graphical representation, this influence takes the form of an additional layer of nodes (or processing elements), as shown in Fig. 2.14.

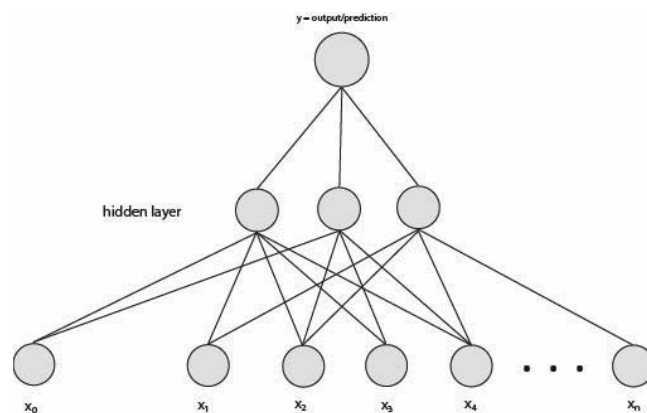


Fig. 2.14 – A multilayer perceptron, where the hidden layer allows for non-orthogonality between input variables.

Learning, in this sort of model, is a process of adjusting connection weights to allow the model to best reflect the input datasets.

While the perceptron was initially intended for use in computer vision and image processing (Olazaran, 1996), its perceived utility was quickly extended to other tasks. In an interview with The New York Times (1958), Rosenblatt predicted “Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language.” In retrospect, this seems like a rather lofty set of goals for the single-layer perceptron, which attempts to reduce every desirable output to a simple sum of weighted input variables. These goals also seem out

of reach of the multilayer perceptron, because the hidden layer allows only very limited types of interaction between input variables. This idea will be addressed in more detail in the next section.

The limitations of the single-layer perceptron were expounded upon by Minsky and Papert (1969) who argued, among other claims, that the single-layer perceptron could not represent an XOR relationship. It was thought that the ability to learn all logical operators was essential to a model that purported to perform tasks requiring intelligence. In logic, an XOR relationship is an “exclusive or”. “A XOR B” is true if A is true or if B is true, but not if both are true (or if both are false). Single-layer perceptrons are only capable of linear separation of data (Rosenblatt, 1958). Because there is no interaction between input variables, the model cannot respond differently to the firing of one input depending on the state of another input. The multilayer perceptron’s admission of interaction between input variables overcomes this limitation. There is some debate over whether Minsky and Papert were aware that multilayer perceptrons could handle the XOR problem (Olazaran, 1996). Their book proved the XOR limitation only for single-layer perceptrons. Nevertheless, many people understood the argument in Minsky and Papert (1969) to mean that neural network type models represented a dead end in artificial intelligence research, which resulted in a decline in the development of models based on the perceptron (Olazaran, 1996). This trend was reversed with McClelland and Rumelhart’s (1986) introduction of Parallel Distributed Processing.

2.3.2.2 Parallel Distributed Processing

The publication of “Parallel Distributed Processing” (Rumelhart & McClelland, 1986) marked a turning point in artificial intelligence research. Parallel distributed processing (PDP) does not refer to a particular model structure as much as it defines a philosophy of modeling or a modeling framework. This framework is inspired by the structure and function of the brain, in recognition of the idea that differences between brains and computers are not simply a matter of “software”, but of “hardware” as well.

“In our view, people are smarter than today’s computers because the brain employs a basic computational architecture that is more suited to deal with a central aspect of the natural information processing tasks that people are so good at.” (ibid, p. 3)

Eight major aspects of the framework are defined as follows (ibid, p. 46).

1. A set of processing units
2. A state of activation
3. An output function for each unit
4. A pattern of connectivity among units
5. A propagation rule for propagating patterns of activities through the network of connectivities
6. An activation rule for combining the inputs impinging on a unit with the current state of that unit to produce a new level of activation for the unit
7. A learning rule whereby patterns of connectivity are modified by experience
8. An environment within which the system must operate

These eight aspects capture more of the complexity of the brain than previously discussed models. Each of the first seven is related to neurons and their behavior (Medlar, 1998). Processing units are analogous to neurons. The state of activation is loosely analogous to the internal behavior of the neuron. The output of a processing unit represents the action potential of a neuron. The pattern of connectivity evokes the synaptic connections between neurons. The propagation rule evokes the transmission of action potentials through a network of neurons. The activation rule relates to a neuron’s

response to the collective inputs of connected neurons. The learning rule is a method of updating connections weights, which is analogous to the strengthening or weakening of synaptic connections as part of the learning process in a biological neural network. The eighth aspect, “the environment within which the system must operate”, refers to the model’s characterization of the input, the output, and the relationship between them. PDP models “represent the environment as a time-varying stochastic function over the space of input patterns” (Rumelhart & McClelland, 1986, p. 53). While a biological neural network must have some sort of mapping from input to output, it may not be of the type that is adopted in PDP models, leaving room for potential failure at the Implementation level. Nevertheless, the PDP model framework represents an effort to meet the requirements of the Implementation and Algorithm levels that surpasses the models discussed earlier in this chapter.

Specific implementations of PDP-type models have developed over time. Two that will be briefly described here are Convolutional Neural Networks (CNNs) and Long short-term memory (LSTM) networks. CNNs were designed with connection patterns between processing elements that are intended to mimic the neuron connection patterns seen in the visual cortex of animals (Matsugu, 2003). Their architecture is similar to that of a multilayer perceptron, but with multiple hidden layers, each with a specific purpose. These hidden layers include convolutional layers, pooling layers, and normalization layers (Schmidhuber, 2013). The convolutional layer is a type of feature map comprising a set of detectors, each built from multiple processing units, that perform a transformation on the input and pass the result to the next layer. In a phoneme recognition model, each of these detectors acts as a filter on the input, refining its ability to detect a particular

pattern (Murphy, 2012). Each of those patterns could correspond to a particular phoneme, much like the leaves of the binary branching HMM tree in Lin's (2005) model described in section 2.3.1.1. The pooling layers combine the outputs of groups of neurons from one layer and passes the result on as the input to a single neuron in the next layer (Schmidhuber, 2013). These layers can have varied effects, depending on the method of combination. The output of a group could be the average of the outputs of the neurons in that group or the maximum value. The processing effect is to limit the number of computational load on the network by reducing the number of neuron signals to be calculated. The effect on the operation of the network is to selectively reduce the resolution of the data, by discarding information that, in a natural system, would be overwhelmed by neighboring signals (ibid.). The normalization layers mimic the ability of neurons in biological networks to inhibit the activation of their neighbors (Le, 2015).

All of these aspects of CNNs are inspired by the visual cortex, which makes them fare better than some competing models under the Implementation level of analysis. There are notable differences between the structure of the visual cortex and that of the auditory cortex, so there is some question of the applicability of CNNs to speech recognition, if biological realism is part of the goal.

The long short-term memory (LSTM) network is a type of recurrent neural network (RNN) that is made up of LSTM units and was introduced by Hochreiter and Schmidhuber in 1997. RNNs can define a nonlinear dynamic system (Murphy, 2012), which makes them reasonable candidates for modeling brain function, i.e., for satisfying the Implementation level of analysis. The feature that allows this type of network to develop nonlinear dynamic behavior is the system of feedback within and between

processing units (Schmidhuber, 2013). An LSTM network is an RNN in which the processing units are LSTM units. Compared to a simple “fire/don’t-fire” processing element, an LSTM unit is capable of more complex behavior. It has the ability to “remember” information by maintaining its state. This is accomplished by a gate that allows or prevents the acceptance of incoming signals. There is also a “forget gate” that controls when the unit will reset its memory. This arrangement allows an LSTM unit to behave in a more complicated way in response to the relationship between the information it has been exposed to and any incoming information (Hochreiter and Schmidhuber, 1997). This makes LSTM networks well-suited to modeling data with a strong sequential component, such as speech. In recent years, LSTM networks have regularly outperformed other types of models on a range of speech-related tasks (Murphy, 2012).

2.4 Summary

This chapter has examined several types of computational models for the purpose of evaluating their suitability as models of phonemic acquisition. All of them potentially suffer from a lack of biological plausibility because of the methods of sound pre-processing that are commonly in use. Although aspects of this pre-processing are roughly analogous to what the human hearing apparatus does, there remains a question of whether the differences in processing result in meaningful differences in model performance. In the absence of clear information on this issue, it might be prudent to more precisely model the actual functioning of the human auditory system, to the extent possible.

We reviewed the concept of clustering algorithms, focusing on a particular one known as k-means. With regard to its suitability as a model of phonemic acquisition, the k-means algorithm has one strong point in its favor. The development of centroids is strongly reminiscent of the concept of perceptual magnets in acquisition (Iverson and Kuhl, 1995). However, this type of model fails in a number of other respects. It has no real connection to the physical structure of the brain, making it rather weak on the Implementation level of analysis. It also employs algorithms that require either a priori information from the researcher or built-in methods for stopping the partitioning process, both of which are unavailable to the infant learner, rendering the model unsatisfactory under the Algorithm level of analysis. Furthermore, k-means models are not typically designed to deal with time-series data, which makes them inappropriate for clustering speech sounds, which are temporal sequences of acoustic data.

HMMs were also reviewed, with mixed results. They are capable of handling sequential data and, like infants, they exploit statistical regularities, but they fall short on ecological validity, both in structure and in process. As mentioned before, the mirror heuristic may be satisfied by the transformation of input to output, but models of acquisition should include a focus on the process and the way it derives from the structure. A specific implementation of an HMM model of acquisition (Lin, 2005) was explored. It performs a hierarchical clustering process and generates a binary branching tree of HMMs, with the leaves representing phonemes and the intermediate nodes representing acoustic features. Apart from issues at the Implementation and Algorithm levels, the results of this model are largely built in from the start, limiting what it has to say about how infants learn to categorize phonemes.

Computational models that are loosely inspired by the human brain were also explored, starting with the single layer perceptron and going through convolutional neural networks and long short-term memory networks. One trend that emerges from this sequence is of significant interest. The single layer perceptron implements a logistic regression, which is far too simple to serve as a model of acquisition. The multilayer perceptron offers a slightly more complex model by including a middle layer and rejecting the orthogonality of input variables. As much as this improves the ability of the perceptron to deal with more complex data, it is worth noting that the interaction between input variables is still tightly constrained. Subsequent models continue this trend of incremental ornamentation of existing models. With the multiple goals of performing a language task, minimizing the use of computational resources, and trying to maintain the analyzability of the model by the researcher, it makes sense to start with something simple – even too simple – and progress from there. However, the structure of the human brain and the knowledge gained from acquisition experiments suggest that simplicity of structure and process is incompatible with modeling the process of acquisition. Even with the development and reported success of very impressive models like LSTM networks, they still fall short on the Implementation and Algorithmic levels of analysis. In too many cases, the performance of the model is dependent on decisions made by the researcher, and the structure of the model is, at best, roughly analogous to the human system at every level.

2.4.1 The Mirror Heuristic Revisited

In an effort to put a final nail in the coffin of the mirror heuristic, let us recall that Bechtel (2018) defines it as the assumption that, if two processes take the same input and

produce the same output, they are performing equivalent mathematical functions and are therefore equivalent in a meaningful way. This issue arises whenever a model of a natural process is created. If the purpose of the model is to help the researcher understand the natural system, one hopes that the model and the natural system are similar enough for the model to offer some insight into the operation of the natural system. In terms of Marr's three levels of analysis, the question of sufficient similarity depends to a great degree on the Computation level – the definition of what the model is intended to mimic. If we are modeling nothing more than the movement of a heavier-than-air object from one point to another, then an airplane is a reasonable model of a bird. However, we would be remiss in then claiming that an airplane offers insight into a bird's maneuverability, musculature, energy usage, or reproductive habits. To gain a greater understanding of these aspects of birds through modeling, one would have to build a model that includes the structures and processes that are relevant to what one is trying to understand. If this analog seems overly cartoonish, it is only because we have ready access to actual birds that we can examine and test directly. The situation is significantly different when it comes to modeling neural processes. Computational neuroscience is rife with models of vision, especially as compared to models of language behavior. One reason for this is that the human eye is a physical object whose function is extremely similar to the eyes of other animals, giving researcher greater ability to experimentally discern its response to clearly measureable physical input. The situation with language is noticeably different. Hearing is well understood. The structures and functions of the tympanic membrane and the cochlea are similar across mammals, and a great deal of experimentation has clarified the details of their operation. What happens in the auditory

cortex and beyond to turn sounds – but only some sounds – into the multivalent abstractions of language is more difficult. With regard to the discrimination of speech sounds, the situation is not quite so bleak. We have animal models that can serve to illuminate the response of the auditory cortex to acoustic stimuli. There are claims that some animals, most notably chinchillas (Kuhl, 1972) exhibit categorical perception in the way humans do when exposed to voiced/voiceless stop pairs, which would certainly be an exciting way to explore how phonemic perception might work in the human brain. However, there are concerns about the experimental design behind these claims about chinchillas that will be explored in Chapter 4.

The differences between humans and other animals in the processing of language pose a conundrum for the modeler who wishes to satisfy Marr’s analysis, even with the relatively simple goal of modeling phonemic acquisition. A great deal is known about the general brain regions where aspects of sound processing occur. Less is known about the fine-grained detail of how individual neurons in those regions behave, connect, and interact to produce an apparent emergent understanding of linguistically relevant distributional information from acoustic input. Even so, a modeler can approach the Implementation level armed with significant information about the types of neurons, connections, and behaviors that exist in brains, and refrain from oversimplifying those details. In the absence of this type of effort, appeals to the mirror heuristic look like efforts to mask the inadequacy of some models.

Chapter 3 will examine some models that approach the Implementation level from the perspective of mimicking details known from neuroscience. Computational cognitive neuroscience models attempt to bridge the gap between computational neuroscience

models, which tend to be focused on single neurons or small networks of neurons, and higher level cognitive behavior. The underlying assumption is that cognitive behavior is driven by the details of lower-level neuronal organization and behavior.

CHAPTER 3

Computational Cognitive Neuroscience Models

This chapter introduces Computational Cognitive Neuroscience (CCNS) models. This type of model aspires to a greater degree of biological plausibility than the models discussed in Chapter 2, especially in low-level structure, while still modeling higher-level cognitive behaviors. This plausibility is approached by building on Computational Neuroscience (CNS) models, which attempt to mimic the remarkably complex behavior of neurons and small networks of neurons. This chapter begins with an exploration of the structure and behavior of neurons and biological neural networks, and builds to CNS and CCNS models. Both types of models are evaluated under Marr’s three levels, particularly the Implementation and Algorithm levels. Unsurprisingly, they do better than the models in Chapter 2 under this analysis, since they are designed under the same philosophy that drives Marr’s three levels.

3.1 Neurons

Santiago Ramón y Cajal, in 1899, was the first to propose that neurons are contiguous, rather than continuous (Finger, 2000), meaning that, rather than forming a continuous structure with neighboring cells, they act as independent bodies. Unlike the cells in other tissues of the body, which behave largely in concert with the cells they are adjacent to, neurons form a more complex set of connections with cells that may be centered some distance away. This discreteness of neurons is known as the “neuron theory” (ibid.) and is reflected in computational neuroscience models, as well as in the structure of Rosenblatt’s (1958) perceptron and its conceptual descendants.

The anatomical structure of the neuron can be simplified into four primary components – the soma (or cell body), dendrites, the axon, and presynaptic terminals (Kandel, 2000) (See Fig. 3.1).

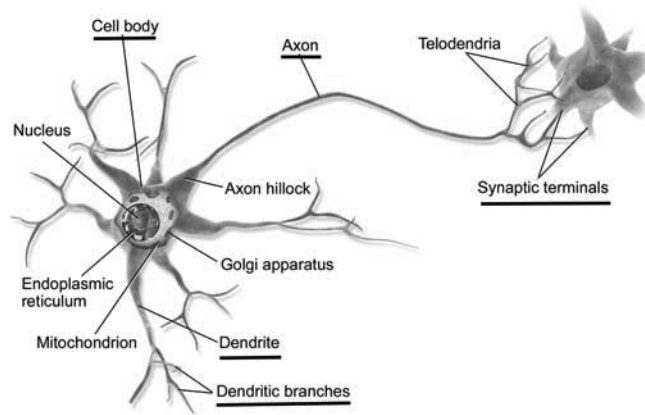


Fig. 3.1 – Parts of a neuron

The soma contains the type of nucleus and organelles that are present in all cells. While the soma is obviously important to the neuron, the dendrites, axon, and presynaptic terminals are the structures that play a primary role in the signaling between neurons (ibid.). A typical neuron has one axon and multiple dendrites, structures that extend away from the soma. The axon sends signals to other neurons, through multiple terminals that branch away from the main body of the axon, while the treelike branching dendrites receive information from other neurons (Gazzaniga, 2009). The places where the transfer of information occurs are called synapses, which are small gaps between the axon terminals of one cell and the dendrites of another. Because information flows across the synapse from axon to dendrite, axon terminals are described as *presynaptic*, and dendrites are described as *postsynaptic* (ibid.). Also of interest is the membrane of the neuron, which contains large numbers of *ion channels*, specialized protein structures

that allow or prevent the flow of ions across the membrane, i.e., between the inside and outside of the neuron, in either direction (see Fig 3.2).

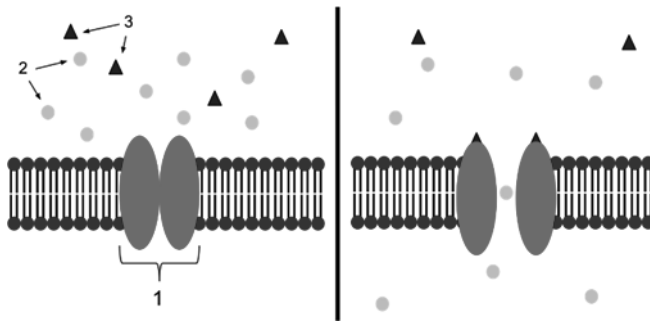


Fig. 3.2 – Ion channel closed (left) and open (right).

A single neuron may have thousands of ion channels. Some are passive, allowing ions to pass freely. Others are gated, allowing or preventing the passage of ions depending on electrical or chemical changes in their immediate environment (Kandel, 2000).

Differences in the relative concentration of ions inside and outside the neuron cause a voltage difference across the membrane. A sufficient voltage difference can cause ion channels to open, allowing ions to move across the membrane. This can lead to the opening of more ion channels. As this process continues, it is possible for the voltage across the membrane to reach a neuron-specific threshold value, which causes the neuron to produce an action potential – a voltage spike. This action potential travels down the axon to the presynaptic terminals, where neurotransmitters are released into the synapse. On the postsynaptic side, the neurotransmitters bind with receptors in the dendrites of other neurons. These bound neurotransmitters can influence the membrane voltage of the neurons they are bound to, resulting in the opening of ion channels and the possibility of generating an action potential (Gazzaniga, 2009). Because the typical neuron has many dendrites, receiving signals from many other neurons, it acts as an “integrator”,

combining all the signals it receives and responding accordingly. This process is known as “integrate and fire” (Anderson, 2014). After a neuron fires, it goes through a refractory period, during which it cannot fire for a brief period of time, and its membrane potential returns to its resting state (See Fig. 3.3).

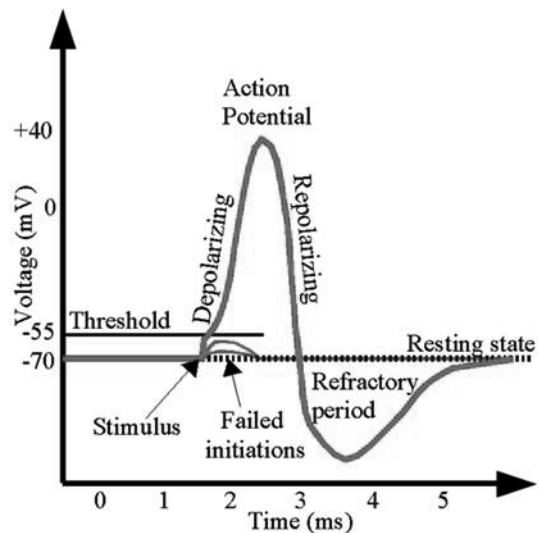


Fig 3.3 – Action potential followed by refractory period.

If a neuron receives sufficient continuous input, it will alternate between firing and experiencing its refractory period. This results in a series of action potentials, known as a “spike train” (Brette, 2007)(See Fig. 3.4).

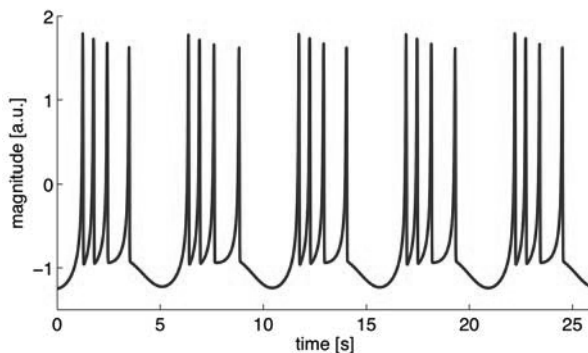


Fig. 3.4 – Spike train

Consider a hypothetical simplified network consisting of two neurons, A and B, each releasing enough neurotransmitter with an action potential to cause the other to fire. If the time between receiving a signal at the dendrite and firing an action potential down the axon is the same for both neurons, and they have the same refractory period, one would expect the firing pattern between A and B to be as regular as the swinging of a pendulum. However, if the neurons vary in any parameter – length of refractory period, dendritic length, membrane potential threshold, number of ion channels, length of axon, etc. – the firing pattern between A and B will be uneven. If the network is extended to include a third neuron, the potential for irregular spike patterning increases dramatically. Extend this system to include tens of thousands of neurons, all with different characteristics, and the potential complexity of the spike patterning of such a network is staggering.

The foregoing explanation of the function of neurons and small networks of neurons is deliberately simplistic for the sake of clarity. Even in this simplified version, the inherent complexity of small biological neural networks is evident. In reality, neurons are considerably more complicated. There are three different ions – sodium (Na^+), potassium (K^+), and calcium (Ca^{2+}) – whose complex interactions are implicated in creating and changing the membrane voltage potential (Gazzaniga, 2009). Over three-hundred types of ion channels have been discovered (Gabashvili, 2007), with each neuron having varied proportions of multiple types that are permeable to different ions. There are scores of different neurotransmitters, some of which serve to inhibit the generation of impulses in the postsynaptic neuron, as well as variations in number and length of dendrites, in threshold voltage, in axon length, in spike rate, and in a range of other

parameters (Kandel, 2000). With each axon connecting to the dendrites of up to one-thousand other neurons, the combinatorial possibilities for the behavior of a biological neural network are seemingly limitless.

3.1.1 Simplifying models

Naturally, a primary focus when modeling neurons is finding elements that can plausibly be simplified. Ideally, any simplification will still produce a model that is capable of reproducing some of the behavior of the natural object being modeled. In some cases, an apparent over-simplification can be subject to an ecological justification, even if the model falls short on other grounds. For example, consider computational models based on the work of Pitts and McCulloch and, later, Rosenblatt, described in Chapter 2. In those models, the analogs of neurons have a probability of firing that is described by a simple logistic curve (See Fig. 3.5), although some other mathematical functions are also in common use.

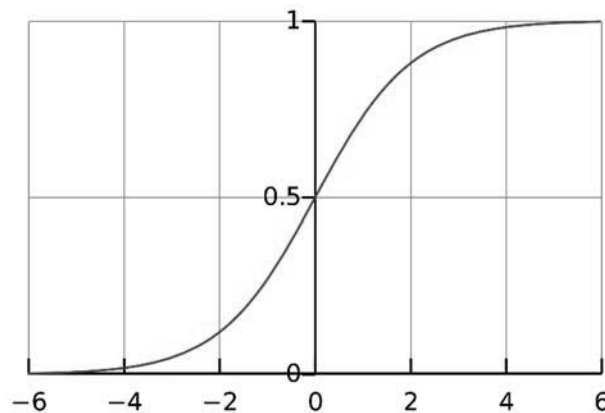


Fig. 3.5 – Logistic curve

A logistic or bounded exponential function is used to represent a variable whose value increases exponentially, but is subjected to a limiting factor. A simple example comes from animal population studies. A given species, with abundant resources and

limited risk of predation, will undergo an exponential increase in population as each breeding generation multiplies itself by the average litter size. At some point, the population becomes large enough to feel the limiting effects of finite resources. If animals do not have access to enough food and water, they will breed less and die more easily. Thus, the exponential increase in the population levels off and approaches an equilibrium point. This process is conceptually related to the behavior of individual neurons, described above, with regard to likelihood of firing. Any neuron is relatively unlikely to fire when in its resting state. As incoming ions cause changes to the membrane potential, that likelihood increases. Ion channels in the membrane open, leading to further depolarization of the membrane, opening more ion channels, and continued increase in the likelihood of firing. The process is exponential in the beginning, but tapers off as it approaches 1 (where 1 represents a 100% chance of firing). The changes in probability in response to changes in ion concentration and membrane behavior follow a logistic curve. This is the inspiration for the use of this curve in Rosenblatt-style models. Related models sometimes use the hyperbolic tangent (\tanh) as an activation function (Fig. 3.6), while others use a rectified linear unit (ReLU)(Fig. 3.7). The \tanh function can be seen as a scaling of the logistic function to allow values on the interval $[-1, 1]$, where the logistic function only allows values on the interval $[0, 1]$. Depending on the structure of the network in question, one of those intervals is more appropriate than the other. The ReLU activation function, introduced by Hahnloser (2000), departs somewhat from the operation of biological neurons, ignoring negative inputs and passing positive inputs through unaltered. The softmax function (Dugas, 2001)(Fig. 3.7) is an approximation of the ReLU function, but with a smoother transition

at the zero point. Softmax is the anti-derivative of the logistic function, which highlights the relationship between the various activation functions in common use. All of these functions can be understood as conceptual variants of a cumulative density function (Fig. 3.8) or a step function with normally distributed error in the transition (Fig. 3.9). The details of these last two functions are not especially important. The point is simply that all commonly used activation functions in perceptron-related models have the same general sigmoid shape (although ReLU is noticeably degenerate).

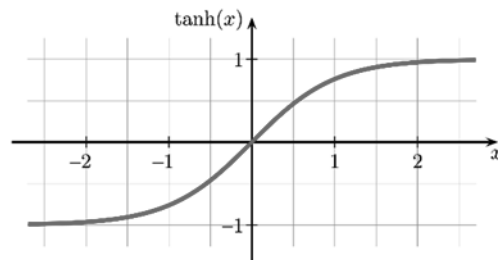


Fig. 3.6 – tanh function

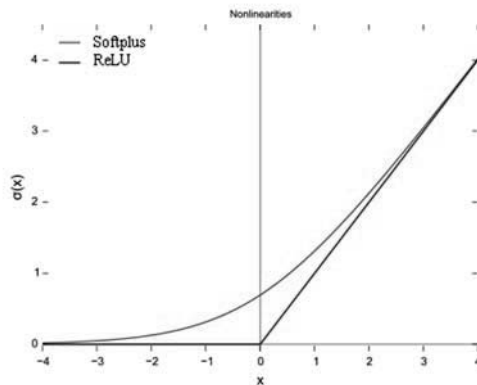


Fig. 3.7 – ReLU and Softmax

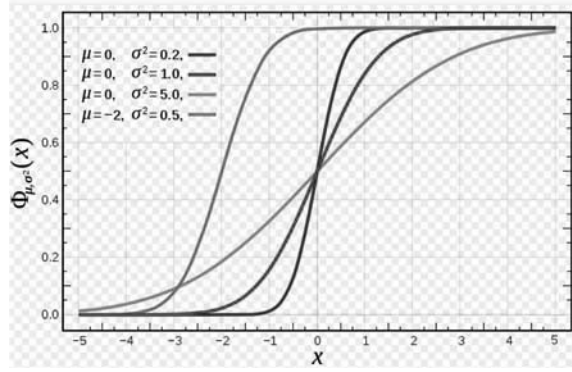


Fig. 3.8 – Cumulative density function

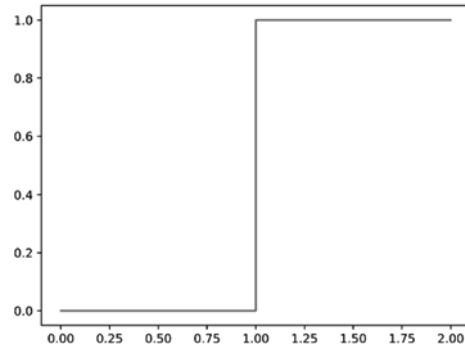


Fig. 3.9 – Step function

Because their intended use is in simplified “integrate and fire” models, they all vaguely follow the firing-probability curve of biological neurons. However, they overlook the fact that, in many cases, the complex behavior of interest in biological neural networks is driven not by the simple firing of neurons, but by the spike *rate* and spike *timing* of neurons (Brette, 2007). As previously noted, simplified models of this type have shown the ability to perform certain language tasks adequately, but modeling the process of acquisition requires a model that is closer to the actual biological system.

3.2 Neuroscience models

Computational neuroscience is a branch of neuroscience that uses computational techniques and models in an effort to develop an understanding of how the brain works.

“Although computational neuroscience is theoretical by its very nature, it is important to bear in mind that models must be gauged on experimental data; they are otherwise useless for understanding the brain. Only experimental measurements of the real brain can verify “what” the brain actually does. In contrast to the experimental domain, computational neuroscience tries to speculate “how” the brain operates. Such speculations are developed into hypotheses, realized into models, evaluated analytically or numerically, and tested against experimental data.” (Trappenberg, 2010, p.2)

This emphasis on experimental verification of computational models is what sets computational neuroscience models apart from the types of models discussed in Chapter 2. The goal in computational neuroscience is to understand the natural system by building models that are biologically plausible in structure and process, and that exhibit the same behavior as their natural counterparts. The reproduction of behavior alone, even with a high degree of precision, is insufficient. This goal aligns with the goal of modeling language acquisition. The development of a model that acquires some aspect of language without biological plausibility, without experimental confirmation, or without adhering to the facts about acquisition that are known to experimentalists is not a model of acquisition. At best, it is a model that illustrates or reiterates a particular theory of language. At worse, it is a model that merely does something interesting.

The goals of computational neuroscience are not without their pitfalls. The overwhelming complexity of the neuron, with its tremendous number of moving parts, requires simplification, either of structure, of process, or of both. But there is a fundamental tension between maintaining biological plausibility and achieving computational feasibility. Much of the behavior of the neuron arises from the interaction

of different parts. If those parts are simplified excessively, the behavior may not arise. A great deal of the art of computational neuroscience modeling is in finding the parts of the natural system that can be simplified without destroying the behavior one hopes to elicit from the model. In keeping with Trappenberg's perspective quoted above, any simplification of neural structure or process should be theoretically motivated and experimentally verified. In other words, a simplification in a model should proceed based on reasonable and supported beliefs about the degree to which the proposed simplification will affect the operation of the model on the dimension of interest, and the resulting model should be verified against experimental data from the natural system, to the extent possible. One approach to simplifying structural aspects of a model is through the use of compartment models.

3.2.1 Compartment models

A compartment model is the result of dividing a continuous system into discrete regions, each made up of numerous elements, with each region producing outputs that represent the average behavior of its constituent elements (Eriksson, 1971). Familiar examples include models in any social science, such as economics, in which a group of people are treated as monolithic, regardless of any differences between individuals. In neuroscience, a compartment model might treat all of a neuron's ion channels as a single unit that effects an aggregate change in membrane potential, rather than modeling individual ion channels and summing their individual impact. A well-established instance of compartment modeling in neuroscience has roots in the study of electronic transmission in undersea cables, which inspired the simplification of the function of

dendrites by modeling them as lengths of cylindrical cable whose physical properties are the average of the properties of the branching structures they encompass. This so-called cable theory has been a mainstay of neuronal modeling since the early 20th century (Trappenberg, 2009). While compartment modeling can be effective in reducing the number of parameters of a neuronal model to a manageable few hundred, a network of such modeled neurons can still place too great a computational burden on the computer used to run the model. A further degree of modeling abstraction results in models that focus on the spiking behavior of neurons, without regard for the details that produce those spikes. These models are conceptual descendants of work by Hodgkin and Huxley.

3.2.2 Spiking models

Although many aspects of the general function of the neuron were explored earlier, Hodgkin and Huxley (1952) developed the first mathematical representation of the electrical behavior of neurons. Their work on the giant squid axon resulted in a system of four differential equations describing how changes to the voltage potential across the cell membrane of a neuron could lead to an action potential (Trappenberg, 2002). Differential equations are used to describe process that include feedback. In the context of a neuron, drawing on the discussion above, changes in ion concentrations around a cell membrane cause certain ion channels in the membrane to open, leading to further changes in ion concentration, which could lead to cascading changes that ultimately produce a spike of electrical transmission. In other words, a change to the environment of a neuron causes the neuron to alter its behavior, which further changes

the environment, leading to further changes in the behavior of the neuron, until this cycle produces an action potential.

A currently popular approach to modeling neurons reduces Hodgkin and Huxley's four continuous-time differential equations to two discrete-time iterated functions, each influencing the other. These radically simplified models are known as neural maps. An example from Rulkov (2002) offers the following equations

$$x_{t+1} = F(x_t, y_t + \beta_t)$$

$$y_{t+1} = y_t - \mu(x_t + 1 - \sigma_t)$$

where σ , μ , and β are adjustable parameters. Note that each equation defines the value of a variable at a discrete time step in terms of its value at the previous time step and the value of the other variable at the previous time step. So, the value of x at time $t+1$ is determined in part by the values of both x and y at the previous time step, t . Similarly, the value of y at time $t+1$ is determined in part by values of x and y at the previous time step. One advantage to using discrete-time iterated functions is that the calculations are greatly simplified, reducing the computational load. Discrete-time functions are analogous to digital sampling of audio. The natural audio signal is continuous; it has a value at every point in time. For the purposes of speech analysis, knowing the amplitude of a waveform at every nanosecond is unnecessary and represents a huge amount of data. Sampling the waveform every 0.0000625 seconds (a sample rate of 16kHz) reduces the size of the resulting data set by many orders of magnitude while still delivering all the information needed for the intended analysis. (For other purposes, different sample rates are recommended. For example, commercial music is typically sampled at 44.1kHz.) Similarly, a discrete-time function as a neural map allows for the

“sampling” (generation, actually) of the neuronal output at enough points to give a clear picture of overall behavior without the computational burden of calculating a continuous function. The time step (analogous to sample rate) can be selected to give the desired temporal resolution. (As the size of the time step approaches zero, the discrete-time function approaches the continuous function. Too large a time step will omit useful information about what the neuron is doing between samples.)

It has been demonstrated that Rulkov-type maps, depending on the choice of parameters, can generate many of the behaviors seen in biological neurons, including steady spike trains, spike bursts, and chaotic behavior (Rulkov, 2002)(Fig. 3.10).

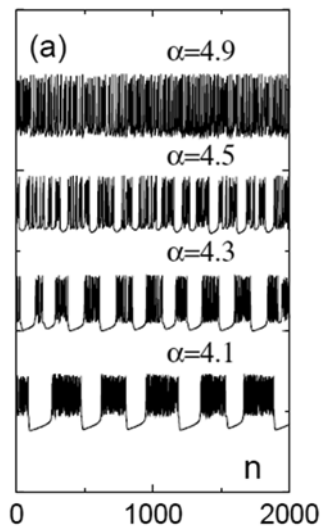


Fig. 3.10 – Rulkov map spiking behavior for different values of α

Networks of neural maps – small numbers of neural models that are interconnected so that each affects the behavior of the others – have the ability to synchronize their outputs and to settle into regular oscillations, both of which behaviors are seen in biological neural networks (Rulkov, 2004, 2008).

At the level of the neuron, the unit of currency, so to speak, seems to be the spike rate; while in networks of neurons, relative spike timing plays an important role (Brette, 2007). For these reasons, as well as reasons of computational tractability, modeling neurons at the level of abstraction achieved by neural maps is appealing. The fine-grained details of membrane depolarization and neurotransmitter release are subsumed by the simpler model of spiking behavior at the axon terminal.

In terms of Marr's three levels of analysis, Rulkov-type neural maps fare quite well, because their goals and methods are so narrowly defined. At the Computation level, their goal is simply to model the spiking behavior of individual neurons and networks of neurons. At the Implementation level, not much can be said about the individual neuron. Networks of neurons are connected in biologically plausible ways, with each neuron typically being connected to all the others in a small network (Rulkov, 2008). At the Algorithm level, if we accept the validity of Hodgkin and Huxley's mathematical representation of the generation of action potentials, then a simplified set of equations that produce the same output should also be accepted as valid. In this type of model, the details of the natural system that are either aggregated or disregarded in the model design (for example, the specifics of the cascading voltage along the cell membrane or the number of neurotransmitter receptors in the dendrites) do not affect the behavior that is being modeled. In principle, the spike train will behave identically whether there are a thousand ion channels in a given neuron or nine-hundred. The simplicity of the goal of this type of model almost guarantees its success under Marr's analysis. Whether the details of neuron behavior that the model ignores are relevant to

the emergence of higher level cognitive behavior from the low-level interactions of neurons is an issue that will arise in the next section.

3.3 Cognitive neuroscience modeling

Computational cognitive neuroscience (CCNS) is a relatively new discipline that attempts to bridge computational neuroscience and cognitive neuroscience (Zednik, 2018). As discussed above, computational neuroscience attempts to create models of neural structures that are biologically plausible in their structure and in their function. Cognitive neuroscience focuses on developing explanations of higher-level behavior that are rooted in attested neuronal structures and function. Computational cognitive neuroscience models are intended to exhibit higher-level cognitive behavior while abiding by the constraints of neuroscientific information about low-level neuronal behavior. Several researchers have laid out standards for CCNS models.

O'Reilly (1998, pp. 455-456) offers six principles to guide the design of CCNS models.

- (1) Biological realism
- (2) Distributed representations
- (3) inhibitory competition
- (4) bidirectional activation propagation
- (5) Error-driven task learning
- (6) Hebbian model learning

The concept of biological realism undergirds the entire enterprise of computational cognitive neuroscience modeling. The goal is to understand how the brain gives rise to cognition, so the model should be constrained by neuroscientific knowledge about the brain.

Distributed representation refers to the well-established notion that representations in the brain rely not on single neurons, but on some number of neurons acting in concert. Any neuron can participate in many representations.

Inhibitory competition refers to the tendency for some active neurons to suppress activity in other neurons. This is part of the learning process, in that it promotes the activity of the most strongly activated neurons.

Bidirectional activation propagation refers to simultaneous top-down and bottom-up communication. It reflects the fact that, if neuron A can send an action potential to neuron B, neuron B can usually send an action potential to neuron A. A strict hierarchical structure with information moving in only one direction violates the requirement of biological plausibility. This also has implications for the emergence of chaotic behavior, which will be addressed in the next chapter.

Error-driven task learning is essentially supervised learning. On the surface, this requirement is problematic, because it seems to require an external process that has access to information about expected behavior. However, O'Reilly posits a mechanism by which a network can settle into a pattern that reflects an expectation, and then respond to the actual input pattern in a way that allows the difference between the two to surface as an "error" pattern. Minimizing this error means, in effect, learning to align expectations with observations.

Hebbian model learning refers to unsupervised learning, which is simply responding to the distributional patterns of the input. In a biological network, this type of learning increases the synaptic strength between neurons that are regularly activated together. (Preceding explanations adapted from O'Reilly, 1998.)

Meeter (2007, pp. 760-761) offers a shorter set of criteria for CCNS modeling.

- (1) Sparse models
- (2) Binding to Biology
- (3) Ontological clarity

The sparse model requirement suggests that unjustified assumptions in a model make it more difficult to analyze and interpret. This is related to the concept of biological plausibility in that a model bound by the constraints of known neuroscience will incorporate knowledge about the brain. Any additional assumptions are not based on knowledge about the brain and are thus in violation of the sparse model requirement.

Binding to biology is essentially the mirror image of the sparse model requirement, as noted in the previous paragraph. It requires as many of the model assumptions as possible to be evidence based.

The requirement of ontological clarity is a constraint on the researcher, rather than on the model. It requires the modeler to be clear and unambiguous about the intent of the model, the nature of the model's algorithms, the model's level of representation, and the plan for testing and evaluating the model. (Preceding explanations adapted from Meeter, 2007.)

Ashby (2011, pp. 276-276) offers four requirements for building a good CCNS model.

- (1) The neuroscience ideal
- (2) The simplicity heuristic
- (3) The set-in-stone ideal
- (4) The goodness-of-fit ideal

The neuroscience ideal, similar to O'Reilly's first criterion of biological realism, says that a model should not make any assumptions that contradict known facts about

neuroscience. Since biological plausibility is one of the goals, a CCNS model should be rooted in current neuroscience knowledge.

The simplicity heuristic echoes Meeter's first criterion of having few assumptions that are not rooted in known neurobiological fact. Ashby allows an exception for unsupported assumptions without which the model would not function, but these are clearly dispreferred.

The set-in-stone ideal requires that model structure be fixed. Ashby points out that neural connections and network response to stimuli do not change from one task to the next. The notion of establishing a learning machine that can be programmed for specific tasks violates the neuroscience ideal.

The goodness-of-fit ideal requires that the model make predictions at both the behavioral and neuroscience levels. Without addressing both of these levels, a model cannot properly be called a computational cognitive science model. (Explanations adapted from Ashby, 2011.)

Whichever standards one adopts to guide the development of a computational cognitive neuroscience model of phonemic acquisition, a number of challenges stand out. Poeppel (2017) raises the paired questions of whether linguistics has anything to tell us about the nature of computation in the brain and whether improved knowledge of the workings of the brain can offer any insights into the structure of language. It is not immediately clear that either question can be answered in the affirmative. Part of the problem is that linguistic definitions and processes tend to be of a different type than descriptions in neuroscience. Linguistic analyses are typically couched in the paradigm of the electronic computer – datasets are acted on by a process to produce more data.

But, in the brain, the software and hardware are not only inseparable, they are the same thing (Jacobs, 1992). The very nature of neural computation is at odds with common notions of linguistic computation. This seems like less of a problem with the acquisition of sub-meaning units, like phonemes. Part of the phonemic acquisition process is simply a hearing task, which is relatively straightforward. The learner must attend to acoustic differences and then, through the use of statistical reasoning, begin to disregard some of those differences and pay attention to others. But the heart of phonemic acquisition is in the development of a sound system, not just a collection of sounds. Just what this would look like in the brain is an open question that will be taken up in Chapter 4.

Another difficulty in modeling acquisition lies in the relative dearth of empirical evidence at the requisite level of analysis. Models of individual neurons are assisted by experimental evidence obtained from actual neurons. Although there are many types of neurons, their most basic behavior is consistent across types (Kandel, 2000).

Furthermore, the types of neurons that exist in humans are also found in other animals (Ullman, 2001), facilitating experimentation that cannot be done on humans. Although single-electrode measurements in human subjects are not unheard of, they are restricted to cases in which a person's brain is being operated on for other reasons. The bulk of the information we have about how individual neurons work comes from animal studies (Kandel, 2000).

The same is true for higher functions that we share with other animals. Our understanding of the cortical networks that are involved in motor control or vision or hearing is greatly enhanced by studying those system in non-human brains (Gazzaniga, 2009). Even emergent functions – those that arise from the interaction of simpler

systems – can be studied in animal models and can offer some insight into how specific emergent properties might arise in certain parts of human brains. For behaviors that do not have direct physical analogs to be measured, such experimentation is less useful. For a cognitive behavior like language that is unique to humans at least in degree and scope, animal models cannot tell us much, so we must rely more heavily on measurements taken from humans. On the subject of phonemic acquisition, there is a line of research suggesting that chinchillas can develop categorical perception in certain phonemic contrasts, like humans do (Kuhl, 1975). If this were true, then chinchillas would provide a useful avenue of investigation for the neural correlates of certain phonemic distinctions. However, there are problems with this line of research that will be addressed in Chapter 4.

Brain imaging techniques have improved tremendously and have deepened our understanding of how human brains operate. Unfortunately, their resolution is too low to tell us anything beyond the average behavior of a large number of neurons. Magnetic Resonance Imaging (MRI) is a brain imaging process that relies on the magnetic properties of water. Water is a polar molecule, meaning it has different charges at different ends of the molecule, causing it to behave like a miniscule magnet. In MRI, the subject is inserted into the center of a super-cooled magnet capable of generating a very strong magnetic field. This field causes the water molecules to align with the external magnetic field. When the magnet is turned off, the water molecules return to random positions. The changes in the magnetic fields generated by groups of water molecules can be measured by the MRI system to give an indication of the water content of specific small volumes of the body. Water content correlates with density, which allows the

machine to create a detailed map of the part of the body it is focused on. Functional MRI (fMRI) is an adaptation of this process that measures not only relative quantities of water, but changes in water content over a short time periods. As a portion of the brain becomes more active, blood flow to that area is increased. With increased blood flow comes increased water content. In this way, fMRI is able to indirectly measure blood flow to an area of the brain, and researchers infer increased neural activity in that area. This allows experimenters to determine what areas of the brain are implicated in various cognitive tasks. One serious limitation of this process stems from the low resolution of an fMRI. Spatial resolution of fMRI is between 1 and 5 cubic millimeters, a cortical volume that can contain a few million neurons and tens of billions of synapses (Huettel, 2009). Clearly, this technique does not allow us to correlate higher cognitive behavior with the activity of small numbers of neurons. A cognitive neuroscience model based on fMRI data would have compartments consisting of millions of neurons, because no information is available at a smaller scale. Temporal resolution of fMRI is on the order of 1 to 6 second (ibid.). For the purposes of examining the neural correlates of language behavior, this low temporal resolution presents insurmountable difficulties. Times spans of interest in phonetic research are on the order of tens of milliseconds. Even syntactic research focuses on structures that can start and end within a second, and certainly within six. This limits the utility of fMRI in gathering experimental data to inform the construction of computational cognitive neuroscience models of language. Other brain imaging techniques, such as electroencephalogram (EEG) have a much better temporal resolution than fMRI, but considerable lower spatial resolution. Combinations of techniques can

offer slight improvements to spatial resolution, but not to the level that would dictate the construction of neural models.

From the perspective of Marr's three levels of analysis, computational cognitive neuroscience modeling suffers from a conflict between the Implementation and Algorithm levels, even though the nature of CCNS models drives an effort to satisfy both. Realism at both of those levels results in models that are computationally intractable. Simplifying the model at either the Implementation or Algorithm level introduces the risk of reducing overall model complexity to the extent that it cannot model the behavior of interest. A balanced approach is indicated, with both Algorithm and Implementation reduced only as far as necessary to achieve computational feasibility while pursuing the goal specified by the Computation level.

The fundamental problem in developing neutrally plausible computational models of cognition lies in the conceptual distance between cognitive behavior and individual neurons. Cognitive behavior necessarily involves the interaction of large enough numbers of neurons to make direct and accurate small-scale modeling impractical. But simplifying the model in terms of structure or function presents the risk of eliminating whatever interactions give rise to the behavior of interest, while ignoring the structural details of the natural system. This creates the problem that the model of the behavior is based on the modeler's theoretical understanding of the behavior, which might not align with the reality of how the natural system operates. Given that one purpose of modeling is to test and refine theoretical assumptions, developing a model based primarily on those theoretical assumptions restricts the falsifiability of the model.

Chapter 4 explores some ideas that could lead to more effective models of phonemic acquisition. The potential role of animal studies is examined in some detail, and the goal at the Computation level is questioned. Finally, a reexamination of Marr's three levels suggests a slightly different perspective on the evaluation of acquisition models.

CHAPTER 4

Proposals for a Model of Phonemic Acquisition

4.1 Model Complexity

Every model that is expected to achieve a goal needs an appropriate degree of total complexity, although that complexity can be distributed differently across varied parts of the model. No model should be as complex as the system it is intended to model. The most obvious reason is that the added complexity is unnecessary to the task, as evidenced by the existence of a simpler system that performs that same task. Another reason is that models are intended to give the modeler some insight into the operation of the system being modeled. It is easier to see patterns in and draw conclusions from a model that is relatively straightforward in its details. But focusing on the analyzability of the model to the extent that it is simplified beyond its ability to perform its function defeats the purpose of modeling. The total model complexity must be sufficient to complete the task. Model complexity can be divided into complexity of structure and complexity of process, as intimated in Chapter 3. Simplifying a model on one of these dimensions requires making it more complex on the other. Take as an example a binary branching tree with three nodes labeled “A”, “B”, and “C” (Fig. 4.1a). Imagine that the goal is to generate all six possible permutations of those three labels. With a very simple flat, binary branching structure, permutation of the labels requires a number of processes. There must be processes to insert a new node (Fig. 4.1b), move a label from its position to the new node (Fig. 4.1c), and then delete the old node (Fig. 4.1d).

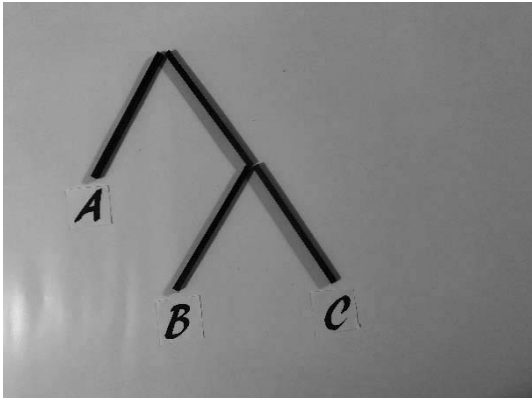


Fig. 4.1a – Binary branching tree with nodes A, B, and C.

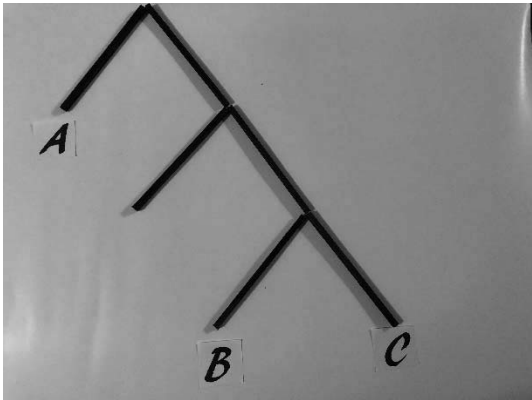


Fig. 4.1b – Inserting a node.

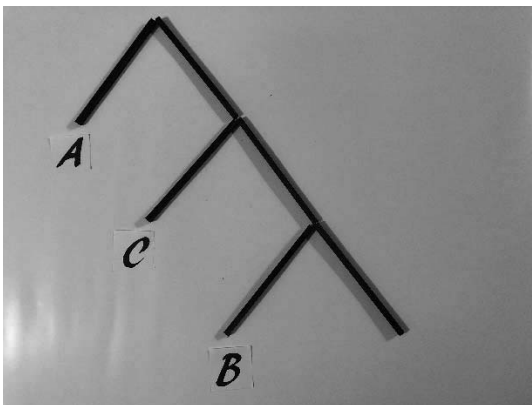


Fig. 4.1c – Moving C from one node to another.

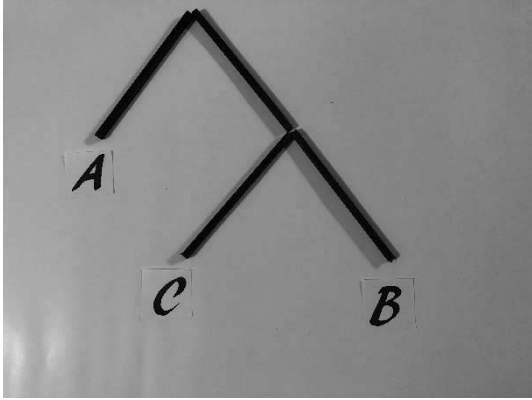


Fig. 4.1d – Deleting the empty node

There might need to be a process for determining how nodes can cross each other. Finally, there must be a governing process that tracks the movements of the labels to avoid duplicate permutations and ensure that the complete set is generated. (Compare this to the stopping problem discussed in Chapter 2 – the difficulty, in a clustering algorithm, of deciding when to stop forming new clusters.) The structurally simple model is paired with a fairly complex process, and the total complexity is sufficient to complete the task. What if, instead, we move some of the total complexity from the process to the structure? Consider a tetrahedron – a triangular-based pyramid – suspended from its top point. Depending from the remaining vertices are labels, “A”, “B”, and “C”. The sequence is to be read from left to right at the level of the labels. In one possible arrangement, one would see “A” hanging from the vertex that is close and to the left, “B” hanging from the vertex that is far and centered, and “C” hanging from the vertex that is close and to the right, for the sequence “ABC” (Fig. 4.2a).

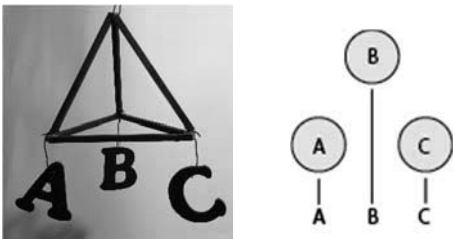


Fig. 4.2a – A tetrahedral permutation-generator in physical and schematic versions.

This is a more complex structure than the binary-branching tree, but generating the full set of permutations can be accomplished with a much simpler process. The only step that is needed is to rotate the tetrahedron 60 degrees in one direction. Rotating the above arrangement 60 degrees moves the front left “A” to the center back position, the front right “C” to the front left position, and the center back “B” to the front right position, giving a new sequence of “ACB” (Fig. 4.2b).

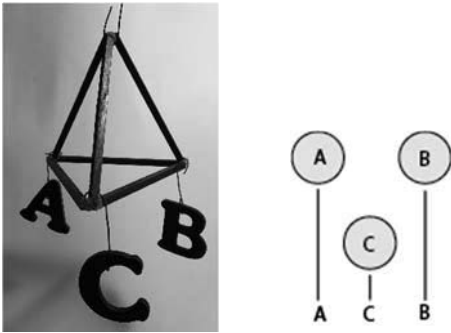


Fig. 4.2b – 60 degree rotation of ABC yields ACB.

Repeating the simple rotation generates all six permutations (Figs. 4.2c-4.2f).

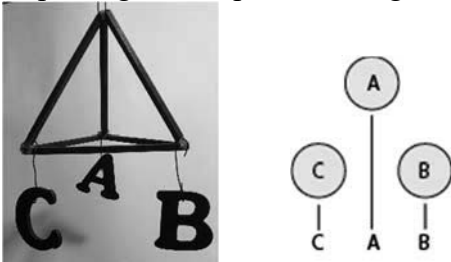


Fig. 4.2c – 60 degree rotation of ACB results in CAB.

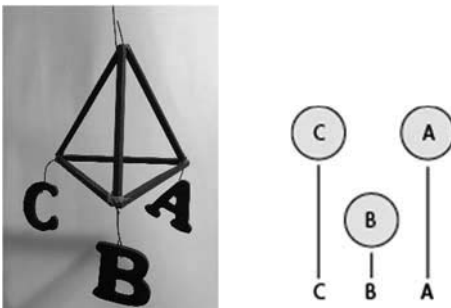


Fig. 4.2d – Next rotation gives CBA.

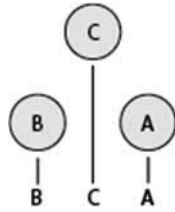
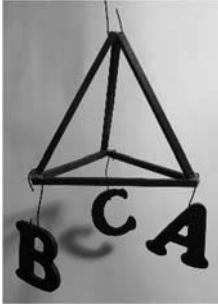


Fig. 4.2e – Then BCA.

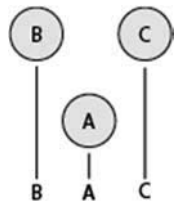


Fig. 4.2f – And, finally, BAC. One more rotation will return to ABC.

The concept of balancing complexity of structure and process provides a useful perspective when analyzing computational models. Note that structure and process are roughly analogous to the Implementation and Algorithm levels of Marr's style of analysis.

A question that must be answered in the process of designing any model is what degree of total complexity is necessary. This is intended as a philosophical question with no suggestion that complexity needs to be explicitly quantified. With a well-defined goal, one can, in principle, approach the question empirically from one end or the other. One could start with a simple model and incrementally make its parts more complex until it manages the desired goal. One could also start with a complex model and incrementally simplify its parts until it loses the ability to perform its task. In reality, practical constraints make both of these approaches untenable, and model details are

driven by theoretical considerations. The primary differences between the models discussed so far are based on differing theoretical concerns.

4.2 Quick Review

Chapters 2 and 3 examined both information theoretic and neural-based models in terms of Marr's three levels of analysis. To achieve the goal of developing a computational process that learns to categorize and label phonemes quickly and accurately, very well-established and successful models have been developed based on the information theoretic approach of Shannon (1948) and the style of neural network developed from Rosenblatt's (1958) work. These models reduce the complexity of the natural neural process sufficiently to remain computationally feasible while still maintaining enough complexity to produce acceptable results. However, for the very different goal of modeling the acquisition process that infants go through, rather than simply producing similar end results, a greater degree of similarity to the natural process is desirable. For this goal, existing information theory models fall short on the Implementation and Algorithm levels. They involve structures that are related to the structure of the target system only in the most abstract way. Their algorithms are also far removed from the functioning of a natural neural system.

Computational cognitive neuroscience models fare much better under analysis at the Implementation and Algorithm levels. They are designed with actual biological neural networks in mind, so they adopt simplifications of structure and process only to the extent necessary to allow for computational execution. The goal of these models is not to simply reproduce the output of the natural system. Instead, the goal is to reproduce, as faithfully as practical, the structure and functioning of the natural system

with success measured by how closely the behavior of the model matches that of the system on all levels. Whereas the information theoretic models focus on the Computation level of analysis and accept broad deviation on the Algorithm and Implementation levels, cognitive neuroscience models attempt to succeed on all three levels.

4.3 Modeling Goals

For the purposes of this dissertation, the goal of developing a model of phonemic acquisition in infants is to model the infant's acquisition process. Modeling the end result of the process of acquisition is a different goal, more in line with the models that are intended to provide some language-related functionality. Our goal is to model the process that the infant goes through. The first departure from the models previously discussed comes from a recognition that those models fail at the Computation level. Marr's first level of analysis is intended to define the goal, the input and output, the reason for modeling. The behavior of existing models of acquisition suggests that the goal is to convert an acoustic input into a phonemic label of some sort. Naturally, the acoustic input is intended to be analogous to what the infant hears. Determining the analog of the phonemic label requires some discussion. In toy models, one imagines the output being an actual label, such as [a] or [z], flashed on the screen. The label is clearly intended to give evidence of the model's success in a way that is easily interpreted by the researcher. The infant merely recognizes that two tokens belong to the same class. The infant is performing a categorization task, which implies the formation of categories, not the principled labeling or analysis of those categories. But is this a valid task to assign an infant? Do infants actually form phonemic categories? This may seem like a silly

question, since infants eventually learn to distinguish phonemes, but there is evidence that they learn to distinguish features during this same process. White (2008) showed that nineteenth-month-olds presented with mispronunciations of familiar object names showed sensitivity to the degree of mispronunciation. Degree of mispronunciation was measured in the number of features of the word-initial consonant that were changed – place alone; place and voicing; or place, voicing, and manner. It seems plausible that, while infants are learning to distinguish [f] from [g], they are also necessarily learning to distinguish [labiodental] from [velar], [-voice] from [+voice], and [+fricative] from [+stop]. The problem with computational models of phonemic acquisition is that they have the hypothetical infant perform only one of these tasks. Either the infant learns to transform acoustic information in a way that identifies phonemes, or the infant learns transformations to identify acoustic features and then learns to combine those features into bundles representing phonemes. I propose that both of these perspectives are flawed, in that they posit an unrealistically complex task. The task of learning to categorize phonemes can be broken down into a sequence of simpler tasks that are much more in line with what dynamic neural networks excel at.

4.3.1 Supervision

The process of converting an acoustic input into a label that can be right or wrong implies some sort of supervised learning. One could argue that no learning is wholly unsupervised, since one learns in the context of an environment that provides both the motivation for learning and the corrective mechanisms for learning expected categories and rejecting infelicitous possibilities. However, this implies too much knowledge on the part of the infant. Supervised learning requires a directed interaction between the learner

and the supervisor, whether it is an explicit instructor or simply the learning environment. The infant must know to react to corrective information from the environment in a way that spurs a reevaluation of the information the infant is trying to learn. This is a variant of the labeling problem in syntactic acquisition, in which the learner cannot know to apply a label to a category without an understanding of what that category is, but that understanding eliminates the need for the label (Landau, 1985; White, 2017). I propose that the infant learner initially has the much more straightforward task of simply modeling the acoustic input.

4.3.2 Revising the Goal

This concept is easier to understand when one abandons the notion that the learner acts on the input and considers that the learner initially simply interacts with the input. A set of connected neurons with some number of parameters including connection weights, activation thresholds, and spike rates can be analyzed as a multi-dimensional surface – a state space – with each point on the surface representing a possible state of the system. Because the elements of this system are interconnected – because there is feedback – any input applied to the system will propagate through the system and cause changes to the behavior of the elements. In any system that can be described with a set of differential equations, the dynamics of the system can run to infinity, reach an equilibrium, oscillate between two or more values, or behave chaotically. Because actual neurons are subject to physical constraints like real-time movement of ions, time from baseline to threshold, and timing of the post-spike refractory period, they are prevented from exceeding certain extremes of behavior. A neuron simply cannot fire faster than its maximum spike rate, regardless of what input it gets. The natural, physical limiting factors prevent a network

of neurons from running to infinity. This leaves equilibrium, oscillation, and chaos as the only possible behaviors. It is well understood that learning in neurons takes place through a facilitating effect in synapses that have recently fired (Kandel, 2000). In other words, connections that are used are strengthened, while connections that are not used are weakened and eventually culled. The state space – the multidimensional surface representing the states of the set of neurons – will find itself more frequently conforming to the shapes that correspond to frequent inputs. One could extend the state space by one dimension and, for each point in the state space, i.e., for each possible state, set the new dimension equal to the probability that the state occurs. The result would be a probability density map over the state space. Over time, as the system responds to the inputs it receives, propagates those signals through the network, and strengthens the participating connections, the probability density map over the state space will begin to correlate strongly with a probability density map over the input. This process, without supervision, without direction, without any goal beyond each neuron behaving as neurons behave in response to differential input, results in the system creating a model of the input. The configurations of the model may not be interpretable to the researcher, but the system has essentially developed a way to translate acoustic input into its “neural language”. At this point, the infant “knows” something about input categories in the sense that brain states vary predictably in response to different inputs. This is when the infant has enough information to begin acting as if categories exist, because there exist different brain states that are correlated with the categories of interest.

4.3.3 Secondary Goal

Once this initial process is completed, the model is free to engage in the supervised process of generating labels in response to inputs. From the model's perspective, that process is not "Learn to associate this label with this input." Instead, it is "When you find yourself in the state that results from receiving this input, produce this label." That is a much easier task, because it does not connect unknown labels with unknown inputs, but rather associates unknown labels with *known* states. One effect of this two-part process is that the modeler does not have to choose whether the model will learn phonemes or features. The model learns acoustic distributions and can be taught any relevant category. The model can be taught to produce the labels [m] and [n] as appropriate. It can also be taught that the states associated with those inputs can produce the label [+nasal]. We can't necessarily know if, when exposed to inputs [m] and [n], the model enters states that have meaningful similarities on some dimension that are effectively the neural correlate of the label [+nasal]. The fact that the network has learned to model acoustic space – more to the point, it has *become* a model of acoustic space – would suggest that this might be the case, but it is conceivable that the model would be nudged into entirely different states for [m] and [n] and would simply learn that both of those states suggest the label [+nasal]. This should be the case with allophones.

Allophones are acoustically distinct speech sounds that count as the same phoneme, i.e., they are never the sole segment distinction between words with different meaning. Alveolar [t], dental [t], aspirated [t], and glottalized [t] have significant acoustic differences, which places them at different locations in a multidimensional acoustic space. For a clustering algorithm intended to define or discover phonemic clusters, these four distinct groups of acoustic tokens present a problem. They will

appear as four separate clusters because of their acoustic differences but, as variants of the same phoneme, they should be assigned to the same cluster. If the model somehow decides to disregard the acoustic dimensions that correspond to “alveolar” and “dental”, it might lose the ability to distinguish other phoneme classes. In any case, the contextual disregard for a particular acoustic dimension requires knowledge of the context – the phonemic class. If this information is made available to a clustering model, then it is essentially being told which clusters to find. The model being elaborated here sidesteps this problem in a simple way. It should recognize that the four allophones listed above are acoustically distinct tokens, given that it has essentially become a model of acoustic space. However, in the second round of training, the model would learn to produce the label [t] for each of those acoustic events. The state that it finds itself in after being exposed to an alveolar [t] input would cause it to produce the label [t]. The different state it finds itself in after being exposed to a dental [t] would also cause it to produce the label [t]. It recognizes the difference between allophones, because of their acoustic distinctness, but it produces the appropriate phonemic label, because the specifics of the acoustics have been decoupled from the labeling process. This departs from other models of phonemic acquisition, in which the system is expected to develop a single mathematical transformation from several acoustically distinct allophones to a single label. The proposed model is much more satisfying in its ability to recognize the differences between allophones but still label them appropriately.

4.3.4 Overfitting

Overfitting occurs when a learning model is too precise in its treatment of a training dataset to the detriment of its performance on test data. Fig. 4.3 illustrates overfitting in the context of separating two sets of points.

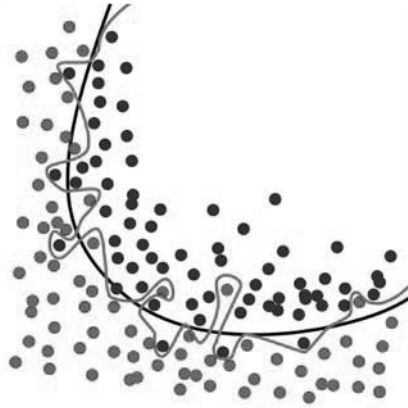


Fig. 4.3 – Overfitting (green line) vs more conservative curve (black line)

The red and blue dots overlap at the boundary, suggesting that the categories are not wholly distinct or that there is a missing dimension that could disambiguate the tokens near the boundary. The black curve separates the blue and red dots with a small degree of error – some dots end up in the wrong group. The green line reflects an attempt to separate the blue and red dots without error. It meanders through the boundary area with the sole goal of having all the blue dots on one side and the red on the other. This scheme can cause problems in dealing with later data. Assuming this training dataset is representative of the complete set of data, any randomly selected subset of sufficient size will have some overlap near the boundary. The black curve accepts that there will be some error, imposes a separation that reflects the general trend in the data, and probably produces the same degree of error in any subset of the data. The green line tries to reduce the error to zero but fails to accommodate for the fact that other subsets of data will have different overlap at the boundary, rendering the green line at least as error prone as the black one, but with an unjustifiably complex shape. As Chicco (2017, p. 17) explains:

“Overfitting happens as a result of the statistical model having to solve two problems. During training, it has to minimize its performance error. But during testing, it has to maximize its skills to make correct predictions on unseen data. This ‘double goal’ might lead the model to *memorize* the training dataset, instead of *learning* its data trend, which should be its main task.”

Because the proposed model’s first task is not to identify clusters, but to learn to model the input, there is no opportunity for overfitting, because there is no fitting to begin with. Any ambiguity that exists in the input will be reflected in the model of the input. In the second stage, when the model is learning to generate labels in response to its internal states, there will be some degree of error, reflecting the noise in the input.

4.4 Empirical Evidence

In determining the structure of a model, it would be useful to look to animal models for neural correlates of speech perception. As mentioned previously, the question of whether non-human animals use anything like language presents problems for the use of animal models. Nevertheless, chinchillas and monkeys have been shown to be sensitive to certain speech contrasts, so there is some potential for animal brain studies to provide useful information. One of the strongest areas of research in this vein involves voice onset time and categorical perception.

4.4.1 Categorical Perception

Let us define “categorical perception” as the inability to distinguish tokens that vary along a continuum if the tokens fall on the same side of some perceptual midpoint on that continuum, coupled with the ability to distinguish the tokens if they fall on opposite sides of that perceptual midpoint. Let us further define “continuous perception”

as the ability to distinguish two tokens that vary sufficiently on the measured dimension, regardless of their position on the continuum.

One classic example is VOT – voice onset time – as the primary distinguishing characteristic between voiced and voiceless plosives followed by vowels. For example, in English, the bilabial plosives /b/ and /p/ are said to differ in the voicing feature, with /b/ considered voiced and /p/ considered voiceless. Since the production of each of these phonemes consists of blocking airflow by pressing the lips together, allowing air pressure to build up behind the lips, and releasing the pent up pressure in a burst, there is no direct involvement of the vocal folds, voicing does not come into play in this part of the articulatory gesture. English phonotactics do not allow a voiceless phoneme to immediately follow a syllable-initial bilabial plosive. Since speech consists not of individually and discretely articulated phonemes, but rather of a continuous stream of gestures altering the acoustic signal, the time between the release of the bilabial closure of /b/ or /p/ and the initiation of voicing of the next phoneme provides a distinguishing characteristic between /b/ and /p/. This characteristic is known as voice onset time (VOT) and is measured in milliseconds. The syllable /ba/ will typically have a space of 0 to 10 milliseconds between the release of the bilabial closure (at the end of /b/) and the onset of voicing in the vowel, /a/ -- a VOT of between 0ms and 10ms. For a /pa/ syllable, more time is left between the two events, yielding a VOT in the neighborhood of 80-90ms.

Since the physiological mechanisms underlying the release of the bilabial closure and the initiation of voicing are independent of each other, and it is possible to produce VOTs of 0ms and of 90ms, it is clearly possible to produce VOTs between these

extremes. It is also possible to artificially alter tokens to have a desired VOT, by inserting or deleting space between the release of the stop and the onset of timing. If, for the sake of argument, we define the canonical /ba/ syllable as having a VOT of 10ms and the canonical /pa/ syllable as having a VOT of 90ms, a question arises regarding the identification of non-canonical tokens. If an adult speaker of English hears a token with a VOT of 30ms, will she perceive it as a slightly anomalous production of canonical /ba/ or will she subconsciously disregard the slightly longer VOT and perceive the token as a canonical /ba/, even though it is imperfect? One can easily grasp the benefit of having a speech perception system that suppresses conscious awareness of small deviations from the norm, relying instead on recognizing productions that are “close enough”. There is significant evidence suggesting that, at least on the dimension of VOT, English speakers have categorical perception, i.e., they perceive /ba/ or /pa/, even when exposed to intermediate tokens that are, at best, mediocre versions of canonical /ba/ or /pa/. This categorical perception was demonstrated in an experiment (Lieberman, 1967) using a discrimination task. Subjects heard pairs of tokens, differing in VOT by 10ms. So, one pair might have VOTs of 10ms and 20ms, while another had VOTs of 70ms and 80ms. The subjects did not know that the tokens varied in VOT and were simply assigned the task of saying whether the two tokens were the same or different. There is a sharp increase in the number of subjects who could distinguish the tokens whose VOTs straddled a perceptual transition point of 40ms. This result shows that English speakers perceive measurably distinct tokens as identical, as long as they sit on the same side of a continuum.

A different experimental design has also been used to demonstrate categorical perception, but flaws in the design render the results inconclusive. This experiment design relies on an identification task. The subject is exposed to tokens of the type described above, with VOTs ranging from 0ms to 90ms. The task is to press one key or lever if the token is perceived as /ba/ and another if it is perceived as /pa/. The percentage of tokens identified as /pa/ as a function of VOT yields a familiar sigmoid curve. The tokens with lower VOT values are identified as /pa/ virtually none of the time, while the tokens with higher VOT values are identified as /pa/ virtually all of the time. The transition is at around 40ms and is relatively sharp. It is best understood as a step function with some degree of variability around the transition point, yielding a sigmoid graph. This result is interpreted as demonstrating categorical perception in humans. Although humans do show categorical perception, as demonstrated by the discrimination experiment, this identification task neither supports nor refutes the existence of categorical perception.

The design of this experiment contains a fundamental flaw that renders its results uninterpretable. The hypothesis is that the subjects have categorical perception. The test seems designed to confirm this hypothesis. However, experiments should be designed to confirm the null hypothesis – the logical statement that is necessarily false in every case where the hypothesis is true. In this case, the null hypothesis is that the subjects have continuous perception, i.e., that they can distinguish tokens with different VOT, with some degree of precision. Imagine a subject that is instructed to press the left lever whenever it hears a canonical /ba/ token with a VOT of 0ms and the right lever whenever it hears a canonical /pa/ token with a VOT of 90ms. Imagine further that the subject is

perfectly capable of distinguishing tokens with VOTs of 0ms, 10ms, 20ms, and so forth, all the way up to 90ms. If, during the experiment, the subject hears a token with a VOT of 30ms, what should the subject do? The token is recognized as somewhat anomalous, but there are only two lever choices. Presumably, the subject would select the lever corresponding to the canonical token that is closest to the anomalous token. In this way, all tokens that seem closer to /ba/ will be identified as /ba/, and all tokens that seem closer to /pa/ will be identified as /pa/, simply because there is no alternative. The result for a subject with continuous perception should look exactly the same as the result for a subject with categorical perception – all /ba/ with a fairly sharp transition at the perceptual midpoint, followed by all /pa/.

This type of experiment is useful in determining the perceptual midpoint, which is a factor of interests to compare between humans and other animals. Kuhl (1978) tested chinchillas' response to VOT differences with a forced choice experiment, and wrongly concluded that chinchillas have categorical perception on this measure. The better conclusion is that chinchillas seem to perceive the midpoint between /ba/ and /pa/ at around 40ms, just like humans.

4.4.2 Other Animal Experiments

Other animal experiments demonstrate that non-human animals can discriminate a variety of speech sound contrasts. Kuhl and Padden (1982, 1983) used a discrimination task to test macaques, a species of Old World monkeys, for categorical perception. They appear to show similar perception to humans, suggesting that the categorical perception of VOT might stem from natural limitations in the response of the auditory system to simple acoustic differences.

However, although macaques may show categorical perception, it appears that monkeys are generally less sensitive to VOT differences than humans (Sinnot & Adams, 1987). This could be due to different physical constraints in the perceptual system of humans, as compared to monkeys. It could also be that humans and monkeys are working with very similar perceptual systems, but humans, as language users, undergo a language-specific refinement of the underlying system.

The complexity of the speech code is another issue of interest in comparing humans to other animals. Monkeys are considerably less sensitive than humans to differences in pure sine wave tones, but they perform similarly to humans in discriminating vowels (Sinnott and Kreiter, 1991). While the reason for this is unclear, it suggests that the speech signal, in interaction with the perceptual system, is more than the sum of its parts.

The preceding suggests that non-human animals can perceive many of the differences that are important to speech, and that they sometimes resemble humans very closely in their perceptual abilities. A related question is whether non-human animals can be trained to recognize phonemic contrasts. Hienz and Bradley (1988) trained baboons to distinguish five synthetic steady-state vowels (/a/, /æ/, /ɔ/, /u/, and /ε/), which they could eventually do with 95% to 100% accuracy. Sinnott (1989) trained monkeys to distinguish ten synthetic steady-state vowels. The monkeys performed similarly to humans on front vowels, but struggled comparatively with back vowels.

All of these experiments demonstrate that non-human animals, especially other primates, are able to discriminate some speech contrasts without training, and are also able to learn some other speech contrasts. It is important to recognize that every auditory

system is intended to discriminate sounds, and speech is made up of sounds. The ability of a monkey to discriminate sounds that humans use in speech is not necessarily the same as discriminating speech sounds. However, for the purposes of the proposed model, finding neural correlates of sound discrimination can help.

4.4.3 Neural Correlates

In a series of experiments on monkeys, small neural complexes have been implicated in the discrimination of sounds from speech. Steinschneider (1982) measured multiple unit activity (MUA), the average activity of a small set of neurons, in the primary auditory cortex and thalamocortical fibers of awake monkeys. Thalamocortical fibers are connections between the thalamus and the cortex, and have been implicated in the processing of sound. The subjects were exposed to synthetic syllables /ba/, /da/, and /ta/, as well as a click noise, and the activity of the two types of brain cells was recorded. Fig. 4.4 shows the differential responses to the various inputs.

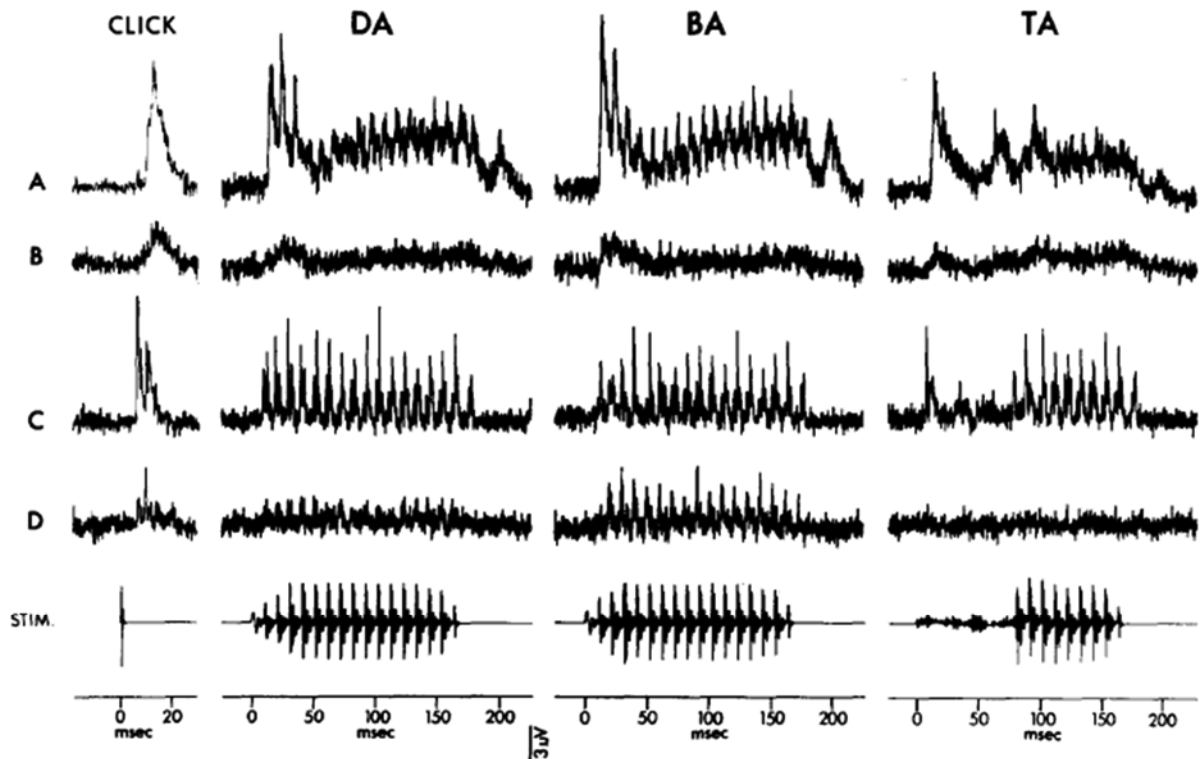


Fig. 8. MUA at 4 progressively deeper sites in an electrode pass through posterior auditory koniocortex illustrates response pattern differences between thalamocortical fibers and cortical cells (see text for description). Thalamocortical axon responses at site D reflect consonant place of articulation, as the early response to /ba/ is larger than those to /da/ and /ta/. Activity during the later acoustically identical portions of /a/ also differs across syllables, illustrating effects on following vowel responses by preceding consonants. Note double-peaked response at site C to clicks and to each pitch period of the syllables.

Fig. 4.4 – From Steinschneider (1982, p361).

Thalamocortical responses correlate with place of articulation, while the cortical response encodes other aspects of the speech signal. “Stimulus parameters that play a role in the differential perception of stop CV syllables are expressed in the temporal patterning of activity within the auditory radiations and cortex, sites necessary for speech decoding. Perceptually significant parameters that are reflected in the neural responses include fundamental frequency, VOT, place of articulations, and voiced formant transition duration” (ibid., p.362).

Steinschneider (1989) tested monkeys’ responses to the formants in /da/, /ba/, and /ta/ syllables. There was a difference in response to steady state formants and to formant

transitions indicating the place of articulation of the onset consonants. Steinschneider posits that the interaction between the different patterns in the auditory cortex and the thalamocortical cells could provide a stronger contrast in the response to complex stimuli that differ only in subtle ways. Later research in this sequence (Steinschneider, 1993, 1995) demonstrates specific neuron responses that correlate with the perceptual boundary in VOT between /ba/ and /da/ syllables, and with the consonant burst and the onset of voicing. Steinschneider (1999), this time working with human epilepsy patients, suggests that the categorical perception of stop consonants might be driven by temporal processing limitations within the auditory cortex. This is reinforced by experiments on monkeys exposed to two-tone complexes with variable tone onset time (TOT). These are simply pairs of tones that vary in their temporal relationship to each other. The first tone stands in for the plosive burst in a VOT experiment, while the second tone represents the onset of voicing (Steinschneider, 2005). The signal recorded from the monkeys' brains fails to show a difference in the response to tokens with TOT smaller than 20ms.

4.5 Designing the Model

In order to design a model of the type that has been hinted at in this chapter, several steps are necessary. The basic model element should be something along the lines of a Rulkov map – a relatively simple construct that can exhibit much of the spiking behavior of actual neurons. In light of Steinschneider's demonstration that different types of neurons are implicated in auditory processing, the Rulkov maps should be adjusted to reflect the behavior of the type of cell they are mimicking – auditory cortex neurons or thalamocortical fibers. The number of cells accessed by the multiple unit activity (MUA) process used by Steinschneider should be duplicated in a network of

Rulkov maps, with their interconnectedness determined by what is known of the network structure of the auditory cortex and the relevant thalamus neurons. At this point, the model network can serve as a monkey in one of Steinschneider's experiments. With known stimuli and (roughly) known structure, one need only tweak the Rulkov map parameters until the network behavior is similar to that seen in monkey subjects. This process of refinement should continue, along with any necessary structural extensions to the model, until it accommodates all relevant experimental data from direct recordings of neural responses to speech sounds.

The cited research focuses primarily on VOT in categorical perception, but similar experiments could, in principle, be performed for any acoustic feature that distinguishes speech sounds. Eventually, a single model should exhibit all the behavior discovered from animal studies.

Further animal studies can illuminate the next step. Chinchillas or monkeys should be trained to distinguish those speech contrasts that do not appear to be solely byproducts of perceptual limitations. The differences in neural behavior before and after training should give an indication of the way biological neural networks encode learned speech differences. This should, of course, inform an extension of the structure of the model, including a scheme for updating model parameters.

4.6 Marr Revisited

In terms of Marr's three levels of analysis, this type of model satisfies the Implementation level as well as a computational cognitive neuroscience model, because it is that type of model. It is also identical to that type of model at the lower end of the Algorithm level. The neurons, or their analogs, behave exactly as they would in any

other type of dynamic model. The model diverges from other dynamic cognitive models at the higher end of the Algorithm level. This is because of the interaction between the Computation and Algorithm levels and the change to the Computation level. The change to the Computation level involves limiting the model's task to what infants are logically capable of doing, and this has an influence on the operation at the Algorithm level. The perspective adopted in some other models of phonemic acquisition focuses too much on the knowledge of the researcher or the expected knowledge of the adult language user, and not enough on the limitations of the infant. One alternative resolution to the excess focus on the researcher is to claim that the infant has access to the relevant knowledge of the researcher. This is essentially a nativist position that claims that the infant is born with an innate understanding of the categories to be formed, and is only required to find boundaries and apply labels. This is an unsatisfying approach, because it answers the question of why things work the way they do by simply asserting that they do.

One aspect of Marr's three levels to bear in mind during any analysis of a model is that they are not levels of the model; they are levels of analysis (Bechtel, 2015). They do not represent a requirement that models have discrete Implementation and Algorithm levels. Instead, Marr's three levels offer a systematic way of looking at a model. While the three levels offer distinct perspectives on a model, they are also interrelated (*ibid.*, p. 320). As illustrated at the beginning of this chapter, model complexity can be distributed across model structure and model process in various ways. In biological neural networks, while it is easy to talk about the physical structure of a neuron as a set of facts distinct from the generation and transmission of action potentials, the reality is that these sets of facts are intimately interrelated. A strict separation between structure and process would

suggest that the existence of an ion is intrinsically distinct from the fact that the neuron has a charge. At this low level, the melding of structure and process is clear, but it persists at higher levels. In a biological neural network, as well as Rulkov-style models of these networks, the behavior of the set of neurons is determined by their individual and collective structure which, in turn is altered by their behavior. This circularity can break down the distinctions between the Implementation and Algorithm levels. Just like linguistic models should model processes rather than the researcher's theories about the processes, Marr's three levels should be used to guide the analysis of the model, but model itself should be based on knowledge of the actual system being modeled.

4.7 Summary

This chapter has offered some sources of empirical evidence to inform neural models of acquisition, and suggested some reassessment of the goals of such models. The fifth and final chapter will briefly review the ideas presented in this dissertation.

CHAPTER 5

Conclusion

5.1 Overview

This dissertation has reviewed two very different classes of models that can be used to illustrate linguistic behavior, especially acquisition. One class has its roots in Information Theory (Shannon, 1948) and is further divided into those inspired by mathematical functions intended to lead to quick and accurate results, and those inspired by the behavior of neurons, but not their physics (Rosenblatt, 1958). This Information Theory approach focuses on the flow of relevant information through the system and its ultimate application to a categorization decision (Beer and Williams, 2015). The other class is intended to mimic the physical behavior of neurons, with some necessary degree of simplification (Ashby, 2011), and inspires an analysis of geometrical and temporal relationships underlying model behavior (Beer and Williams, 2015).

An analysis of these different model types under Marr's (1982) three levels suggests that the Information Theory models do not fare well at the Implementation of Algorithm levels, although the perceptron types do somewhat better with Implementation. It is argued that all of these models fail under the Computation level, because they identify a mistaken goal of mapping acoustic input to labels. Infants do learn to do this, but the process is broken down into simpler steps. I propose that models should adopt the biological plausibility of computational cognitive neuroscience models, but adapt the training paradigm to reflect the realities of what infants are logically capable of doing.

5.2 Future Directions

Clearly, any discussion of model adjustments or innovation should be followed by the construction and testing of the proposed model. Selection of a computational cognitive neuroscience model on which to base the new model will itself be the subject of much exploration and empirical testing. The Rulkov-style maps discussed in Chapter 3 offer a promising starting point. The animal research of Steinschneider and others presented in Chapter 4 offers a practical low-level target for the development and refinement of a neural model. This notion of mimicking the actual system at the lowest practical levels can be applied at other stages.

Given the thesis that a model of phonemic acquisition should closely follow the neural behavior of an infant, it would be prudent to revisit the acoustic processing that the input undergoes before it is fed to a model. The methods of preparing acoustic data for presentation to a model are well-established, but that does not mean they should not be changed. We certainly know enough about the function of the inner ear to model it more directly than the mathematical approximations we currently use.

Eventually, an acquisition model must express the fact that phonemic acquisition, like acquisition of any aspect of language, does not occur in a vacuum. Moulin-Frier (2014) describes the role of intrinsic motivation or curiosity in vocal development. The interaction of production and perception or motor and sensory modalities has an impact on the development of phonemic categories (Oudeyer, 2002, 2005). The types of models considered here could eventually be extended to include the influence of production, but that is clearly a task for later. These models would be hard pressed to include curiosity as a parameter, but it should be remembered that every process relies in part on extrinsic information and a model will experience some error in the absence of that information.

An incremental approach is warranted, perhaps including well-studied sensory modalities, like vision. For example, a model that can learn the phonemes of color words while getting regular visual input that corresponds to the audio input would begin to tie together the supervisory aspects of non-linguistic data over linguistic cognitive tasks.

In the end, the hope is that the ideas presented here will inform more valid models that will help us develop our understanding of the near miracle of language acquisition.

REFERENCES

- Alderete, J. and Tupper P. (2017). Connectionist approaches to generative phonology. In *The Routledge Handbook of Phonological Theory*. Routledge, NY, NY.
- Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *The Journal of the Acoustical Society of America*, 106(4), 2031-2039.
- Anderson, B. (2014). *Computational neuroscience and cognitive modelling: a student's introduction to methods and procedures*. London: Sage.
- Andres-Barquin, P. J. (2001). Ramón y Cajal: a century after the publication of his masterpiece. *Endeavour*, 25(1), 13-17.
- Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research. In: *Advanced methods in marketing research*. Blackwell. Oxford. 160-169.
- Ashby, F. G. Computational Cognitive Neuroscience. In *New Handbook of Mathematical Psychology, Volume 2*. New York: Cambridge University Press.
- Ashby, F.G. and Helie, S. (2011). A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition. *Journal of Mathematical Psychology* 55(4). 273-289.
- Baldi, P., & Hatfield, G. W. (2011). *DNA microarrays and gene expression: from experiments to data analysis and modeling*. Cambridge University Press.
- Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, 59, 617-645.
- Bechtel, W. and Richardson, R. (1993). *Discovering Complexity*. Princeton University Press. Princeton, NJ.
- Bechtel, W. and Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science* 7. 312-322.
- Becker, S. and Daw, N.D. (2009) Computational cognitive neuroscience. *Brain Research* 1299. 1-2.
- Beer, R. and Williams, P. (2015). Information processing and dynamics in minimally cognitive agents. *Cognitive Science* 39. 1-38.
- Békésy, G. V. (1947). The variation of phase along the basilar membrane with sinusoidal vibrations. *The Journal of the Acoustical Society of America*, 19(3), 452-460.

- Bickle, J. (2015). Marr and reductionism. *Topics in Cognitive Science* 7. 299-311.
- Bock, H. (2008). Origins and extensions of the k-means algorithm in cluster analysis. *Journal Electronique d'Histoire des Probabilités et de la Statistique Electronique Journal for History of Probability and Statistics*, 4(2).
- Boersma, P., Benders, T. & Seinhorst, K. (2013). Neural network models for phonology and phonetics. Manuscript, University of Amsterdam. [<http://www.fon.hum.uva.nl/paul/papers/BoeBenSei37.pdf>]
- Bower JM (1991) Relations between the dynamical properties of single cells and their networks in piriform (olfactory) cortex. In: McKenna T, Davis J, Zornetzer S (eds) *Single neuron computation*. San Diego: Academic, pp 437–462.
- Bower, J. M. (2013). *20 years of computational neuroscience*. New York: Springer.
- Brette, R. (2015). What is the most realistic single-compartment model of spike initiation?. *PLoS computational biology*, 11(4), e1004114.
- Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J. M., ... & Zirpe, M. (2007). Simulation of networks of spiking neurons: a review of tools and strategies. *Journal of computational neuroscience*, 23(3), 349-398.
- Brown, K., Miller, J. (2013). *The Cambridge Dictionary of Linguistics*. Cambridge University Press.
- Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature* 10. 186-198.
- Cao, R., Pastukhov, A., Mattia, M., and Braun, J. (2016). Collective activity of many bistable assemblies reproduces characteristic dynamics of multistable perception. *The Journal of Neuroscience* 36(26). 6957-6972.
- Chen, G. (2016). A gentle tutorial of recurrent neural network with error backpropagation. *arXiv preprint arXiv:1610.02583*
- Cheour, M., Ceponiene, R., Lehtokoski, A., Luuk, A., Allik, J., and Naatanen, R. (1998). Development of language-specific phoneme representations in the infant brain. *Nature Neuroscience* 1(5). 351-353
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData mining*, 10(1), 35.
- Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.
- Chomsky, N. (1959). *Reviews: Verbal behavior by B. F. Skinner. Language* 35(1). 26-58.

- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. New York: Harper & Row.
- Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science*, 242(4879), 741-745.
- Ciarelli, P., Oliveira, E. and Salles, E. (2012). An incremental neural network with a reduced architecture. *Neural Networks* 35. 70-81.
- Clements, G. N. (1985). The geometry of phonological features. *Phonology*, 2(1), 225-252.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple correlation/regression analysis for the behavioral sciences*. UK: Taylor & Francis.
- Curtin, S. and Zamuner, T. (2014). Understanding the developing sound system: interactions between sounds and words. *WIREs Cognitive Science*. 5(5), 589-602.
- Davis, A. (2015). *The Interaction of Language Proficiency and Talker Variability in Learning*. Ph.D. thesis, University of Arizona.
- Dayan, P., & Abbott, L. F. (2003). Theoretical neuroscience: computational and mathematical modeling of neural systems. *Journal of Cognitive Neuroscience*, 15(1), 154-155.
- DeCaspar, A. and Spence, M. (1986). Prenatal maternal speech influences newborns' perception of speech sounds. *Infant Behavior and Development* 9. 133-150.
- Destexhe A, Rudolph M, Fellous J-M, Sejnowski TJ (2001) Fluctuating synaptic conductances recreate in vivo-like activity in neocortical neurons. *Neuroscience* 107:13–24
- Dhillon, I. S., Guan, Y., & Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 551-556). ACM.
- Dodge FA, Cooley JW (1973) Action potential of the motor neuron. *IBM J Res Dev* 17:219–229.
- Dudai, Y. and Evers, K. (2014). To simulate or not to simulate: What are the questions? *Neuron* 84. 254-261.
- Dugas, C., Bengio, Y., Bélisle, F., Nadeau, C., & Garcia, R. (2001). Incorporating second-order functional knowledge for better option pricing. In *Advances in neural information processing systems* (pp. 472-478).

Dupoux, E., Beraud-Sudreau, G., and Sagayama, S. (2011). Templatic features for modeling phoneme acquisition. *Proceedings of the 33rd Annual Cognitive Science Society*.

Eimas, P., Siqueland, E., Jusczyk, P., and Vigorito, J. (1971). Speech Perception in Infants. *Science* 171(3968). 303-306.

Eliasmith, C. and Kolbeck, C. (2015). Marr's attacks: On reductionism and vagueness. *Topics in Cognitive Science* 7 323-335.

Érdi, P. (2015). Teaching computational neuroscience. *Cognitive neurodynamics*, 9(5), 479-485.

Eriksson, E. (1971). Compartment models and reservoir theory. *Annual Review of Ecology and Systematics*, 2(1), 67-84.

Etymonline.com. (2018). *orthogonal* / Origin and meaning of orthogonal by Online Etymology Dictionary. [online] Available at: <https://www.etymonline.com/word/orthogonal> [Accessed 04 Aug. 2018].

Finger, Stanley (2000). "Chapter 13: Santiago Ramón y Cajal. From nerve nets to neuron doctrine". *Minds behind the brain: A history of the pioneers and their discoveries*. New York: Oxford University Press. pp. 197–216.

Fisher, J. (2017). *Representations of Spectral Differences Between Vowels in Tonotopic Regions of Auditory Cortex*. Ph.D. thesis, University of Arizona.

Fodor, J. (1997). Connectionism and the problem of systematicity (continued): why Smolemsky's solution still doesn't work. *Cognition* 62. 109-119.

Frank, M., (2011). Computational models of early language acquisition. *Author manuscript*. [<https://langcog.stanford.edu/papers/F-underreview-b.pdf>]

Friston, K. (2011). Functional and effective connectivity: A review. *Brain Connectivity* 1(1). 13-36.

Friston, K. (2012). The history and the future of the Bayesian brain. *Neuroimage* 62. 1230-1233.

Gabashvili IS, Sokolowski BH, Morton CC, Giersch AB (September 2007). "Ion channel gene expression in the inner ear". *Journal of the Association for Research in Otolaryngology*. 8 (3): 305–28.

Gazzaniga, M.S., Ivry, R.B. y Mangun, G.R. (2009). *Cognitive neuroscience: The biology of the mind*. New York: Norton.

Gervais, R. (2015). Mechanistic and non-mechanistic varieties of dynamical models in cognitive science: explanatory power, understanding, and the 'mere description' worry. *Synthese* 192. 43-66.

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review* 98(2). 254-267.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013* (pp. 6645-6649). IEEE

Griffiths, T., Vul, E., and Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Models of Cognition* 21(4). 263-268.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454), 746-774.

Harnad, S. (2003) Categorical Perception. *Encyclopedia of Cognitive Science*. Nature Publishing Group/Macmillan.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569-79.

Helie, S., Chakravarthy, S., & Moustafa, A. A. (2013). Exploring the cognitive and motor functions of the basal ganglia: an integrative review of computational cognitive neuroscience models. *Frontiers in computational neuroscience*, 7, 174.

Hendrickson, E. B., Edgerton, J. R., & Jaeger, D. (2008). A general method for creating realistic reduced compartmental models from electrophysiological traces. *BMC Neuroscience*, 9(S1), P83.

Hernandez, A. E., & Li, P. (2007). Age of acquisition: its neural and computational mechanisms. *Psychological bulletin*, 133(4), 638.

Hickok, G., Okada, K., Barr, W., Pa, J., Rogalsky, C., Donnelly, K., Barde, L., and Grant, A. (2008) Bilateral capacity for speech sound processing in auditory comprehension: Evidence from Wada procedures. *Brain & Language* 107. 179-184.

Hienz, R. D., & Brady, J. V. (1988). The acquisition of vowel discriminations by nonhuman primates. *The Journal of the Acoustical Society of America*, 84(1), 186-194.

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

Hochstein, e. (2016). One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese* 193. 1387-1407.

Hochstein, E. (2017). Why one model is never enough: a defense of explanatory holism. *Biological Philosophy* 32. 1105-1125.

Hockett, Charles F. (1968). A Note on Design Features. Indiana University Press.

Hodgkin, A.L. and Huxley, A.F. (1952). Propagation of electrical signals along giant nerve fibres. *Proceedings of the Royal Society of London. Series B, Biological Sciences* 140(899). 177-183.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4. 251-257.

Huettel, S. A.; Song, A. W.; McCarthy, G. (2009). *Functional Magnetic Resonance Imaging* (2 ed.), Massachusetts: Sinauer.

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, 97(1), 553-562.

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons?. *IEEE transactions on neural networks*, 15(5), 1063-1070.

Jacobs, B., & Schumann, J. (1992). Language acquisition and the neurosciences: Towards a more integrative perspective. *Applied Linguistics*, 13(3), 282-301.

Jain, A. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*. 31, 651-666.

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., & Seltzer, M. (2013, May). A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 8111-8115). IEEE.

Jessell, T., Siegelbaum, S., & Hudspeth, A. J. (2000). *Principles of neural science*. E. R. Kandel, J. H. Schwartz, & T. M. Jessell (Eds.). New York: McGraw-Hill.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014, November). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 675-678). ACM.

Johnson, M. (2017). Marr's levels and the minimalist program. *Psychonomic Bulletin Review* 24. 171-174.-

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing*. London: Pearson.

- Kandel, E., Schwartz, J., and Jessell, T. (2000). *Principles of Neural Science*. McGraw-Hill Medical. NY, NY.
- Kaplan, D. (2015). Moving parts: the natural alliance between dynamical and mechanistic modeling approaches. *Biological Philosophy* 30. 757-786.
- Kasabov, N. (2014). NeoCube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data. *Neural Networks* 52. 62-76.
- King, S. and Taylor, P. (2000). Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language* 14(4). 333-353.
- Knill, D. and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neuroscience* 27(12). 712-719.
- Kohonen, T. (1988). An introduction to neural computing. *Neural Networks* 1. 3-16.
- Kording, K. (2014) Bayesian statistics: relevant for the brain?. *Current Opinion in Neurobiology* 25. 130-133.
- Kuhl, P. (2010). Brain mechanisms in early language acquisition. *Neuron* 67. 713-727.
- Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, 63(3), 905-917.
- Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, 32(6), 542-550.
- Kuhl, P. K., & Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, 73(3), 1003-1010.
- Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044), 606-608.
- Kuhl, P., Ramirez, R., Bosseler, A., Lin, J., and Amada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceeding of the National Academy of Sciences of the United States of America*, 111(31). 11238-11245.
- Kuzmanovic, B., Bente, G., von Cramon, D. Y., Schilback, L., Tittgemeyer, M. and Vogely, K. (2012). Imaging first impressions: Distinct neural processing of verbal and non-verbal social information. *Neuroimage* 60. 179-188.

- Kwisthout, J., Wareham, T., and van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science* 35. 779-784.
- Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: II. Clustering systems. *The computer journal*, 10(3), 271-277.
- Landau, B., Gleitman, L. R., & Landau, B. (1985). *Language and experience: Evidence from the blind child* (Vol. 8). Harvard University Press.
- Le, Q. V. (2015). A tutorial on deep learning part 2: autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*, 1-20.
- Lee, H. and Kang, I.S. (1990). Neural algorithm for solving differential equations. *Journal of computational physics* 91. 110-131.
- Li, X., Wang, W., Xue, F., and Song Y. (2018). Computational modeling of spiking neural network with learning rules from STDP and intrinsic plasticity. *Physica A*(491). 716-728.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431.
- Lin, Y. (2005). *Learning Features and Segments from Waveforms: A Statistical Model of Early Phonological Acquisition*. Ph.D. thesis, University of California, Los Angeles.
- Lotto, A. (2000). Language acquisition as complex category formation. *Phonetica* 57(24). 189-196.
- Maass, W. (1996). Networks of spiking neurons: The third generation of neural network models. *Neural Networks* 10(9). 1659-1671.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1(14), 281-297.
- Macwhinney, B. (2010). Computational models of child language learning: an introduction. *Journal of Child Language* 37(3). 477-485.
- Mallot, H. A. (2013). *Computational Neuroscience: A First Course*, volume 2. Switzerland: Springer.
- Maloney, L. and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience* 26. 147-155.
- Mannel, C. and Friederici, A. (2013). Accentuate or repeat? Brain signature of developmental periods in infant word recognition. *Cortex* 49. 2788-2798.

Matsugu, M., Mori, K., Mitari, Y., & Kaneda, Y. (2003). Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16(5-6), 555-559.

Maye, J. (2002). The development of developmental speech perception research: The impact of Werker and Tees (1984). *Infant Behavior & Development* 25. 140-143.

Maye, J., Werker, J., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82. B101-B111.

McClelland, J., Rumelhart, D. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, 95(2), B15-B26.

Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys*, 1, 61-101.

Meeter, M., Jehee, J., & Murre, J. (2007). Neural models that convince: model hierarchies and other strategies to bridge the gap between behavior and the brain. *Philosophical Psychology*, 20, 749-772

Milkowski, M. (2016). Explanatory completeness and idealization in large brain simulations: a mechanistic perspective. *Synthese* 193. 1457-1478.

Miller, G. A. (2003). The cognitive revolution: a historical perspective. *Trends in cognitive sciences*, 7(3), 141-144.

Minsky, M., & Papert, S. (1972). *Research at the Laboratory in Vision, Language, and Other Problems of Intelligence: Progress Report*. Massachusetts Institute of Technology, AI Laboratory

Montavon, G., Braun, M., and Muller, K. (2011). Kernel analysis of deep networks. *Journal of Machine Learning Research* 12. 2563-2581.

Moore, B. C. (2004). *An introduction to the psychology of hearing*. Elsevier.

Moses, W. (2011). Fundamental limits of spatial resolution in PET. *Nucl Instrum Methods Phys Res A*. Aug 21; 648 Supplement 1: S236-S240

Moulin-Frier, C., Nguyen, S. M., & Oudeyer, P. Y. (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology*, 4, 1006.

Moulin-Frier, C., Nguyen, S.M. and Oudeyer, P.Y. (2014). Self-organization of early vocal development in infants and machines: the role of intrinsic motivation. *Frontiers in psychology*, 4, p.1006.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press.

New York Times (1958, July 8). New navy device learns by doing. *The New York Times*. p25.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2, 455–462.

Olazaran, M. (1996). A Sociological Study of the Official History of the Perceptrons Controversy. *Social Studies of Science*, 26(3), 611-659.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., and Contributors (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition. URL: <http://ccnbook.colorado.edu>

Oudeyer, P. Y. (2002, August). Phonemic coding might result from sensory-motor coupling dynamics. In *Proceedings of the seventh international conference on simulation of adaptive behavior on From animals to animats* (pp. 407-416). MIT Press.

Oudeyer, P. Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3), 435-449.

Oudeyer, P.Y. (2002). Phonemic coding might result from sensory-motor coupling dynamics. In *From animals to animats: Proceedings of the seventh international conference on simulation of adaptive behavior* (pp. 407-416). MIT Press.

Oudeyer, P.Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3), pp.435-449.

Oudeyer, P.Y. (2006). *Self-organization in the evolution of speech* (Vol. 6). Oxford University Press.

Paré D, Shink E, Gaudreau H, Destexhe A, Lang EJ (1998) Impact of spontaneous synaptic activity on the resting properties of cat neocortical neurons in vivo. *J Neurophysiol* 79:1450–1460.

Pater, J. (2017). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Author's manuscript*. [<https://people.umass.edu/pater/pater-perceptrons-and-syntactic-structures-at-60.pdf>].

Peebles, D. and Cooper, R. (2015). Thirty years after Marr's *Vision*: Levels of analysis in cognitive science. *Topics in Cognitive Science* 7. 187-190.

Perez T, Garcia GC, Eguiluz VM, Vicente R, Pipa G, et al. (2011) Effect of the Topology and Delayed Interactions in Neuronal Networks Synchronization. PLoS ONE 6(5): e19900. doi:10.1371/journal.pone.0019900

Poeppel, D., & Embick, D. (2017). Defining the relation between linguistics and neuroscience. In *Twenty-first century psycholinguistics* (pp. 103-118). Routledge.

Polomé, E. (1967). *Swahili Language Handbook*. Center for Applied Linguistics. Washington, DC.

R Hahnloser, R. Sarpeshkar, M A Mahowald, R. J. Douglas, H.S. Seung (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*. **405**. pp. 947–951.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257-286.

Rall W (1959) Branching dendritic trees and motoneuron membrane resistivity. *Exp Neurol*. 1:491–527.

Rall W (1962a) Electrophysiology of a dendritic neuron model. *Biophys J* 2:145–167.

Ramón y Cajal, S. (1899, 1904) *Textura del Sistema Nervioso del Hombre y de los Vertebrados*, Madrid, Spain: Moya.

Rasanen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication* 54. 975-997.

Rivera-Gaxiola, M., Silva-Pereyra, J., and Kuhl, P. (2005). Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants. *Developmental Science* 8(2). 162-172.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6). 386-408.

Rulkov, N. (2000). Regularization of synchronized chaotic bursts. *Physical Review Letters* 86(1). 183-186

Rulkov, N. (2004). Oscillations in Large-Scale Cortical Networks: Map-Based Model. *Journal of Computational Neuroscience* 17, 203–223.

Rumelhart, D., McClelland, J. (1986). *Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

- Saffran, J. (2003). Statistical language learning: mechanisms and constraints. *Current Directions in Psychological Science*. 12(4), 110-114.
- Sanborn, A., Griffiths, T., and Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review* 117(4). 144-1167.
- Saporta, S. and Contreras, H. (1962). *A phonological grammar of Spanish*. University of Washington Press. Seattle.
- Sardi, S., Vardi, R. Sheinin, A., Goldental, A. and Kanter, I. (2017). New types of experiments reveal that a neuron functions as multiple independent threshold units. *Nature, Scientific Reports* 7. 1-17.
- Schmidhuber, J. & Hochreiter, S. (1997). Long short-term memory. *Neural Computing* 9(8), 1735-1780.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61. 85-117.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Schreiner, C. and Winer, J. (2007). Auditory cortex mapmaking: Principles, projections, and plasticity. *Neuron* 56. 356-365.
- Shagrir, O. (2010). Marr on computational-level theories. *Philosophies of Science* 77(4). 477-500.
- Shaham, U., Cloninger, A. and Coifman, R. (2018). Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis* 44. 537-557.
- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal* 27. 379-423, 623-656.
- Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *The Journal of the Acoustical Society of America*, 106(2), 1078-1083.
- Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8), 888-905.
- Singleton, B., & Alkahby, H. (2011). The structural role of the basilar membrane in the hearing process. *International Journal of Evolution Equations*, 6(3), 309-318.

Sinnott, J. M. (1989). Detection and discrimination of synthetic English vowels by Old World monkeys (*Cercopithecus*, *Macaca*) and humans. *The Journal of the Acoustical Society of America*, 86(2), 557-565.

Sinnott, J. M., & Adams, F. S. (1987). Differences in human and monkey sensitivity to acoustic cues underlying voicing contrasts. *The Journal of the Acoustical Society of America*, 82(5), 1539-1547.

Sinnott, J. M., & Kreiter, N. A. (1991). Differential sensitivity to vowel continua in Old World monkeys (*Macaca*) and humans. *The Journal of the Acoustical Society of America*, 89(5), 2421-2429.

Skinner, B.F. (1957). Verbal behavior. New York: Appleton-Century-Crofts.

Smolensky, P. (1999). Grammar-based connectionist approaches to language. *Cognitive Science* 23(4). 589-613.

Smolensky, P. (2012). Symbolic functions from neural computation. *Philosophical Transactions of The Royal Society A*(370). 3543-3569.

Sonoda, S. and Murata, N. (2017). Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis* 43. 233-268.

Steinhaus, H. (1957). Sur la division des corps materiels en parties. *Bull Acad. Polon. Sci. IV. (C1.III)*, 801-804.

Steinschneider, Arezzo, & Vaughan. (1982). Speech evoked activity in the auditory radiations and cortex of the awake monkey. *Brain Research*, 252(2), 353-365.

Steinschneider, Arezzo, & Vaughan. (1990). Tonotopic features of speech-evoked activity in primate auditory cortex. *Brain Research*, 519(1), 158-168.

Steinschneider, M., Fishman, Y. I., & Arezzo, J. C. (2003). Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. *The Journal of the Acoustical Society of America*, 114(1), 307-321.

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan, H. G. (1994). Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Clinical Neurophysiology*, 92(1), 30-43.

Steinschneider, M., Schroeder, C. E., Arezzo, J. C., & Vaughan, H. G. (1995). Physiologic correlates of the voice onset time boundary in primary auditory cortex (A1) of the awake monkey: temporal response patterns. *Brain and language*, 48(3), 326-340.

Steinschneider, M., Volkov, I. O., Fishman, Y. I., Oya, H., Arezzo, J. C., & Howard III, M. A. (2004). Intracortical responses in human and monkey primary auditory

cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter. *Cerebral Cortex*, 15(2), 170-186.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard III, M. A. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of neurophysiology*, 82(5), 2346-2357.

Stemmer, N. (1973). An empiricist theory of language acquisition. The Hague: Mouton.

Stromswold, Karin. (2000). The cognitive neuroscience of language acquisition. *The New Cognitive Neurosciences*.

Sun, R. (2008) Introduction to computational cognitive modeling. *The Cambridge Handbook of Computational Psychology*. Cambridge University Press. 3-19.

Sussman, E., Steinschneider, M., Gumenyuk, V., Grushko, J., & Lawson, K. (2008). The maturation of human evoked brain potentials to sounds presented at different stimulus rates. *Hearing research*, 236(1-2), 61-79.

Syrett, K. (2016). Acquisition of comparative and degree constructions. *In The Oxford Handbook of Developmental Linguistics*. Oxford University Press.

ten Bosch, L., Boves, L., Can Hamme, H. and Moore, R. (2009). A computational model of language acquisition: the emergence of words. *Fundamenta Informaticae* 90. 229-249.

Thakur, A. K., Rescigno, A., & Schafer, D. E. (1972). On the stochastic theory of compartments: I. A single-compartment system. *The bulletin of mathematical biophysics*, 34(1), 53-63.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411-423.

Trappenberg, T. (2002). *Fundamentals of Computational Neuroscience*. Oxford University Press. Oxford, UK.

Tsoulos, I., Gavrilis, D., and Glavas, E. (2009). Solving differential equations with constructed neural networks. *Neurocomputing* 72. 2385-2391.

Tupper, P., Smolensky, P. and Cho, P. (2018). Discrete symbolic optimization and Boltzmann sampling by continuous neural dynamics: Gradient Symbolic Computation. *Author's preprint*.

Turner, B., Forstmann, B., Love, B., Palmeri, T., and Maanen, M. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology* 76. 65-79.

Van Hout, A. (2016). *Lexical and grammatical aspect*. In *The Oxford Handbook of Developmental Linguistics*. Oxford University Press.

Venturelli, A.N. (2016). A cautionary contribution to the philosophy of explanation in the cognitive neurosciences. *Minds & Machines* 26. 259-285.

Vihman, M., & Keren-Portnoy, T. (Eds.). (2013). *The Emergence of Phonology: Whole-word Approaches and Cross-linguistic Evidence*. Cambridge: Cambridge University Press.

Vincent, B. (2015). A tutorial on Bayesian models of perception. *Journal of Mathematical Psychology* 66. 103-114.

Werker, J., Tees, R. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology* 50. 509-535.

White, A. S., Hacquard, V., & Lidz, J. (2017). Main clause syntax and the labeling problem in syntactic bootstrapping. *Semantics in Acquisition*. *TiLAR*. Amsterdam: John Benjamins.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, 59(1), 114-132.

Willems, R. (2011). Re-appreciating the *why* of cognition: 35 years after Marr and Poggio. *Frontiers in Psychology* 2(244). 1-5.

Wintner S. (2010) Computational Models of Language Acquisition. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2010. Lecture Notes in Computer Science, vol 6008. Springer, Berlin, Heidelberg

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks* 84. 103-114.

Yu, Q., Tang, H., Tan, K., and Yu, H. (2014). A brain-inspired spiking neural network model with temporal encoding and learning. *Neurocomputing* 138. 3-13.

Yuste, R. (2016). From the neuron doctrine to neural networks. *Nature, Neuroscience* 16. 487-497.

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science* 78(2). 238-263.

Zednik, C. (2018). Computational cognitive neuroscience. In *The Routledge Handbook of the Computational Mind*. New York: Routledge.