A Framework for Modeling Non-Supervised Learning of Phonemes from Acoustic Input

Márk Jelasity

Master's thesis in general linguistics,
University of Szeged, Hungary
May 2, 2001

Abstract

This thesis describes a self organizing learning algorithm that learns phonemes from acoustic input. For generating acoustic input a highly simplified articulatory model is used and the phonemes are expressed in these articulatory terms. The simplification allows us to concentrate on specific issues like the effect of articulatory effort on learning and sound change without having to deal with rather complex realistic models that only hinder these phenomena and make extensive computational simulations impossible. The thesis is also concerned with methodological issues in general linguistics. The approach taken here is motivated by arguing against structurally inclined ways of looking at linguistic problems that still dominate the field.

Contents

1 Introduction

This thesis describes a self organizing learning algorithm that learns phonemes from acoustic input. Our approach makes use of the rapid development of computer hardware: the performance of computers has been exponentially increasing for decades. Furthermore the possibilities of existing and also exponentially growing computer networks are just beginning to be exploited. This makes it possible to refresh the methodology of linguistics trough the application of extensive computer simulations. We can capture features and test theories that were completely impossible before. This possibility can put several aspects in a new light and can help understand linguistic phenomena better.

On applying computers I do not mean natural language processing. My short term goal is not to solve the practical problem of e.g. machine translation or human-machine communication. The approaches there tend to be quite pragmatic and therefore probably unrealistic since human and machine capabilities differ remarkably. My intention is to apply computers in "pure linguistics". This would help understand language better but in the short run it would not necessarily help building talking machines. So the goal must be capturing some basic features that still allow us to build realistic models of a simplified but complete world (as opposed to building unrealistic models of isolated parts of the real world and then trying to figure out how these isolated models fit together).

In the present model for generating acoustic input a highly simplified articulatory model is used and the phonemes are expressed in these articulatory terms. This model has every component of a real situation but these components are simplified. The simplification allows us to concentrate on specific issues like the effect of articulatory effort on learning and sound change without having to deal with rather complex realistic models that only hinder these phenomena and make extensive computational simulations impossible. Unfortunately realistic sound processing is still out of the scope of computer simulations in spite of the rapid improvement of technology.

As the above ideas already suggest the thesis is also concerned with methodological issues in general linguistics. The approach taken here is motivated by arguing against structurally inclined ways of looking at linguistic problems that still dominate the field. I will suggest that viewing language as an essential part and product of the human mind and culture is of essential importance. Conducting research with this in mind is a great step towards a unified theory of human intelligence.

2 Background

In this chapter the methodology of the present approach is explained and an overview of the recent relevant literature is given.

2.1 Methodology

Even though the motivation of the researcher is normally not as relevant as the results themselves, I felt it necessary to elaborate on placing this work in the branch of scientific fields that somehow have to do with linguistics. The ideas are presented here without much argumentation since the point is to sketch my background only and not to convince the reader. It would need at least a book, or maybe even a whole life's work. Maybe it is not even necessary or possible to convince anyone about such questions and most likely even my own position will change over time.

Everyone trying to tackle problems in linguistics has to have a clear idea about the subject of linguistic investigation. Those who do not make a choice explicitly do it implicitly and I prefer to keep things explicit. Methodology has always been in the focus of attention in the field of linguistics. The most far-reaching idea in the XX. century was probably Saussure's *structural* approach (de Saussure, 1939). His basic idea is that when studying language we must differentiate between two essentially different subjects. The first is the *parole* which refers to an implementation of the second, the *langue*. Thus *langue* is abstract and therefore cannot be observed directly, and *parole* is everything that we say, hear, write or read. In fact this is just a variant of the frequently recurring idea of making a difference between important an unimportant features, between design and implementation, between higher and lower levels of description. Every natural science uses such level structures for organizing the accumulated knowledge. Even everyday knowledge is organized this way. Most people have no idea about engines, yet everyone can drive any (standard) car. It is true even for mathematics since it is perfectly possible to understand and apply theorems without understanding why they are true.

To summarize: the separation of *parole* and *langue* is not original in itself. In fact it is trivial that for a scientific investigation we have to ignore some aspects of reality if we do not want to get lost in its infinite richness. The real question lies in the choice of features to ignore. Saussure's standpoint is rather radical. We have to ignore everything that has to do with any other existing science: psychology, history, sociology, biology, physics, etc. The rest we can call linguistics. In Saussure's view this residue is an abstract *structure* in which the *relationship* and function of elements are the valid topics of research. Furthermore this structure can be studied *synchronously* without any reference to historical changes.

At the beginning of the XX. century many schools adopted a similar view and tried to describe structure based on strict methodological principles, trying to do *clean* linguistics. Chomsky's *generative* approach (Chomsky, 1979) grew out of the American branch of structuralists. It was an answer to the emerging problems of the structural methodology in connection with the *way* structuralists tried to describe structure, their main subject (see e.g. (Harris, 1951)). Without going into the details, Chomsky discovered that it would be more natural to give a method to *generate* the linguistic data collected instead of giving rules that can only filter incorrect forms. In other words, he discovered that the structure that can be observed within the actual linguistic data is not the direct target of the research. Instead, it is only a consequence of deeper and more general structures that are responsible for these surface structures, as he called them. He lifted the target of linguistic research to an even more abstract level.

This turn placed linguistics into the domain of cognitive science which was founded just before Chomsky's first influential works. By suggesting that people perform rather difficult symbolic computations the theory was fresh and interesting at that time. Unfortunately the relatively strictly defined methodological framework of concentrating only to strictly symbolic computational problems results in pushing the huge amount of work that aims to involve other aspects to "interfaces" like socio-, psycho-, computational linguistics, pragmatics, etc. Even though the generative theory has undergone quite radical modifications during the last decades the basic methodological approach did not change, only the implementational details. The desirable goal of unification of knowledge about human cognition certainly requires another approach.

This leads us to the motivation of the present work. My hypothesis is that language is a *skill*, in fact very similar to any other skills, like walking, playing chess, playing the piano, etc. This approach emphasizes that it is something that has to be *learned*, has to be *able to be learned* and its very essence is that it has to be able to be *used in certain situations to solve certain problems*, just like any other skill. It is also important that the actual organization of this skill should be described just like other skills. A theory of language would solve the more general problem of skills too, and a lot more. A lot more since it is the most distinguished skill that we have which has an effect on almost everything in our life. This approach does not mean that we have to focus on Saussure's *parole*. This is still *langue*, the only modification is that the abstraction has to contain the abstraction of the environment too, in which the language exists, namely the purpose of the language in the human life and the psychological limitations and capabilities which are *at least as common* between people as words and syntactic structures.

Such an approach will always have to face the rather difficult problem of having to explain how symbolic capabilities of people come to existence. I must say I do not know the answer to this question but I am still convinced that if lower level *non-symbolic* modeling of language is done right and the important aspects are modeled in sufficient detail—especially consciousness, and intentional models—it will emerge automatically. But I do not expect this to happen in the near future. The above argument also implies that trying to combine e.g. artificial neural networks (that contain a couple of dozens of nodes) with symbolic representation in order to explain symbolic processing does not seem to be a fruitful approach.

A very important aspect of a skill is its layered organization in which lower level processes provide input to higher level layers. This organization is very common even in perception in every modality (Sekuler and Blake, 1994). In this sense vision itself is already a skill: our lower level vision system "knows" what is worth to watch and how to adapt to certain situations. And in fact this skill even has elements that are learned during development based on the input from the environment. Difficult skills have of course more complex structure. Language can be thought of as a combination of many different skills that may even develop independently and that are combined only later. For example language certainly assumes the skill of communication, the motor control of articulation and perceptual categorization, and certain findings suggest that these are indeed developing independently. For example (Tomasello, 2000) assumes that modeling of other people's minds is a prerequisite for word learning. Other works suggest that phonological development does not depend on communication (see Section 2.2). The temporal order of the development of certain aspects of language also involves the emergence of layer structure within each major component.

The research presented here fits in the above framework. We try to model the very first phase of phoneme learning, when the continuous input is clustered into discrete categories forming the very first layer (after perception) needed for the acoustic aspects of language. This makes it possible to step to the next level of learning words. This involves extracting structure from the varying and continuous acoustic input. The framework sketched above also suggests that the learning process and the nature of the learned knowledge is similar in different domains which gives us hope that investigations in this direction could be useful in other domains as well.

2.2 Early Infant Phonological Development

Due to lack of space only a summary of some important aspects will be given mostly without reference to the rich literature of the field. The interested reader is kindly asked to refer to (Vihman, 1996, pp. 50–97) for a thorough introduction.

The perceptual capacities of infants are not the same for consonants and vowels. Even though this work is concerned with (an abstraction of) vowels a summary on consonants is also included.

2.2.1 Consonants

With this respect we have a relatively clear picture of the capabilities of infants. The two most important properties of infant perception here are categorical perception and perceptual constancy (which also occur in other modalities like color vision).

Categorical perception. It is widely accepted that infants are not sensitive to within-category differences but they are sensitive to between category differences. For example they can differentiate between /pa/ and /ba/ but they cannot between /pa/$_1$ and /pa/$_2$ even if the acoustic difference between the stimuli is the same for both pairs. Another very important fact is that this effect is observable with practically all known phonological contrasts, not only with those of the infant's ambient language. This means that "within-category" should be understood in this universal sense since a category in a given language is often composed of more "natural categories".

Perceptual constancy. Another striking property of infant perception is perceptual constancy. Infants can be trained to perform some action as a function of the linguistic quality of a stimulus. That means that they perceive the category of a sound independently of the person who produced it.

It is very likely that the above two properties are language independent and they are consequences of the more general properties of the auditory system, i.e. the physical structure of the ear and probably further neural information processing (feature extraction). It is supported by findings about different mammalian species which showed the same categorical perception and perceptual constancy effects. This results tell us that it is the sound structure of language that is adapted to the auditory system and not vice versa. Given that in automatic speech recognition the "normalization" of the rough acoustic input signal (i.e. the extraction of speaker and noise independent linguistic features) is still lacking a satisfying solution we can learn to respect the (innate) auditory system.

Moving on to learning we find that the infants gradually start to loose their sensitivity for the phonological contrasts that do not exist in the ambient language during their first year of life. With this respect their performance is getting close to adults who have problems with perceiving the difference between sounds that are not contrastive in their language (but belong to one of the phonemes) even if they are contrastive in another language. We can conclude that w.r.t. consonants the learning is more like forgetting, i.e. becoming insensitive to contrasts that are not present in the ambient language.

2.2.2 Vowels

The case of vowels is similar to the consonants, the difference is that here we cannot observe the perceptual categorical perception effect. Even adults can discriminate different samples taken from a vowel of their language quite easily. The perceptual constancy effect can be observed however.

The exact way of learning is less clear. One strong hypothesis is that categories are defined by prototypes and if a sample is close to this prototype then it is perceived as belonging to the given category. An influential work is (Kuhl, 1991) where the author suggests that the prototype vowels work like a "perceptual magnet" warping the perceptual space around them. Thus differentiation between samples with equal acoustic distance would be gradually harder as they are getting closer to a prototype. Adults show a stronger and infants a weaker effect. Monkeys show no effect. It has to be noted though that this theory is the subject of ongoing debates.

The general properties of the auditory system have another influence on vowels by determining the perceptual distance between the members of the continuous vowel space. The regularities of existing vowel inventories support this observation (Boë et al., 1995). All the vowel inventories are such that the vowels are distributed evenly over the vowel space with respect to this perceptual distance.

2.2.3 Babbling

So far only perception has been discussed. The infant's sound production capabilities turn out to be much less developed than their perception. By the end of the first year the infant already shows clear perceptual adaptation to the ambient language while the speech production still does not show language specific effects. According to neuropsychological results about general motor control (Whiting, 1984) and its application to phonological development (Kornev, 2000) the infant is not able to control the speech production organs voluntarily until up to 8 months of age. After that voluntary control is possible but no patterns (*engrams*) are created for storage in long term memory. The memorization and recall of motor-sequences begins even later. However the babbling of 12 month old infants already shows weak perceivable effects, i.e. adults are able to recognize the child's ambient language above chance (Engstrand et al., 1998). We can conclude that the motor functions (e.g. imitation) play no role in the early development of perceptual categories, at least not in the first 8 months.

2.3 Sound Change

Sound change can provide an excellent possibility for testing models of phonological development if they are sufficiently global to account for perception and production as well. It is possible to simply iterate the model, i.e. after the acquisition of a phoneme set it can be used as input to a new learning step. But care should be taken when applying this tool because sound change can happen for many different reasons and for testing purposes only the so called *organic* sound change is appropriate when no external factors like social layers, prestige, etc. are present. On the possible non-organic reasons of sound change see (Wardhaugh, 1992).

An example of connecting sound change and speech perception in one theory is (Ohala, 1993). The author suggests that sound change is a result of errors in speech perception. Errors may happen because speech sounds always vary to some extent either randomly or due to their context. Normally we can correct these errors using categorical perception or some other way of categorization. When there is a situation in a language that the possibility of misunderstanding is large, sooner or later the misunderstandings are built in the phonological structure of the language.

Independently of the reason of sound change there are different theories about the exact way the change takes place. We can identify two quite radically different views. The first is represented by Saussure and also by Labov (Labov, 1980). They basically say that the structure of a language (including phonological structure) is always a consistent system at every time step even if changes are going on. This means that e.g. if a phoneme changes then it changes in every word that contains it. However some results suggest that this is not the case and changes "diffuse" into the language. That is first only a few words contain the change and the lexical diffusion speeds up occupying almost all the words then slows down again before the process is completed (Bailey, 1973). The latter view gives us a rather different and much more complicated picture. It becomes virtually impossible to describe a grammar without reference to ongoing diffusions.

2.4 Related work

This section summarizes three recent works (actually all of them are PhD thesis). they offer a model of phonological development from rather different perspectives using different methodology. Although it is impossible to reconstruct these models here in sufficient detail this summary still helps the reader put the present work in a somewhat broader perspective.

2.4.1 Functional Phonology (Boersma, 1998)

In his thesis Boersma works within the framework of *optimality theory* (Prince and Smolensky, 1993). Genetically optimality theory belongs to the branch of generative linguistics. Its aim is to give a formal model that is able to generate correct surface forms but does not generate incorrect ones. The theory moves one step closer to psychological reality by giving a large emphasis to the learnability of the grammars (Boersma and Hayes, 2001). In the case of phonology it means that if we are given a set of underlying representations with the corresponding correct surface form then the grammar should be able to be learnt iteratively. Iterative learning means that we show the learner every example one by one. Each example results in a correction step in the grammar, and after iterating trough the complete teaching set a number of times the grammar should be correct.

The basic structure of a grammar in optimality theory is a set of constraints which are partially ordered. That means that for every pair c_1, c_2 of constraints either c_1 dominates (outranks) c_2 or the other way round or they can be freely ranked (free variation). The grammar also assumes a *generator* that generates candidate surface forms based on a given underlying form. Every surface form will violate a constraint. For every candidate we determine the maximal rank of the violated constraints and the candidate with the minimal maximal rank is the winner, i.e. the generated surface form. Table 2.1 gives a somewhat more visual version of the explanation above. Of course when making a decision we have to produce a linear ordering of the constraints. The table representation already assumes such a linear ordering.

/underlying form/	c_1	c_2	c_3	c_4	c_5	c_6
candidate 1		$*$	$*$		$*$	$*$
candidate 2	$*$	$*$	$*$	$*$		$*$
→ candidate 3			$*$	$*$	$*$	$*$

Table 2.1: The scheme of surface form generation in optimality theory. A $*$ in the column of a constraint indicates violation. Candidate 3 is the winner, i.e. the generated surface form.

In his thesis Boersma suggests a differentiation between perception and production. In the case of production the constraints are indeed related to speech production and there is another grammar that describes speech perception. We can only agree with such a decision as far as separation of perception and production is concerned but let us make a remark about a methodological issue.

As already mentioned is Section 2.1 generative schools work with models of sentence generation. In fact they don't care about the psychological reality of their models in that they examine sentences when suggesting the models and never people. But interestingly enough they care about the actual structure of the model. This latter concern is supported neither by practice nor by theory. A generative grammar is simply an algorithm in a mathematical sense. Turing machines and unrestricted grammars are equivalent. That means that for generation in fact every algorithm that generates sentences is equally acceptable from a theoretical point of view *if* we take the structuralist principles seriously and we are prepared to ignore psychological and sociological relevance (degrading psycho-, and sociolinguistics to a verification of *already existing models*).

For the above reasons Boersma's effort for creating a grammar along the lines of optimality theory for *perception* is a very good example of the "magnet effect" of a paradigm. While reading the rather absurd discussion of expressing perception using an infinite number of constraints like $*$CATEG(400) (which actually expresses that the perceived input does *not* belong to a category with a center with an F1 formant of 400Hz) we can get a fresh feeling about the similar absurdity of generation, the above mentioned philosophy of generative linguistics of putting the cart before the horse.

Boersma does not suggest a workable learning algorithm for the perception grammar, only for the production grammar. He assumes that learning the perception grammar precedes learning the production grammar (this is supported by empirical evidence, see Section 2.2) and refers to other works that suggest an actual solution.

2.4.2 Emergence of Vowel Systems (de Boer, 1999)

This thesis represents a rather different point of view. The author was Luc Steel's student and since Steels has rather original and ambitious ideas about language it is worth to devote a paragraph or two to him.

His background is artificial life, a field which seeks to understand the self-organizing behavior of complex systems. Steels applies the conceptual framework of this field—which was originally developed to model complex systems like societies, ant colonies, co-evolution of species, etc.—for modeling language. This framework among other things involves the frequent application of the so called multi-agent systems. The common property of these systems is that they are composed of entities (agents) that can make their own decisions, can interact with each other, and can learn. A simulation with a multi-agent system usually involves designing agents, i.e. determining the way they can interact, designing the algorithm that they use for planning their actions and the learning method. After this agents are put in a virtual world, and left alone. The results of the simulation are the (hopefully) emerging patterns, and organization.

Steels' idea is that language can be looked at as a pattern of organization that is *emerging* in such multi-agent systems (Steels, 1998). (Furthermore he thinks the nature of intelligence in general can be described this way (Steels, 1996)). The basic method he applies to develop e.g. emergent common vocabularies and phoneme systems is the so called *language game*. The behavior of agents consists of a series of games that they play with randomly chosen other agents. One possible game for example is the discrimination game where the goal is to differentiate a topic object (seen by both parties) form a set of other objects, the context (also seen by both parties). The initiator tries to create a description for the topic based on its available features and then communicates this description based on its vocabulary that assigns words to these features. The set of features and the vocabulary are both emergent, not *a priory* fixed. The other agent then tries to understand the description using its own vocabulary and based on the success of the game they both update their knowledge.

Of course the whole approach is extremely sloppy and uses many unrealistic assumptions as admitted by Steels. But the point is that in such a decentralized multi-agent framework we can see emerging vocabularies trough which agents can "understand" each other (or they develop a common phoneme repository, as we will see soon) is in itself worth a look. A model that tries to capture language (and intelligence) as a whole putting it into a unified framework is useful even if it is far from modeling the exact linguistic data. Especially in the present world of scattered, isolated and incompatible models of different specific phenomena.

Another important hypothesis is that according to Steels learning the language is not essentially different from the emergence of language. As he puts it

> . . . no separate mechanism for language acquisition is necessary because the mechanisms that explain the origin of language also explain how it is acquired be new agents entering the community. (Steels, 1996)

Thus the model of the emergence of language and language evolution is also a model of language acquisition. Keeping this in mind let us take a look at how de Boer applies Steels' approach for phoneme acquisition.

First of all we have to note that de Boer uses only standalone vowels. His decision was based on preliminary simulation with more difficult words but it proved to be too complex to get conclusive results. Nevertheless de Boer intended to keep pronunciation and recognition as realistic as possible to achieve realistic phoneme inventories. Thus the phonemes are internally represented as feature vectors with the features *position, height* and *rounding*. Note however that these were continuous valued features. A synthesizer was applied to produce formants based on the feature vector. Perception was implemented based on the already known vowels. When hearing a vowel the agent finds the vowel from its repository that is closest (when pronounced) according to a realistic distance measure.

The multi-agent framework is the same as described above, so we have to give only the kind of language game that the agents are playing. This game is the *imitation game*. This involves two agents, the *initiator* and the *imitator*. The initiator chooses one of its phonemes and pronounces it (with a little noise added). The imitator analyses the vowel and pronounces its closest match. The initiator then analyses the answer and if the closest match is the original vowel then the game is successful. Otherwise it is a failure.

In case of success the imitator moves its closest match closer to the input to help convergence in the population. In case of failure the agent adds a new vowel that matches the input closely if the closest match was used many times successfully in other games (since shifting the closest match would probably ruin performance in future games). Otherwise the bad vowel is shifted towards the input. In a simulation every agent starts with an empty inventory and fills it in trough playing games.

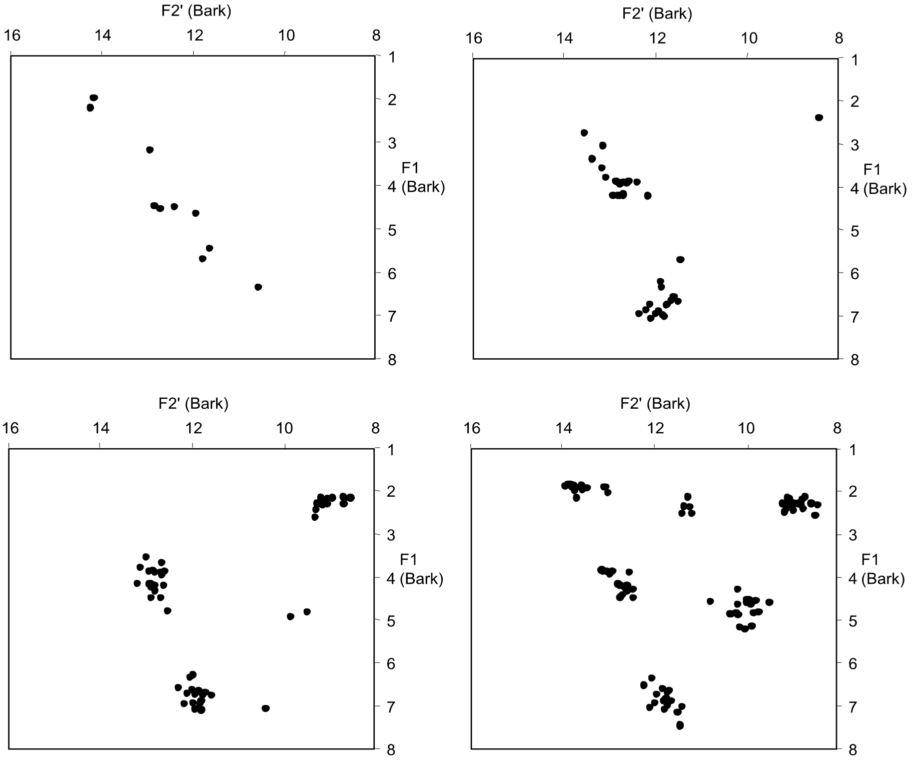

Figure 2.1: Vowel system in de Boer's simulation after 20, 200, 1000, 2000 games, 10% noise.

Finally let us show the result of de Boer's system with 20 agents in Figure 2.1. All vowels of all agents are plotted on top of each other. As a result of using a realistic distance function in the perceptual space the emerging vowel system is realistic without any central intervention.

2.4.3 Acquisition of Vowel Systems (Behnke, 1998)

Behnke's work represents a third completely different approach. He applies machine learning, in particular artificial neural networks for modeling phoneme acquisition.

The field of artificial neural networks is large and only loosely defined. Behnke's thesis applies self-organizing maps which form a subfield within artificial neural networks (Kohonen, 1989). The most important characteristic of a self-organizing map is that the learning happens completely without feedback. Thus the types of problems where these networks can be applied is limited. Normally they are used to learn some kind of *mapping* between the input space (the space where the learning examples come from) and the output space which is the neural network itself. The output space has structure too since the neurons are normally connected to each other which defines neighborhood relations and also a kind of distance. The input space must also have a structure in terms of a distance function.

The point is that this structure of the input space is unknown and we want to learn it. So we apply a self-organizing map, teach it the input space and then we look at the resulting map to learn something about the structure of the input space. Structure often means two things: clusters and topology. Clusters are spots in the input space where the elements are particularly dense, i.e. they form a sort of grain. In this context topology is the relation between the elements. For example three elements in the input space may form a specific kind of triangle. We can talk about *topology preserving maps* which guarantee that (if some assumptions about the input space hold) the topology of the resulting map will reflect the topology of the input space. Other maps form only clusters without preserving topology, these kind of methods are usually referred to as *vector quantization* methods.

Behnke developed a model for phoneme acquisition which contains a phoneme map as a filter for higher level processing. In his thesis he concentrated only on the phoneme map. His motivation was to develop a model which is not necessarily topology preserving but which has *local* representations. Locality means that once a representation is developed it is not disturbed by newly emerging representations. Behnke's greatest problem with Kohonen's algorithm is that there representations are not stable trough the learning process since it is topology preserving so new clusters result in the reorganization of the map. This way the map cannot act as a filter that provides the same output for already learned categories for higher level processing. To solve this he suggests a vector quantization method.

For testing only Dutch long vowels were used. The data was preprocessed to make sure that only continuous vowel segments are input to the map. According to the final result

- The system did not work if non-preprocessed input was presented to it.
- Overlapping vowels like /u/ and /o/ are mapped onto the same cluster.

As mentioned above, the interaction of the phoneme map with higher level processing like lexicalization was not examined.

2.5 Summary

After comments on my methodological point of view a brief overview of early phonological development in infants was given. We have seen that the mammalian auditory system is probably responsible for the infant's categorical perception of consonants, thus that language is adapted to the auditory system and not the other way round. During the first year, infants start to "forget" contrasts that are not present in their ambient language. Vowels do not show the categorical perception effect, instead they are probably organized according to a prototype structure where prototypes work as "perceptual magnets".

Sound change was mentioned emphasizing the complexity of using it for validating phoneme acquisition models. Nevertheless if done with care sound change data may offer a good possibility for such tests.

Three recent relevant works were also briefly discussed to illustrate the existing methodological diversity of tackling the same problem, phoneme acquisition. The first example used an optimality theory approach which in fact belongs to the more general category of generative approaches. It is based on a formal grammar that is build according to the optimality theory standards. We saw that probably this approach is the least satisfying as nor learning nor representation is especially effective. I have to admit that the author's main efforts are devoted to the generation grammar, not perception.

The second approach used a multi agent architecture and emphasized the role of emergence in both the origins of language and language acquisition. It suggests that acquisition and the origins of language should be explained on the same grounds. The third approach used machine leaning techniques, in particular self-organizing feature maps to model phoneme learning. They all attempted to model realistic sound features and as a result they had to make strong simplifying assumptions on the nature of input which turned out to be crutial in each case.

3 The Model

This chapter describes the model that was used to perform the simulations in Chapter 4. First the conceptual model is described in Section 3.2 then we discuss implementational decisions that are based on preliminary experiments with earlier versions of the model in Section 3.3.

3.1 Overview

The actual outline of the model (Figures 3.1 and 3.2) and its implementation are based on several assumptions that are enumerated here.

Simplification. It is unrealistic to expect a phoneme acquisition and lexicalization model that uses realistic sound data. Such a model has to account for not only phoneme learning but auditory perception, motor organization, the development of these and their effect on each other. Any mistake in any of these sub-models makes learning of phoneme representations unrealistic. Furthermore having too many variables in a model makes it very hard to separate the effects of the individual parameters. On top of that a correct model would provide us with a perfect speech recognizer and synthesizer. Yet research in machine learning and machine pattern recognition shows us that both tasks are extremely hard. And apparently the most successful pattern recognizers have little psychological relevance (although not much research is devoted to explore this possibility).

The above arguments lead us to the conclusion that it might be a luckier choice to introduce simplifications to smooth out the details to get a bird's eye view of the overall picture. Note that there is the another possibility of focusing on little isolated subfields and keeping as many details as possible. Our choice here is closer to the "nothing about everything" than to the "everything about nothing" approach. That means the sound is a continuous one dimensional curve. One can think of it as having only one feature with a continuous value.

Self-organization. For the development of phoneme learning the principle of self organization seems to be more appropriate then supervised learning. There are many reason for this. First of all perceptual capabilities of infants develop earlier than their motor capabilities so there is no way to reinforce successful encodings as they remain latent until the child is actually able to control the speech organs. Beside of this, it is well known that the early development of other perceptual modalities also happens in a self-organizing way, vision for example.

According to this the model is based on a map that adaptively creates clusters based on observed frequencies of its components. The map eventually converges to a phoneme set.

Layer structure. I assume that the layer of words is one layer above phonemic organization as words are basically fixed recurring sequences of phonemes. This means that we can not expect a lexicon until the phonemic layer is sufficiently stable for providing a basis for lexicon buildup. In other words first the phonemic layer is formed without storing things in long term memory then—when the stable phonemic layer makes it possible to detect stable sequences—lexicon building gradually begins. This layer structure is present everywhere around us from living organizations trough knowledge representations and science to culture. Thus this feature is not language specific either. Rather it is a property of complex self-organization.

In the model this situation is simplified by separating two phases: phoneme learning and lexicalization which happen disjunctly one after the other. Actually modeling the emergence of new layers is out of the scope of the present work.

Compression. My assumption is that storage in the short term memory involves quite serious preprocessing or encoding. It is not the rough input that is stored, since it would need too much resources. For an effective storage modeling of the input is necessary to compress it appropriately. This is a general feature in every modality for every kind of perception and I assume that in the case of language this is the precursor for phoneme formation. That would mean that—just like with the categorical perception of certain consonant features—it is our mammalian heritage that determines the properties and organization of language, phonemes in this case.

I also assume that in the motor side similar compression is taking place and the perceptual model of a sound sample is not perfectly dissimilar from the motor representation of the given sound (if the sound is a human sound of course) in terms of its organization and units.

3.2 The Conceptual Model

The model consists of two phases, the first is phoneme acquisition and the second is the building of the lexicon.

3.2.1 Phoneme acquisition

The schematic model of one step of phoneme acquisition is shown in Figure 3.1. The process of phoneme acquisition is repeating this step until the child's representations converge. The main motivation was to develop a framework with which we can discover the underlying structure of a continuous environment, sound in this case. Let us proceed according to the model component by component. Only the function of the components will be described, the details of the implementation are in Section 3.3.

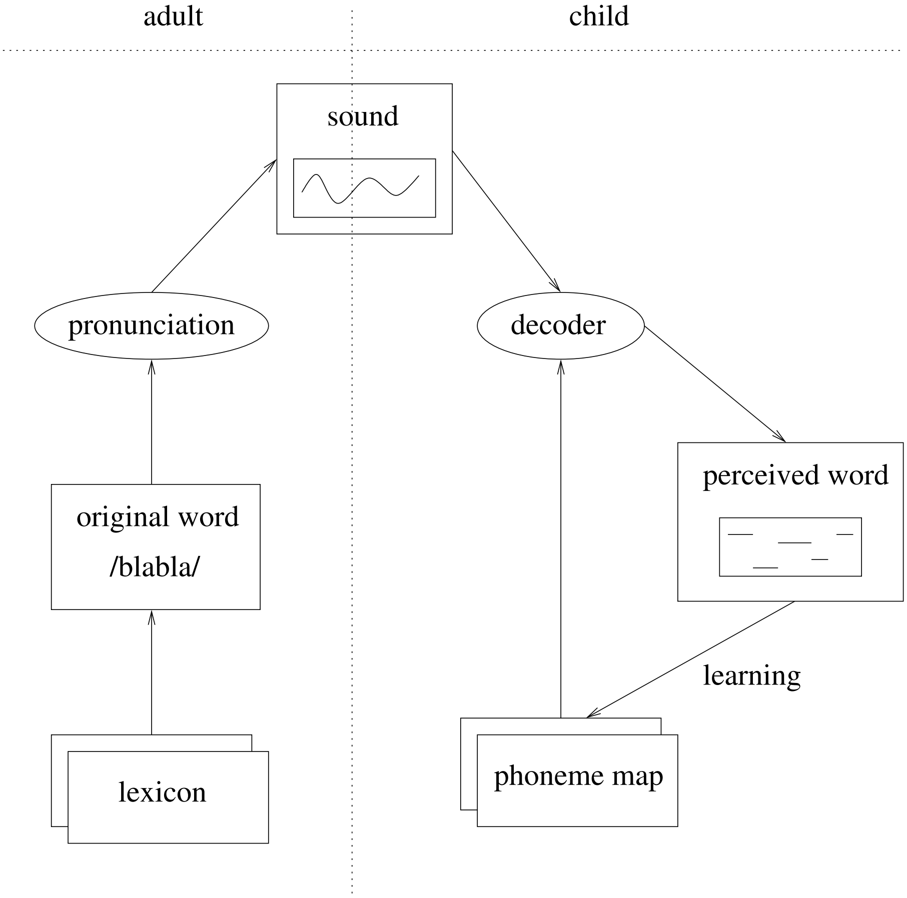

Figure 3.1: The schematic model of phoneme learning before the beginning of lexicon building.

Adult and child. In the framework we have one adult (the teacher) and one child (the learner). Unlike in the case of de Boer's multi-agent approach we assume that these roles are asymmetric and do not change during learning. Observe that there is no feedback for the child to evaluate performance. The infant can rely on unsupervised learning methods only as mentioned previously.

Adult lexicon. We assume that the child is born into an already existing language environment so we assume the existence of a lexicon. This lexicon is basically a set of words only, so no other information is stored there. A word is a list of phonemes. A phoneme is defined by one continuous value and a length. That means that the actual quality of the phoneme is determined only by one number which is far from a realistic feature vector, but this simplification allows us to concentrate on other issues like self organization and the historical dynamics of the framework (sound change). Length introduces an interesting new dimension compared to the other approaches discussed earlier.

Pronunciation. The learning step begins by selecting a word from the adult dictionary and pronouncing it. Since the words are stored in the lexicon as phoneme sequences it is necessary to convert this representation to a sound form which is a continuous signal. The learner can hear only this form, not the underlying representation. (About the implementation of this process see Section 3.3.)

Decoder. This is the key procedure of the learning. As shown in the figure the decoder has two inputs. The first is the continuous signal and the second is the developing phoneme map. The role of the phoneme map is to assign weights to every possible phoneme, i.e. every possible value-length combination (both features are continuous so the number of possibilities is infinite in principle). The decoder tries to find an underlying representation that sounds the same as the input sound if pronounced. More frequent phonemes (as determined by the phoneme map) have a higher chance to make it into the representation but in principle every possible phoneme can get there.

The resulting representation is such that it contains the fewest phonemes that can represent the given pronunciation within a given error threshold. This conforms to the principle of economy, i.e. we want to use as little amount of resources as possible and we want to keep things as simple as possible.

We have to admit that the notion of perceptual encoding and motor encoding is confused and the child has of course no way to check how would a given pronunciation sound like since the motor capabilities are not at that level. However (as mentioned in Section 3.1) one of the assumptions was that motor and perceptual encodings have isomorph structure. Furthermore the child is assumed to be able to "play back" a perceptual (compressed) representation which is a similar process to actually pronouncing a word from our point of view. Thus the confusion of motor and perceptual encodings becomes a further simplification our model makes.

Learning. Learning takes place when the decoder has created its output representation. The phoneme map is modified in such a way that the weights (or frequencies) of the phonemes that are present in the representation are slightly increased.

3.2.2 Lexicon building

The schematic model of lexicon building is shown in Figure 3.2. As clearly seen, it is very similar to the model of phoneme acquisition. What is changed is that the phoneme map is converged to a phoneme set and there is no more learning anymore. Otherwise the decoder does exactly the same as before. Let us see the changed parts in what more detail.

Fixed phoneme set. The fixed phoneme set is almost identical to the converged phoneme map (i.e. the phoneme map where the frequency difference between the frequent and rare phonemes is already so large that the rare phonemes are practically never used). The only difference is that a little "cleanup" is done which is in fact a clustering algorithm that creates clusters around the

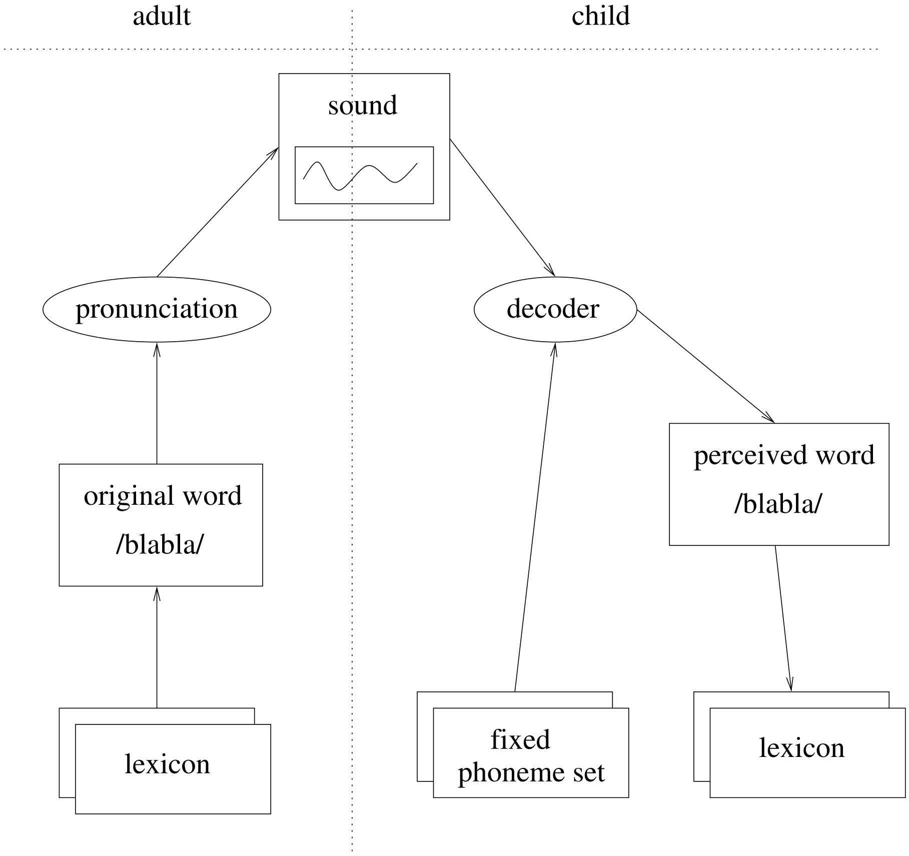

Figure 3.2: The schematic model of lexicalization after phoneme learning is finished.

most frequent candidates. This ensures that we do not get phonemes that are perceptually too close to each other.

Child lexicon. The lexicon will contain the words that the decoder outputs. Note that all the words use the learned phoneme set (which might slightly differ from the original adult set) and every adult word maps to exactly one child word.

3.2.3 Organic Sound Change

The learning model sketched above can be iterated. This means we start from an adult lexicon. The child learns first the phonemes then the lexicon. Then the child becomes an adult and the same process starts with a new child. This iteration allows us to explore the long term dynamics and implications of our model.

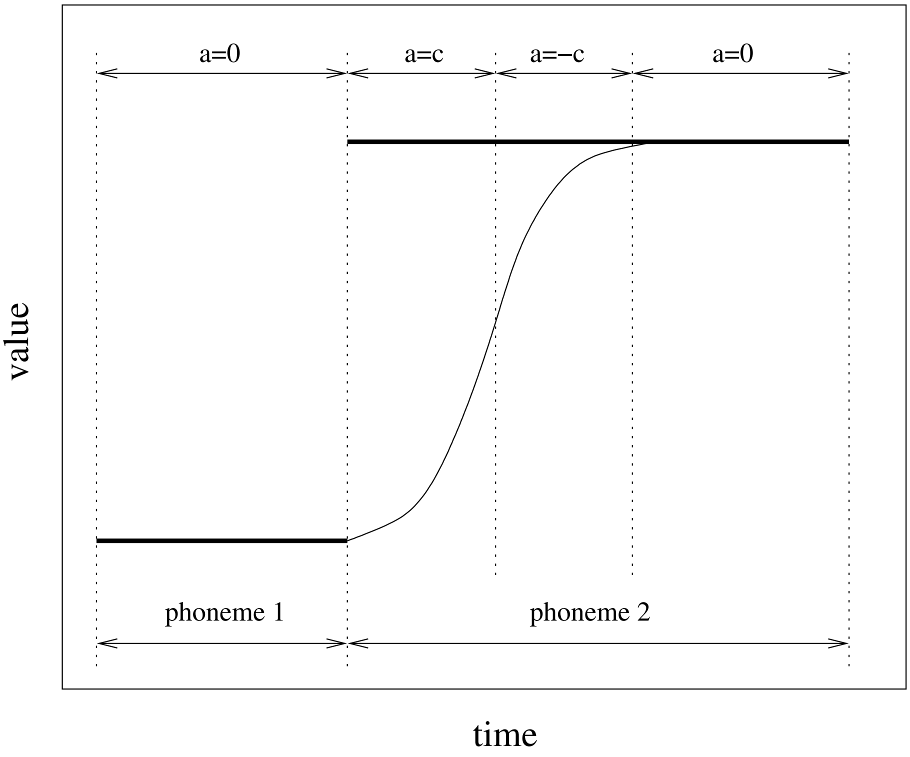

Figure 3.3: Illustration of the pronunciation model.

3.3 Implementational Issues

In this section we describe the implementational decisions that have a serious effect of the behavior of the model.

3.3.1 Pronunciation

As already mentioned, pronunciation generates a continuous curve based on a phoneme sequence. A phoneme is defined by a continuous value and a length. Strictly speaking the set of all possible phonemes was defined as $[-1, 1] \times [0, \infty]$, the second component being the length. The implementation that was chosen is illustrated in Figure 3.3.

One point of the curve represents the place of the hypothetical one dimensional speech organ. The acceleration of the speech organ is always c or $-c$, no intermediate values are taken. When we enter into the area of a phoneme along the time dimension, the phoneme value starts to attract the organ which changes its acceleration to c or$-c$ if necessary depending on its position. The acceleration is adjusted in such a way that no oscillation happens when reaching the phoneme value, the organ stops smoothly. The resulting curve is smooth (i.e. its first derivative exists). It is also very similar to the movement of the formants in a spectrogram of a pronounced word that consists of only vowels.

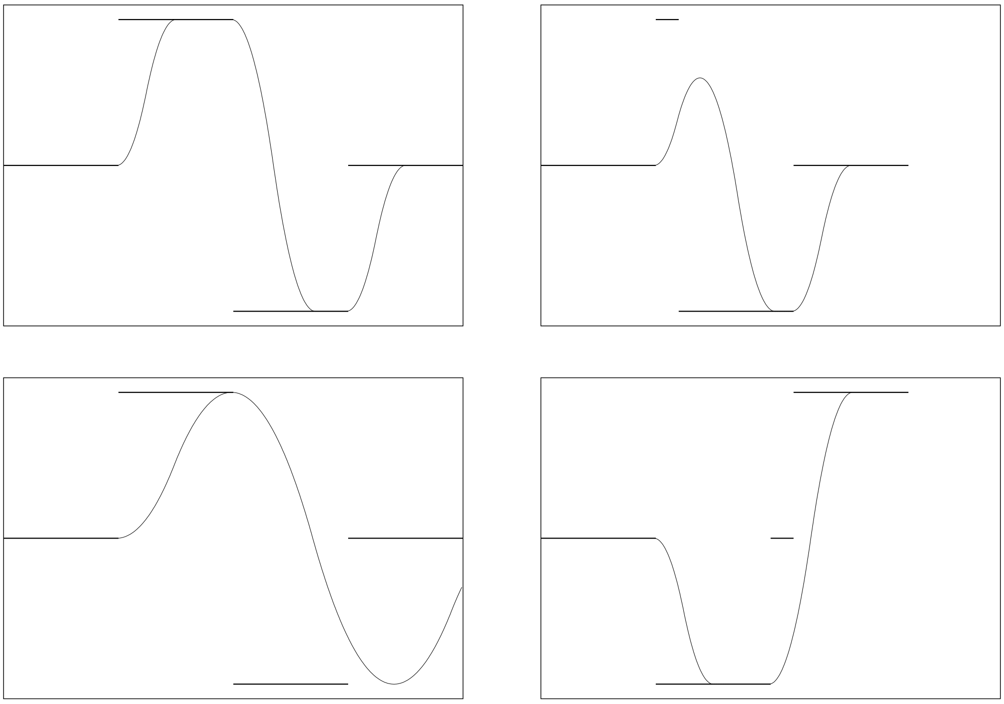

Figure 3.4: Examples for pronunciations. All plots were created with $c = 0.15$ except the lower left one which was created with $c = 0.04$.

Figure 3.4 gives examples of interesting effects of such a model. The upper left pronunciation is the typical case to give a point of reference. The lower left one is the same phoneme sequence but it was pronounced with a slower speech organ. Of course we get the same effect with the same speech organ and faster speech (shorter phonemes). In spectrograms we can see similar effects when the person who speaks is drunk, sleepy or speaks fast. It is very interesting to experiment with slowing down sound samples keeping the pitch. We get a quite clear “drunk-effect”.

The upper right plot shows the case when a short phoneme gets into a context which would require very fast movement to reach the correct value. In such cases the pronunciation results in different phoneme qualities depending on the length of the phoneme. Another interesting remark is that in such cases pronouncing the word which has a lower short phoneme results in the same pronunciation. When decoding the pronunciation there is no way to tell where was the original phoneme based only on this sample. This makes learning of phoneme systems a nontrivial task.

Another tricky case is the “invisible” phoneme in the lower right plot. Completely removing the short phoneme results in almost no change in the pronunciation. The only resulting difference would be that the new pronunciation becomes a little bit shorter. Again, in such contexts it is hopeless to decode the invisible phoneme based only on this sample.

3.3.2 Phoneme Map

The function of the phoneme map is to assign weights (proportional to frequency) to every possible phoneme. It could be implemented many ways, for example it could provide values for every continuous value and length pair if it was defined by a parametrized function. In the current model I chose the simpler solution of representing it the brute-force way, i.e. by a two dimensional grid, which involves discretization. In the map there is no connection between the phonemes; nor inhibitory nor excitatory. Note however that a set of phonemes is still automatically selected due to the decoding mechanism (see Section 3.3.3) and learning.

The clustering that outputs the actual learned phoneme set is done only at the end of phoneme learning by a separate algorithm (see Section 3.3.4). Note that it *would* have been possible to implement a map that automatically maintains and outputs a clustering without a separate clustering algorithm but again: from an implementational point of view this approach is more effective.

3.3.3 Decoder

The decoder uses the pronunciation of the adult and the developing phoneme map as input as shown in Figure 3.1. The problem is finding an underlying representation that results in a pronunciation that is closer to the original than a given threshold. (The distance function that defines closeness is discussed later.) This problem is a very hard search problem if one wants to find an exact solution. For this reason a simple yet effective heuristic was developed with a satisfactory performance.

We have to emphasize that at this point the implementation is not intended to be psychologically realistic at the implementational level. The reason is that computers have a rather different hardware than humans. For example in a highly parallel environment completely different algorithms can be designed which would probably be closer to reality.

In the first step it produces a rough underlying form that contains phonemes only with the smallest possible length. After that there is a “cleanup” phase which outputs the final result. The result of these two phases is illustrated in Figure 3.5. In the following these phases are discussed in detail.

Rough Approximation

In this phase an underlying form is created using only phonemes of a fixed length, the minimal possible length. This minimal length is a parameter of the algorithm. This underlying form is built starting with an empty word and appending newer and newer short phonemes until the length of the original pronunciation is reached.

For deciding which phoneme to append next the penalty is calculated for each possible value. It is possible because we work with a discrete set of values, not the whole continuous $[-1 : 1]$ range. Then those which fall within a given threshold are selected into a sampling pool. From this pool one phoneme is drawn randomly according to a probability distribution which is calculated based on the phoneme map which holds the frequency of every phoneme.

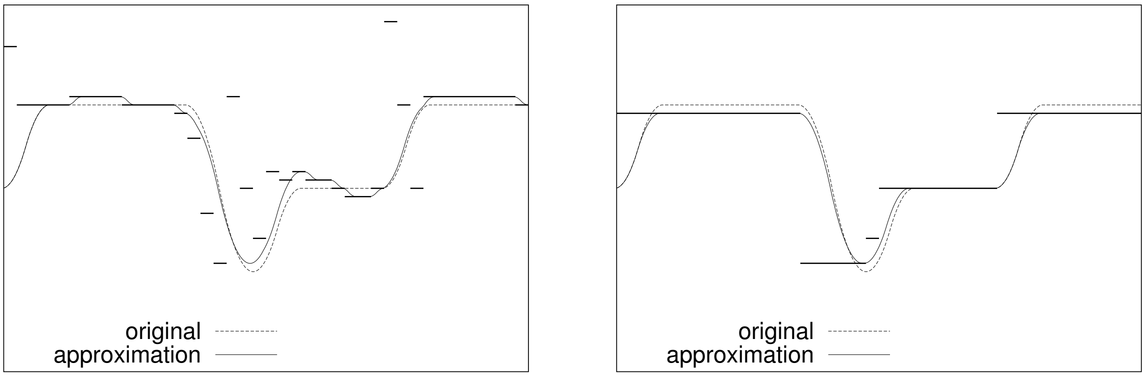

Figure 3.5: The phases of finding an underlying representation.

Distance Function

Penalty is defined by a distance function which is defined over pairs of pronunciations. In the present implementation this function is the average of square distances at equidistant time samples. It is interesting to mention that another penalty term must be added for the algorithm to work. At the end of the compared interval the difference of the derivatives of the two curves. Without this term we get into problems because the limited moving capability of the speech organ results in the accumulation of differences and for longer words the end of the word becomes completely different. Thus we need to keep not only the place but also the moving direction close to the original.

Cleanup

This algorithm works on the rough approximation which contain phonemes only of the minimal length. First neighboring phonemes with the same value are fused into one phoneme using the sum of the length of the fused phonemes as the new length.

Then we try to reduce the number of phonemes further by trying to change the value of every phoneme to either the value of the preceding phoneme or the next phoneme. If this can be done without getting out of the penalty threshold then it is done. This process is iterated until no more movements can be done within the error threshold.

Learning

The output of the cleanup is used to update the weights in the phoneme map. This process is very straightforward: the weight of phonemes which are included in the perceived word are multiplied by a constant which is a parameter of the algorithm. The constant is normally close to one, say 1.05. Due to this method the increase is first slow the it becomes fast. This results in an initial exploration phase followed by the sharpening of the contrasts.

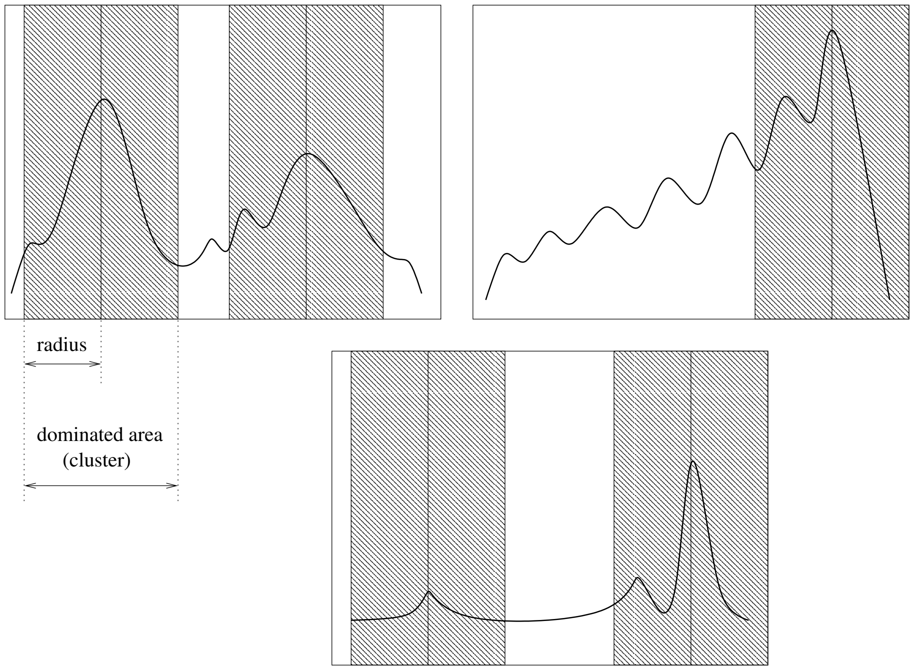

Figure 3.6: Illustration of the clustering algorithm.

3.3.4 Clustering

Clustering takes place when the phoneme learning has finished. First the phoneme values are clustered. The weight of a value is the product of all the weights of the same value with different lengths. After this the lengths are clustered along every resulting value cluster.

The clustering algorithm is illustrated in Figure 3.6. The main idea is that we are looking for *dominant peaks*. A peak is dominant if there are no larger values within a given radius. The centers of the clusters become the dominant peaks while the clusters themselves are the areas around the centers defined by the radius.

This seems simple yet it results in rather intelligent behavior. In the figure we can see three cases. The upper left plot shows a general case for illustration of the concepts. The upper right plot illustrates the smoothing effect. We have many peaks but all of them are dominated by a close larger peak. This results in only one cluster. Note that this behavior is not equivalent with smoothing out the curve. In this case smoothing could have resulted in one peak only but in other cases smoothing can result in information loss: small peaks can disappear. With the present method “smoothing” is done only selectively while small peaks are still found if they are dominant.

The lower plot illustrates this demonstrating that dominance does not depend on the height of the peak directly. It is not a problem if different centers have very different values as long as they are dominant. It is very useful for us since the phoneme map typically has very different weight values in the emerging clusters.

4 Simulations

The detailed description of the experiments discussed in this chapter can be found in Appendix A. Here we shall concentrate on a short illustration of phoneme learning and a somewhat longer discussion of a couple of interesting observed phenomena.

4.1 Phoneme Learning

Before going on to sound change maybe it is worth to give a little illustration of phoneme learning itself. The algorithm has been discussed in detail so let us see what is actually happening during phoneme learning. Figure 4.1 shows the development of the phoneme map during the first iteration trough the dictionary. As it can be seen clearly the overall structure of the phoneme system becomes quite clear by the end as the relative weight of low frequency candidates become less and less.

4.2 Sound Change

This section is concerned with commenting on some phenomena that can be seen on the figures showing the results of the simulations in the Appendix. Let us begin with some general comments over these figures. The first observable fact is that for larger acceleration and smaller thresholds the phoneme learning is very accurate. In fact those settings that have zero deviance in Figure A.3 resulted in an exact learning of the dictionary trough 100 generations. In other words the dictionary of the 100th generation is the same as the original. These cases can be thought of as control cases which show that the phoneme learning model is accurate if the level of noise is low, and changing effects are indeed due to the increased noise and not the inaccuracy of the learning model. Many types of effects that represent a deviation from this correct case will be mentioned in the following.

Finally I have to mention that it is a pity that it is not possible to include animations on paper because the plots in Figure A.2 are especially useful when animated. However in this form they still provide us with a rather holistic view of what happened during the experiment as every relevant aspect is represented in the plots.

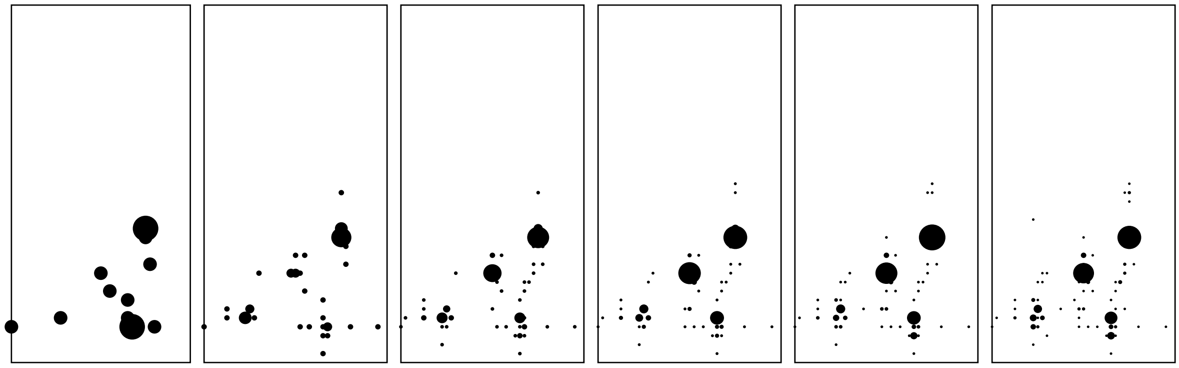

Figure 4.1: Acceleration is 0.04, error threshold is 0.002, the rest of the parameters is the same as in all other experiments. The phoneme map is shown after learning 3, 10, 20, 30, 40 and 50 words respectively. In a plot the horizontal dimension is the interval of phoneme values $[-1, 1]$, the vertical dimension is length, $[0, 4]$ with 0 being the bottom line. The size of points indicates frequency (the smaller the less frequent).

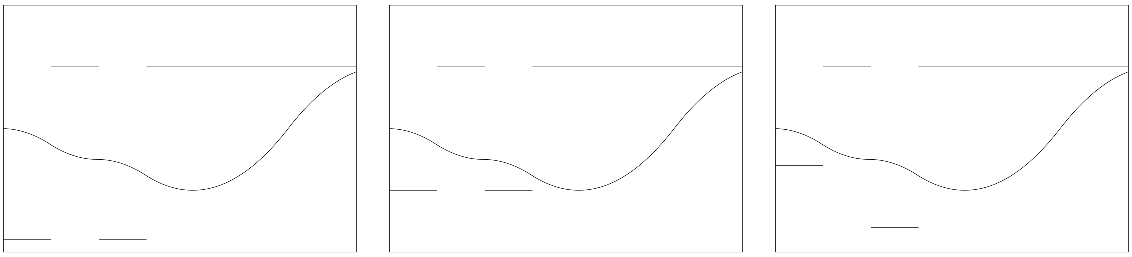

Figure 4.2: Three different encodings of a word for an acceleration of 0.01 and threshold 0.0005.

4.2.1 Random Fluctuation

Under some circumstances it is possible that some phonemes are not uniquely defined even if the threshold is very small. One example is the setting $(0.01, 0.0005)$ (see Figure A.1). In the diagram it can be seen that the left side is noisy. This can happen because most of the phonemes have a positive value so the only phoneme with a negative value $(-0.5, 0.5)$ which is rather short can not be found based on the pronunciations. Several equivalent encodings exist. Figure 4.2 illustrates this effect. As a result the encoding that is found by the learner is random within the allowed range while the pronunciation itself remains the same.

Note that it is not the error of the learning algorithm; it is an inherent property of the model which is based on pronunciation. The representations of an individual can be different as long as the resulting behavior is the same.

4.2.2 Fusion

Another effect is the fusion of phonemes. From Figure A.3 it is clear that in many cases the average length of the words decreased and even if not, there was some deviation which means that in almost all cases shorter words appeared. This obviously means fusion.

There is a different type of fusion however which does not change word size. One example can be seen for the setting $(0.02, 0.002)$. Taking a good look at Figures A.1 and A.2 we can see that the two phonemes $(0.3, 0.5)$ and $(0.5, 1.4)$ united their values while the lengths remained the same. In Figure A.2 we can also see a phoneme which is quite long (in fact 1.9) with the same value. This indicates that in some of the words the sequence $(0.3, 0.5)(0.5, 1.4)$ might have fused. Closer inspection shows that a few words show the signs of such a fusion although not all such sequences fused in this way. Depending on context other changes could take place.

4.2.3 Insertion

In the experiments insertion is common as can be seen from Figure A.3. One example of insertion can be seen in Figure 4.3 in the first three steps. Although the pumping effect does not seem realistic, its first step, when the step is created seems to be acceptable. For instance in Hungarian a similar effect can be observed in the pronunciation of “fiú” which sounds like [fiju:]. A common spelling error is “fijú” which suggests the underlying form /fiju:/ at least in a part of the population.

4.3 Side-Effects of Implementational Decisions

This section discusses effects that can be explained by a particular *ad hoc* implementational decision and which are therefore not predictive. To explore these problems is especially useful for the further improvement of the model.

4.3.1 Jumping

This effect can be seen only with a threshold of 0.005. First let us note that (as mentioned earlier) the distance function applied for calculating the difference between pronunciations includes the average of the squares of differences in sample points (which are dense enough). Since we set the minimal phoneme value distance at 0.05 (see the Appendix) we can see that with thresholds smaller that 0.0025 every phoneme is outside the threshold except the correct one. A threshold of 0.005 already allows matching different phonemes.

The distance function also includes a term which measures the difference of the derivatives of the curves which is in fact essential for successful matching. However these settings cause an interesting jumping effect which means that a phoneme never remains the same in two consecutive iterations (see e.g. setting $(0.01, 0.005)$).

The exact reason is that if the approximation is close to the original then the penalty term allows larger deviations in the derivative which makes it very likely that the approximation gets further away from the original. On the other hand if the curve is already far from the original then it remains so for the very same reason.

Note that for the setting $(0.04, 0.005)$ the two phonemes that actually emerge move completely together while producing the jumping effect as well. This is due to the same problem. Once we are e.g. over the first phoneme of a word it becomes very hard to turn back (and if we re under it the same holds). That means that the phonemes tend to be above or below their respective originals together.

4.3.2 Drifting

For the setting $(0.01, 0.002)$ a very clear tendency to drift towards the value 1 can be seen. Closer inspection reveals that the reason is the following. If the value of the first phoneme is larger than 0 then the pronunciation starts with an S-shape curve. When the approximation follows it exactly there is no problem. Drift upwards can happen only after the inflection point since we allow only two accelerations: c or $-c$. But if we drift upwards we can never go under the original curve until the S-shape finishes for the same reason. And once we are over the constant part of the original it is hard to go downwards because of the derivative penalty. This is not a problem with drifting downwards (it can happen only before the inflection point) since we can always go over the curve again. This results in a tendency of overshooting.

4.3.3 Length Pumping

Probably the most interesting observation is the "length pump" mechanism that can be observed only in the case of $(0.15, 0.005)$. All the other settings result in a relatively conservative lexicon size while in this case the lexicon simply explodes.

After investing some work in analyzing the data in greater detail the reason turned out to be the length pump illustrated in Figure 4.3. The pump works as follows. We start with a word that has at least one wave-like part like in the first plot. Very soon the inner approximation of this word becomes something like the one in the second plot. Recall that inner representation is not connected to any phoneme system. The whole phoneme map is used (taking frequencies into account of course).

It is only in the lexicon building phase when the phoneme set is chosen and the words are encoded using only this phoneme set. Since this set is much more restrictive than the phoneme map (even if weights are highly converged) it always results in some quality degradation. We can think of it as the first half of the U curve that can be observed in many domains, including phoneme learning. This quality degradation can be seen in the third plot. As a result a step is introduced in one of the peaks.

After this phase another type of insertion takes place. If the value of the phoneme that forms the step disappears or simply less frequent than the values of the top and bottom phonemes then the step gets approximated in the pretty surprising way which is shown in the fourth plot. Observe that again the inner representation is correct in the sense that the pronunciation is approximated perfectly just like in the first case. The difference is that in this case the performance degradation is much more drastic. So much so that the resulting pronunciation includes one more peaks.

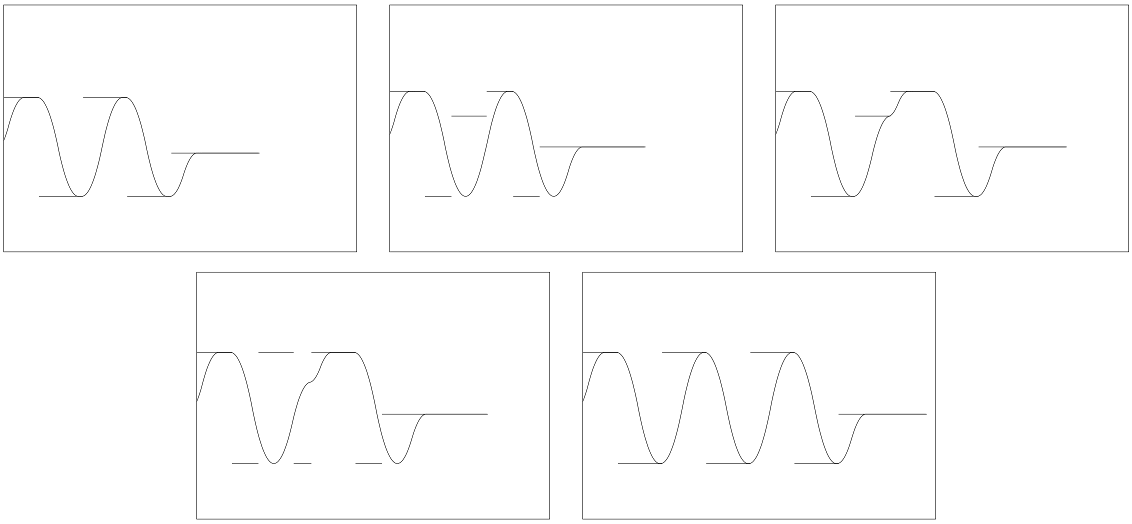

Figure 4.3: The pumping effect. The plots show in time order the first lexical form, its inner approximation (which is correct w.r.t. pronunciation), its lexicalization after length standardization (which is not correct anymore) the next approximation and finally a lexical form after length standardization. The overall result is one more peak.

Compared to the very first form we arrived in a very similar situation only with more possibilities to applying the length pump. In the first generations the growth of length is almost exponential.

4.4 Summary

We can conclude that if there is little noise than model is capable of perfectly learning the lexicon through many generations. This means that the deviation from perfection are due to the increasing amount of noise which is introduced by the slower speech organs (by smoothing out the words) and increasing error threshold (by allowing the acceptance of lower quality underlying representations). We could observe phenomena that naturally occur in natural languages like insertion and fusion. We could observe other phenomena which are caused by *ad hoc* implementational decisions.

Finally let us recall that the original goal was not to give an exact model of phoneme learning or sound change. Rather I wanted to introduce a model which gives us a bird's eye view of the process and allows us to experiment with the dynamics of such an adaptive unsupervised computational approach in the hope that in the future a more complete theory of language can be built along similar lines.

5 Conclusions

In this thesis it was emphasized that modeling language phenomena should be psychologically grounded. Structural approaches fail to fulfill this goal concentrating only on modeling surface forms using rather *ad hoc* model frameworks and hoping (if at all) that their theory will someday meet with the other fields. As Boersma puts it in his thesis:

> This book treats the explanation and description of phonological phenomena, but has little to say about their mental implementation. In other words, the *what* and *why* are accounted for but the *how* is not. (original emphasis) (Boersma, 1998, pp. 465)

However such an approach makes it hard to see the connections between seemingly different fields and—more importantly—it ignores the possibility that what is pushed into the territory of implementation might turn out to be the model itself.

In this work a model was developed with this in mind. Another assumption was that it is worth to try to model as much as possible in a simplified setting because this way we can avoid the problem of having to deal with isolated and often incompatible models of different aspects of the world. The prize to pay is of course the loss of fine grain detail.

5.1 Future Work

As mentioned in previous chapters the actual implementations of model components are not crucial from my point of view. This means that it would be interesting to replace some of them, especially phoneme learning. In earlier phases of this work I experimented with a continuous approach, i.e. the phoneme map was indeed infinite. The results were quite insufficient. The reason is that in that case finding an underlying representation is indeed a multi-constraint continuous global optimization problem. Maybe further exploration of some of the methods for solving the problem would help us see what change does a new learning method introduce in the system.

The simpler possibility of increasing the phoneme value resolution would also be interesting since as we have seen some of the properties of the output can be explained by using the particular phoneme resolution.

Fixing the drifting and the pumping effect is also possible using another interpretation of the threshold value. In the present model every phoneme value within the threshold has an equal probability to make it into the solution. By using a more sophisticated distribution that favors lower error values the results would have probably been different.

In the long run letting the lexicon develop automatically seems to be a very interesting possibility to explore. In the present model it is built after an explicit phoneme clustering so the phoneme learning and lexicon building are strictly sequential (although the phoneme learning is done on the basis of the lexicon of the adult). This solution suffices as a rough approximation but one of the really interesting aspects of human skill (and language) learning is the emergence of new layers on top of older ones. Modeling such an emerging layer would be useful also because it would introduce a new constraint on the formation of the first layer and the overall representation of knowledge so the model of phoneme learning itself could be expected to improve as well.

A Appendix

A.1 Configuration Parameters

The model has a number of configuration parameters that have an effect on the behavior of the simulations. For the sake of completeness the names used in the configuration file of the actual implementation are also included in `typewriter` font. However completely technical configuration parameters like random seed, etc. are not included.

acceleration capability of speech organs (`a`) Described in Section 3.3.1. The notation c was used in the text for this value.

recognition threshold (`rth`) Described in Section 3.3.3.

weight growth ratio (`wgrowthratio`) Described in Section 3.3.3 in connection with learning.

phoneme value resolution (`valres`) See Section 3.3.2. This defines how many equidistant phoneme values are defined in the interval $[-1, 1]$. The values -1 and 1 are always included thus this value is at least 2.

phoneme length resolution (`lenres`) See Section 3.3.2. This defines the step size of the lengths of possible phonemes.

minimal phoneme distance (`minphd`) See Section 3.3.2. This defines the minimal distance between the values of any pairs of phonemes from a phoneme system.

minimal phoneme length distance (`minphlend`) See Section 3.3.2. This defines the minimal distance between the lengths of pairs of phonemes with the same value.

initial phoneme system (`phlist`) This gives a set of phonemes to start with.

initial word list size (`wordlistsize`) This gives the initial size of the lexicon. This makes sense if no previous lexicon exists. In that case words of length 5 are generated.

iterations (`iterations`) This gives the number of iteration steps to learn a phoneme system. An iteration step is defined by going through the lexicon completely.

generations (`generations`) Used in sound change experiments. It gives the number of times a new lexicon is learned based on the previous one.

A.2 Settings and Complete Output of Experiments

The varying parameters of the experiments discussed in Chapter 4 were acceleration (`a`) and recognition threshold (`rth`). The values of `a` were taken from

$$\{0.01, 0.02, 0.04, 0.08, 0.15, 0.30\}$$

and the values of `rth` from

$$\{0.005, 0.002, 0.001, 0.0005\}$$

All combinations were tested in $6 \times 4 = 24$ experiments.

The fixed parameters were the following. The step size for phoneme values was 0.05, for lengths 0.1 with a maximal length of 4. Thus a phoneme map of size 41×40 was used. the initial phoneme set was

$$\{(0.5, 1.4), (0.3, 0.5), (0, 1), (-0.5, 0.5)\}$$

with the first value being the phoneme value and the second the length in the pairs. The first generation started with a randomly generated lexicon of size 50. The words contained different neighboring phonemes only. The rest of the parameter values were `wgrowthratio`=1.05, `minphd`=0.12, `minphdlen`=0.5, `iterations`=20, `generations`=100.

Figures A.1, A.2 and A.3 contain different visualizations of all the experiments.

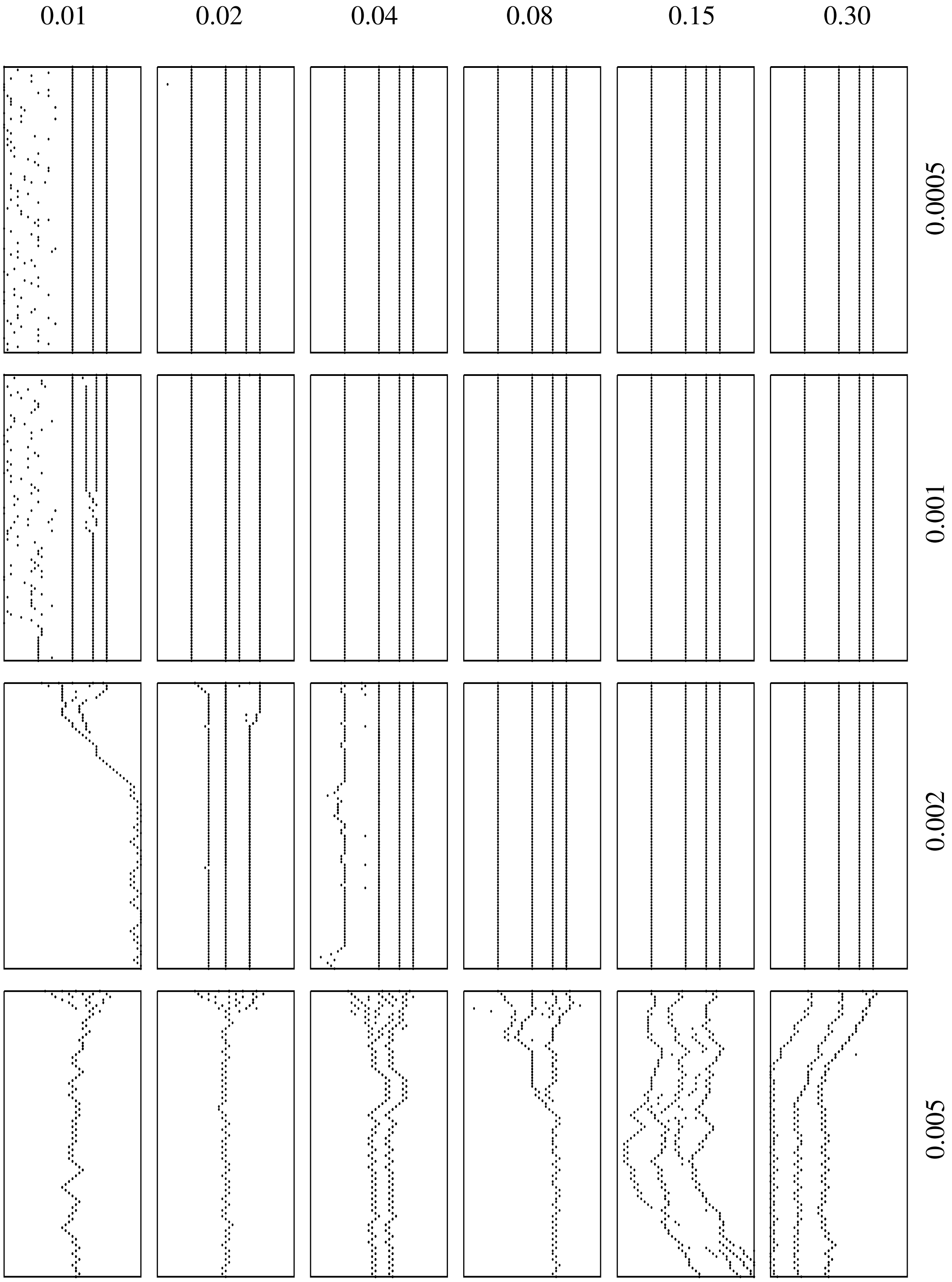

Figure A.1: The plots correspond to combinations of acceleration of speech organ ($0.01 - 0.30$) and error threshold ($0.005 - 0.0005$). In a plot time goes from top to bottom. The horizontal dimension is the interval of phoneme values $[-1, 1]$. Every time slide of a plot corresponds to the phoneme values present at the given time point. In these plots length and frequency information is ignored. However the development of the value system can be clearly followed.

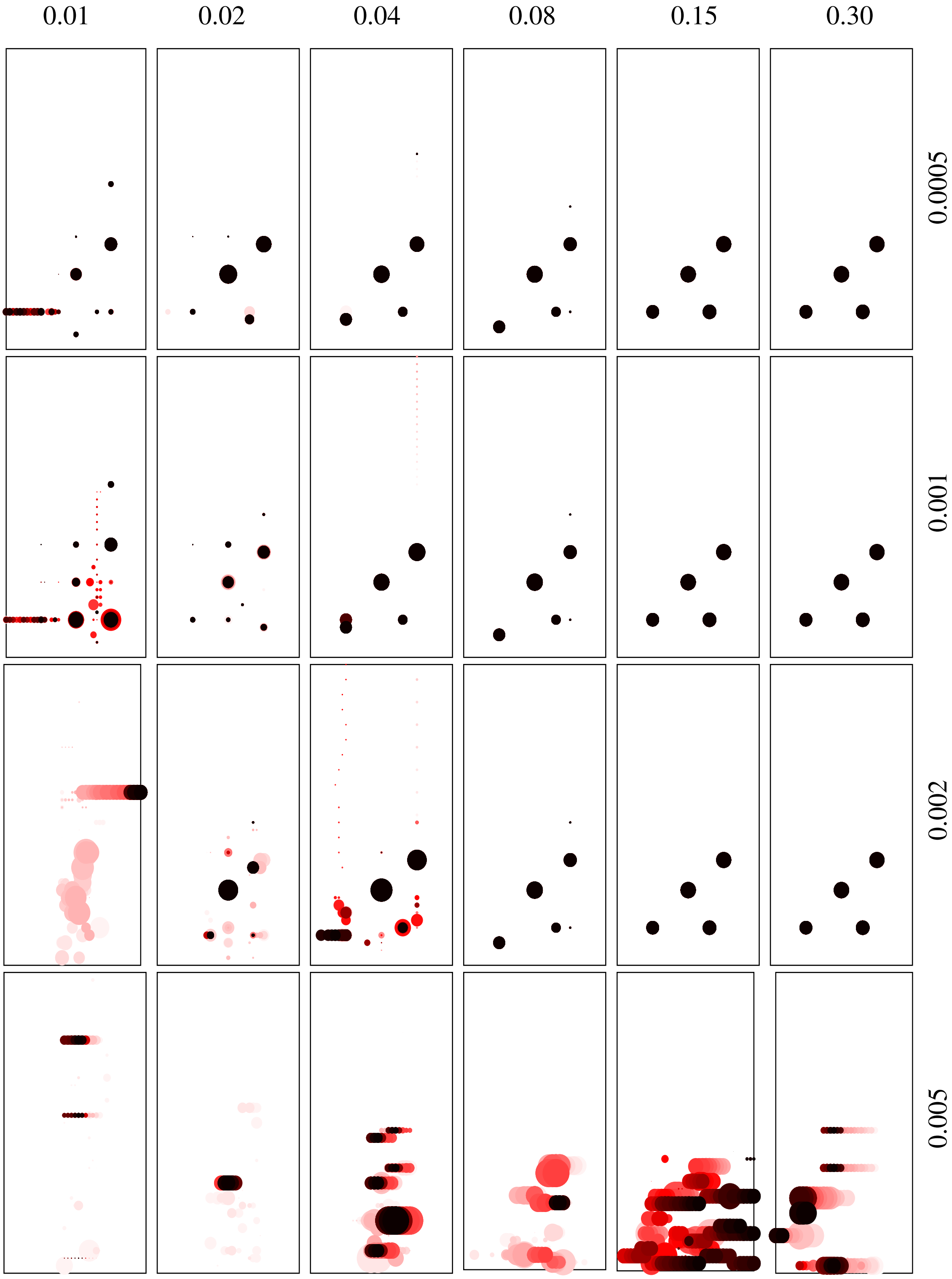

Figure A.2: The plots correspond to combinations of acceleration of speech organ $(0.01 - 0.30)$ and error threshold $(0.005-0.0005)$. In a plot the horizontal dimension is the interval of phoneme values $[-1, 1]$, the vertical dimension is length, $[0, 4]$ with 0 being the bottom line. The shade of points indicates time (the lighter the older) and size indicates frequency (the smaller the less frequent). In these plots every information about the development of the phoneme system is represented.

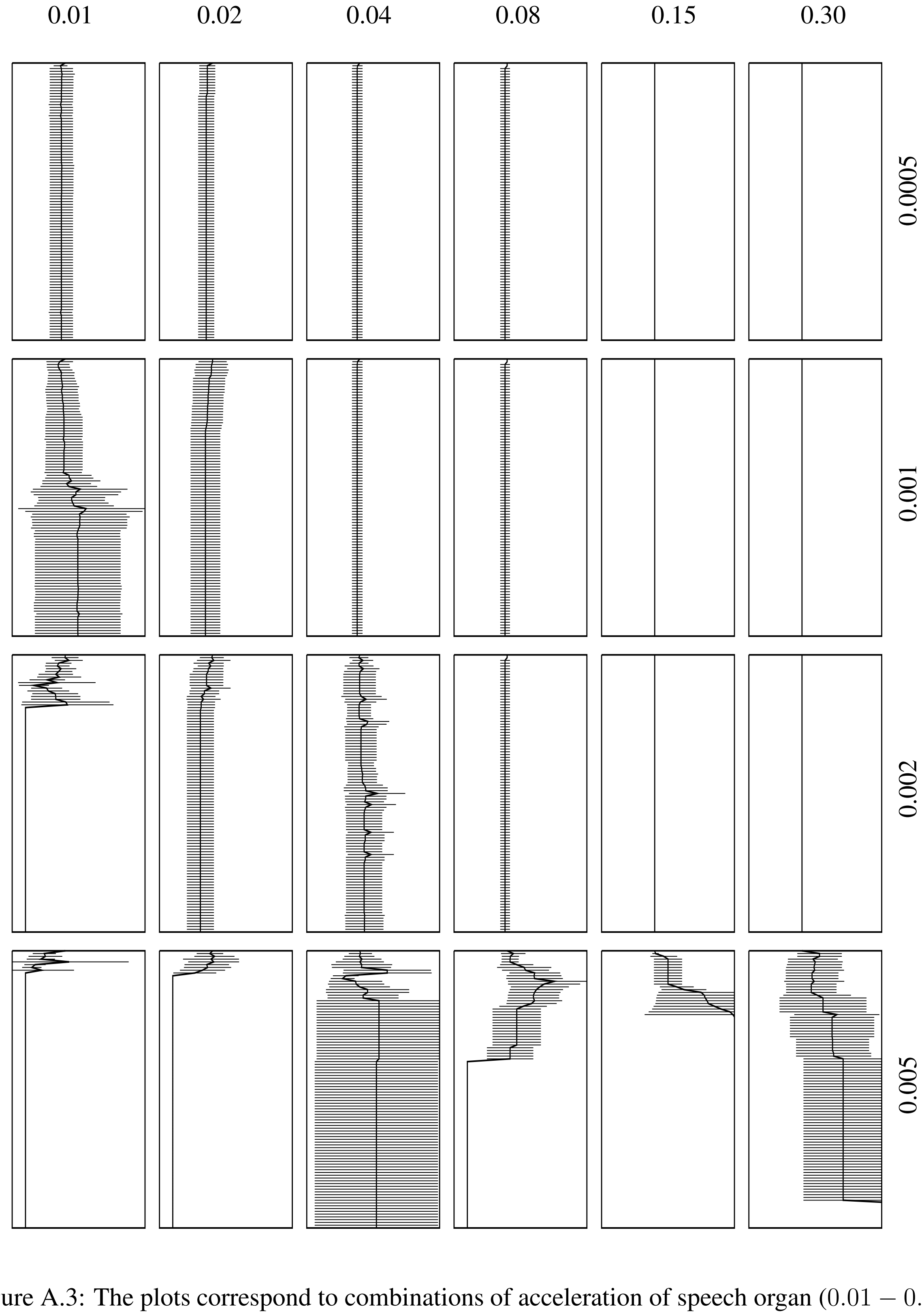

Figure A.3: The plots correspond to combinations of acceleration of speech organ ($0.01 - 0.30$) and error threshold ($0.005 - 0.0005$). In a plot time goes from top to bottom. The horizontal dimension is the average word size ($[0, 10]$). Error bars show the empirical variance of word size in the lexicon. The curve in the plot corresponding to the setting $(0.15, 0.005)$ is out of range (actually it reaches 65) so only the beginning can be seen.

Bibliography

Bailey, C.-J. N. (1973). *Variation and linguistic theory.* Center for Applied Linguistics, Arlington, Va.

Behnke, K. (1998). *The Acquisition of Phonetic Categories in Young Infants: A Self-Organising Artificial Neural Network Approach.* PhD thesis, Universiteit Twente. MPI series in psycholinguistics 5.

Boë, L.-J., Schwartz, J.-L., and Vallée, N. (1995). The prediction of vowel systems: Perceptual contrast and stability. In Keller, E., editor, *Fundamentals of Speech Synthesis and Speech Recognition*, pages 185–213. John Wiley.

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives.* PhD thesis, Universiteit van Amsterdam.

Boersma, P. and Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1):45–86.

Chomsky, N. (1979). *Language and responsibility: based on conversations with Mitsou Ronat*, chapter Generative Grammar. Pantheon Books, New York.

de Boer, B. (1999). *Self-Organisation in Vowel Systems.* PhD thesis, Vrije Universiteit Brussel, AI-lab.

de Saussure, F. (1939). *Cours de linguistique générale.* Payot, Paris.

Engstrand, O., Williams, K., and Lacerda, F. (1998). Is babbling language-specific? A listening test using vocalizations produced by Swedish and American 12- and 18-month-olds. In *Proceedings from FONETIK 98, The Tenth Swedish Phonetics Conference*, pages 118–121, Department of Linguistics, Stockholm University.

Harris, Z. S. (1951). *Methods in Structural Linguistics.* University of Chicago Press, Chicago.

Kohonen, T. (1989). *Self-Organization and Associative Memory.* Springer, New York, 3rd edition.

Kornev, A. N. (2000). Toward a neuropsychological model of phonological development. In Bichakjian, B. H., Chernigovskaya, T., Kendon, A., and Möller, A., editors, *Becoming Loquens. More Studies in Language Origins*. Peter Lang, Frankfurt.

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

Labov, W. (1980). The social origins of sound change. In Labov, W., editor, *Locating Language in Time and Space*, pages 251–266. Academic Press, New York.

Ohala, J. J. (1993). Sound change as nature’s speech perception experiment. *Speech Communication*, 13:155–161.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University.

Sekuler, R. and Blake, R. (1994). *Perception*. McGraw-Hill, New York.

Steels, L. (1996). The origins of intelligence. In *Proceedings of the Carlo Erba Foundation Meeting on Artificial Life*, Milano. Fundazione Carlo Erba.

Steels, L. (1998). Synthesising the origins of language and meaning using co-evolution, self-organization and level formation. In Hurford, J. R., Studdert-Kennedy, M., and Knight, C., editors, *Approaches to the Evolution of Language*, pages 384–404. Cambridge University Press, Cambridge.

Tomasello, M. (2000). Perceiving intentions and learning words in the second year of life. In Bowerman, M. and Levinson, S. C., editors, *Language Acquisition and Conceptual Development*. Cambridge University Press.

Vihman, M. M. (1996). *Phonological Development: The Origins of Language in the Child*. Blackwell.

Wardhaugh, R. (1992). *An Introduction to Sociolinguistics*. Blackwell.

Whiting, H. T. A., editor (1984). *Human Motor Actions: Bernstein Reassessed*. Elsevier Science.