"It's dark and obscure, but intellectual."
F. Dostoevski, The brothers Karamazow

Self-organisation in Vowel Systems

by
Bart de Boer

Vrije Universiteit Brussel
Faculteit Wetenschappen
Laboratorium voor Artificiële Intelligentie

Promotor: Prof. Dr. L. Steels
Academiejaar 1998–1999

Proefschrift voorgelegd voor het behalen van de academische graad van doctor in de wetenschappen.

Table of Contents

Table of Figures

Table of Tables

Abstract

The research described in this thesis tries to explain the origins and the structure of human sound systems (and more specifically human vowel systems) as the result of self-organisation in a population under functional constraints. These constraints are: acoustic distinctiveness, articulatory ease and ease of learning. The process is modelled with computer simulations, following the methodology of artificial life and artificial intelligence. The research is part of a larger research effort into understanding the origins and the nature of language and intelligence.

The emergence of sound systems is studied in a setting called the imitation game. In an imitation game, agents from a population interact in order to imitate each other as well as possible. Imitation is a binary process: it is either successful or a failure. Agents are able to produce and perceive speech sounds in a human-like way, and to adapt and extend their repertoires of speech sounds in reaction to the outcome of the imitation games. The agents' vowel repertoires are initially empty and are bootstrapped by random insertion of a speech sound when an agent with an empty repertoire wants to produce a sound. When the agents' repertoires are not empty anymore, random insertion does not happen anymore, except with very low probability. This low-probability random insertion is done in order to keep a pressure on the agents to extend their number of vowels.

As the agents' repertoires are initially empty and their production and perception are not biased towards any language in particular, the systems of speech sounds that emerge are language-independent and can be considered predictions of the kinds of systems of speech sounds that can be found in human languages.

The main focus of the thesis is on the emergence of vowel systems. It is shown that coherent, successful and realistic vowel systems emerge for a wide range of parameter settings in the simulation. When the vowel systems are compared with the types of vowel systems that are found in human languages, remarkable similarities are found. Not only are the most frequently found human vowel systems predicted, (this could already be done with direct optimisation of acoustic distinctiveness) but also less frequently occurring vowel systems are predicted in approximately the right proportions.

Variations on the basic imitation game show that it is remarkably robust. Not only do coherent, successful and realistic vowel systems emerge for a large number of parameter settings, but they also emerge when either the imitation game or the agents are changed qualitatively. Coherent and realistic systems still emerge when the perception and production of the agents are changed. Even if the rules of the imitation game are slightly changed, coherent and realistic systems still emerge. Of course, there are circumstances under which no systems emerge, indicating that the process is non-trivial.

It is also shown that the vowel systems can emerge and be preserved in changing populations. When old agents are removed from the population, and new, empty agents are added, coherent and realistic vowel systems can still emerge, provided that the replacement rate is not too high. It is also shown in the thesis that vowel systems can be preserved in a population, even though all original agents in it have been replaced. Furthermore, it is shown that under certain circumstances it can be advantegeous to have an age-structure in the population, so that older agents learn less quickly than young ones.

Finally, some experiments with more complex utterances are presented in the thesis. An experiment with artificial CV-syllables is presented and it is shown that, although phonemically coded (as opposed to holistically coded) systems can emerge, this simulation is much harder and much more sensitive to parameter changes than the vowel simulation. This probably has to do with the fact that in the case of CV-syllables multiple independent and partly contradictory constraints have to be satisfied simultaneously, whereas in the vowel simulations, only one constraint (acoustic distinctiveness) is really important. Also, the first attempts at building a system that can produce complex and dynamic utterances without any constraints on their structure are presented, and it is argued that the main obstacle to getting such a system to work is the mapping from acoustic signals back to articulatory commands.

The conclusion of the thesis is that universal tendencies of human vowel systems, and probably of human sound systems in general can be explained as the result of self-organisation in a population of agents that try to communicate as well as possible under articulatory and acoustic constraints. The articulatory and acoustic constraints cause the emerging sound systems to tend towards articulatory and acoustic optimality. However, the fact that the agents communicate in a population forces them to conform to the sound system in the population and causes sub-optimal systems to emerge as well.

Samenvatting

Het werk in dit proefschrift probeert het ontstaan en de universele eigenschappen van menselijke spraakklanken, met name van klinkers, te verklaren als het gevolg van zelforganisatie in een groep taalgebruikers. Elke taalgebruiker is beperkt in zijn vermogen om spraakklanken te produceren, van elkaar te onderscheiden en te leren. Het hele proces wordt met behulp van computersimulaties gemodelleerd, en het werk vormt daarom onderdeel van de onderzoeksgebieden kunstmatige intelligentie en *artificial life*. Het onderzoek is een onderdeel van een groter onderzoeksprojekt dat is gericht op het begrijpen van het ontstaan en de aard van taal en intelligentie.

Het ontstaan van systemen van spraakklanken wordt onderzocht in het vereenvoudigde kader van het imitatiespel (*imitation game*). In een imitatiespel proberen twee leden van de populatie (verder *agents* genoemd) elkaar zo goed mogelijk te imiteren. Imitatie is in dit geval een binair proces. Het is ofwel een succes ofwel een mislukking. Agents kunnen spraakgeluiden produceren en verwerken op een zo menselijk mogelijke manier. Zij passen hun repertoire van spraakklanken aan of breiden het uit aan de hand van de uitkomst van de imitatiespelen waaraan zij meedoen. In het begin zijn hun repertoires leeg. Het imitatiespel wordt op gang gebracht door het lege repertoire van een agent die een imitatiespel wil spelen, van een willekeurig gekozen klank te voorzien. Als de repertoires van de agents niet meer leeg zijn, worden er nauwelijks willekeurige klanken meer toegevoegd. Klanken worden alleen zeer af en toe toegevoegd om druk op de agents uit te oefenen om hun repertoire van klanken uit te blijven breiden.

De repertoires van de agents zijn in het begin leeg, en de manier waarop zij spraakklanken produceren en van elkaar onderscheiden is niet gebaseerd op een specifieke taal, maar slechts op algemene menselijke eigenschappen. Daarom zijn de systemen van spraakklanken die ontstaan taalonafhankelijk en kunnen ze beschouwd worden als voorspellingen van de systemen van spraakklanken die in menselijke talen aangetroffen kunnen worden.

Het grootste deel van het proefschrift houdt zich bezig met het ontstaan van klinkersystemen. Het wordt aangetoond dat coherente, succesvolle en realistische klinkersystemen ontstaan voor een groot aantal waarden van de parameters van de simulatie. Wanneer men de ontstane klinkersystemen vergelijkt met de klinkersystemen die men aantreft in menselijke talen, vindt men dat niet alleen de vaakst voorkomende systemen goed voorspeld worden (dit kon al gedaan worden door rechtstreeks te optimaliseren voor akoestische onderscheidbaarheid) maar dat ook de minder vaak voorkomende systemen voorspeld worden in ongeveer de juiste verhoudingen.

Variaties op het imitatiespel laten zien dat het buitengewoon robuust is. Coherente, succesvolle en realistische systemen ontstaan voor een groot aantal waarden van de parameters van het systeem. Ook fundamentele veranderingen van de agents en van de regels van het imitatiespel zijn mogelijk zonder de uitkomst fundamenteel te veranderen. Als men de productie en de perceptie van de agents verandert, ontstaan er nog steeds coherente en realistische klanksystemen. Ook kleine veranderingen aan de regels van het imitatiespel veranderen niet veel aan de uitkomst. Natuurlijk kunnen de omstandigheden wel zo veranderd worden dat het imitatiespel niet meer werkt en er geen klinkersystemen meer ontstaan. Dit toont aan dat het imitatiespel niet triviaal is.

In het proefschrift wordt ook aangetoond dat klinkersystemen kunnen ontstaan en bewaard kunnen blijven in veranderende populaties. Indien men oude agents uit de populatie verwijdert en jonge (lege) agents toevoegt, kunnen er nog steeds coherente en realistische systemen ontstaan, als men er maar voor zorgt dat de snelheid waarmee de populatie verandert niet te hoog is. Op die manier kan een repertoire van klanken bewaard blijven in een populatie ook al zijn alle originele agents uit die populatie vervangen door nieuwe. Tenslotte wordt er gedemonstreerd dat er omstandigheden zijn waarin het voordelig is als er een leeftijdsstructuur is in de populatie, zodat oude agents minder snel kunnen leren dan jonge agents.

Tenslotte wordt er een aantal experimenten met meer complexe klanken gepresenteerd in het proefschrift. Een experiment met kunstmatige lettergrepen die bestaan uit een medeklinker gevolgd door een klinker wordt behandeld. Het wordt aangetoond dat lettergrepen kunnen ontstaan die fonemisch (in tegenstelling tot holistisch) gecodeerd zijn. Het probleem hierbij is dat dit veel moeilijker is en veel gevoeliger voor veranderingen in de parameters dan het experiment met de klinkers. Dit heeft waarschijnlijk te maken met het feit dat voor het doen ontstaan van lettergrepen er tegelijkertijd aan meer en tegenstrijdige eigenschappen voldaan moet worden. Voor de klinkersimulaties hoefde maar op één eigenschap: het akoestische verschil tussen de klinkers, gelet te worden. Ook worden de eerste pogingen tot het bouwen van een systeem dat kan werken met meer complexe en dynamisch veranderende klanken zonder kunstmatige beperkingen gepresenteerd. Het belangrijkste obstakel om zo'n systeem te laten werken lijkt het omzetten van een akoestisch signaal in articulatorische akties te zijn.

De conclusie van het proefschrift is dat de universele eigenschappen van menselijke klinkersystemen (en waarschijnlijk van systemen van menselijke spraakklanken in het algemeen) verklaard kunnen worden als het resultaat van zelforganisatie in een populatie van agents die zo goed mogelijk trachten te communiceren, maar die articulatorische en akoestische beperkingen hebben. De akoestische en articulatorische beperkingen zorgen ervoor dat er systemen ontstaan die optimaal zijn met betrekking tot akoestische onderscheidbaarheid en articulatorisch gemak. Aan de andere kant zorgt het feit dat de agents moeten communiceren met ander leden van de populatie ervoor dat ze zich zoveel mogelijk moeten conformeren aan de populatie en daardoor kunnen suboptimale systemen ook behouden blijven.

Acknowledgements

It is always difficult to know exactly where the origins of a certain work lay. Was it in October 1995 in the Ardennes when Luc Steels explained his ideas about the origins of language and proposed that they could be applied to speech sounds as well, whereupon I remarked sceptically that this seemed infeasible? Up until then I had been working on learning robots, and Luc was not aware that I knew a couple of things about phonology and phonetics as well. Or were the seeds of the work laid much earlier, in 1989, in Leiden when I followed the introduction into the Nepali language by George van Driem? For the first time in my life I was confronted with speech sounds that were really quite different from the ones that are used in the languages I knew. My knowledge of speech sounds was deepened in the course on articulatory phonetics, given by Thomas Cook at Leiden University.

Perhaps the origins of the research must be sought even earlier, in 1985 when Paul Lemmers introduced me to the fascinating world of computers, with the Apple II and the ZX-spectrum. This made me decide to study computer science at Leiden University instead of physics. Here Ida Sprinkhuizen-Kuyper and Egbert Boers introduced me to the field of artificial intelligence. I decided to write my Master's thesis on the subject of learning classifier systems. After finishing the thesis, I had some time left to do a project in Brussels at the AI-lab of Luc Steels. This was in the summer of 1994. After this project I was invited to do my Ph. D. at the AI-lab of the Vrije Universiteit Brussel, which eventually resulted in this thesis.

First and foremost I must thank Luc Steels for providing the idea, the supervision and the research environment for the research in this thesis. The money for the project has come from the Belgian federal government FKFO project on emergent functionality, (FKFO contract no. G.0014.95) the UIAP 'Construct' project (no. 20) and from the GOA project of the Vrije Universiteit Brussel. This has given me the possibility to pursue my research in an undisturbed way that is unfortunately becoming rarer and rarer with the ever-decreasing budgets for fundamental research.

Next I must thank my friends and colleagues of the AI-lab (in alphabetical order) Tony Belpaeme, Karina Bergen, Andreas Birk, Sabine Geldof, Petra Heidinga, Edwin de Jong, Holger Kenn, Joris van Looveren, Paul Vogt and Thomas Walle for providing the atmosphere for good research, for discussions and for feedback on my work. I also must thank the people of Sony CSL in Paris, especially Frédéric Kaplan and Angus McIntyre for scientific feedback and for giving me the opportunity to so some quiet work at their lab every once in a while. I also wish to thank Björn Lindblom and Christine Eriksdotter for giving me the opportunity to present my work for the first time for an audience of serious phoneticians at Stockholm University. Special thanks go to my friends Egbert Boers, Igor Boog, Petra Heidinga, Stephan de Roode and Maurice ter Beek and to my brother Martin de Boer for discussion of my work and for metaphorically kicking my behind whenever I got stuck or was too lazy.

Of course, I thank my parents. They always stimulated me intellectually and gave me the opportunity to study at my ease so that I could broaden my education oustide the narrow scope of my specialisation and could thus lay the foundations of this interdisciplinary work. This thesis owes as much to them as it owes to me.

Finally, I thank my girlfriend Cécile Dehopre for supporting me, making me feel at home in Brussels and for putting up with my absent-mindedness, my lack of attention for her and all the evenings I was not available when I was working on my thesis.

1. Introduction

Language defines man. It is generally agreed that what distinguishes humans from other animals is their intelligence and their ability to talk. Intelligence however, is often defined in terms of language. The famous Turing test, designed by Alan Turing (Turing 1963) intended for deciding whether a computer program is intelligent, is based on the computer's ability to use language. Ethnic identity is also often defined by language. In Tok Pisin, the lingua franca of Papua New Guinea, the word for referring to one's ethnic group is *wantok*, "one talk", meaning the people that speak the same language. The Slavic peoples refer to the Germanic peoples as *немец* (e.g. Russian *немецкий*) "those who cannot speak". The ancient Greeks called the Persians βαρβαροι, barbarians, because all they heard when they heard the Persian language were unintelligible sounds: "Barbarbar..." Language is essential for man.

If one wants to understand the origins of human intelligence, it is therefore of the greatest importance to understand the origins of language. Man has always speculated on the origin of language. This used to be the domain of religion. Language was usually seen as a gift (or a damnation) of the gods. Since the renaissance, scientists have also started speculating about the origins of language (see e.g. Rousseau 1986, Jespersen 1968). Most of the early speculation was rather impressionistic. More recently, with advances in archaeology, neurology and linguistics, speculation on the origins of language has become more grounded in facts (see e.g. the contributions in Hurford *et al.* 1998).

This thesis attempts to shed light on the origins of one aspect of language—the sounds it uses. This is done within the framework of language as a self-organising system (e.g. Steels 1995, 1997b, 1998b, Kirby & Hurford 1997, Hurford, *to appear*, Kirby 1998, *to appear*) and is put to the test and elaborated with computer models.

1.1 The Aims

This thesis aims to show that the structure of human vowel systems and most likely the structure of human sound systems in general, can be explained as the result of self-organisation under acoustic, articulatory and cognitive constraints. Often, innate distinctive features and markedness constraints have been proposed (e.g. Jakobson & Halle 1956, Chomsky & Halle 1968) as explanations for the occurrence of phonological universal tendencies. However, emergence of these universals as the result of interactions in the population would show that innate features and markedness constraints are not necessary. Rather, some systems of speech sounds are more likely outcomes of the interactions in a population of language users than others are.

In order to show that this happens, it is necessary to build a computer simulation with sufficient realism, so that human perception, production and learning of vowels can be modelled with accuracy. It must then be shown that:

a) Coherent vowel systems emerge from scratch in a population of agents.
b) The systems that emerge are realistic.

In order to fulfil the first aim, it is necessary to construct a simulation that is free from bias. First of all, this means that it should not be based on a specific language. The aim is not to say something about any language in particular, but rather about language in general. The agents should therefore neither be constrained to working with systems with a fixed number of vowels, as was done in previous work, (Liljen-

crants & Lindblom 1972, Schwartz *et al.* 1997b, Glotin 1995, Berrah 1998) nor should they be restricted to working with predefined sets of possible vowels. Secondly, measures that can objectively measure the coherence and quality of the vowel systems will have to be defined.

The realism of the emerging systems will be tested by comparing them to the vowel systems that are found in human languages. There are two possible ways of doing so. One way is to compare the emerged vowel systems in individual populations with vowel systems of groups of people that speak a particular language. In this way the realism of the distribution of vowels in a given population can be checked. This is in fact a purely phonetic comparison. The second way of comparing is to compare the different *types* of emerged vowel systems with the different types of vowel inventories found in human languages. This is a phonological and typological comparison. It can establish the realism of the emerged vowel configurations. The comparison of the artificial vowel systems with human vowel systems will necessarily be qualitative. Fortunately, there are lots of data on the possible vowel phoneme inventories in the world's languages. Unfortunately there is less data on the actual acoustic realisations of vowels for a given language.

Apart from the main aims of the thesis, there are three minor aims as well. These are concerned with showing that self-organisation occurs, independent of the implementation details, with testing Steels' theories (1997b, 1998b) on the origins of language and with extending the work to more complex utterances. They are:

c) Showing that realistic and coherent systems also emerge when the simulation is slightly changed.
d) Showing that vowel systems can be transferred successfully from one generation to the next.
e) Investigating the possibilities of applying the theory to more complex utterances, involving consonants and sequences of sounds.

The first of these aims can be pursued by implementing variations on the production and perception of the agents and on the rules of the imitation games. These variations can be either quantitative—changes in parameter values—or qualitative—changes in the algorithms that are used. When vowel systems emerge that are similar, independent of the parameter settings or the algorithms, this indicates that the emergence of realistic vowel systems is a necessary outcome of the interactions in a population of agents with successful imitation as a goal.

Aim d) is important, because Steels' theory on the origins of language depends on the fact that the mechanism that is responsible for the transfer of language from one generation to the next is also responsible for the emergence of language and vice versa. Also, for cultural evolution to take place there must be a transfer of learned items from one generation to the next. It is therefore necessary to construct a simulation in which the population changes, where old agents (with their phonological knowledge) can be removed from the population and new (empty) agents can be added. It must be measured whether it is possible to preserve a vowel system, even if the whole original population is eventually replaced. It must be investigated under what circumstances transfer of the vowel systems is possible and under which circumstances transfer breaks down. It is also interesting to investigate whether a new vowel system can emerge from scratch in a changing population.

The final aim is to investigate whether and how the simulation should be extended to more complex utterances. After all, vowels are only a part of the sound systems that humans use. All human languages also use consonants and combine

vowels and consonants into complex utterances in non-random ways. This is probably also subject to acoustic, articulatory and cognitive constraints. Actual implementation of a simulation that works with complex utterances falls outside the scope of the research in this thesis, but investigations have been made into how the simulation could and should be extended to more complex utterances.

1.2 The Contributions

This thesis takes its fundaments from two different sources: *artificial intelligence* and *phonetics*. For this reason the contributions it tries to make will also be in these two fields. The contribution in the field of artificial intelligence will be to examine the theory of Steels about the origins of language and ultimately the origins of intelligence. Complex phenomena, in this case the vowel system of a language, can emerge without the need for complex learning mechanisms or complex interactions. The learning mechanism that is used is simple prototype learning. The interactions are simple imitation games. Still, a coherent and realistic system of speech sounds emerges. This is due to the fact that the interactions are iterated a large number of times. Apparently complexity can be derived from iterating simple interactions. This is nothing new. However, the application of these ideas to speech sounds is new.

A second contribution to artificial intelligence is to provide a simple way in which vowel (and probably other sounds as well) systems can be learnt. The problem of learning the sound system of a language is always that it is not clear beforehand which sounds can distinguish meaning and which sounds are just random (or systematic and predictable) variations. The combination of direct imitation and non-verbal feedback about the success of the imitation turns out to be able to learn the distinctive sounds in a sound system without being fooled by the other variations of the sounds. It is shown in the thesis (although only in a preliminary experiment) that the model can be used to learn a human vowel system by connecting it to a loudspeaker and a microphone. These ideas could probably be applied to computer models that learn natural language, for example in adaptation to speaker characteristics or dialects. However, applications are not the topic of this thesis.

The contribution of this research to phonetics is to establish why systems of speech sounds become the way they are. It is already well established why vowel systems in the world's languages are the way the are. This is because they tend to be optimised for acoustic distinctiveness, articulatory ease and articulatory consistency. However, it is not clear who is doing the optimisation. The individual speakers do not optimise their vowel systems. This thesis tries to show that the iterated interactions under constraints of perception and production will inevitably lead to near-optimisation of the sound systems that emerge. Neither innate cognitive structures nor explicit optimisation are necessary. Once a vowel system is established in a population, it is preserved even though it might not be totally optimal. This accounts for the fact that both in the simulation and in human languages different types of vowel systems are found for a given number of vowels. This is not the case in simulations that directly optimise the vowel systems for acoustic distinctiveness. These will generally find only one or two different types. Another contribution to the field of phonetics of the simulation presented here is therefore that it makes more realistic predictions of the vowel systems of human languages than previous computer simulations.

1.3 The Background

The research presented in this thesis is rooted in *artificial intelligence* and *phonetics*. Within artificial intelligence, the most relevant subfield is the one that tries to model the origins of intelligence. Within phonetics it is the subfield that is interested in explaining the structure of sounds that are found in human languages. The methodology of the research—using agent-based simulations for modelling aspects of human intelligence—is that of artificial intelligence. The data on which the simulations were based and the data that were used for verifying the results were taken from the field of phonetics. The research questions were taken from both fields.

The thesis's most important pillar in the field of artificial intelligence is the work by Steels (1995, 1996, 1997a, 1997b, 1997c, 1998a, 1998b) on the origins of language, described in more detail in chapter 2. His work views language as a complex dynamic, open and distributed system. The term *complex dynamic* indicates that the dynamics of language—the way it changes, the way its speakers interact and the way it works—are complex and cannot be predicted by simple rules. Language is an *open system* with respect to both its community of speakers and with respect to what it can express. The population of speakers as well as what a language can express (its words, its constructions) can change without disrupting it. Finally language is a *distributed system* in Steels' view, because none of the speakers has perfect knowledge of the language nor does any of the speakers have central control over the language. Language is to a large extent independent of its community of speakers. According to Steels, coherence is maintained through self-organisation, while changes of the language are caused by cultural evolution. Steels also claims that in a population of speakers that are sufficiently intelligent to learn a system as complex as language, a language will indeed spontaneously emerge. This emergence is driven by the same processes of cultural evolution and self-organisation that drive language change.

The most important phonetic work on which the thesis is based is that on the functional explanation of the regularities that are found in the vowel systems of the world's languages. Phoneticians have also used computer models (see e.g. Liljencrants & Lindblom 1972, Schwartz *et al.* 1997b) for this purpose. However, all these models were based on direct optimisation of functional criteria. This is unfortunate, because humans do not optimise their sound systems. Nevertheless, the models based on optimisation predict the most frequently occurring human vowel systems very well. The hypothesis that is investigated in this thesis is that the optimisation is the result of self-organisation in the interactions between the language users.

1.4 The Model

The computer simulations presented in this thesis are based on a population of agents that can produce, perceive and learn vowels in a human-like way. Each agent maintains a repertoire of vowels. These are represented as acoustic and articulatory prototypes. Whenever an agent perceives a sound, it looks up the vowel in its vowel repertoire whose acoustic prototype is closest to the perceived signal, and considers this vowel as recognised. The use of prototypes is based on the observation that humans tend to perceive speech sounds in terms of prototypes as well (see e.g. Cooper *et al.* 1976, Liberman *et al.* 1976).

The agents' "goal in life" is to imitate each other as well as possible. At the same time they are under pressure to increase the number of vowels in their reper-

toire. Initially their repertoires are empty. They bootstrap the imitation by initially creating random sounds, or by storing imitations of the sounds they hear. The fact that they start out empty and that they are able to produce any (basic) vowel that a human could make implies that their behaviour is independent of a specific language.

They engage in interactions that have been called imitation games in analogy with Steels' use of the Wittgenstein's (1967) term *language game* (Steels 1995). In an imitation game, one agent picks a random sound from its repertoire and the other agent tries to imitate it. Then feedback is given about the success of the imitation. On the basis of this feedback, the agents update their vowel repertoires. The agents cannot look at each other's vowel systems directly. Just as humans, the agents are not capable of telepathy. The only way in which they can interact is through making (and imitating sounds) and through giving the simple (one-bit) feedback about whether an imitation was successful, or not. From the sounds they perceive and the feedback they receive, the agents can improve their vowel systems, so that they can imitate the other agents in the population better.

The interactions between the agents are iterated. Pairs of agents are picked from the population at random. Each agent has an equal probability of either initiating or imitating in an imitation game. Because the imitation games are iterated, because the assignment of roles is random and because the agents all start out empty, all agents in the population are equal. There is no division in "teachers" that already have knowledge on the one hand, and "students" that have no or limited knowledge on the other hand.

The proposed model is sufficiently flexible that different variations are possible. Some of these variations will be investigated in this thesis. An example of such a variation is making the population "open". Just as in human populations, where people get born and die, agents can then enter and leave the population. It will also be shown that small variations in the rules of the imitation game or in the production and perception of speech sounds will not qualitatively change the outcome of the simulations.

1.5 The Results

The results of the research presented in this thesis consist of showing that coherent and realistic vowel systems emerge for many different parameter systems and for some variations on the perception and production of the agents and their way of learning speech sounds. It was also shown that the emerging vowel systems are significantly better than randomly generated ones and that they were close to optimal. Furthermore a comparison of the emerged vowel systems with vowel systems found in human languages showed that the emerged vowel systems exhibit the same universal tendencies as human vowel systems. It has also been shown that realistic vowel systems can emerge in changing populations and that existing vowel systems can successfully be maintained in these changing populations, even though after a certain time all original agents in the population have been replaced. Finally it has been shown that implementing production and perception of agents that work with more realistic signals is computationally and conceptually feasible. The main problem for extending the system to more complex utterances is the mechanism for learning and imitating them, and especially the mapping from acoustic signals to articulatory movements.

The results are presented in two forms: graphical representations of representative vowel systems that emerged from the simulations and numerical representations based on the calculations of certain measures over a large number of runs of the simulations using the same parameter settings. The averages and the standard deviations of these measures are presented and if necessary, the complete distributions are given. The measures that are used most often are the average over the population of the number of vowels in the vowel systems of an agent, of the energy of the vowel systems and of the success of imitation.

1.6 How to Read the Thesis

The best way to read the thesis is "to begin at the beginning, and go on till you come to the end: then stop"[1]. However, not all readers will have the time or the energy to study everything. Therefore, a small overview will be given of what information can be found in which chapters. Chapter 2 contains a more detailed description of the two theoretical pillars of this work: the theory of Steels and others on the origins of language and the phonetic theories about the universal tendencies of vowel systems and their explanations. It also contains some reflections on the use of computer simulations in gaining more insight into complex phenomena. At the end of the chapter the research questions of the thesis are restated.

Chapter 3 is of great importance to the thesis. The history of the research and the history of the simulation are described here. But most importantly, the architecture and behaviour of the agents as well as the basic imitation game are described. In order to appreciate the results in the rest of the thesis, this chapter is essential.

Chapter 4 contains most of the results of the simulations with the basic imitation game. In this chapter it is shown what happens in the populations when the imitation game is played and how the simulations reacts to different parameter settings. The most realistic settings of the basic parameters are determined. Quantitative measures for measuring the quality and realism of the vowel systems are defined. The values of these measures are determined for both random systems and systems that are optimised in the same way that systems in earlier work were optimised, so that the values that emerge in the simulations can be put in perspective. Also an articulatory view of the system is presented (most of the representations of the vowel systems in this thesis, are acoustic). It is shown that although the agents did not use distinctive features, markedness or rules, their vowel systems can nevertheless be described in terms of features, rules and markedness.

Chapter 5 presents a number of qualitative changes to the simulation. First "open" populations and the transfer of vowel systems from generation to generation are investigated. This transfer, and its dependence on a number of parameters, such as the speed with which a population is replaced, is investigated into some detail. The second part of the chapter studies a variation on the imitation game that does not use non-verbal feedback. The non-verbal feedback is considered to be the least realistic aspect of the simulations. It is therefore investigated what happens when it is not used. It turns out that a little bit of non-verbal feedback will probably always be necessary. The last part of the chapter is concerned with changing the perception and production of the agents so they can work with more realistic signals. An attempt is made to learn a human vowel system. The implications of doing this are discussed.

---

[1] Lewis Caroll, *Alice's Adventures in Wonderland*.

Chapter 6 is the most interesting chapter for phoneticians. This chapter contains a detailed comparison of the artificial systems that emerge with human vowel systems. It turns out that the similarities are striking, both for the most frequent systems and for the less frequently found systems. Only for small vowel systems (3 and 4 vowels) the similarities are less. This is probably due to an unrealistic aspect of the perception function that was used.

Chapter 7 describes and discusses ongoing and future work. It describes the work that has been done with complex utterances so far. The scientific necessity of extending the simulation to complex utterances is put forward. Design decisions that have to be made when implementing imitation games that are to work with complex signals are presented and possible solutions are discussed. In this chapter and in appendix F technical details for building a complex model are presented. Chapter 7 and appendix F are essential for everybody who wants to continue the work in this thesis.

Finally, in chapter 8 the conclusions of the thesis are presented. Of course, these are interesting to everybody who is reading this thesis.

The thesis also contains eight appendices. Appendix A contains a list of symbols that are used in different places in the thesis. Appendix B contains a more comprehensive calculation of the quality measures for random and optimal systems. Appendix C contains a mathematical analysis in of the imitation success of randomly generated vowel systems. Appendix D contains a description of the signal processing that was used for implementing real vowel imitation games. Appendix E contains some data on measurements of consonants in different contexts. These have been done for implementing imitation games with consonant-vowel syllables. As has been mentioned above, appendix F contains a detailed description of models for production and perception of realistic speech signals, as well as the presentation of an experiment that has been done with this simulation. Appendix G contains short descriptions of languages that are referred to in the text and their inventory of vowels. Finally appendix H contains the tables of the International Phonetic Alphabet for those of the readers that are not quite familiar with it.

As it is impossible to reproduce sounds in print the next best way to present vowel systems is in figures. For this reason, this thesis contains a great many figures and graphs. The thesis therefore looks slightly like a comic book in places. The reader is kindly requested to forgive the author and enjoy the thesis as if it *were* a comic book.

2. The Theoretical Background

The theoretical background of the work presented here consists of two main bodies of work. The most important one is the work by Steels and co-workers (Steels 1995, 1996, 1997a, 1997b, 1997c, 1998a, 1998b, Steels & Kaplan 1998, Steels & Vogt 1997, for related work see e.g. Hurford *et al.* 1998, part III) on the origins of language. In this work computer simulations and robotic experiments are used to model the origins and the dynamics of language. The idea to use computer simulations to model the origins and dynamics of human sound systems was derived from this work. The general architecture of the computer simulations was also based on this work.

The second body of work is the work on the universal tendencies of sound systems of the world's languages. Universal tendencies of human sound systems are among the best-researched universal properties of language. The linguistic questions that are addressed here were taken from this research and it was also used for checking whether the computer simulations actually produced results that are compatible with what is known about human languages.

2.1 Universal Tendencies of Human Sound Systems

Most human languages use sound as their primary medium for conveying meaning. Only sign languages use vision. The stream of speech sounds is usually analysed as consisting of a sequence of separate speech sounds that are called *phonemes*. Phonemes are defined as minimal speech sounds that can make a distinction in meaning. In English, for example, /ɛ/ and /æ/ are phonemes, because the words /bɛt/ "bet" and /bæt/ "bat" have different meanings. In Dutch or French, for example, these words would be indistinguishable, so these languages are analysed to have only have one phoneme /ɛ/. In making a description of a language, one first has to make an inventory of which sound distinctions can make a distinction in meaning, i.e. which phonemes the language uses. Usually a rather unambiguous analysis of the set of phonemes of a language is possible.

However, there are some complications. The most important one is that it is not easy to separate the actual physical speech signal into phonemes. This is because the human articulators do not produce phonemes separately, but already start producing new phonemes when they are not yet completely finished producing the previous ones. This effect is called *co-articulation*. Co-articulation causes phonemes to be realised differently in different contexts. This is called allophonic variation. However, not all allophonic variation can be explained as the effect of co-articulation. For example, the fact that the phoneme /l/ in English is produced quite differently at the beginning of a word than at the end of a word can not easily be explained by co-articulation effects. Rather this variation is something that must be learned by a speaker. This variation can assume rather extreme forms, especially in languages with small phoneme inventories. For example in the language Rotokas, with an inventory of only 11 segments, the phoneme /ɾ/ has allophones [ɾ], [n], [l] and [d] all of which are apparently in free variation (Firchow & Firchow 1969). Linguistics therefore makes a distinction between the abstract elements that can distinguish meanings of words, called phonemes, and their physical realisation, which is called *phonetic* realisation. In parts of this thesis, frequent reference will be made to the phoneme inventories of languages, without making reference to their actual phonetic realisations. The reader should therefore be aware that in the use of these

inventories of phonemes one should always ask the question: "What about allophonic variation?"

2.1.1 Regularities of systems of speech sounds.

The phoneme inventories of the world's languages show at the same time remarkable diversity and remarkable regularities. In the $UPSID_{451}$, the UCLA Phonological Segment Inventory Database (Maddieson & Precoda 1990, the first version is described in Maddieson, 1984), a database that contains the phoneme inventories of a representative sample of 451 of the world's languages, a total of 921 different segments occurs. Of these, 652 are consonants, 180 are vowels and 89 are diphthongs. Apparently the human vocal tract is capable of producing an amazing diversity of sounds. Still, any single language only uses a small subset of these possible sounds. In the $UPSID_{451}$, the smallest inventories are those of the East-Papuan[1] language Rotokas (Firchow & Firchow 1969) and the South-American language Múra-Pirahã (Everett 1982, Sheldon 1974) both with only 11 phonemes. The language with the largest inventory is the Khoisan language !Xũ (Snyman 1970) with 141 phonemes. The typical number of phonemes, according to Maddieson (1984), lies between 20 and 37.

The phonemes that a language uses are not chosen randomly from the possible sounds the human vocal tract can make. In fact some sounds appear much more frequently than others do. In the case of vowels, [i], [a] and [u] appear in 87%, 87% and 82% of the languages in $UPSID_{451}$, whereas the vowels [y], [œ] and [ɯ] appear only in 5%, 2% and 9% of the languages. This is also true for consonants. Some consonants, like [m] (94%), [k] (89%) or [j] (84%) appear almost universally, while others, such as [ʀ] (1%), [ʃ'] (1%) or [ʔ] (1%) appear very rarely. According to Lindblom and Maddieson (1988) the possible sounds of the world's languages can be divided into basic articulations, elaborated articulations and complex articulations. Apparently languages with small inventories only use basic articulations, while for larger inventories elaborated and complex articulations are used.

Also, phoneme inventories tend to be symmetric. If, for example a language has a front, unrounded vowel of a certain height, e.g. [ɛ] (which occurs in 41% of the languages in $UPSID_{451}$) it tends to have a corresponding back rounded vowel of the same height. In this case this would be [ɔ], which occurs in 36% of all languages in the sample, but in 73% of the languages that have [ɛ]. Symmetries can also be observed in the consonant inventories of languages. If a language has a voiced stop at a given place of articulation, for example a [d] (appearing in 27% of the languages in $UPSID_{451}$), it usually also has the corresponding voiceless stop with the same place of articulation. In the example this is [t], which appears in 40% of the languages of the sample, but in 83% of the languages that have a [d]. In general, languages use the full range of possible combinations of place of articulation and manner of articulation (voiced, voiceless, plosive, fricative, etc.) rather than a subset of these.

This implies that some systems of speech sounds will occur more frequently than others. In fact, this is even more strongly the case than would be predicted from the above mentioned symmetries. In principle a three vowel system consisting of [i], [a] and [u] would seem only slightly more likely than a system consisting of [ɛ], [a] and [ɔ] (considering the a priori probabilities of the different segments). However,

[1] Details of genetic affiliation and location of languages have been taken from Grimes (1996).

in a previous version of the UPSID with 317 languages, the former system occurs ten times, while the latter system does not occur at all (Vallée 1994, Annexe 2). The most common vowel system is the one consisting of [i], [e], [a], [o] and [u]. This occurs in 34 of the 317 languages, (Vallée 1994) much more often than any other system. Certain systems seem to be favoured, while others seem to be avoided.

2.1.2 Regularities of speech sound sequences.

Further regularities can be found in the way languages combine sounds into syllables and words. All languages have syllables that consist either of a single vowel (V) or of a consonant followed by a vowel (CV). Syllables that end in a consonant (VC or CVC) are rarer and so are syllables with clusters of consonants (CCV, VCC etc.) When a language has clusters of consonants, some of them are more frequent than others (Vennemann 1988). At the beginning of a syllable for example, a cluster consisting of a plosive followed by a nasal, such as [gŋ] is much more common than a nasal followed by a plosive. At the end of a syllable however, the reverse is true. The preferred sequence of the different types of consonants in a cluster is sometimes described with the sonority hierarchy. The hierarchy is approximately as follows: voiceless stop ≤ voiceless fricative < voiced stop ≤ voiced fricative < nasal < semivowel. This hierarchy means that at the beginning of a syllable a sequence of voiceless stop followed by a semivowel (for example [pl], like in English "please") is possible, but the inverse sequence [lp] (*"lpease") is not, whereas at the end of the syllable the reverse is more likely ("help" as opposed to *"hepl").

2.1.3 Explanations of regularities based on features.

Apparently sound systems of languages show great regularities. One can now ask the question where these regularities come from. Traditionally explanations have been based on innate properties of the human language capacity. These explanations (see e.g. Jakobson & Halle 1956, Chomsky & Halle 1968) assume that there are (innate) features in the human brain that determine which distinctions between sounds can be learned. These features are usually binary. An example of a feature is nasality. A sound can either be nasal or not. Some of the features and some of their values are more marked than others. This means that certain distinctions are preferred over others, so that, for instance, a language would prefer to use the distinction high/low for vowels before it would use the distinction nasal/non-nasal. Non-marked values of the features are preferred over the marked ones. For example, nasality for vowels is considered to be marked. Nasal vowels will thus be rarer than non-nasal ones. In general, sounds with unmarked features and unmarked values for these features will be more frequent than ones with marked features and values.

Although the theory of distinctive features is quite useful as a tool for describing sound systems of languages, it does not work very well for explaining the observed patterns. First of all, it is not quite clear which features should be used or even how many features there are. There are many ways in which languages can make phoneme distinctions (Ladefoged & Maddieson 1996). Some of these distinctions are only used in very few languages. Furthermore, languages make subtle sound differences that are not used to distinguish meanings. For example, the English word "coo", the French word "cou" (neck), the German word "kuh" (cow) and the dutch word "koe" (cow) are all pronounced differently and perceived as recognisably different by speakers. It is not clear, however, how these subtle differences would have to be represented or explained in a distinctive feature framework. Also, there is no clear markedness hierarchy. This can be seen from the fact that phoneme inventories of languages can differ in one segment. If there would be an unambiguous

markedness hierarchy, languages with the same number of phonemes would have to have the same phoneme inventories. Apparently the markedness of the features cannot predict the sequence in which phoneme inventories grow. Furthermore, if innate features and markedness play a role, it still remains to be explained why and how these particular features became innate, preferably in an evolutionary framework. Finally, and most importantly, features and their markedness are derived from observation of linguistic data. It is therefore circular to "explain" this linguistic data with innate features and markedness, which have been derived from the very same data. Rather, one would like to have a theory that is based on independent, preferably physical, physiological or psychological data (see Lindblom *et al.* 1984, Lindblom *in press*.)

Several attempts have been made to build a theory that explains the structure of human sound systems based on physical and psychological properties of human speech production and perception. The work of three researchers, Kenneth Stevens, René Carré and Björn Lindblom will now briefly be discussed. They have proposed different independent factors for predicting the sound systems of human languages.

2.1.4 Stevens' quantal theory of speech.

Stevens' *quantal theory of speech* (Stevens 1972, Stevens 1989) is based on the observation that for certain positions of the articulators, a small change in position results in a small change in acoustic perception, while for other positions, an equally small change of articulator position results in a much larger change in acoustic perception. Thus the space of possible articulations can be divided into plateaux of relative stability and regions of rapid transition. According to Stevens, distinctive features can be predicted, or at least explained from the positions of the plateaux and transitions. The two plateaux of stability correspond to the two values of the distinctive feature, while the transition region is avoided. The continuous space of possible articulations is thus divided into discrete, so-called *quantal* states.

The quantal theory of speech does not predict which vowels and which consonants will appear in systems of speech sounds of a given size. It is rather a theory of distinctive features. It explains from independent physical, physiological and psychological arguments why certain distinctive features are expected in natural languages. It does not explain in which sequence these distinctive features will appear, nor does it explain why certain features would be more marked than others. Another problem is that some articulator positions are quantal relative to movements in one articulatory dimension, but not relative to an independent other dimension. Quantal theory does not provide an explanation why certain articulatory and acoustic dimensions are preferred over others. Although the theory is incomplete in certain respects, it is a good attempt at finding an independent explanation for the distinctive features one finds in human languages.

2.1.5 Carré's distinctive region model.

Another theory for explaining the structure of sound systems is the *distinctive region model* developed by René Carré (Carré 1994, 1996, Carré & Mrayati 1995). This theory considers human speech communication as a near-optimal solution to the physical problem of producing communication over an acoustic channel using an acoustic tube that can be deformed. The theory assumes that an optimal communication system can produce maximal acoustic differences with minimal articulatory movements. Minimal articulatory movements are defined as linear and orthogonal deformations of a uniform acoustic tube. Carré uses a computational model with which he calculates the deformations of the uniform tube that result in maximal

acoustic distinctions. This model finds deformations that result in an acoustic space that corresponds to the vowel space of human sound systems. The uniform tube is divided into four *distinctive regions* that correspond to the regions of the vocal tract that are used in vowel production. The model can be extended to predicting places of articulation of consonants by looking at maximal changes in formant frequencies. The uniform tube is then divided into eight distinctive regions, each corresponding to different places of articulation for consonants.

This model is able to predict the possible places of articulation, as well as the available vowel space from purely physical principles and from the assumption that speech communication is a near-optimal solution to the problem of communicating with acoustic signals produced by a deformable acoustic tube. However, this model does not directly predict which of the possible articulations will be chosen for building a sound system (although see: Carré 1996). Note also that there seems to be a discrepancy between the Stevens' theory and Carré's theory. Given a certain articulatory movement, Stevens seems to favour minimum acoustic change whereas Carré seems to favour maximal acoustic change.

2.1.6 Predicting sound systems as a whole.

Lindblom and Engstrand (1989) have pointed out (in a reaction to the quantal theory of speech, but their comments hold equally well for Carré's work) that for explaining the sound systems one finds in human languages, one should not look at the qualities of individual sounds and features alone. Rather, one should look at the role of each sound in the sound system as a whole. It should be sufficiently distinct from all the other sounds in the sound system. A sound might have very salient acoustic properties, but if the sound system already contains a sound very much like it, it is not going to be a very good candidate for extending the sound system. If one wants to explain the sound systems of the world's languages one should therefore look at systems as a whole, rather than at the merits of individual speech sounds.

A first attempt to predict sound systems as a whole, without looking at the qualities of the sounds that make up the system was undertaken by Liljencrants and Lindblom (1972). They predicted vowel systems with a given number of vowels by minimising an energy function of the total system. The energy function is defined as:

$$2.1) \qquad E = \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} \frac{1}{r_{ij}^2}$$

where E is the energy, n the number of vowels and r_{ij} the perceptual distance between vowels i and j. The function adds the inverse square of all the distances between all the vowels in the system.

The minimisation procedure effectively spreads the vowels as evenly as possible over the available vowel space. The procedure starts with a predefined number of vowels scattered randomly near the centre of the available acoustic space. It then makes modifications to the positions of the vowels (within an acoustic space that is limited by what can be produced by the human vocal tract) and checks whether the energy function becomes less. If it does, the new state is kept and the procedure is repeated until the energy cannot be lowered anymore. This procedure amounts to the vowels repelling each other within the limited acoustic space. It can also be modelled, as Liljencrants and Lindblom (1972) point out, by repelling magnets floating in a basin of water with the required shape.

For systems with limited numbers of vowels their simulation produced realistic results. The systems that were generated correspond with the vowel systems that are frequently found in the world's languages. With improved ways of calculating the perceptual distances between different vowels, (Crothers 1978, Lindblom 1986, Vallée 1994, Schwartz *et al.* 1997b) the optimisation's predictions match even better with the observations of real languages.

The same method can be applied for predicting systems of consonants, although it is much more difficult to build computer simulations of this. Work has been done on predicting repertoires of consonant-vowel syllables (Lindblom et al 1984). Here the criterion of minimal articulatory complexity has to be added. In the sound systems of human languages one can observe that not only acoustical distinctiveness is maximised, but also economy of articulatory movements. If making sounds more distinctive requires much more complex articulations, it is preferred to use less distinctive, but less complex sounds. As Maddieson (1984, §1.5) observed, the most frequent vowel system is /i, e, a, o, u/, not /i, ẽ, a̤, o̤, ṳˠ/. The latter is more distinctive acoustically, but much more complex articulatory. A repertoire of basic articulations is used first, and only when the number of segments in the sound system becomes large, more complex (and more acoustically distinctive) articulations will be used (Lindblom & Maddieson 1988). Of course, this introduces many more parameters in a computer simulation, and thus makes them much harder, as well as more controversial to build.

The point remains that sound systems of languages can be considered as the result of an optimisation of acoustic distinctiveness and articulatory ease of a complete system of sounds. The available articulatory gestures, the acoustic distinctiveness as well as the articulatory ease of sounds relative to the other sounds can be determined by (among others) the models of Stevens and Carré. This provides a quite detailed account of *why* human sound systems are the way they are.

2.1.7 How sound systems have become optimised.

However, this account is not quite complete. Apparently, more or less optimised sound systems are found in the world's languages, but it is not clear *how* they have become optimised. Clearly, the individual language users and language learners do not do an explicit optimisation. On the contrary, they try to imitate their parents and peers as accurately as possible. This can be observed from the fact that people make and observe much finer distinctions in their sound systems than are necessary for successful communication. This makes it possible that speakers of slightly different dialects of a language can understand each other perfectly, but still perceive that the other speaks a different dialect. The question why this is the case is very interesting, but falls outside the scope of this thesis. This fact will just be accepted as a given.

Apparently, the sound system of a language is optimised to a certain extent, even though the language users themselves do not do any explicit optimisation. However, as has been pointed out in the first section of this chapter, there are individual variations of the language that tend towards ease of production, understanding and learning. Apparently then, there is a global optimisation in the language, due to local interactions. This is an example of self-organisation. In order to investigate this phenomenon and in order to check what exactly is its role in explaining the structure of sound systems, one has to abandon the point of view of language as a purely individual behaviour and assume the point of view of language as a collective, complex dynamic behaviour. Due to the complexity of self-organising phenomena, the best way to investigate them is by building computer simulations.

2.1.8 Glotin's AGORA model.

The first model to be described that used a simulation of a population in order to explain the properties of vowel systems was the AGORA-model by Hervé Glotin (Glotin 1995, Glotin & Laboissière 1996, Berrah *et al.* 1996). It is based on a community of talking "robots" called *carls* (Cerveau Analytique de Recherche en Linguistique/Cooperative Agent for Research in Linguistics). Each *carl* has a repertoire of vowels, that are represented both articulatory and acoustically. It is equipped with an articulatory model, based on Maeda's model (Maeda 1989) with which it can produce acoustic signals consisting of formant patterns. Initially, for each *carl* a fixed number of vowels is chosen at random near the position of the neutral vowel. In the simulations, two *carl*s are selected from the population at random, and they both produce a vowel that is randomly chosen from their repertoire. They then find the vowel in their repertoire that is closest to the sound they hear. They shift this vowel, so that its acoustic signal will be closer to the sound they heard, and shift all the other vowels in their repertoire away from this signal.

Depending on the amount of shifting a *carl* does, a fitness is calculated. The less shifting a *carl* does, and thus the more it confirms to the sound systems in the other *carl*s, the fitter it will be. After a number of interactions between *carl*s, the least fit *carl*s are removed from the populations, and the fittest are used to calculate a replacing *carl*, in the way of a genetic algorithm (for an introduction, see Goldberg 1998). The vowel systems of the replacing *carls* are initialised with a cross between the vowel systems of the parent *carls*.

After a while the population usually converges to a common vowel system that looks like the most common vowel system in the world's languages for the given number of vowels (usually four or five). However, convergence was not guaranteed.

There are a number of disadvantages to the AGORA-model. The first is that, due to the complexity of the Maeda articulatory model, the simulations are very calculation intensive. This made it impossible to use populations of any realistic size. The population size in most of Glotin's experiments was limited to five *carls* only. Also the number of vowels was limited to four or five. Furthermore, the model had great difficulties to converge. The genetic component was added in order to get the model to converge more rapidly. However, this genetic component confuses the simulation (is the driving force natural evolution or cultural self-organisation?) and makes it quite unrealistic. Apparently a new *carl* can inherit a sound system, something which obviously does not happen in humans. Glotin is aware that this is unrealistic (Glotin, personal communication) but considers it a simplification of humans learning the sound system of their parents. He says that it does not influence the outcome of the experiments much, except for making them converge more rapidly. Another problem of his model is that the agents push the vowels in their vowel systems away from each other. This makes the model equivalent to Liljencrants and Lindblom's (1972) original simulation. As the agents do a local optimisation of their vowel systems, the interactions between them are not crucial for the shape of the emerging vowel systems. An agent talking to itself would get the same results.

2.1.9 Berrah's ESPECE model.

Similar criticisms apply to the work of Berrah, (Berrah 1998) which is a continuation of the work of Glotin. Berrah's model, called ESPECE, is a simplified version of Glotin's model. The agents' vowels are now only represented in the acoustic domain, and the "genetic" component has been made much simpler. This speeds up the simulation and makes its behaviour more transparent, so that more varied experi-

ments, with larger populations and larger vowel inventories could be done. Berrah describes a large number of experiments for many different parameter settings. The vowel systems that appear in the population of agents match very well with the most frequent vowel systems that are found in the world's languages. In the second part of his thesis, Berrah explores the effects of the maximum use of available distinctive features (MUAF) principle. He shows that by using a slightly different distance metric, one can predict the phenomenon that available distinctions will be used maximally before other distinctions will be used. For example with vowels, extra features such as length or nasalisation will only be used if there is already a minimum number of ordinary vowels. However, a discussion of these results falls outside the scope of this thesis.

The results of Berrah's experiments make it quite clear that in his model, too, the shape of the vowel systems is determined by the repulsion between the vowels of an individual agent, not by the interactions between the agents. He describes an experiment with a single agent that results in the same optimal vowel system as the ones with multiple agents (Berrah, 1998 p. 72). Berrah's model is essentially the same as the Liljencrants-Lindblom model. Berrah realises this himself: "*Remarquons, finalement, que le cas extrême où la société n'est composée que d'un seul agent revient, en réalité, à effectuer uniquement des répulsions. Par conséquent, le principe simulé dans ce cas n'est autre que le principe de dispersion globale.*"[2] (Berrah, 1998, p. 72). Both Glotin's and Berrah's models do therefore not capture the essential aspect of self-organisation: global optimisation without local optimisation. The optimisation is still caused by local actions of the individual agents. The only effect of the interactions is that all agents end up with the same system, although this also has to be boosted by replacing non-conforming agents with copies of conforming agents.

2.2 Steels' Work

This work has been influenced quite directly by Steels' work on the origins of language (Steels 1995, 1996, 1997b, 1998b). Although others (see e.g. the contributions in Hurford *et al.* 1998, part III) are working in the same direction, the influence of their work is not as direct as that of Steels' work. The focus of this section will therefore be on his work. But in order to understand his ideas, the views on language of two eminent linguists have to be taken into account.

Ferdinand De Saussure, in his *cours de linguitsique générale* (reprinted, 1987) stressed that there are two aspects of language: the imperfect language that individual speakers actually produce, with speech errors, reductions, interruptions etc. and language as a convention in a population that is more abstract and idealised and of which all speakers know an imperfect version. He introduced the term *parole* for the first variety and the term *langue* for the second. He considered only the second form to be worthy of scientific study, thus effectively viewing language as a macroscopic, social phenomenon. Later linguists, most notably Noam Chomsky (1965, 1972, 1975, 1980) have taken a different view. Chomsky calls the speech that people actually produce *performance*, and supposes that underlying *performance* is a more abstract *competence*. Competence is the linguistic knowledge of an idealised individual in an idealised, homogeneous population. As the idealised population is homogene-

[2] Let us remark finally that the extreme case of a population that consists of a single agent boils down to performing repulsions only. Consequently, the principle that is simulated in this case is none other than the principle of global dispersion.

ous, it effectively asserts no influence on the individual language users. In Chomsky's view the fact that language is a social convention is therefore not relevant to its study. One could say that Chomsky views language at the microscopic level. The division in performance and competence is useful when one wants to write a grammar of a language. However, if one wants to understand the actual dynamics of language, including the way it originated, the way it is learnt and the way it changes, this division turns out to be unnatural.

Steels returns to the view of language as a social system, but stresses that the *parole* and *performance* are as important in understanding language as the *langue*. The imperfect *parole* is the only kind of language people can observe and produce and is therefore a fundamental basis for the *langue*. Transfer of language from one individual to another is subject to noise and speech errors. Speakers have their own *idiolect*, or personal and incomplete knowledge of the language. Steels says that the macroscopic behaviour of language can be seen as the emergent outcome of the microscopic interactions between the speakers, just as temperature and pressure of a gas can be seen as the macroscopic emergent result of the interactions between the individual molecules.

Steels also does not adhere to the view that the basis of grammar is innate. He does admit that the language input that language learners receive is too little for purely inductive learning. However, he proposes selectionistic learning mechanisms that are more powerful and that are able to learn quickly from limited stimuli. Describing his learning mechanisms in detail falls outside the scope of this thesis. If one does not accept that much of linguistic structure is innate, it remains to be explained why the world's languages show universal similarities. The work in this thesis will show that innate structures are not necessary for explaining universal (phonological) properties of language. Self-organising interactions in a population are sufficient to explain the emergence of structure.

2.2.1 Language as an open, complex dynamic system.

Language is an open system because it does not depend on individual speakers. Speakers can enter (be born or migrate into) and leave (die or migrate out of) a speech community without affecting the language. Language is not only an open system for its speakers, it is also an open system with respect to the things it can express. New words, expressions and grammatical constructions can enter the language and be adapted by the speech community, and obsolete words, expressions and constructions can disappear.

Steels realised that in order to understand these phenomena, and to understand such things as the origins of language and language change, language needs to be viewed as an *adaptive, complex dynamic* system. First it will be discussed what it means for a system to have complex dynamics and then what it means to be adaptive. A complex dynamic system in this context is a system in which there are a large number of elements that interact only on a local scale in a non-linear and non-hierarchical way. This means that the behaviour of the whole system is not predictable in any straightforward way. Most notably, in these systems there is the possibility of organisation on a global scale without global interactions. This global organisation is said to be *emergent*. An example of such a system is a colony of bees building a honeycomb. Although there is no central supervision, and every bee can only perceive and act on its local environment, organisation on a large scale (the hexagonal pattern of the honeycomb) emerges. A car is a counterexample of a complex dynamic system. It has many locally interacting parts, but they are organised in a hierarchi-

cal way. Iron filings following magnetic field lines are not a complex dynamic system either, as here there is the non-local organising force of the magnet.

Language is a typical complex dynamic system. The interacting elements are the individual language users. Their local interactions consist of talking to each other and learning the language from each other. There is no central authority controlling the language. The fact that the language is and remains coherent must therefore be an emergent property.

2.2.2 Language as an adaptive system.

A system is adaptive if it can change itself (or its behaviour) in reaction to its environment in order to optimise certain internal criteria. An example of an adaptive system is an animal that can change its behaviour in order to exploit a new food source and thus to optimise the internal criterion of satisfaction the need to eat. Language is also an adaptive system. It changes in order to optimise at least three criteria: communicative efficiency, communicative effectiveness and ease of learning. Maximisation of efficiency amounts to minimising the amount of effort necessary to produce speech. This can be illustrated, for example, by the reduction of pronouns or by the reduction of complex consonant clusters. Maximisation of effectiveness is achieved by maximising the success of the communication. This is illustrated by fixing word order for expressing grammatical roles if a case system has been lost in a language or by the emergence of fixed words and expressions for frequently used objects. Ease of learning is maximised by using as few items (words, sounds and grammatical rules) as possible. This can be illustrated by the loss of words for infrequently used cultural items or by the loss or reduction of complex case systems or verb paradigms (for a critical discussion of the role of these phenomena see e.g. Hopper & Traugot 1993).

No individual speaker, however, actively performs these optimisations. The global organisation is the result of local interactions. Speakers tend to have many registers of speech. Some of these registers are more formal than others. In formal slow speech, speakers produce all aspects (sounds, words, grammatical structures) of their language to the fullest. In fast, informal speech, however, the utterances are reduced. As informal speech is more frequent than formal speech, children will tend to learn the informal variants before the formal ones. They will therefore have a slight tendency towards reducing the language. Also, the things that are easiest to learn and that are most frequent will be learnt first. On the other hand, things that are hard to learn or infrequent are most likely to be lost from the language. New words and new fixed expressions can spread through the community of speakers by a process of *positive feedback*. In the example of a new cultural item, many new words or expressions will be created at first. However, some of these words or expressions will be used by larger groups of speakers than others. Speakers that use the most frequent word will have fewer problems in communication than ones that use less frequent words. Speakers will therefore tend to switch from less frequent words to more frequent words, and children will tend to learn the most frequent words first. Frequent words will therefore become more frequent at the expense of less frequent words. This process will eventually lead to one dominant word.

2.2.3 Mechanisms of language origins.

The process by which order on a global scale emerges from local interactions is called *self-organisation*. According to Steels, self-organisation plays an important role in explaining the emergence and the change of language. However, according to him, there are other mechanisms that play a role as well. These mechanisms—that

play a lesser role in the research described here—are *cultural evolution*, *co-evolution* and *level formation*.

Evolution is a process in which optimisation takes place by selecting the fittest individuals from a population and replacing less fit individuals by individuals that are like the selected individuals. The best known variant of this process is Darwinian or genetic evolution, (Darwin 1859) the process by which biological species arise. For evolution to take place, three things are needed, apart from a population of individuals. The first is preservation of information from one generation of individuals to the next. The second is a selection criterion. The third is the introduction of variation. In genetic evolution, information is preserved by the DNA, selection is based on the fitness of an individual in its environment and variation is introduced through mutation. In cultural evolution, the population undergoing evolution does not consist of biological individuals, but rather of ideas or knowledge, or memes as Richard Dawkins (1976) has called them. In the case of language this will be a person's knowledge of the language. Information is preserved by the learning of the language from old speakers by new speakers. Selection takes place on the ground of the criteria of communicational effectiveness, efficiency and learnability. Variation is introduced through imperfect production and perception, but also by conscious innovation (e. g. invention of new words.) Thus language is subject to cultural evolution.

Whenever multiple species evolve in the same environment, there is the possibility of co-evolution. A standard example of co-evolution in nature is that between the cheetah and the gazelle. The cheetah hunts the gazelle by trying to outrun it. This means that faster gazelles have a higher chance of surviving (and producing offspring) than slower ones. Over time gazelles will evolve towards running faster. However, faster cheetahs will be able to catch more gazelles, and therefore to raise more offspring. Cheetahs will thus become faster over time as well. Both species exert pressure on each other to become faster and faster. Co-evolution generally speeds up evolutionary change by increasing the pressure on species. Therefore it is sometimes called an "evolutionary arms race".

According to Steels, co-evolution can take place in language as well. In this case different parts of language (the sound system, the lexicon, the grammar etc.) are comparable to the different species. They exert pressure on each other, because they make use of each other. The lexicon uses the sound system in order to form words. The grammar uses the lexicon for grammatical function words. The lexicon uses the grammar to determine which new words can be formed by combining elements from the lexicon. The sound system uses the grammar in order to determine which combinations of sounds are allowed, etc. Because all systems are separately subject to evolution, they will also co-evolve. According to Steels (Steels 1997, Steels 1998b), this speeds up the emergence of complexity in the language.

The third mechanism that Steels proposes for the emergence of complexity in language is called level-formation. In level formation elements that are subject to evolution (and possibly co-evolution) join together in a bigger entity that then becomes subject to evolution as a whole. An example of this is the formation of multicellular organisms from single cells. The originally individual cells joined together in a larger organism that became itself subject to evolution. This evolution does not only take place on the level of the individual cells anymore, but on the higher level of the individual cells and their interactions that make up the behaviour of the larger organism, hence the name level formation.

Steels conjectures that in language, level formation has been responsible for the combination of sounds into words, and for words into sentences. Both have made it possible for more complex and more varied messages to be coded in speech. The process of level formation has to be understood if one wants to explain how grammar and phonemic coding emerged.

2.2.4 Arguments against innateness of language.

Steels' view of language as a cultural phenomenon is at odds with the more accepted view that the basis of language is innate. Noam Chomsky is the best-known proponent of the innateness of language (see e.g. Chomsky 1965, 1972, 1975, 1980). He argues that all people have an innate *universal grammar* with which they can learn and use language. Because he was more interested in finding out what this universal grammar consists of, Chomsky never paid much attention to how it originated. Other researchers, most notably Steven Pinker (Pinker & Bloom 1990, Pinker 1995) have made a case for the possible evolution of universal grammar. However, there are a number of problems with the notion of innateness and the evolution of something as specific as a universal grammar. After many years of linguistic research, there is still no consensus at all about what the universal grammar would look like. Every possible theory is either too general, so it has no predictive power, or is too restricted, so it is refuted by actual linguistic data. Also, the mechanisms by which brains grow are too unspecific to code for something as precise as a universal grammar. This is not to say that there can be no specialisation at all in the brain, but it can not be so specific as is required to code for most proposed models of universal grammar. Sometimes data from language loss after lesions is used in order to support the innateness of language. When certain areas of the brain are damaged, certain linguistic functions disappear. Therefore, it is sometimes concluded, language must be innate. This reasoning is unsound, because there are also certain lesions (Kandel *et al.* 1991 pp. 849–850) that cause problems with reading and writing only (alexia and agraphia). However, nobody would want to argue that reading and writing are innate. Furthermore, evidence from brain lesions in young children indicates enormous flexibility. If the areas that are responsible for language in most adults (e.g. Broca's area and Wernicke's area in the left brain hemisphere) are damaged in very young children, these children will grow up to learn language almost perfectly (Stiles & Thal 1993, Johnson 1997 §6.2). This indicates that other regions can take over language functions, which is at odds with a very specific innate universal grammar.

Evolution of a very specific mechanism for language poses problems as well. How would a universal grammar evolve? In order for evolution to take place, there has to be a population of individuals with universal grammars, on which variations exist. Some of these variations should enable the individuals that have them to produce more offspring. Now it is not clear how individuals with a slightly more sophisticated universal grammar would benefit from this. Language is a population phenomenon, so any variation can only be beneficial if more individuals possess it. It is very unlikely that there would be many individuals with the same variations in their universal grammars.

Still, there is ample physical evidence that humans are adapted for language. The shape of the vocal tract, the unique control over breathing and the fine motor control over tongue and lips all indicate specialisations for language. However, the hypothesis that will be defended here is that cultural evolution of language is primary and drove biological evolution. Only after language became more complex

through cultural evolution and the other mechanisms proposed by Steels, it became beneficial to get biological adaptations to language. This way (proto-) humans could use it better and learn it faster. This is an example of the Baldwin effect (Baldwin 1896). Cultural mechanisms will therefore be considered as the primary factor in explaining the origins of language, and this is what will be focused on in the rest of this thesis.

The mechanisms proposed by Steels imply extremely complex dynamics. It is very hard to understand the implications of any theory that uses them. A crucial part of Steels' work is therefore the implementation of these theories as computer models. A large number of experiments have already been performed, most notably in the area of lexicon and concept formation (Steels 1995, 1996, 1997a, Steels & Kaplan 1998, Steels & Vogt 1997).

2.3 The Use of Computer Simulations

What does it mean to investigate linguistic phenomena with computer simulations? It can be argued that computer simulations can never capture the full complexity of human language and this would be right. However, one does not need to capture all complexity in order to get interesting results. In order to get an understanding of a phenomenon as complex as natural language, it has to be broken up into more manageable parts.

In fact this is what is being done in all natural sciences. To investigate a certain phenomenon, one first observes its behaviour. Then one makes a theoretical model of the phenomenon, including which parameters should influence its behaviour and one makes hypotheses about how the behaviour would change when the parameters are manipulated. Experiments can then be designed to manipulate the different parameters independently, in an artificial, controlled setting. In this way the hypotheses and the theory can be tested and refined iteratively. However, for phenomena as complex as language, it is quite difficult to do experiments in the way experiments can be done with simple physical systems. It is difficult to identify which parameters play a role. It is usually not possible to manipulate parameters, even if one suspects that they exist. Also, the interactions between different parameters can be so complex, that it does not make sense to manipulate them independently.

One of the big problems of linguistics research is therefore that it is relatively easy to make theories, but that it is very hard to test them. Traditionally, the only way to test linguistic theories was to make linguistic predictions and test with data from natural languages whether the theory held, or whether it was falsified. This is a very complex and time-consuming process and the results of linguistic observations can often be interpreted in different ways. Also, linguistic theories can become so complex that it is not always easy to see what linguistic behaviour they predict.

Computer models are therefore a useful tool for testing the implications of linguistic theories. They do not have difficulties with complex theories. Also, parameters in computer models can easily be adjusted. Linguistic theories can be implemented on computers and tested with corpora of real linguistic input. If the behaviour of the computer model corresponds with human behaviour, the underlying theory is not refuted. However, if there is a discrepancy between human behaviour and behaviour of the model, it is clear that the theory needs revision. In this way, competing theories can also be compared. The theory that has the best performance is probably the best model of actual human behaviour.

Especially theories of language as a collective behaviour, such as the one of Steels (Steels 1997b, Steels 1998b) described in the first part of this chapter and work by among others Batali (1998) and Kirby (1998), benefit from computer implementations. The results of repeated interactions in a large population of agents are almost impossible to predict without actually modelling them. In recent years, in the field of artificial life (see e.g. Langton 1989, Langton *et al.* 1990), a lot of successful research has been done on similar modelling of biological systems. Results have been achieved that would have been impossible without computer models. It is therefore justified to build computer simulations of linguistic phenomena.

2.4 The Research Questions

The main question that will be addressed in this thesis is whether the shape of human sound systems can be explained as the result of self-organising interactions in a population of language users. As has been pointed out above, the functional criteria of acoustic distinctiveness, articulatory ease and learnability can explain why human sound systems are the way they are, but they do not explain how they become this way. It has also been pointed out that it is unlikely that individual humans do the optimisation, and that optimisation is probably also not the result of biological evolution. Therefore, the only remaining explanation is that the optimisation is the result of self-organisation in the dynamic system consisting of the language users and their interactions. This is a system that is so complex that it has to be investigated by computer models.

Another question that will be addressed is whether the same mechanisms that explain the learning of language can also explain the emergence of language. It has been observed, for example in the emergence of sign language (Kegl & Iwata 1989, Kegl 1994, Senghas 1994, Senghas & Kegl 1994) in populations of deaf children, that language will only emerge in a sufficiently large population of people that are cognitively capable to use language. The computer model has to be able to generate a sound system (the basis of language) in a sufficiently large population. But the very same system has to be able, once a sound system has established itself, to transfer the sounds from one generation to the next.

The third question is in what way the model is sensitive to changes in parameters. What would happen, for example, if the noise level in the environment were changed? Or what happens if the properties of the auditory system of the agent were changed? Or if the population size changes or the rate with which new agents are added to the population, or are removed from the population? All such questions are extremely hard to investigate for a real language situation, but are quite simple once a reasonably realistic computer simulation exists.

This thesis does not pretend to provide final and definitive answers about human language. However, it is a first attempt to investigate the dynamics of sound systems as adaptive, complex dynamic systems.

3. The Simulation

Simulating the development of a system of speech sounds in a potentially large population of agents requires a computer model that is at the same time realistic and fast. These are two contradictory requirements. Realism can only be increased by doing extra calculations, which reduces speed. Consequently, it is necessary to make a compromise. The properties that are really essential for getting results that are comparable with human speech will have to be kept, and the rest will have to be sacrificed. However, if realism is sacrificed in a sensible way, the results of the simulation will still be comparable with observations of data from real human languages. Of course, in doing this, one has to keep in mind which parts of the simulation were realistic and which ones were artificial. This chapter describes the computer model that was used for investigating the formation of vowel systems in a population of agents. It also describes and defends the choices that were made between realism and speed. In order to understand the reasoning behind these choices better, a short history of the simulation is first presented.

3.1 The History of the Simulation

A simulation as complex as the one described in this chapter is not built or designed in one day. The present model is the end result of a number of different attempts to build a model that simulates the emergence of a system of speech sounds in a population of agents. All of the previous simulations were discarded for a number of reasons. However, the results obtained with these simulations have influenced the design of the present one as well as the way it is used. For this reason an overview of the history of the simulation is given here.

3.1.1 A first complex model

The idea of building a simulation that models interactions with human-like speech sounds was first put forward in a discussion with Luc Steels (Ardennes, October 1995). It was immediately clear that building a simulation that captures all complexity of human speech production and perception was impossible and way outside the scope of a PhD project. However, the first simulation that was tried was quite ambitious. It was equipped with a simple articulatory synthesiser, with dynamically moving articulators and with a model of perception based on formant frequencies.

The articulatory synthesiser was inspired by Ladefoged's observations (Ladefoged 1981 ch. 8, fig 12) of the dependence of formant frequencies on the positions of the articulators. He observed that first formant frequency seems to increase with decreasing height of a vowel and that the distance between the first and second formant frequency seems to increase with increasing frontness of a vowel, while lip rounding results in lowered first and second formant frequencies. But Ladefoged warns: "...measurements of formant frequencies are not so simply related to the other traditional labels high—low, front—back and spread—round." Nevertheless a crude synthesiser was based on these observations, together with data on the absolute formant frequencies of vowels.

The synthesiser worked as follows:

$$\begin{aligned} F_1 &= 900 - 500 \cdot h - 100 \cdot p - 100 \cdot r \\ F_2 &= 800 + 1000 \cdot h + 600 \cdot p - 200 \cdot r \\ F_3 &= 2200 + 1000 \cdot p \end{aligned} \qquad 3.1)$$

where h is the height of the tongue, p is the position of the tongue in the front-back dimension and r is the rounding of the lips. All parameters are supposed to lie be-

tween 0 and 1, where 0 is the lowest height, the most front and the least rounded, respectively. As can be learned from a cursory examination of the formula, it is not very realistic.

An acoustic signal consisted of 32 frequency bins. The frequency bins whose centre frequency was close to that of a formant frequency were filled with a higher value than those whose centre frequency was far away from the formant frequencies. The value in a frequency bin corresponded to the energy of that frequency. A number of different methods for distributing energy over neighbouring frequency bins were used, all of which were found to give similar results.

The agents could produce dynamic utterances; they could move their articulators while producing sound. This means that they could, in principle, produce simple consonant-like sounds. They could not produce noise, so fricatives and plosive bursts could not be produced. However, rapid formant transitions that gave a consonant-like impression could be produced. Dynamic movement of articulators worked by approaching target articulatory values in the following way:

3.2) $$p_t = p_{t-1} + \alpha(g - p_{t-1})$$

where p_t is the position at time t, α is a constant indicating the speed with which an articulator moves and g is the goal value.

The complex utterances were built up of smaller units, equivalent to phonemes. The agents therefore had to store items at two levels. They had to store a list of phonemes as well as a lexicon of possible words. Whenever an agent heard a certain sound sequence, it would analyse it in terms of the phonemes it knew. This would then result in a sequence of phonemes, which could correspond to an existing word or form a new word. Previously unheard words could be added to the lexicon with a certain probability.

Perception of sounds by agents was based on a heuristic to split up the speech stream and on a neural network. The input of the agent consisted of an essentially continuos speech stream. This stream had to be split up into discrete units (phonemes). It was split on the basis of the detection of certain events. The first of these events was the transition from presence of signal into silence or from silence into presence of a signal. The reasoning behind this was that presence of silence is what characterises stop consonants. The second event was triggered when the formant pattern was stable. This is characteristic of a vowel. For each event, the values for all 32 frequency bins were used as input for a neural network. This network was a simple perceptron (Hertz *et al.* chapter 5) with an output for every phoneme. The phoneme that corresponded to the output with the highest value was recognised. The network was trained by making the agents listen to themselves while they were speaking. The inputs to the network were calculated in the same way as when the agent listened to another agent. The response to be learnt was to make active the output node corresponding to the phoneme that was said at that moment, while making all other outputs inactive.

The interactions between the agents worked in the same way as the imitation game, which is described in more detail below. One agent would produce a word; the other would listen to it, analyse it in terms of its own phonemes and produce an imitation. The first agent would then check if the word it heard as imitation was the same as the word it originally said. If this was the case, the imitation was successful. If not, the imitation was a failure. Depending on the outcome of the interaction, the agents could add phonemes or words to their repertoire. They also kept score of the success of phonemes and words, and could occasionally throw away bad ones.

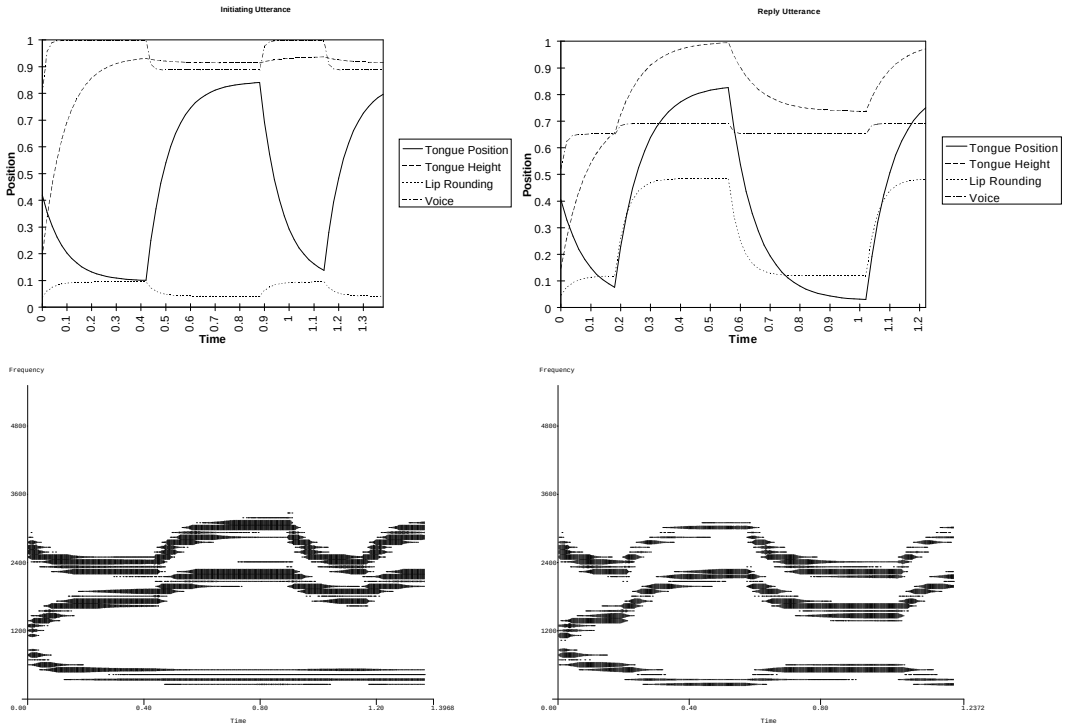


Figure 3.1: A sample conversation of the complex simulation.
Top left is the gestural score of the initial utterance, bottom left the acoustic image, top right is the gestural score of the imitating reply, bottom right is its acoustic image. Note that apparently the *timing* does not matter, only the position of the formant frequencies.

3.1.2 Results of the first complex model

This model, which was tested only with a "population" of two agents, occasionally produced good imitations (see figure 3.1) and the average success score of phonemes was around 30% (with inventories of up to five phonemes and up to a hundred words). This seems to indicate better than random performance, as with a repertoire of a hundred words one would expect a success rate of one percent when randomly selecting words. However, because of the way words of different length were compared, random performance would result in approximately 30% success, which made the simulation's performance no better than random. Any good imitations were mostly due to coincidence.

This observation teaches that not all results should be trusted at face value. It is essential to analyse the performance of the agent in the case that it chooses its behaviour without paying attention to the other agent's behaviour (resulting in either random or simple systematic behaviour). The experiments can only be considered successful if agent behaviour is observed that is significantly better than random or simple systematic, input-less behaviour.

3.1.3 A Feature-based model

The complex simulation did not work satisfactorily. It did not result in a shared system of speech sounds. As no stable systems of sounds were reached, nothing could be said about the structure of these systems, either. There were several reasons for this failure. The model was too complex, because it worked with dynamic utterances, which had to be represented at multiple levels. Also the perception part, with its neural network, introduced too many extra parameters. At the same time, the sound production was too simplistic and unrealistic. A simpler, more controllable model was needed.

			[–fric]		[+fric]	
			[–v]	[+v]	[–v]	[+v]
[–back]	[–sh]		p/m̥	b/m	f/m̥	v/m
	[+sh]		t̪ n̥	d/n	s/n̥	z/n
[+back]	[+sh]		c/ɲ̊	ɟ ɲ	ʃ/ɲ̊	ʒ/ɲ
	[–sh]		k/ŋ̊	g/ŋ	x/ŋ̊	ɣ/ŋ

Table 3.1: Example of [+cons] phonemes. Per entry [–nasal]/[+nasal] are shown.

This model was completely based on distinctive features. Phonemes were coded as strings of binary features. Each phoneme has eight features. For consonants they are the following: [+consonantal], [+/–nasal], [+/–back], [+/–sharp], [+/–voice], [+/–fricative], [+/–long] and [+/–high tone]. For vowels they are: [–consonantal], [+/–nasal], [+/–back], [+/–sharp], [+/–low], [+/–rounded], [+/–long] and [+/–high tone]. These features were not chosen to be realistic from a linguistic point of view, rather they were chosen to make it easy for humans to interpret the phonemes that were generated by the agents. The possible consonants are shown in table 3.1, the possible vowels in table 3.2.

			[–round]		[+round]	
			[–low]	[+low]	[–low]	[+low]
[–back]	[+sh]		ɪ̫ ˜	ɛ/ɛ̃	y/ỹ	œ̫/œ̃
	[–sh]		e/ẽ	a/ã	ø/ø̃	œ̫/œ̃
[+back]	[–sh]		ɤ/ɤ̃	ʌ/ʌ̃	o/õ	ɔ/ɔ̃
	[+sh]		ɯ̫/ɯ̃	ɑ/ɑ̃	u/ũ	ɒ/ɒ̃

Table 3.2: Example of [–cons] phonemes. Per entry [–nasal]/[+nasal] are shown.

In this model, as in the previous one, agents produce utterances consisting of multiple phonemes. In real speech, phonemes in sequence influence each other. This can either happen because of language dependent rules, such as the ones that determine that final stops and fricatives in German and Dutch words will be devoiced, or because of the dynamic nature of the movements of the articulators (Browman & Goldstein 1995). In the feature-based simulation, a subset of the features of any phoneme determined how neighbouring phonemes would influence each other. This subset of features of a phoneme was called the subset of crucial features of a phoneme. These were the features that were considered to be essential to the recognition of this phoneme. If they were present, the phoneme was recognised. In English, for example, the crucial feature for nasal consonants is often only their nasality, so that they assimilate in place of articulation ("impossible" becomes [ˌɪmˈpʰɔsəbl̩], but "incorrect" [ˌɪŋkəˈɹɛkt]) or they loose their consonantal character completely and can only be observed through the nasalisation of the preceding vowel ("can't" becomes [kʰæ̃t]). In this model, crucial features are not only crucial for the recognition of phonemes, they can also influence neighbouring phonemes. If a phoneme occurs in between two phonemes (or in between a phoneme and a word boundary) and the neighbouring phonemes both have the same crucial feature with the same value, the phoneme takes over their value for this feature. So for example in the sequence: /mam/ where both /m/'s have the feature [+nasal] marked as crucial, the/a/ would become [ã].

3.1.4 Results of the feature-based model

This model resulted in successful imitations when used in a population of two agents. Words of average length three were created. This is not very long, but suffi-

cient for co-articulation to play an important role. Lexicons up to 600 words and phoneme sets of up to 15 phonemes were generated. Success of word imitation went up to 70%, which in this case was much better than random behaviour. Good similarities between the sound systems of the agents are found if one looks at the most used (minimally 5 times) and most successful (minimally 70% successful of all the times it was used) phonemes. For example [ɵː], [ɛ́], [ɔ́], [ɲ̀], [ŋ́ː], [ɟ́ː] and [ɔ́ː] for the first agent and [ɵː], [œ́], [ɔ́], [ɲ̀], [ŋ́ː], [ɡ́ː] and [ỳː] and [ɤ̀ː] for the second agent. The diacritics [́] and [̀] stand for [+high tone] and [-high tone] respectively. The similarity is not perfect, but [ɛ́] and [œ́] differ only in one feature and so do [ɟ́ː] and [ɡ́ː]. The case of [ɔ́ː] in the first agent and [ỳː] and [ɤ̀ː] in the second agent is less clear-cut. They probably correspond to multiple phonemes in the other agent, depending on how they are modified by the context.

This simulation thus resulted in shared sound systems in agents that started out with no sound system at all. Unfortunately, due to the arbitrary nature of the signals, which were essentially bit strings, and due to the arbitrary nature of the way in which signals could influence each other, the resulting sound systems had no relevance whatsoever to understanding human sound systems. One possible way of remedying this situation is to introduce a set of more realistic features and more realistic rules of how sounds can influence each other. However, there are a number of drawbacks to this approach. First of all, representing speech sounds as features is not without its problems (see the previous chapter). Secondly, choosing a set of features, even though they are linguistically motivated, still depends on more or less arbitrary decisions. And finally, determining rules of how sounds influence each other is extremely difficult. Speech sounds can influence each other in a large number of ways, some of which are language dependent and some of which seem to be universal. Linguists do not have a complete picture of these influences, yet. Therefore, any set of rules to model this influence will be incomplete and to some extent arbitrary. A different approach to modelling speech sounds was therefore adopted.

This approach was inspired by the work of Hervé Glotin (Glotin 1995, Berrah *et al.* 1996, Glotin & Laboissière 1996) at the *institut de communication parlée* in Grenoble. In his work a population of agents develops a set that consists of vowels only. Vowels can be modelled as single utterance, so there is no need for a separate word-level. Also, vowels are easy to synthesise. Furthermore, a lot is known about the universal tendencies of vowel systems of the world's languages. Therefore it was decided to first build a simulation to investigate the development of vowel systems in a population of agents. When this would work, and the dynamics were well understood, an extension to more complex utterances could be undertaken.

3.2 Purpose of the Simulation

The purpose of the simulation is to investigate the emergence of a vowel system in a population of agents that learns to imitate each other as successfully as possible with an open system of vowel sounds. The agents' production, perception and learning of speech sounds should be as human-like as possible. Each agent should be able to produce, perceive and remember a set of realistic vowels. It should be able to engage in interactions with other agents and to learn and adapt its system of vowels from these interactions. The number of vowels it knows or their positions should not be determined beforehand. Once an agent has developed a vowel system that works, it should keep this system, without altering it too much. In a group, the agents

should be able to generate such a system from scratch. The aim is not to model the exact way in which human vowel systems emerge and change historically, but to investigate whether a population is in principle able to develop a coherent set of vowels from scratch, and whether the sets of vowels that emerge show the same universal tendencies as human vowel systems.

In order to keep the simulation tractable a number of things should *not* be modelled. First of all, the utterances of the agents do not have any meaning. They are just sounds. The goal of the agents is to imitate the other agents as well as possible. This is considered to be basic to language; only if you are capable of identifying and imitating the other speaker's sounds, can you begin to learn the meaning that is attached to the sounds. Other researchers (Steels 1997a, Steels & Vogt 1997, Gasser 1998) are investigating the origins of meaning and the way in which meanings can be coupled to words.

The question why agents would want to communicate with language, and thus to imitate, is not posed either. In the work presented here, the need for communication with language is assumed as a given. Other researchers (Hauser 1997, De Jong 1998) are investigating the origins of communication with language. Having the agents develop the need to imitate would complicate the model needlessly; this need is therefore pre-programmed.

The drive to add new sounds to the inventory is also pre-programmed. It is needed, because the agents start out with empty sound systems, but still have an urge to imitate. It is therefore necessary to add new sounds every once in a while, in order to get the imitations started. In a natural language one can imagine that addition of new sounds would be driven by the need to distinguish as many meanings as possible, while keeping the length of utterances low. In order to make more distinctions, more sounds and an effective use of the available acoustic space is necessary. This is an example of a case where one part of language, the lexicon, exerts pressure on another part of language, the sound system.

The many subtle social factors that determine the use and change of human speech sounds will also not be modelled for two reasons. First of all, they are extremely hard to model. There is no clear picture of which factors are important, nor how they influence the use of sounds in human languages exactly. Secondly, social factors are important for explaining the specific historical sound changes that particular human languages have undergone, but that for the sake of modelling it is assumed that these factors can be considered random fluctuations. As far as is known, social factors can determine which one of a number of variations will be chosen. However, the variation itself will be random (but biased by factors such as articulatory ease and perceptual distinctiveness).

Finally, only utterances of single vowels will be modelled. This has the advantage of being easily implementable, but the disadvantage of being unrealistic. However, vowel systems are often investigated without taking into account the contexts in which the vowels can appear. This is true for much of the work on explaining the universals of vowel systems (Liljencrants & Lindblom 1972, Schwartz *et al.* 1997). The predictions of these models are quite accurate, so it can safely be concluded that for predicting vowel systems the context in which the vowels can appear does not play a very important role. However, the context *does* influence the possible ways in which vowels can change historically. In order to model realistic historic sound change, sequences of sounds will have to be modelled. This means that al-

though the model presented here can in principle be used to predict vowel systems, it can not be used to model *change* of vowel systems accurately.

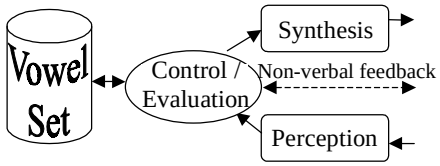


Figure 3.2: Agent architecture.

3.2.1 Agent architecture

The agents should be equipped with an accurate articulatory vowel synthesiser, a realistic model of perception and an associative memory to store vowel prototypes (see figure 3.2). Vowel prototypes consist of an articulator position with an associated acoustic signal. The articulatory prototype of a vowel consists of the three major vowel parameters, position, height and rounding. The acoustic signal consists of the first four formant frequencies of the vowel. The reason for storing articulatory representations of vowels and for using an articulatory synthesiser to produce acoustic signals, rather than to just store acoustic representations of vowels, (such as Berrah (1998) does) is that this is probably the most realistic way. People have control over the way they move their articulators, not over the sounds they perceive. It would therefore seem likely that children have to learn which sounds are produced for which articulatory movements. As the mapping between acoustic signals and articulatory gestures is complex, this is a non-trivial learning task. Also, when an unknown sound is heard, it is unlikely that it can be reproduced exactly, because it has to be analysed first in terms of articulator movements. Finally, if one uses an articulatory synthesiser, one does not have to worry about the limits of the acoustic space available for communication. The limits of the acoustic space are automatically determined by the limits of the articulators.

The acoustic signal associated with the articulatory prototype is used for recognition only. Incoming signals are compared with the stored acoustic signals, instead of with signals that are generated from the articulatory prototypes with the articulatory synthesiser. This is done to reduce the amount of computation needed. Every time the agent produces a vowel, however, a new acoustic signal is generated with the articulatory synthesiser, with noise added.

Depending on the outcome of the interactions with other agents, the agent can either add or remove prototypes from its memory, or shift existing prototypes. The exact mechanisms for producing and perceiving vowels, as well as the exact interactions between the agents are described in the sections that follow.

The reason imitation was chosen as a model for investigating the emergence of sound systems in a population is that it is the simplest way to capture the complexity of learning of a sound system. It does involve recognising and distinguishing sounds, but there does not have to be meaning to the sounds. There is also no need for explicit optimisation of the vowel system. But as (near-) optimal vowel systems are easier to learn and imitate than non-optimal sound systems, they have the advantage and will be adopted by the agents more easily. The human-like perception and production of speech will take care that systems that are optimal for the agents are also optimal for humans. Self-organisation thus takes care that near-optimal systems are found more frequently than sub-optimal ones. Optimal systems could be considered *attractors* of the dynamic system that is formed by the agents (their perception and their production) and the interactions between the agents.

3.3 The Articulatory Model

The articulatory model of the agents maps the articulatory representation of vowels to an acoustic representation. The articulatory representation consists of three parameters, *position*, *height* and *rounding*, corresponding to the three major vowel parameters (Ladefoged & Maddieson 1996, ch. 9). They are real numbers between zero and one. For position, corresponding to the position of the tongue in the front-back dimension, zero means most front, while one means most back. For height, corresponding to the height of the highest point of the tongue, and thus the openness of the vowel, zero means most open, while one means most closed. Rounding corresponds to the rounding of the lips. Zero means most spread, while one means most rounded. Thus the vowel [a] has values (0, 0, 0) for its parameters (position, height, rounding) while [i] has (0, 1, 0) and [u] has (1, 1, 1).

The acoustic signals that are exchanged between agents are represented by the first four formant frequencies (F_1, F_2, F_3, F_4) of the vowel. The first three or four formant frequencies are usually considered to be sufficient to represent a vowel. As calculating just the formant frequencies is much faster than calculating an actual signal, this increases the speed of the simulation. Also, the perceptual distance metric that was used (see next section) was based on formant frequencies. Obviously, a real vowel signal contains much more information than just the formant frequencies. Other properties of the vowel signal are, for example, volume and frequency contours, duration, voicing characteristics and formant bandwidths. Although some of these properties do have linguistic relevance, they do not influence the perception of the vowel *quality* much. Vowel quality is the only property that is investigated in this thesis.

The articulatory model has to be realistic as well as fast. It has to be realistic so that the results of the simulations can be compared with observations of real languages and it has to be fast so that the simulations can be run interactively. A full articulatory synthesiser, such as Maeda's (1989) or Mermelstein's (1973) was therefore out of the question. Modelling the vocal tract's area function from the large number of degrees of freedom of these models and then calculating the formant frequencies for the area function would be too computationally intensive. Also, it was not clear how to map the rather abstract parameters of position, height and rounding to the degrees of freedom of these models.

It was therefore decided to calculate the formant frequencies from the articulator positions directly. This can be done with an interpolation function. The interpolation was based on information on formant frequencies of a large number of vowels in Vallée's thesis (Vallée 1994, pp. 162–164). These vowels had been artificially generated using the Maeda articulatory synthesiser. A subset of these vowels, for three degrees of position and height and for two degrees of rounding was used. The values of the articulatory parameters were assigned to the vowels according to the phonetic symbols that were used in Vallée's list. The data points that were used, with their articulatory and acoustic representations are shown in table 3.3. Note that two data points, the [œ] and the [ɐ̟] did not appear in Vallée's list. Also, the notation used for the low front vowels is different in this thesis than in Vallée's thesis. The original notation is given between parenthesis.

Vowel	p	h	r	F_1(Hz)	F_2(Hz)	F_3(Hz)	F_4(Hz)
[a] ([a̟])	0	0	0	708	1517	2427	3678
[ɶ]*	0	0	1	670	1400	2300	3500
[ɐ] ([a])	0.5	0	0	742	1266	2330	3457
[ɐ̟]*	0.5	0	1	658	1220	2103	3200
[ɑ]	1	0	0	703	1074	2356	3486
[ɒ]	1	0	1	656	1020	2312	3411
[e]	0	0.5	0	395	2027	2552	3438
[ø]	0	0.5	1	393	1684	2238	3254
[ə]	0.5	0.5	0	399	1438	2118	3197
[ɵ]	0.5	0.5	1	400	1267	2005	2996
[ɤ]	1	0.5	0	430	1088	2142	3490
[o]	1	0.5	1	399	829	2143	3490
[i]	0	1	0	252	2202	3242	3938
[y]	0	1	1	250	1878	2323	3447
[ɨ]	0.5	1	0	264	1591	2259	3502
[ʉ]	0.5	1	1	276	1319	2082	3118
[ɯ]	1	1	0	305	1099	2220	3604
[u]	1	1	1	276	740	2177	3506

Table 3.3: Data points for articulatory synthesiser.

$$
\begin{aligned}
F_1 &= \left((-392+392r)h^2+(596-668r)h+(-146+166r)\right)p^2 + \\
&\quad \left((348-348r)h^2+(-494+606r)h+(141-175r)\right)p + \\
&\quad \left((340-72r)h^2+(-796+108r)h+(708-38r)\right) \\
F_2 &= \left((-1200+1208r)h^2+(1320-1328r)h+(118-158r)\right)p^2 + \\
&\quad \left((1864-1488r)h^2+(-2644+1510r)h+(-561+221r)\right)p + \\
&\quad \left((-670+490r)h^2+(1355-697r)h+(1517-117r)\right) \\
F_3 &= \left((604-604r)h^2+(1038-1178r)h+(246+566r)\right)p^2 + \\
&\quad \left((-1150+1262r)h^2+(-1443+1313r)h+(-317-483r)\right)p + \\
&\quad \left((1130-836r)h^2(-315+44r)h+(2427-127r)\right) \\
F_4 &= \left((-1120+16r)h^2+(1696-180r)h+(500+522r)\right)p^2 + \\
&\quad \left((-140+240r)h^2+(-578+214r)h+(-692-419r)\right)p + \\
&\quad \left((1480-602r)h^2+(-1220+289r)h+(3678-178r)\right)
\end{aligned}
$$

Figure 3.3: Synthesiser equations.

* This vowel does not appear in Vallée's (1994, pp. 162–164) list.

As there were three degrees of position and height, a quadratic interpolation had to be used for these dimensions. There were only two degrees of rounding in the data set, so here a linear interpolation was used. The resulting three-dimensional, quadratic-linear interpolation function is given in figure 3.3.

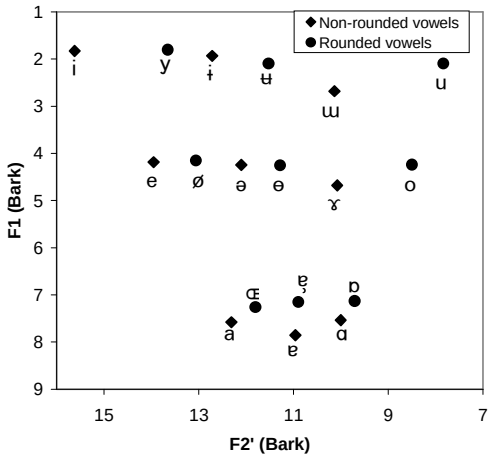


Figure 3.4: Vowels in F1-F2' space.

Although this is a rather crude way of solving the problem of articulatory synthesis, it is nevertheless an effective way. The formant patterns that can be generated sound natural to human ears if synthesised. Even vowels that were not used as data points, such as [ɛ] or [ɔ] sound natural. All formant frequencies that can be generated lay within the formant space that can be generated by humans (figure 3.4). The main advantage of the method, however, is that it is fast. Only 70 multiplications and 68 additions are required to calculate the four formant frequencies of one vowel; 17 multiplications and additions per formant and two more multiplications for calculating p^2 and h^2. This makes the articulatory model very well suited for the kinds of simulations that have to be done, in which very many interactions between agents have to be modelled in a limited time.

3.3.1 The addition of noise

In order to make the simulation more realistic the synthesis of the signal will have to be made noisy. In human communication, no signal will ever be exactly the same as a previously generated signal, due to slight perturbations in articulation, speaker differences and environmental noise. Two sources of noise were therefore added to the synthesiser: articulatory noise and acoustic noise. Articulator noise was modelled by adding a small random value to all three articulatory parameters. This value is taken from the uniform distribution with the range: $\left[-\frac{\psi_{art}}{2}, \frac{\psi_{art}}{2}\right)$, where ψ_{art} is the amount of articulatory noise present, a parameter of the simulation. Acoustic noise is added by shifting formant frequencies up or down a random amount. As human perception of pitch is logarithmic, this means that in order to have the same auditory effect, higher frequencies will have to be shifted more than lower frequencies. This is done in the following way. For every formant i, a random value v_i is taken from the uniform distribution: $\left[-\frac{\psi_{ac}}{2}, \frac{\psi_{ac}}{2}\right)$, where ψ_{ac} is the amount of acoustic noise present, also a parameter of the simulation. Now every formant is modified in the following way:

3.3) $\quad \tilde{F}_i = F_i(1 + v_i)$

where F_i is the originally calculated formant frequency (expressed in Hertz) and $\tilde{F}_i$ is the formant frequency with noise. This causes the formants to be shifted proportionally to their frequency.

With this articulatory model, agents are able to produce vowels in a continuous acoustic and articulatory space. The mapping between articulatory parameters and acoustic parameters is both fast and realistic. The addition of noise to either

articulatory or acoustic parameters makes sure that no two signals that are exactly the same will ever be produced. Although the random noise is uniformly distributed, it is probably a good model of variation of human speech, nevertheless. This makes the articulatory synthesiser very well suited for the research goals that are pursued.

3.4 The Perception Model

Agents should not only be able to produce vowels in a realistic way; they should also perceive vowels in a human-like way. Humans have the tendency to analyse speech sounds they hear in terms of sounds they already know. Speech sounds that are unlike, but close to sounds they already know will be interpreted as familiar speech sounds. A speaker of English, for example, will confuse the French vowel /y/ with either /i/ or /u/. In French these three sounds contrast: /ny/ "nu" (naked), /ni/ "nid" (nest) and /nu/ "nous" (we) are different words, but in English the /y/ does not appear, whereas /i/ and /u/ do. Whenever (linguistically naïve) English speakers hear /y/, for example in the French word "nu" they will think they heard either /i/ (as in "knee") or /u/ (as in "gnu") depending on the context.

Research into the perception of consonants (Cooper *et al.* 1976, Liberman *et al.* 1976) has shown that when acoustic signals, consisting of artificially generated consonant-vowel sequences, are changed continuously, subjects will perceive one consonant for one range of parameter settings, while other consonants are perceived for other settings of the parameter. The perception changes abruptly when the parameter changes from one range into the other. Which consonants are perceived for given parameter settings depends on the native language of the subjects. Apparently humans perceive speech in terms of *prototypical* sounds. In other parts of language, such as syntax and semantics, prototypical perception seems to take place as well (see e.g. Comrie 1981, Lakoff 1987).

The agents should display similar behaviour. They have a list of vowel prototypes. Whenever they perceive an acoustic signal, they find the (acoustic) prototype that is closest to the signal they perceived. This prototype is then considered the phoneme that was heard, even though in fact it can sound quite different, especially if the agent has very few prototypes. The crucial part of recognising acoustic signals is therefore calculating their distance to the acoustic prototypes. Whenever the distance function is realistic, the perception will be realistic.

3.4.1 Calculating the distance between vowels.

The distance function is based on a weighted Euclidean distance between representations of vowels consisting of the first formant frequency and the effective second formant frequency, measured in Barks[1]. The notion of the *effective second formant* (see for a particular insightful explanation of how the effective second formant can be derived from properties of the human auditory system Lindblom *in press*) is inspired by the way human perception of vowels works. If human subjects are asked to approximate vowel sounds as closely as possible, using only two formant frequencies, it is found that the first formant frequency that is chosen corresponds closely with the first formant frequency of the vowel. However, the second formant frequency that is chosen does not always correspond to the position of the second for-

[1] The Bark scale is an (approximately) logarithmic frequency scale that models the human perception of pitch. Pairs of sounds that are perceived to have equal distance in pitch have equal distance in Barks, no matter what their absolute frequency is.

mant frequency in the actual signal. Sometimes it is located between the second and the third formant frequency, sometimes even between the third and the fourth (Carlson *et al.* 1970, Lindblom *in press*). The perception of a four formant pattern as a pattern of two formants is caused by the fact that in the high-frequency range, human hearing is not able to resolve peaks with narrow bandwidth. The signals at different frequencies effectively merge into each other. If there are two or more peaks close together in the actual signal, they will actually be observed as one wide peak, located approximately in between. It should be noted, however, that not all four formant patterns are perceived as two formant patterns. If the higher formant peaks are sufficiently far apart, they will be perceived separately. The effective second formant model is therefore not a totally accurate model of human perception. However, it does work most of the time.

There are different ways to calculate the effective second formant frequency. The one that was adopted here (see Lindblom *in press* for a different one) is the one that is also used at the *institut de communication parlée* of Grenoble (Mantakas *et al.* 1986, Boë *et al.* 1995). This particular model was adopted because it gives natural results and because it makes it easier to compare the results of the simulations presented here with the Grenoble simulations (Vallée 1994, Boë *et al.* 1995, Schwartz *et al.* 1997, Berrah 1998).

The effective second formant, F_2' is calculated as a non-linear, weighted sum of the second, third and fourth formants. It is based on a critical distance between formant peaks. This critical distance models bandwidth of the human perception at the higher frequencies (and therefore the confusion of formant peaks). The critical distance is taken to be 3.5 Bark. If the distance between F_2 and F_3 is more than the critical distance, the actual F_2 is used as F_2'. If the distance between F_2 and F_3 is smaller than the critical distance, but the distance between F_2 and F_4 is more than the critical distance, then F_2' is taken to be a weighted average of F_2 and F_3. If the distance between F_2 and F_4 is also less than the critical distance, then F_2' is taken to be the weighted average of either F_2 and F_3 or F_3 and F_4, whichever are closer together.

The weights in the original formula were determined by the strengths of the formant peaks. As the articulatory model of the agents does not calculate the strengths of the formants, the weights are calculated depending on the distance between the formants (as in general, formants that are close to other formants tend to be stronger). This is done as follows:

3.4) $$w_1 = \frac{c - (F_3 - F_2)}{c}$$

3.5) $$w_2 = \frac{(F_4 - F_3) - (F_3 - F_2)}{F_4 - F_2}$$

where c is the critical distance. The calculation of F_2' can then be expressed in the following formula:

3.6) $$F_2' = \begin{cases} F_2, & \text{if } F_3 - F_2 > c \\ \frac{(2 - w_1)F_2 + w_1 F_3}{2}, & \text{if } F_3 - F_2 \le c \text{ and } F_4 - F_2 > c \\ \frac{w_2 F_2 + (2 - w_2)F_3}{2} - 1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 < F_4 - F_3 \\ \frac{(2 + w_2)F_3 - w_2 F_4}{2} - 1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 \ge F_4 - F_3 \end{cases}$$

and the distance between two vowels can be calculated as follows:

3.7) $$D = \sqrt{\left(F_1^a - F_1^b\right)^2 + \lambda\left(F_2^{a\prime} - F_2^{b\prime}\right)^2}$$

In this formula, the parameter λ represents the factor with which the effective second formant is weighted relative to the first formant. As F_1 is proportional to vowel height, and F_2' proportional to vowel position, this factor tends to determine the accuracy with which agents can distinguish between different vowel positions and different vowel heights. The higher λ is, the more distinctions an agent will be able to make in the front-back dimension as opposed to the high-low dimension. Experiments with determining maximally dispersed vowel systems (Vallée 1994, Schwartz *et al.* 1997) as well as independent data from human production of vowels (Lindblom & Lubker 1985) seem to indicate that this factor should be approximately 0.3

The way in which the weights w_1 and w_2 are calculated and used in this work was recently found to introduce some irregularities in the perceptual space. Because for different configurations of the four formant frequencies, different functions are used for calculating the effective second formant, discontinuities sometimes arise in F_2' when the formant pattern itself changes continuously. Most simulations had already been performed when this was discovered, and it was found that the influence on the qualitative outcome of the experiments was not important. Therefore no attempt was made to correct this problem. Whenever it has influenced the results of the simulations this will be noted. However, for future research it is advisable to use a weighting function that does not have discontinuities in order to avoid all problems.

The distance function is used to calculate the distances between the perceived signal and the acoustic prototypes of all the vowels. The agent recognises the vowel that is closest to the perceived signal. As the distance function assigns large distances to signals that are perceived very differently by humans and small distances to signals that are perceived to be very similar by humans, it is a good model of human perception. The distance function is also used in approximating a new, unknown signal. The distance between the unknown signal and a newly added vowel is minimised using an algorithm that will be described in more detail in the next section.

3.5 The Imitation Game

The articulatory model and the perception model determine the kinds of sounds the agents can produce and perceive. The dynamics of the model, however, are determined by the way the agents use these sounds. In the simulations presented here, sounds are used in so-called *imitation games*. Imitation games are played between two agents whose goal is to imitate the other agent as well as possible. The imitation game is based on the idea of language games as introduced by Steels (1995, 1997b, Steels 1998b). Language games are interactions between two (or more) agents that follow definite rules and of which it can be determined unequivocally whether they were successful or not. Depending on the outcome of the language games, the agents update their knowledge of the language. Steels was inspired by Wittgenstein's philosophical theories (Wittgenstein 1967) as well as by research by Suzuki and Kaneko (1994) on artificial bird songs when developing the idea of language games. The "rules" of the imitation game, as well as the agents' reaction to them will be described in this section.

For an imitation game, two agents are randomly chosen from the population. One of these will get the role of *initiator;* the other will get the role of *imitator*. The initiator randomly chooses a vowel from its inventory. It then synthesises the acoustic signal that corresponds to this vowel and noise is added to it. The imitator, who hears the perturbed signal, analyses it in terms of the vowels in its inventory. It finds the closest one (using the distance measure described in the previous section) and then synthesises the acoustic signal that corresponds to this vowel, also adding noise. Note that even without the presence of noise, the imitation can sound quite different from the sound produced by the initiator, especially when the vowel inventories of the two agents are different. The initiator in turn listens to this sound and analyses it in the same way. If it turns out that the closest match to the sound it heard is the same vowel prototype as the one it originally used to initiate the game, the imitation game is successful. If the vowel the initiator perceives is a different one from the one it produced, there is confusion and the imitation game is a failure. The initiator then communicates the success or the failure to the imitator using non-linguistic communication. This might seem unrealistic, as humans do not learn language by being told that their utterances are right or wrong all the time. However, the non-verbal feedback is only an abstract way of letting the imitator know whether its imitation was right. In actual human communication, this feedback could be whether the intended goal of the communication has been achieved. If someone does not use the right sounds in an utterance, he or she will not be understood and it will be clear that the communication was a failure. The appropriate reaction would then be to update his or her knowledge of the sounds. The steps of the imitation game are given in pseudo-code in table 3.4.

The goal of the agents is to imitate each other as well as possible. For this they need to develop repertoires of sounds that are similar to the ones of the other agents. The only way they can learn about the other agents' repertoires is through the imitation games. They should therefore use the outcome of the imitation games for improving their vowel systems. First of all, both the imitator and the initiator keep track of the number of times each of their vowels has been used and the number of times it has been used in successful imitation games. These are called the *use* and *success* counts, respectively. The ratio of these two counts is a measure of the successfulness of a vowel. The successfulness of a vowel is mainly a measure of how well it is shared by all agents in the population. If not many agents have it, it will not be imitated successfully very often. On the other hand, if many agents do have it, it will be imitated successfully most of the time.

The imitator makes the most important changes to its vowel system in response to the imitation game. If the imitation game was successful, it shifts the vowel that was used so that its acoustic signal will match the observed acoustic signal more closely. This is done in order to improve coherence in the population. Changing pronunciation in order to match others more closely is behaviour also observed in humans. As the agents can only directly manipulate the articulatory representations of vowels, and not their acoustic ones, they have to use a trick in order to optimise the acoustic signal of the vowel. The trick is similar to the way people learning a new language try out small variations on an unfamiliar sound in order to improve its pronunciation. It consists of trying out the six neighbours of the vowel in articulatory space and keeping the one that most closely matches the acoustic signal. The six neighbours of a vowel are the vowels that differ from it in each of its three articulatory parameters by a specific small amount in either the negative or

positive direction. The value of the small amount with which vowels are shifted is a parameter of the simulation, which is called ε. In the section on results, it will be investigated how this parameter influences the vowel systems that are found.

If the imitation game was unsuccessful, there are two possible actions the imitator can undertake. If the vowel it used is successful (it has a high success/use ratio) this means that there must be other agents that are using it. Changing it too much might therefore not be a very good idea. The best assumption to make about why the imitation game failed is that the other agent has two vowels in the same space where this agent only has one. The course of action to take is hence to add a new vowel to the inventory that closely matches the acoustic signal that was observed. It is not possible to find the right articulatory parameters from an acoustic signal directly. A new vowel is therefore added in the middle of the articulatory space (all parameters are set to 0.5) and the vowel is shifted closer to the acoustic signal repeatedly in the same way as described in the previous paragraph. This is repeated until no more improvement is possible.

The other possible case is that the vowel was unsuccessful in previous imitation games. Then it is assumed that it is not shared by other agents, and it is shifted towards the acoustic signal that was observed, in the hope of bringing it close enough so that next time it will be successful. The threshold above which vowels are considered successful enough, so that a new phoneme will have to be added is another parameter of the simulation, θ_s.

The agents can also make modifications to their vowel inventories that are not directly driven by the outcome of a particular imitation game. The first modification is a cleanup of the vowel system. In this cleanup all vowels that have been used a minimum number of times, so they have had the occasion to be tested, and of which the successfulness is still lower than a threshold, are removed from the inventory. Both the threshold θ_c and the minimum number of times a phoneme has to be used, θ_u are parameters of the simulation. The whole population is cleaned up with a probability of p_c every imitation game. Another modification is that two vowels in the inventory will be merged if they come so close in either articulatory or acoustic space that they will always be confused by the noise that is added to the articulations. This was found to be necessary to prevent large clusters of unsuccessful vowels clustering around the positions where only one good vowel was necessary. Merging is done by throwing away the worst one of the two vowels that are too close. The use and success counts of the vowel that is kept are increased with the use and success counts of the vowel that was thrown away. The last modification that agents can make to their inventories is to add a new random vowel. This is done with a low probability, p_i that is also a parameter of the simulation. It is done in order to keep a pressure on the agent to utilise the acoustic space maximally. The ways in which agents modify their vowel inventories and the way in which acoustic signals are approximated are described in pseudo-code in tables 3.6 and 3.5.

All of these actions use only local information. Agents only use information about the signals they perceive and the information they have about their own vowel systems. The modifications to the vowel system do not use information of the vowel system as a whole, only about individual vowels or about neighbouring vowels that are close together in the case of merging. The modifications the agents can make do not require an unrealistic amount of computation, either. So, even though the simulation that will be used in this thesis is and does not pretend to be an accurate

model of human language learning, no completely unrealistic hat-tricks are used to make the sound system appear, either.

Table 3.4: Basic organisation of the imitation game.

initiator	imitator
If ($V = \emptyset$) Add random vowel to V	
Pick random vowel v from V $u_v := u_v + 1$ Produce signal $A_1 := ac_v$	
	Receive signal A_1. If (V = $\emptyset$) Find phoneme(v_{new}, A_1) $V := V \cup v_{new}$ Calculate v_{rec}: $v_{rec} \in V \wedge \neg \exists v_2 : (v_2 \in V \wedge D(A_1, ac_{v2}) < D(A_1, ac_{v\,rec})$ Produce signal $A_2 := ac_{vrec}$

Receive signal A_2.
Calculate v_{rec}:
$v_{rec} \in V \wedge \neg \exists v_2 : (v_2 \in V \wedge D(A_2, ac_{v2}) < D(A_2, ac_{v\,rec})$
If ($v_{rec} = v$)
Send non-verbal feedback: *success*.
$s_v := s_v + 1$
Else
Send non-verbal feedback: *failure.*	

Do other updates of V.	Receive non-verbal feedback. **Update V according to feedback signal**. Do other updates of V.

Table 3.5: Other updates of the agents' vowel systems.

Merge(v_1, v_2, V)
{
If ($s_{v1}/u_{v1} < s_{v2}/u_{v2}$)
$s_{v2} := s_{v2} + s_{v1}$
$u_{v2} := u_{v2} + u_{v1}$
$V := V - v_1$
Else
$s_{v1} := s_{v1} + s_{v2}$
$u_{v1} := u_{v1} + u_{v2}$
$V := V - v_2$
}
	Do other updates of V
	{
	For ($\forall\ v \in V$) // Remove bad vowels
	If ($s_v/u_v <$ *throwaway threshold* $\wedge\ u_v >$ *min. uses*)
	$V := V - v$
	For ($\forall\ v_1 \in V$) // Merging of vowels
	For ($\forall v_2$: ($v_2 \in V \wedge v_2 \neq v_1$))
	If (D(ac_{v1}, ac_{v2}) < *acoustic merge threshold*)
	Merge(v_1, v_2, V)
	If (Euclidean distance between ar_{v1} and ar_{v2} <
	articulatory merge threshold)
	Merge(v_1, v_2, V)
	Add new vowel to V with small probability.
	}

Table 3.6: Actions performed by the agents.

Shift closer (v, A)
{
$v_{best} := v$
For (all six neighbors v_{neigh} of v) do:
If (D(ac_{vneigh}, A) < D(ac_{vrec}, A))
$v_{best} := v_{neigh}$
$v := v_{best}$
}
	Find phoneme (v_{new}, A)
	{
	ar_v = (0.5, 0.5, 0.5)
	ac_v = S(ar_v)
	$s_v = 0$
	$u_v = 0$
	Do
	$v_{new} := v$
	Shift closer(v_{new}, A)
	Until($v = v_{new}$)
	}
		Update according to feedback signal
		{
		$u_{vrec} := u_{vrec} + 1$
		If (feedback signal = *success*)
		Shift closer(v_{rec}, A_1)
		$s_{vrec} := s_{vrec} + 1$
		Else
		If($u_{vrec}/s_{vrec} >$ *threshold*)
		Find phoneme(v_{new}, A_1)
		$V := V \cup v_{new}$
		Else
		Shift closer(v_{rec}, A_1)
		}

"Numquam nescis quid simius edit donec simius cacat."
Proverb.

4. Results

Running the simulation described in the previous chapter with the right parameter settings did result in natural vowel systems. Shared sets of vowels were obtained, so that imitation was highly successful. The vowel systems that emerged from the simulations were similar to the ones that are most frequently found in the world's languages. However, it is difficult to analyze the outcome of the simulations in a statistical way. It is easy to get an impressionistic idea of how well the agents' inventories agree by making a scatter plot of all agents' vowels, using the first and second formant as coordinates. This is the way the outcomes of the simulations are usually plotted in this thesis. But the number of vowel clusters that emerge is not fixed, even for a given parameter setting, and not all agents in the population have a vowel in all clusters. Cluster analysis of the outcome of a single run of the simulation at a given point in time is therefore possible, (although hard) but comparing positions of clusters from many different runs of the simulation in order to get an idea of the average system that emerges and the deviation around these clusters is not.

However, there is a possibility to compare the quality of systems that result from different runs of the simulation: Liljencrants and Lindblom's (1972) energy function, that was already presented in chapter 2, equation 2.1, but that is repeated here for reference:

$$E = \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} \frac{1}{r_{ij}}$$

where E is the energy of the system, n is the number of vowels in the system and r_{ij} is the distance (according to the distance measure from chapter 3) between vowels i and j. This function gives a high value if vowels are close together, and a low value if vowels are far apart. Its minimum value is reached when the vowels are maximally dispersed. Liljencrants and Lindblom (1972) found that minimising this energy function results in realistic (and frequently occuring) vowel system. Therefore, a value of the energy function close to its minimum indicates that the vowel system is realistic. Instead of doing a statistical analysis of the complete vowel systems that are found, one can therefore do a statistical analysis of the energy values of these systems.

This is done in this chapter. But first, a first impression of the vowel systems that emerge is given. From this the reader can form an intuitive picture of what the vowel systems that emerge look like. Then the statistical analysis of the systems that emerge, as well as an analysis of the random case and the optimised case are presented. Finally, extensive examples of systems that result from changing parameter settings are presented and discussed.

4.1 A First Example

The aim of the simulations is to see whether a population of agents is able to generate a shared system of speech sounds and whether these systems will resemble human sound systems. Therefore, when the simulation is started, the agents' vowel inventories are empty. In the course of the imitation games, they will fill their inventories and update them according to the result of the imitation games they play with

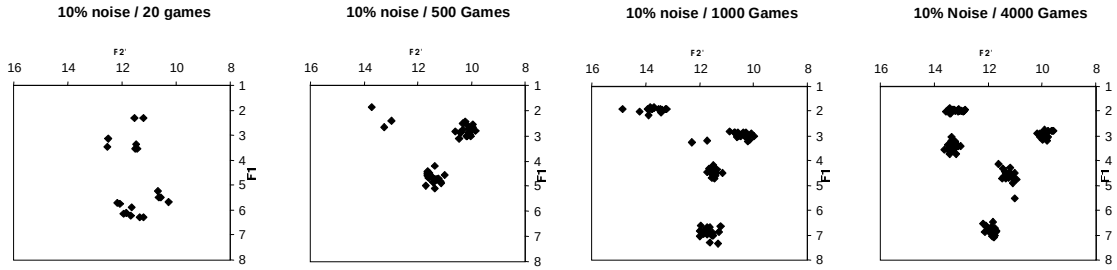


Figure 4.1: Development of a vowel system.

the other agents. An example of the emergence of a vowel system in a population of twenty agents is shown in figure 4.1.

The frames of this figure have been made by plotting all vowels of all agents on top of each other. The frequency of the first formant determines the y-coordinate of each point and the frequency of the effective second formant determines the x-coordinate. Both frequencies are expressed in Barks. The values increase from top to bottom and from right to left, respectively, so that the positions of the points in the graph correspond to the positions the corresponding vowels are traditionally given in phonetic literature. Clusters in the graph generally mean that all agents of the population have a matching vowel in this region of the acoustic space. The space between clusters indicates that these vowels are distinct from the other vowels in all agents' repertoires. Apparently, the agents do not all have exactly the same realisation of a given vowel. This is similar to the case with human speakers, who all pronounce a given vowel slightly differently as well (see figure 4.2). One also has to keep in mind that the points in the graphs are the acoustic representations of the vowel prototypes to which no noise is added. Whenever agents play an imitation game, noise is added to their utterances, so that the actual realisations of the vowels might be shifted somewhat from the point represented in the graph.

The first frame shows the system after 20 imitation games with 10% acoustic noise. The most important process so far has been the random addition of vowels. The agents that initiate the imitation game mostly had empty vowel inventories, and therefore had to invent random vowels. For the imitator there are two possibilities. It either had an empty repertoire and had to create a new vowel that is a good imitation of the vowel it heard, or it had only one vowel in its repertoire which was necessarily used. The first case causes pairs of close points in the graph. The second case will lead to a successful imitation game, because both agents have only one vowel in their repertoire, even though the sounds that are used might sound quite different to human ears. The successfulness of the imitation game will cause the imitator's vowel to be shifted slightly towards the initiator's one. Thus vowels are expected to clus-

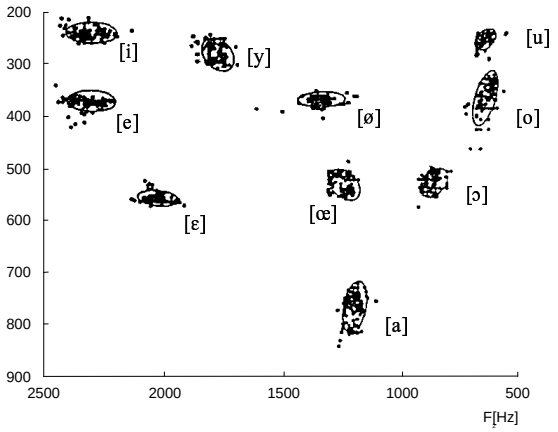


Figure 4.2: Vowel system of French, from (Rober-Ribes 1995) through (Glotin 1995).

ter.

Because only few imitation games have been played, the sounds have not clustered in the first frame, yet, but in the second frame one can observe emerging clusters. This happens after some 500 imitation games. At this point the most important process is the moving closer together of the different agents' vowel prototypes. Usually, imitation games will be successful, and the result will be that vowel prototypes are shifted closer together. More clusters have formed in the mean time, due to the random addition of vowels that continues to take place with low probability. Therefore, some agents will have more vowels than others and sometimes imitation games will fail, forcing other agents to add vowels as well.

The third frame shows the vowel system after 1000 imitation games. The most important process in this phase is the random addition of new vowels. Every once in a while, a random vowel is added to one of the agents and if it is sufficiently different from already existing vowels, the other agents quickly create corresponding vowels in their vowel inventories. If it is not sufficiently different from the other vowels, its success/use ratio will drop quickly and it will be removed from the agent's inventory. New vowels can thus only be successful if there is sufficient room in the acoustic space. After a certain number of vowels have been added, there is no more room, and the vowel inventories of the agents do not change anymore.

This has happened in frame four, after 4000 imitation games. Here one observes a natural looking vowel system, where there are a number of compact clusters with sufficient space in between. All agents have a vowel in all the clusters, meaning that imitation will almost always be successful. The most important process in this phase is shifting of vowels. As vowels are always generated with some noise, no two agents will exactly produce the same signal, ever. Therefore agents will always shift their vowels a little bit in response to an imitation game. This implies that the resulting vowel system is not completely static. Clusters can still shift, and if they shift in such a way that new room is opened, a new vowel might be added. If they shift in such a way that two clusters come close together, they might merge. However, clusters will not disperse over time, as would be expected if movement was totally random. Vowel prototypes are always attracted to each other, so they cannot move away too far.

The size of the clusters is determined by the noise that is added to the acoustic signal. In the simulation that was described above, the acoustic noise parameter ψ_{ac} was 10%. This corresponds to a maximal shift of 0.6 Bark in the graph. The other parameters of the simulation[1] were: $\psi_{art} = 0$, $\lambda = 0.3$, $\varepsilon = 0.03$, $\theta_s = 0.5$, $\theta_c = 0.7$, $\theta_u = 5$, $p_c = 0.1$ and $p_i = 0.01$. The population size was 20 agents. These parameters will be the same for all other experiments presented in this thesis, except where indicated.

[1] Meaning of symbols used in the thesis can be found in appendix A

In order to get an idea what different systems can be obtained with the simulation, another vowel system is given in figure 4.3. The only difference between this run of the simulation and the previous one is that the acoustic noise, ψ_{ac}, was set to 20% instead of 10%. Due to the higher noise level the acoustic realisations of vowels are spread out over a larger part of the acoustic space, so that vowels are confused more easily. This means that there will be fewer clusters, and the clusters are farther apart. Because of the less perfect realisation of vowels, the imitations of the agents will be less perfect as well and the clusters will be bigger. Both phenomena can be observed in the figure.

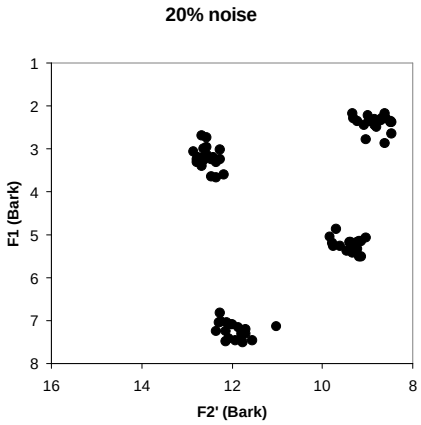


Figure 4.3: System obtained with 20% noise.

4.2 Analysis of Simulation Results

What do these results mean? The vowel systems that form do look like vowel systems that one finds in human languages. Figure 4.2 shows measurements of the vowels of French produced by a single male speaker. They are plotted in the same way, except that the axes are linear and not logarithmic, which causes the picture to become stretched in the horizontal dimension towards the [i], and in the vertical dimension towards the [a]. The similarities between the simulation results and the real system are striking. This indicates that the model is sufficiently realistic with respect to the distribution of realisations of vowels in acoustic space. However, a somewhat less impressionistic measure is needed in order to make an objective evaluation of the results of the simulations. For such an evaluation, a quantitative comparison with completely random systems on the one hand and optimal systems on the other is in order.

Such a comparison could be done with cluster analysis. One could compare the size of the clusters that are found with the distance between the clusters. However this is slightly more difficult than it might seem at first glance for two reasons. The first reason is that it is hard to find the actual clusters, and to determine which vowel belongs to which cluster, because not all agents have the same number of vowels. Also the number of clusters that emerges after a fixed number of imitation games cannot be determined exactly beforehand. The second reason is that one can hardly use cluster analysis for evaluating random systems, nor for optimal systems. In random systems the agent's vowels will be so dispersed that cluster analysis is meaningless. Optimal systems (calculated in the same way as Liljencrants and Lindblom's (1972) models) will be generated one at a time. There will be no such things as clusters to investigate.

4.2.1 Energy of a vowel system

However, one can define interesting quantitative measures that describe the quality of the resulting vowel systems in an objective way. The first of these is Liljencrants and Lindblom's (1972) energy measure (as illustrated in equation 2.1). It measures the dispersion of the agents' vowels through the acoustic space. The lower the energy, the more dispersed the vowel systems. Apparently maximising dispersion results in realistic vowel system. Therefore the energy can be used as a measure of

realism. The lower the energy, the more realistic the vowel system. Of course, in reality one will rarely find human vowel systems that are completely optimally dispersed and that consequently have minimal energy. However, their energy will tend to be low. It will therefore be assumed that low energy means a realistic vowel system. It must be noted, however, that only energies of systems with equal numbers of vowels can be compared. The energy function sums over a number of distances that increases with the square of the number of vowels in the systems. The more vowels there are in a system, the higher the energy of the system will be, even if the distances between all the vowels are equal (see also appendix B for an investigation of random and optimal vowel system energy). This was not a problem in Liljencrants and Lindbloms' (1972) simulation, or in subsequent work that optimises vowel systems with a similar energy function (Vallée 1994, Schwartz *et al.* 1997). It should also not be a problem here, but one should exercise caution when using the energy function to compare runs of the simulation, because they usually contain agents with different numbers of vowels. Also, energies of systems with different settings of the parameter λ, which determines the influence of F_2' relative to F_1, should not be compared, as λ is used in calculating the energy. Calculating the energy values for the same system with different values of λ will result in different values.

4.2.2 Success of imitation

The second measure is the success of imitation the agents can achieve. This is calculated by checking for every vowel of every agent in the population, whether it will be imitated correctly by the other agents in the population. The number of correct imitations relative to the number of possible pairs of agents then gives a measure of the coherence of the agents' sound systems.

One could imagine a hypothetical vowel system that has either low energy and low imitation success, or high energy and high imitation success. These systems will not be realistic. The first will have dispersed vowels, but clusters which are about as big as the distance between the clusters and the second will have small clusters, but these clusters will reside in only a small part of the acoustic space. Both systems would be very sensitive to noise. However if a system has both a high imitation success and a low energy, it is a realistic vowel system, as its vowel clusters will be dispersed (for low energy) and compact (for high imitative success).

4.2.3 Analysis of emerged systems

First the systems that are generated by the simulation will be investigated. For this the same parameter settings as for the simulations in the previous section will be

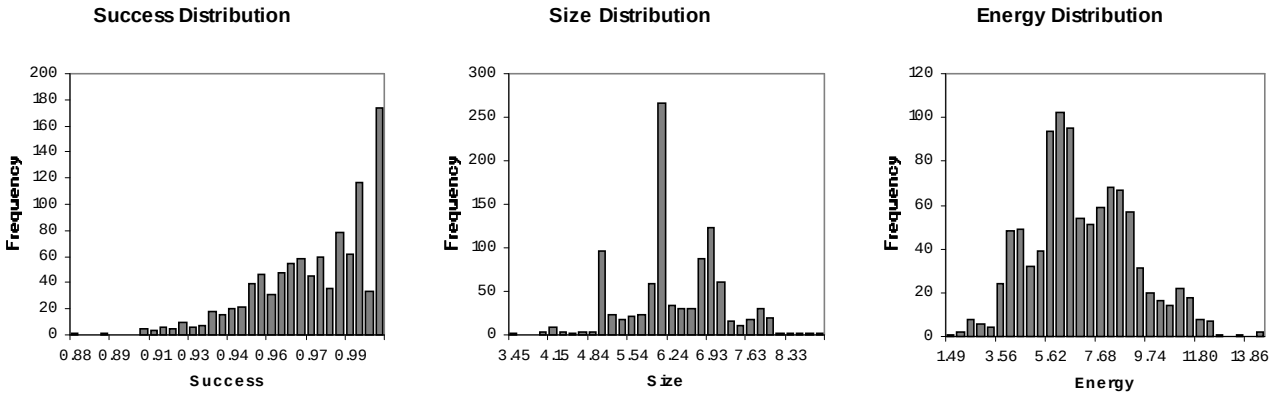


Figure 4.4: Success, size and energy of 10% noise system.

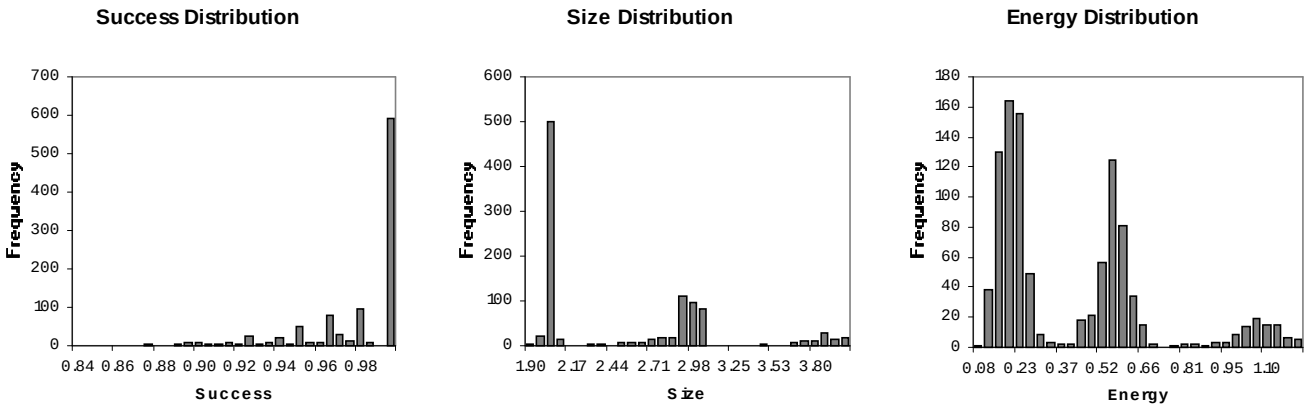


Figure 4.5: Success, size and energy of 20% noise system.

used. Systems with acoustic noise parameter ψ_{ac} of 10% and 20% will be investigated. As it was found that systems in simulations with higher amounts of acoustic noise developed slower, a higher value for the insertion of new phonemes p_t of 0.1 was used for the case of 20% noise. The results for the simulation with 10% noise are presented in figure 4.4. In this figure, the distribution of the success, the number of vowels and the average energy over 1000 runs of the simulation are presented. The success was calculated as the running average over the imitation games and was calculated as follows:

4.1) $s_{av} \leftarrow 0.99 \cdot s_{av} + 0.01 s_t$

where s_{av} is the running average of the success and s_t is 1 if the imitation game that was just played was successful and 0 if it was a failure. This gives a good estimate of the success of imitation (same average and standard deviation) and requires less calculations than exhaustively calculating the successes of all possible interactions between all agents. The number of vowels is the average number of vowels in the agents in a population. The average energy is average over the energies of the vowel systems of all the agents in the population. The population size was 20 agents.

One can see that the success of the imitation games is 0.973 on average with a standard deviation of 0.023, but that complete success appears most often. The average of the average vowel system size of the populations is 6.21 with standard deviation 0.82 but it is not distributed normally. The distribution has peaks at integer sizes five, six and seven. Apparently the simulation has a tendency to converge towards systems where all agents have an equal number of vowels, either five, six or seven in this case. The average energy has an average of 6.75 with a standard deviation of 2.11. The energy also does not follow the normal distribution. One can observe several peaks. These probably indicate different possible configurations for systems consisting of five, six and seven vowels. The different possible "minimal" configurations for systems consisting of six vowels will be investigated in more detail in the section on optimal systems, below and in chapter 6.

The case of 20% noise is illustrated in figure 4.5. The average success is 0.982 with standard deviation 0.027. As can be seen in the figure, complete success occurs most frequently. The average size of the agents' vowel inventories seems to cluster around three peaks, one for populations sharing two, three and four vowels respectively. Half of the simulations end up with two-vowel systems and half of the simulations end up with three- or four-vowel systems. Again integer numbers are preferred, indicating that all agents in the population have the same number of vow-

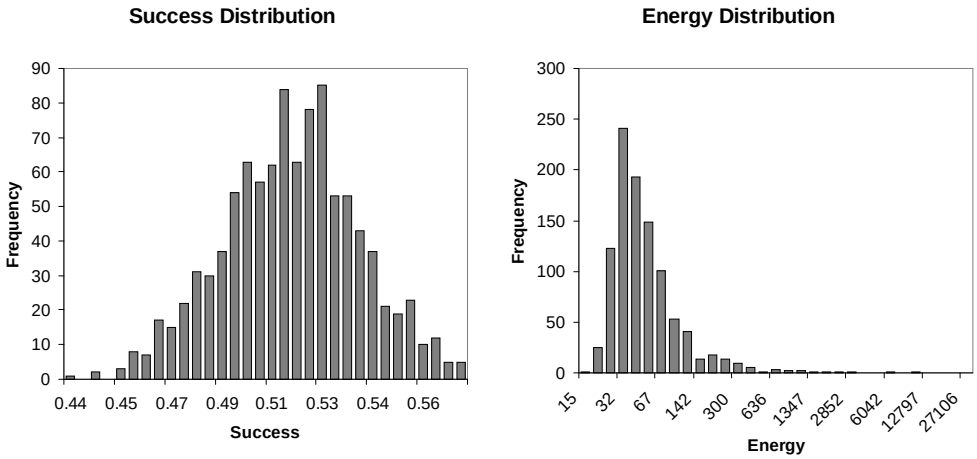


Figure 4.7: Success and Energy of random systems with 5 or 6 vowels.

els. The same peaks appear in the average energy distribution. There are three peaks. One with the lowest energy of 0.19 is for populations consisting of agents with mostly two vowels, as systems with fewer vowels have inherently lower energy. The other peak, at energy 0.55 is for populations with agents that have three vowels, while the peak at energy 1.06 is for systems with four vowels.

The performance of the emerged systems will now be compared with that of random systems. Because these systems were not generated with an imitation game, the number of vowels was determined beforehand. It was chosen to investigate systems with three and with six vowels, because these correspond most closely to the systems that are found with 20% and 10% acoustic noise ψ_{ac}, respectively.

4.2.4 Comparison with random systems

In order to evaluate how good the emerged systems are, their energy and their success should now be compared with the energy and success of random systems. They will first be compared with random systems. The results of a simulation with random systems in which the agents have two or three vowels (as in the 20% noise experiment) are shown in figure 4.6. The results were obtained from calculating the energy and success of 1000 populations of twenty randomly initialised agents. Note that the x-axis of the figure is logarithmic. This was necessary to accommodate the much higher values of the energy that were found. The average of the average energy of the systems was 47, and the standard deviation of the energy was 1102. The distribution is extremely skewed towards high energy values. In any case, it is clear that the energy of the random vowel systems is significantly higher than the energy of the vowel systems that were obtained by the simulation. The success values were calculated in a randomly initialised population of 20 agents with two or three vowels each (with the same distribution as in the simulations). Imitation games with 20% noise added to the acoustic signals were played for all vowels of all agents, and the overall success was taken to be the ratio between the number of successful games and the total number of games. They have an average of 0.57 and a standard deviation of 0.035, which makes them significantly lower (at the 1% level) than the success values of the systems that were obtained in the simulation. It can therefore be concluded that in the case of acoustic noise ψ_{ac} = 20% the vowel systems are both more dispersed and more coherent than random.

The results for the random system with approximately five or six vowels per agent, as in the 10% noise case, are presented in figure 4.7. It can be seen that the

success score of the random system is 0.51 with standard deviation 0.024. The energy of the emerged systems is extremely high, on average 112 with standard deviation 1162. Again the energy distribution is very skewed, and therefore a logarithmic scale was used for displaying it. Apparently the average energy of the random systems is much higher than that of the systems that were obtained with the simulations. Also the success score of the random systems is significantly (at the 1% level) lower than the success score of the vowel systems obtained in simulation. The simulation with acoustic noise ψ_{ac} = 10% therefore also performs better than random. It is perhaps amazing that the success score for random systems remains relatively high, above 0.50 in both cases. It can be shown with a mathematical argument that success scores of random systems will not drop below 0.50. This argument is presented in appendix C.

4.2.5 Comparison with optimal systems

Apparently the imitation games result in systems that show better than random performance, something which could already be suspected from looking at the plots of the obtained vowel systems. But how close do the vowel systems come to the optimally dispersed vowel systems? Comparing the results of the research into optimally dispersed systems (e.g. Liljencrants & Lindblom 1972; Vallée 1994; Schwartz *et al.* 1997b) directly with the results obtained here is only possible in a qualitative, subjective way. The model used here differs in an important aspect from the models researched elsewhere. In the work presented here, vowels are represented by their articulatory parameters and can only be optimised by shifting these articulatory parameters, whereas in the other work, vowels were represented by their acoustic signals only, which could be directly manipulated in order to minimise the energy of the system.

The Liljencrants and Lindblom (1972) model was therefore re-implemented, using articulatory representations for the vowels and using the synthesis function and perception function that have been described in the previous chapter. The energy of the vowel systems was minimised by a gradient-descent method. First the system was initialised with vowels at random positions scattered throughout the articulatory space. Then for all vowels in the system in turn, it was calculated whether a small shift (either a decrease or an increase) in one of the three articulatory dimensions would reduce the total energy. If this was the case, the shift in vowel position was kept. If not, the vowel remained at its old position. This procedure was repeated until no more decrease in energy was possible. Note that it does not always have to end with the same vowel system. It could get stuck in different (local) minima, depending on the initial conditions and on the sequence followed by the minimisation procedure. This will be apparent from the figures.

The systems that resulted from running the minimisation procedure with three vowel prototypes are presented in figure 4.8. This figure can be compared with figure 6.3. The most frequent of these systems (type 2) is the canonical three vowel system with vowels [i], [a] and [u]. The less frequent type (type1) is a "vertical" vowel system consisting of [i], [e] and [a]. Although vowel systems of this particular composition probably do not appear in human languages, there are "vertical" vowel systems with three elements, such as the Caucasian language Kabardian (Choi 1991, Ladefoged & Maddieson 1996, pp. 286–288) but these are usually more centralised: [ɨ], [ə] and [a] in the case of Kabardian. In cases like these, other factors, such as articulatory ease or historical processes probably played a role.

In any case, the most interesting part of these simulations is not the actual systems in which they result, but their energy. This is presented in figure 4.10. It can be observed that there are several peaks in the energy. The highest peak can be observed at energy 0.2. This peak corresponds to the most frequently occurring vowel systems. The peaks around energy levels 0.30–0.34 correspond to the vowel systems of type 2 that have front high vowels that are a bit more back than [i], causing the somewhat dispersed cluster in the left plot of figure 4.8. The peaks around 0.42–0.50 correspond to vowel systems of type 1. The clusters in the plot of this vowel system type are slightly more dispersed than the ones of type 2, so the energies are also more dispersed.

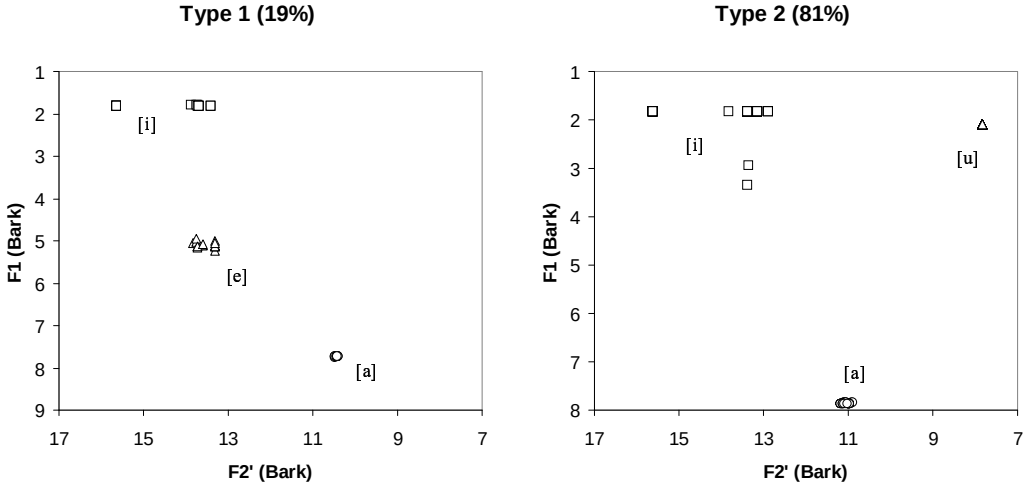


Figure 4.8: Optimised systems with three vowels.

The optimal systems with six vowels are presented in figure 4.9. It can be compared with figure 6.6. There are more different optimal systems for six vowels than there are for three. Except for type 4, all these systems are realistic, and can be found in human languages. Type 4 is a typical example of a case where the minimisation process got stuck in a local minimum, probably because the random initialisation created too many vowels towards the front. The split of the high front vowel [i] that can be observed in almost all of these graphs is probably caused by the discontinuity of the perception function (described in the previous chapter).

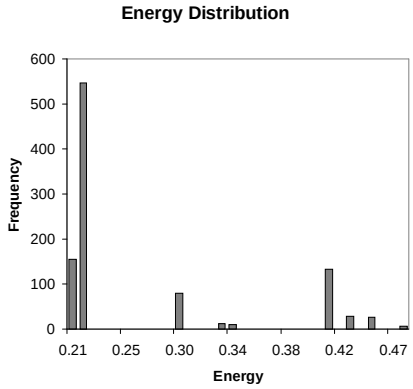


Figure 4.10: Energy of optimal three vowel system.

Based on the data in Nathalie Vallée's thesis (Vallée 1994, Annexe 2) an example of a language with a system like type 1 would be Saami, (Hasselbrink 1965) of a language of type 2 would be Chamorro (Seiden 1960) and of a language of type 3

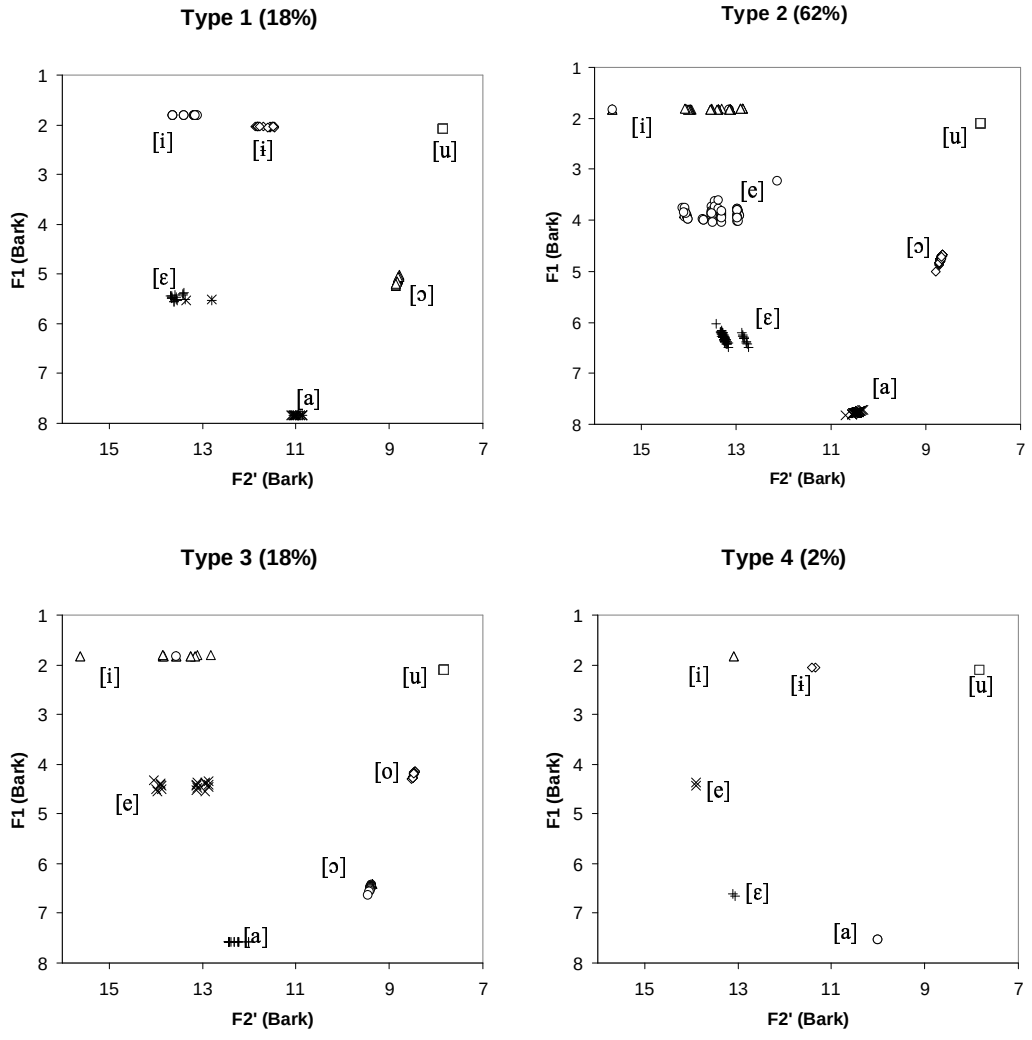


Figure 4.9: Optimal systems with six vowels.

would be Hakka (Hashimoto 1973). These types are not the most frequent types of systems with six vowels in the world's languages, but they do agree quite well with the results Berrah (1998) obtained.

However, reproducing the most frequent vowel systems of the world's languages was not the aim of these simulations. For comparison of emerged vowel systems with real vowel systems, see chapter 6. The aim was to calculate the energies of (near-) optimal vowel systems that could be obtained with the synthesis- and perception function that are used in the imitation games. The energies that were obtained are given in figure 4.11. The energy distribution has a large peak near 2.8, corresponding to the six-vowel systems of type 2 in figure 4.9. The smaller peaks between 2.95 and 3.54 correspond to vowel systems of type 1 and type 3, respectively, while the two very small peaks around 3.94 correspond to the vowel systems of type 4.

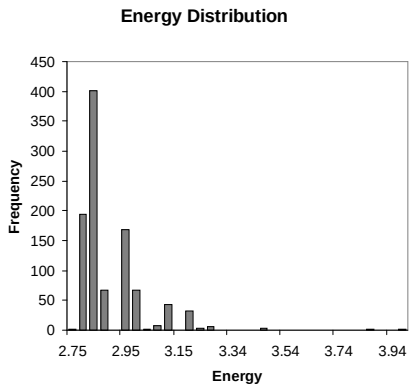


Figure 4.11: Energy of optimal six vowel systems.

4.2.6 Conclusion of comparison

The reason to calculate these optimal energies was to get an idea of how close to optimal the energies of the systems obtained through the imitation games are. The distribution of the average energies of the systems with acoustic noise ψ_{ac} = 20% can be found in figure 4.5. Here there are peaks at 0.17 and 0.45, and the range is 0.08 to 1.04. This seems to be lower than the optimal energy, but note that there are also agents with only *two* vowels. In any case, these energy values compare very favourably with the optimal case, especially if one takes into account that the most frequent energies in the random systems were around 6.

The energies of systems with acoustic noise ψ_{ac} = 10% can be found in figure 4.4. Here there are peaks around energy values 4.99 and 5.88 and the range is between 2.54 and 9.44. This also compares favourably with the energy values that were found in the optimal systems, although here the systems seem to be removed a bit further from the optimum. This is understandable, because there are many more possible configurations with five to six vowels than with only two or three.

The overall conclusion from this analysis is that the systems obtained from the imitation games are indeed realistic systems and that they make successful imitation possible. They are realistic, because they have a low energy. Their energy is much lower than the energies of the random systems, and it comes close to the energies of the optimal systems. They enable successful imitation, because their success of imitation is much higher than in the case of random systems. One can also conclude that the energy and success are good for quantitatively measuring the quality of vowel systems. However, one should not underestimate the success that can be obtained by random vowel systems, which tends towards 50% success for large numbers of vowels. In order to allow a comparison of random and optimal systems with the results that will be presented below, appendix B contains histograms for the success and energy for random and optimal systems with (exactly) two to ten vowels.

4.3 Changing Parameters: a Sensitivity Study

So far, only systems resulting from two different parameter settings have been investigated. The main difference was the acoustic noise level ψ_{ac}, which was 10% in one case and 20% in the other. It was shown that these parameter settings resulted in realistic vowel systems that allowed the agents to imitate each other successfully. The main difference between the two parameter settings was that with the larger noise value, the number of vowels in the system was smaller and clusters that were formed were bigger. This was to be expected, because if the acoustic noise increases, the vowels get more easily confused. In order to maintain successful communication, the agents will therefore adopt less vowel prototypes and space them further apart. Because the prototypes are generated with more noise, and therefore in a wider part of the acoustic space, the agents will have a less focused target to move their vowels to and thus the clusters will become bigger.

Still, the simulation has many other parameters of which it is not always intuitively clear what their influence is on the agents' behaviour and the vowel systems they form. If one wants to understand the behaviour of the simulations more fully, it is necessary to run them with many different parameter settings. This is also necessary in order to get a better understanding of the connection between the vowel systems obtained by the simulation and the vowel systems observed in real human languages. Only after the influences of the different parameters of the simulation on its performance are uncovered, can there be any discussion of the relation between the simulation and its parameters on the one hand and the phenomena observed in human vowel systems on the other.

As there are many parameters in the simulation, their influences cannot all be investigated simultaneously without creating chaos. Instead, the parameters should be changed with small steps, while keeping the other parameters constant, in order to get an idea of how they influence the outcome of the simulations. For every parameter setting the average energy, the average success and the average number of vowels, calculated in the way described in the previous section, should be measured. In order to save computation time, the averages will be taken over one hundred runs of the simulation instead of one thousand. These values can then be plotted, together with example plots of the resulting vowel systems, (for an intuitive idea of what a typical resulting vowel system looks like) in a two-dimensional table in which one parameter varies in the horizontal direction and another parameter varies in the vertical direction. Thus the influences of two parameters can be examined at a glance.

Not all parameters of the simulation will be investigated. Only the parameters that have direct linguistic relevance are interesting. For this reason the different thresholds in the simulation, θ_c, θ_s, θ_u and the probability of cleanup, p_c are not investigated. These are just details of the implementation of the agents' learning mechanism. They have been fine-tuned to their default values in preliminary experiments. Changing these parameters does not influence the structure of the vowel systems that emerge. Setting them to the wrong values just results in slower or no emergence of coherent systems. The performance of the simulations is not very sensitive to the setting of these parameters. They can be changed from their default values without changing performance much, but some relations between them should be observed, such as $\theta_c > \theta_s$.

The other parameters: the step size with which articulatory vowel prototypes can be improved ε, the ratio λ between the influence of F_1 and F_2' in the perception

of vowels, the probability of inserting new phonemes in an agent's repertoire p_i, the acoustic noise ψ_{ac} and the articulatory noise ψ_{art} do have a direct effect on the vowel systems that are obtained by the simulations. They all have a direct linguistic relevance because they influence perception or production or because they directly influence the way the vowel systems change over time. These are therefore the parameters whose influences on the simulation's behaviour have been investigated.

For all parameters a default value was determined by preliminary experiments. The default values were chosen to guarantee realistic looking vowel systems to emerge. A short description of the parameters of the simulation as well as their default values can be found in appendix A.

A range of possibly interesting values for these parameters was determined by preliminary experiments. Parameter settings that resulted in total chaos on the one hand or only one vowel cluster on the other hand, were not considered to be interesting. It will be clear from the figures that are presented that vowel systems falling outside the parameter range are not interesting as well as which values of the parameters result in the most realistic vowel systems. Within the ranges of interesting parameter values, a number of six values, evenly spread over the total range of the parameter, were determined. These values were then used in the experiments.

From the twenty possible pairs of parameters, only a subset was investigated. The reason for this was that some parameters give more interesting results when changed together than others. Certain parameters have the same effect on the vowel systems, no matter what the settings of the other parameters, while others have interesting interactions. Some parameters, such as ψ_{ac} and ψ_{art} that both determine a noise level, belong together conceptually, while others do not. For every pair of parameters that is presented, the reason for taking it as a pair is given in the discussion of the results.

Some parameters were always fixed in the experiments. The population size was twenty agents for all experiments. The number of imitation games done before the results were measured and the example vowel systems was taken to be a fixed 5000. When optimising vowel systems, as done by (Liljencrants & Lindblom 1972, Boë *et al.* 1995, Schwartz *et al.* 1997) there is a clear criterion for stopping. Sooner or later, a minimal energy will be reached, because the energy function is always positive, because the energy decreases at every step and because the step size is fixed and finite. On the contrary, there is no clear stopping criterion in the imitation games. Even though the agents' vowel systems will stabilise after a while, they will never become completely static. This is a general problem of agent simulations, and Berrah (1998, pp. 51–52) had to use a more or less arbitrary limit of 3000 interactions[2] to the number of interactions between his agents as well. Fortunately, as it could be already deduced from figure 4.1, where the system after 1000 games is already quite similar to the system after 4000 imitation games, the vowel systems do not change much after some 5000 imitation games. This number depends a bit on the settings of the parameters, but from preliminary experiments it turns out to be a good limit. In a sense, having a limit like this is completely arbitrary, but it is realistic in the sense that it is also impossible to define a limit on the evolution of human vowel systems. These are also always changing, and one can not find a moment or a

[2] In Berrah's thesis each interaction consists of five exchanges of signals between agents. His 3000 interactions are therefore equivalent to 15 000 imitation games.

criterion by which one can determine when the system has reached a final state. One can only take snapshots of a system as it is at a given moment. And as long as all snapshots are taken after an equally long period, this is fair.

4.3.1 Articulatory and acoustic noise

In figure 4.20 on page 68, articulatory noise ψ_{art} and acoustic noise ψ_{ac} are changed. Articulatory noise changes on the horizontal axis and acoustic noise changes on the vertical axis. Values for both parameters are: 0, 0.05, 0.10, 0.15, 0.20 and 0.25 from left to right and from top to bottom, respectively. The figure shows four items for every parameter setting: the average success, energy and number of vowels over hundred runs of the simulation, together with their standard deviations, and a graph showing a typical example of a vowel system that emerged for this parameter setting. The graph has been generated in the same way as the graphs of vowel systems that have already been shown. On the vertical axis of each graph is F_1 (in Bark) and on the horizontal axis is F_2' (also in Bark). Success, Size and Energy are not usually distributed normally (as could be seen in figures 4.4, 4.5, 4.6, 4.7, 4.10 and 4.11). The standard deviations should therefore be taken with a grain of salt. This is especially the case for the energy, as this tends to have outliers towards extremely high energy values. If its standard deviation is about the same size as or bigger than the energy itself, this indicates that the energies of the different vowel systems obtained for this parameter setting fluctuated wildly, which means that the resulting systems are not very stable and will be very different between runs.

One can observe in the figure that the resulting vowel systems become less stable if the articulator noise ψ_{art} increases, while the number of vowel clusters seems to decrease with increasing acoustic noise ψ_{ac}. The higher the articulator noise, the less successful the imitation becomes. On the contrary, the higher the acoustic noise, the more successful the imitation becomes, but this is caused by the decreasing number of vowel clusters. This decrease is caused by the fact that vowels are more easily confused when there is a lot of noise, and therefore neighbouring vowel prototypes are more easily merged. It can also be observed that the vowel clusters get bigger if the articulatory or the acoustic noise increase.

The resulting vowel systems seem to be realistic for a small number of parameter settings only. Apparently articulator noise should be low, because its only effect seems to be to destabilise the resulting vowel systems. When the articulatory noise is too high, vowel clusters become so spread out that they overlap with other vowel clusters, (especially if the number of clusters is high, as in the upper right corner of the figure) thus hindering successful imitation.

This does not necessarily indicate that in the formation of human vowel systems articulatory uncertainty does not play a role. The deteriorating effect of articulator noise in these simulations could as well be an artefact of this particular implementation of the imitation game. Acoustic noise on the other hand, should not be zero, but rather between 0.10 and 0.20, approximately. If the acoustic noise is lower than 0.10 the agents' perception becomes unrealistically precise and therefore the number of vowels becomes unrealistically high. Acoustic noise higher than 0.20 tends to produce vowel systems that contain only two vowels. Systems like these do occur in human languages (Ladefoged & Maddieson 1996, p. 286) but are very rare.
If one runs systems with an acoustic noise level of 0.25 for long enough, they will eventually form triangular systems with three vowels, though.

4.3.2 Acoustic noise and formant weighting

The next two parameters that are investigated are the parameter that weights the relative influence of F_1 and F_2' (λ) in the perception of distance between two acoustic signals and the acoustic noise ψ_{ac}. This pair of parameters has been chosen, because both parameters influence the number of vowel clusters that will emerge in the population. The results are shown in figure 4.21 on page 69. The values of the acoustic noise change along the horizontal dimension and are 0, 0.05, 0.10, 0.15, 0.20 and 0.25 from left to right. The values from λ change in the vertical dimension and are 0.1, 0.2, 0.3, 0.5, 0.7 and 1.0 from top to bottom.

In this figure it can again be observed that the higher the articulatory noise, the lower the number of vowel clusters. The parameter λ determines the accuracy with which agents can make distinctions in the effective second formant, relative to the first formant. The lower rows in the figure show more distinctions along the effective second formant axis than the upper rows for the same amount of articulatory noise. The total number of vowel clusters is therefore determined through the interaction of the acoustic noise ψ_{ac} and λ. The acoustic noise determines the number of distinctions that can be made on the first formant, while λ determines the ratio between the number of distinctions on the second and the first formants. The highest number of clusters can thus be found in the lower left corner of the figure (15.95 ± 1.98), where the acoustic noise is low and where λ is high. The lowest number of clusters (1.99 ± 0.08) can be found in the opposite corner, where noise is high and λ is low.

The number of distinctions on the first formant is related to the number of distinctions the agents' vowel systems make in the articulatory parameter tongue height. The number of distinctions on the effective second formant is related to the number of distinctions in tongue position and lip rounding. This indicates that purely acoustic parameters can determine what seem to be articulatory properties of the vowel systems. One should note however, that these acoustic parameters do not directly determine the number of vowels or the distinctions that will be used in the vowel systems either. The parameters only influence the agents' perception. Through the self-organising interactions between the agents, vowel systems are favoured with a larger or smaller number of distinctions along any of the acoustic or articulatory parameters.

Not all parameter settings result in vowel systems that are realistic in the sense that they could also be found in human languages. This was already noticed for acoustic and articulatory noise, but it is equally true for λ. In human vowel systems the maximum number of distinctions in the first formant (or vowel height) is (probably) five (Ladefoged and Maddieson 1996, pp. 289–290). The maximum number of distinctions in the effective second formant (resulting from tongue position and lip rounding) is approximately four. However, with increasing numbers of vowels in a vowel system, more distinctions in the first formant will be used than distinctions in the effective second formant.

For the given number of distinctions in the first formant, the number of distinctions in the effective second formant is too high for λ higher than 0.5, and too low for λ lower than 0.2. Similar results were obtained for vowel systems that resulted from optimising the energy function of vowel systems for given numbers of vowels (Vallée 1994, ch. V, Schwartz *et al.* 1997). This is the reason that the standard value for λ was chosen to be 0.3 in the other experiments. Whether this parameter corresponds to acoustic or articulatory properties in humans is not clear. It

does correspond nicely with the finding of Lindblom & Lubker (1985) that humans are able to sense distinctions in the tongue height dimension approximately three times better than distinctions in the tongue position dimension.

4.3.3 Step size

The next parameter that will be investigated is ε, the size of the step with which vowel prototypes are shifted towards the acoustic signals that the agents observe. The values of this parameter that were investigated were 0.005, 0.01, 0.03, 0.05, 0.10 and 0.15. It was varied in combination with the acoustic noise ψ_{ac} and with λ. For these parameters the same range of values was used as in the previous experiments. The step size ε was varied along the vertical axis with the smallest step size at the top and the largest step size at the bottom.

The combination of ε and ψ_{ac} is shown in figure 4.22 on page 70. The main thing that can be observed is that the larger the step size, the more chaotic the resulting vowel system. This can be quantitatively measured by the success- and energy values of the resulting vowel systems. The success values are lowest and the energy values are highest for the systems at the bottom of the figure. The system with maximum noise and maximum step size (at the bottom left of the figure) even has infinite energy, indicating that some of the vowel prototypes came so close together that the energy calculation overflowed.

The number of vowel clusters decreases with increasing acoustic noise, as was expected from the results of the previous experiments. The number of vowel clusters also seems to decrease slightly with increasing step size (from 16.23 ± 1.82 to 13.81 ± 2.98 for zero acoustic noise value and from 2.05 ± 0.22 to 1.96 ± 0.20 for the highest acoustic noise of 0.25). This is probably caused by the fact that the agents are not able to imitate the acoustic signals as well and as readily if the step size is large than if the step size is small. Therefore vowels will be confused slightly more often and the vowel systems will not grow as big with a large step size as with a small step size. Vowel inventory size also varies with step size in figure 4.23 on page 71 in which λ is varied along the horizontal axis and step size ε is varied along the vertical axis (from 4.79 ± 0.78 to 4.36 ± 0.70 for λ=0.1 and from 10.56 ± 1.12 to 7.72 ± 1.32 for λ=1). The influence of step size on the number of vowel clusters is small, but significant at the 1%-level according to the Kolmogorov-Smirnov test. The more clusters there are, the more step size influences the number of clusters that emerge.

The size of the vowel clusters also varies consistently with step size, which causes the resulting vowel systems to look more chaotic and the success ratio to be lower. The success rate varies from 0.904 ± 0.028 to 0.695 ± 0.038 for ψ_{ac} = 0 and from 0.999 ± 0.0047 to 0.978 ± 0.021 for ψ_{ac} = 0.25 in figure 4.22 and varies from 0.994 ± 0.012 to 0.931 ± 0.032 for λ = 0.1 to 0.946 ± 0.031 to 0.766 ± 0.039 for λ = 1 in figure 4.23. The success rate stays high as long as the clusters are so small that they do not overlap. As soon as the cluster size becomes greater than the distance between the clusters, the success rate drops dramatically towards the one expected for random systems, which is approximately equal to $n/(2n–1)$ for systems of n vowels (see appendix B and appendix C).

Energy seems to be higher for systems with smaller step size. The higher number of vowels alone can not explain this. If one compares systems with almost equal numbers of vowels, such as the ones with acoustic noise ψ_{ac}=0.15 and ε = 0.03 respectively ε = 0.15 (in rows three and six, respectively of the fourth column in figure 4.22) one finds that the one with smaller step size has a higher energy (1.52 with standard deviation 0.62) than the one with the larger step size (1.23 with standard

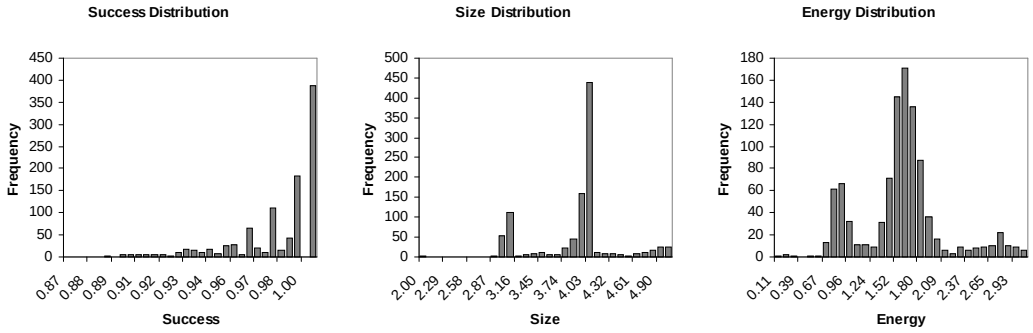


Figure 4.12: Success, size and energy distribution of vowel systems with step size = 0.03 .

deviation 0.54). These values are significantly different at the 1% level on the basis of the Kolmogorov-Smirnov test for thousand runs of the simulation on which they are based. There seems to be a definite tendency for systems with smaller step size to have higher energy. This phenomenon can be understood by comparing figures 4.12 and 4.13. In the first figure the distributions of success, size and energy of the parameter setting ψ_{ac} = 0.15 and ε = 0.03, in the second figure the distributions for parameter setting ψ_{ac} = 0.15 and ε = 0.15 are presented. Although both the average and the standard deviation of the vowel system size are almost equal for the two parameter settings, the two distributions are in fact quite different. In the first figure, with the smallest step size, three distinct peaks at (average) vowel system sizes three, four and five can be distinguished. In the second figure two much wider peaks can be observed, around average vowel system sizes three and four. This means that vowel systems that emerge in populations of agents with large step sizes are not only more confused with respect to the positions of the vowel prototypes, but also with respect to the number of vowels in each agent. The smaller energy of the systems with larger step size has two reasons. First of all the energy of a random vowel inventory increases faster than linear with increasing inventory size (see appendix B for more details). Secondly, the agents that use a larger step size tend not to converge to systems with larger numbers of vowels. In figure 4.13, for example there is no peak at systems with five vowels, whereas in figure 4.12 there is a clear peak. These two reasons combined cause systems with small step size to have lower energy even if the average number of vowels is equal. This stresses the fact that an average and a standard deviation do not say everything about the distribution of measurements of systems that emerge for a given parameter setting. Ideally the whole distribution should be studied.

Systems with a small step size have a higher success rate than systems with a larger step size. This is the case even though agents can never reach complete accuracy in imitation because of the presence of acoustic noise on the signals they try to imitate. This would indicate that it is advantageous to choose the smallest step size possible. However, the smaller the step size, the longer it takes for an agent to reach the given acoustic goals. In the limit of step size zero, an agent will never reach the acoustic goal. The number of practice steps needed to approximate a sound the agent has perceived increases inversely proportional to the size of the steps. In the implementation as it is now, there is no penalty on the number of practice steps. In reality, on the other hand, an agent never has unlimited time to practice, and can therefore never reach 100% accuracy in imitating sounds it hears.

Each practice step needs to use the articulator. If one assumes that using the articulators consumes energy, it follows that the number of practice steps should be limited. One then gets a trade-off between the maximum attainable accuracy (with a minimal step size) and the distance that can be covered in the fixed number of steps. A certain finite step size should emerge from these experiments that results in the most successful imitations. The optimal step size in these cases will probably also depend on other factors, such as the setting of λ and the acoustic noise. The question then arises which maximum number of steps will be allowed. In the experiments presented here, a fixed step size of 0.03 was assumed without a limit on the number of steps. The step size of 0.03 was found to result in vowel systems that looked most like the vowel systems found in human languages. This means that the maximal number of steps should be somewhere near a value for which the step size of 0.03 is the optimum.

The influence of the step size on the quality of the vowel systems that emerge is illustrated in figure 4.14. In this figure the average of size, energy and success of hundred runs with a population of twenty agents are shown for four different step sizes (0.01, 0.03, 0.05 and 0.10). The maximum number of practice steps was ten and the acoustic noise ψ_{ac} was 10%. Also shown are the error bars for the standard deviations of the data points. Note that the size of the systems is shown on a different scale than the energy. As can be seen in the figure, the imitation success decreases for increasing step size. However, the size of the vowel systems (and therefore their energy) increases from step size 0.01 to the step size 0.03 (significant at the 1-% level for the Kolmogorov-Smirnov test). This means that agents are apparently able to learn more vowels with this larger step size. The limited number of steps and the small step size limits them to being able to imitate sounds only in a small part of the available vowel space. The optimal step size in this case is therefore not the smallest possible step size, but the step size that is the best compromise between the quality of imitation and the fraction of the available vowel space that the agents can use.

Step size is thus, next to articulatory noise, an articulatory parameter of the simulation. Both influence the accuracy with which agents can produce vowels. With increasing articulatory noise the ability of accurately producing any vowel decreases. With increasing step size the agents' ability to accurately find the correct articulator positions to imitate sounds they hear will decrease. It is therefore not surprising to find that the influence on the resulting systems of both articulatory noise and step size is similar.

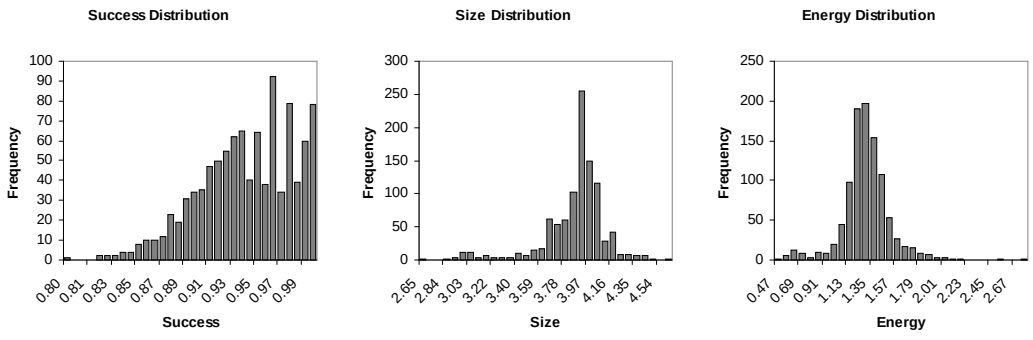


Figure 4.13: Success, size and energy distribution of systems with step size = 0.15 .

4.3.4 Population size

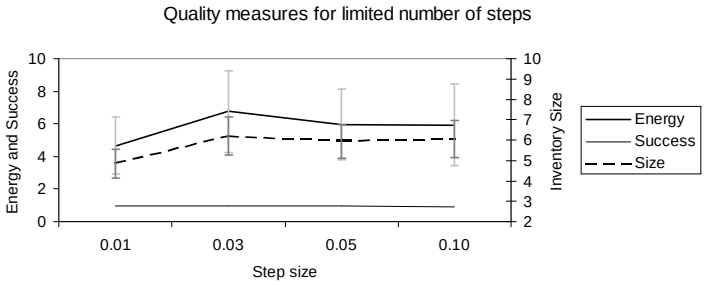


Figure 4.14: Energy, Success and Size of systems with limited number of practice steps.

The population size has always been twenty agents in the experiments so far. What would happen for different population sizes? The results of changing the population size will now be presented, but in a slightly different way than the previous changes in parameters were presented. It was found that the population size influences stability of the emerging vowel systems over time. In order to present the results in a two-dimensional graph, it was decided not to vary population size together with any other parameter of the simulation, but to show the evolution of vowel systems over a (relatively) short period of time for different population sizes.

In order to make a fair comparison between populations of different sizes, one should not run the simulations for an equal number of imitation games for all population sizes. In a smaller population, agents will have a higher probability to participate in any imitation game. They will change their vowel inventories quicker than agents in a larger population do. If one does not keep the number of imitation games constant, but rather the number of games per agent the comparison should be fairer (in Kaplan *et al.* 1998 the need for using the number of games per agent rather than the absolute number of games is also stressed in relation to naming games). In the experiments with different population sizes, all populations were therefore simulated for an equal number of games per agent. As all other results presented so far have been obtained from simulations of 5000 imitation games for populations of twenty agents, the number of games per agents was set to 250.

Another parameter that has to be changed for a fair comparison between populations of different sizes is the probability p_i with which new random vowel prototypes are added to the agents. This parameter determines the speed with which vowel systems grow. As this parameter represents the probability that a new vowel is added in every *game*, the simulations with small populations, and thus with a small total number of games, will result in smaller vowel repertoires. In order to compensate for this, the probability of adding vowels should be higher for small populations and lower for large populations, so that the total number of random insertions is equal for all simulations. As the previous simulations were run for a population of twenty agents for 5000 games with a probability p_i of 0.01, so that the expected total number of insertions was 50, it was decided that this should be the case for all experiments, so the probability became $0.2/N$, where N is the number of agents. The influence on emerging vowel systems of changing the probability of randomly adding new vowels will be investigated below.

The population sizes were chosen to be 2, 5, 10, 20, 50 and 100. The population size of two is an absolute minimum, and the population size of 100 is a reasonable maximum, as this is probably about the size of the group with which humans have close enough contact to have their language influenced (Dunbar, 1996 pp. 69–79). Larger populations of humans would break apart in smaller subgroups so that most of the interactions are within the subgroup and there are fewer interactions

between the subgroups. The results of running the simulation with these population sizes are presented in figure 4.24 on page 72. In this figure the population size increases from above to below. The first column shows the agents' vowel systems after 250 games per agent, and the next four columns show them after 300, 350, 400 and 450 games per agent respectively.

As can be seen in the figure, the larger population size does not influence the vowel systems that emerge very much. Obviously, the vowel clusters become spread over a larger portion of the acoustic space for larger populations of agents. Because of the larger number of agents with which any agent could have interactions and because of the acoustic noise of 10% the agents had a less fixed target to move their vowel prototypes towards. A more remarkable finding is that the larger populations of agents stabilise the agents' vowel systems. In the smallest possible population of two agents, the positions and number of the vowel clusters change easily. In the larger populations the positions of the vowel clusters become more and more stable. This can be explained by the fact that vowel clusters move because of the random shifts of vowel sounds due to the acoustic noise. Whenever an imitation game is successful, the imitating agent shifts the vowel prototype it used so that it becomes more similar to the sound it heard. But because of the acoustic noise this is not a fixed targets and vowel prototypes will therefore be subject to some random drift. The positive feedback provided by the imitation game can then cause the whole vowel cluster to move to a different location. However, if the population is large, the extent of this random drift is much smaller, because the movements of an agent's vowel prototypes will be averaged out in the interactions with the many other agents.

Pop. Size	Success	Energy	Size
2	0.975 ± 0.016	6.06 ± 2.83	5.99 ± 1.16
5	0.971 ± 0.021	6.95 ± 2.40	6.36 ± 0.89
10	0.969 ± 0.026	6.72 ± 2.25	6.23 ± 0.81
20	0.978 ± 0.020	6.61 ± 2.25	6.18 ± 0.85
50	0.974 ± 0.022	7.68 ± 2.44	6.53 ± 0.81
100	0.975 ± 0.023	7.85 ± 2.67	6.51 ± 0.97

Table 4.1: Quality measures for different population sizes.

Table 4.1 illustrates that the lower stability of the vowel systems of smaller populations does not mean that these systems are necessarily of lower quality. It shows the average success, the average energy and the average vowel inventory size for different population sizes. The standard deviation is also given, although this should be taken with a grain of salt, because the distributions are not known. These numbers have been calculated over hundred runs of the simulation, consisting of 250 games per agent. The numbers therefore correspond to the systems shown in the leftmost column in figure 4.24. The success stays relatively constant, always above the 96% level. The energy increases somewhat for vowel systems emerging from larger populations, but this has to do with their slightly larger number of vowels. The reliability intervals of all the measures are small, meaning that the evolution of the vowel systems always proceeds along roughly the same lines. The slightly higher numbers of vowels for the systems of the larger populations is probably caused by their greater stability, so that new vowels have more opportunity to be learned by all agents in the population, without being confused with the constantly moving existing vowel clusters.

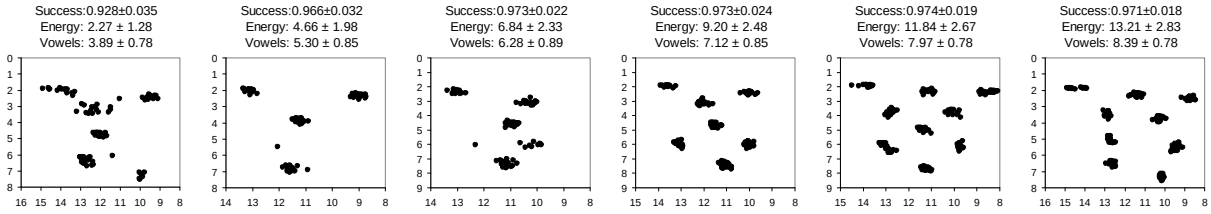


Figure 4.16: Vowel systems after different numbers of imitation games.

For all the population sizes that have been investigated, the vowel systems that emerge have the same general properties. The vowels cluster in relatively small clusters and the clusters themselves are more or less evenly spaced in the available acoustic space. Apparently the size of the population does not influence the possible vowel systems very much.

4.3.5 Adding vowels

The previous experiments already hinted at the influence of the probability of adding new vowels p_i. This value had to be adapted to the total number of games the agents played in order to get approximately the same number of vowel clusters in the emerging vowel systems. The question now arises if the only influence of this parameter is the speed with which the vowel clusters emerge or whether it also has an influence on the quality of the vowel systems.

Figure 4.15 shows the vowel systems that emerge as well as the three quality measures success, energy and size for different settings of p_i. The values shown are, from left to right: 0.002, 0.005, 0.01, 0.02 and 0.05. The other parameters were given their default values and the acoustic noise ψ_{ac} was set to 10%. The simulations were run for 5000 imitation games. As usual, the quality measures calculated over 100 runs of the simulation are given with their standard deviations, which should be taken with a grain of salt, because the measurements do not come from an unknown distribution.

It can be seen that the number of vowel clusters increases with increasing phoneme insertion probability. However, there is a limit to this. If the vowel space is already maximally occupied, no new vowel clusters can be successfully added. So increasing the phoneme addition increases the speed of growth of the vowel systems rather than their final size. The other measures: energy and system size, do not seem to be influenced very much by the speed with which new phonemes are added to the agents' vowel systems. Apparently the way in which the agents adapt vowels that are good in the imitation games and discard vowels that are not good in the imitation games is sufficiently efficient to prevent randomly added vowels from disrupting their vowel systems. However, in order to make sure that the different rates of adding vowels really do not have much influence, the results for different values of the rate with which new vowels are added should be compared with the results obtained from running the imitation games with a fixed rate of vowel addition, but for different numbers of games. It can then be seen whether the differences between the vowel systems in figure 4.15 have to do with the rate of random vowel addition

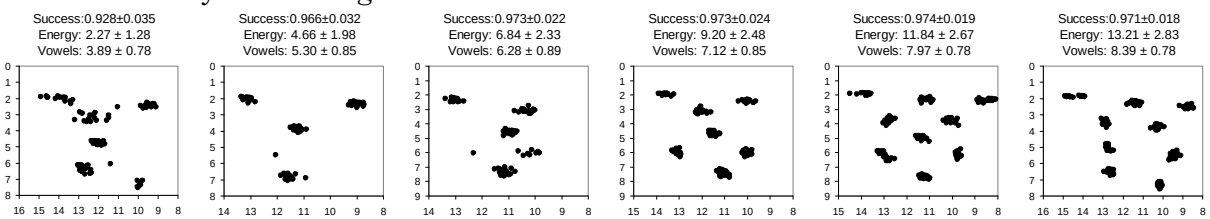


Figure 4.15: Influence of different rates of adding new vowels.

or rather with the absolute number of random vowel additions.

Vowel systems that resulted from the same parameter settings, but for different numbers of imitation games are shown in figure 4.16. In this figure, example vowel systems that emerged after 1000, 2500, 5000, 10 000, 25 000 and 50 000 imitation games are shown from left to right. The probability of adding a random new vowel p_i was always equal to 0.01, so that the (average) number of added vowels is equal to the corresponding graph in figure 4.15 for each graph. It can be seen that only for the smallest number of imitation games, the resulting vowel system is appreciably different from the corresponding (leftmost) vowel system in figure 4.15. The clusters of the vowel system did not quite have the time to converge fully. In figure 4.1 this could already be observed. It can be concluded that the vowel rate of addition of new vowel prototypes does not influence the vowel systems very much. The main cause of the differences in figure 4.15 are thus the total numbers of added vowels, so that they are in fact in different stages of their development.

In fact, figure 4.16 shows the continuation of what is shown in figure 4.1, which stopped at 4000 imitation games. After 4000 imitation games, the number of clusters still increases until the available acoustic space is filled with vowel prototypes. However, the structure of the vowel systems does not change very much anymore. After 25 000 games, it seems as if the maximum number of vowel clusters is reached. From then on the vowel system remains relatively stable, although the vowel prototypes continue to shift position.

4.4 An Articulatory View of the Systems

So far all the systems have been shown in acoustic space. This was done because the agents have to distinguish vowels in acoustic space. The distribution in acoustic space is therefore the most relevant information for evaluating the quality of the emerging vowel systems. Moreover, the two-dimensional plot of the acoustic space is much easier to interpret than a three-dimensional plot of the articulatory parameters. However, it would be interesting to know whether the vowels also form clusters and whether they are also evenly distributed in articulatory space. This is not directly obvious. Acoustic space is two-dimensional (the first formant and the effective second formant), whereas articulatory space is three-dimensional. This means that many different possible articulations will map on the same acoustic signal. Vowel clusters in articulatory space could therefore be much more dispersed than vowel clusters in acoustic space.

This is illustrated in figure 4.17. This figure shows a vowel system that was obtained with the default values for the parameters, after 25 000 imitation games in a population of twenty agents with the acoustic noise ψ_{ac} set to 20%. This figure is an attempt to plot three-dimensional information on a two-dimensional plane. In the lower left corner is a perspective plot of the three-dimensional system. The grey circles represent the agents' vowel prototypes. The projections of these prototypes in the rounding-position, the rounding-height and the position-height planes are plotted as little dots on the walls of the plot. These points can be considered as "shadows" of the vowel prototypes. For reference these planes are also plotted as flat squares around the perspective plot. The same vowel system, but plotted in the acoustic space is given in the left part of figure 4.19.

It can be seen that the clusters are quite concentrated in the position and height dimensions, but that they are rather dispersed in the rounding dimension, except for the cluster representing [u]. However, the clusters are well apart in the acoustic space as a whole, so they are easily kept apart. In a system of only three vowels it appears not to be necessary to control all articulatory parameters equally precisely. One could say that there is considerable allophonic variation in the agents' vowel systems of this population. The two parameters that are most carefully controlled are vowel position and height, as these influence the first and effective second formant most directly.

In figure 4.18 on the other hand, most of the clusters are much more compact. Just as in human vowel systems, back vowels are rounded. The only exception to this is the low back vowel, which is also often unrounded in human languages. The high and low front vowels of the agents are unrounded as well, which agrees with the observation that front vowels are usually unrounded in human languages. Apparently there is an emergent rule: $[\pm \textbf{front}] \rightarrow [\mp \textbf{rounded}]$. However, the mid front

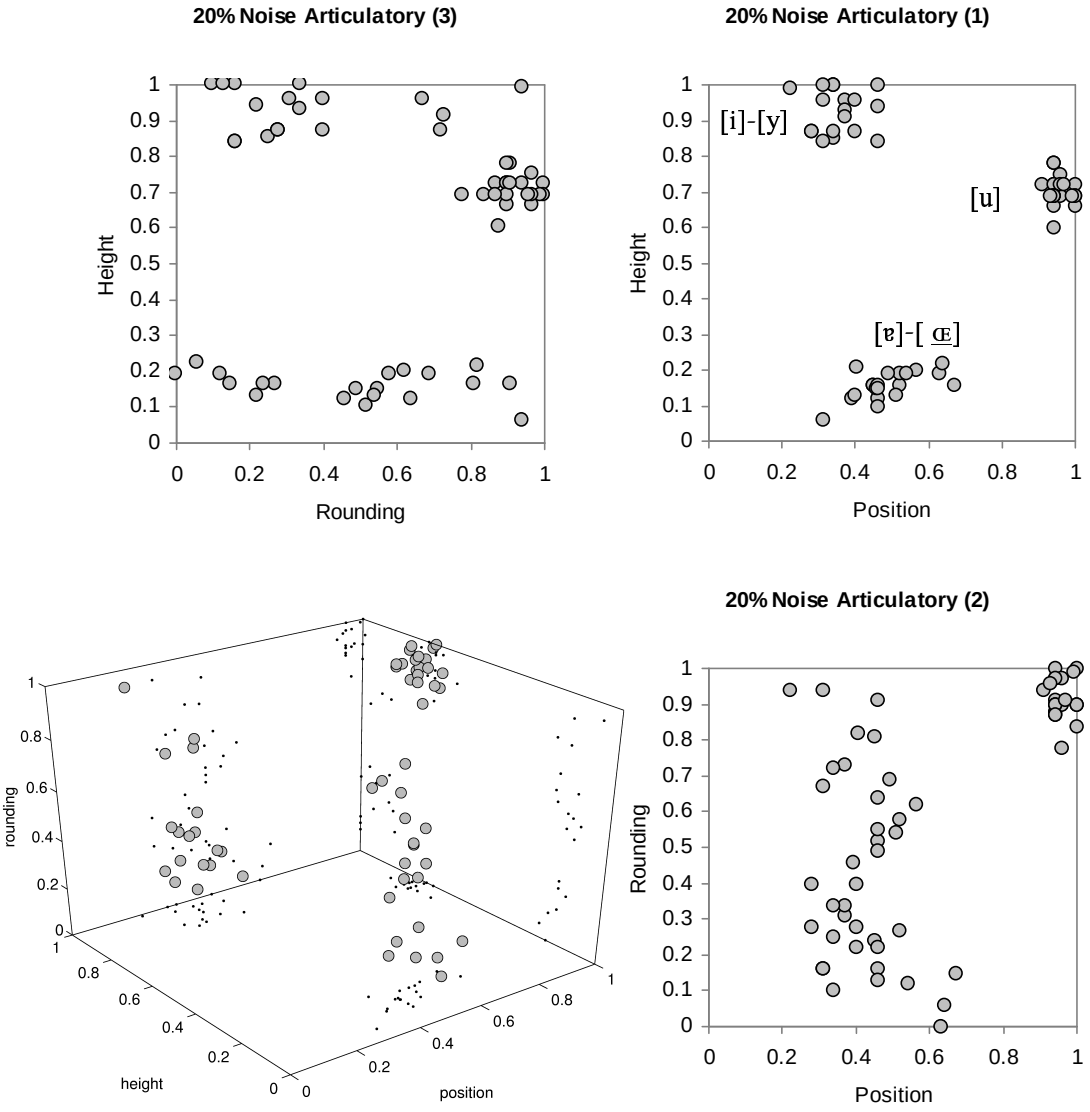


Figure 4.17: Articulatory representations of 20% noise system.

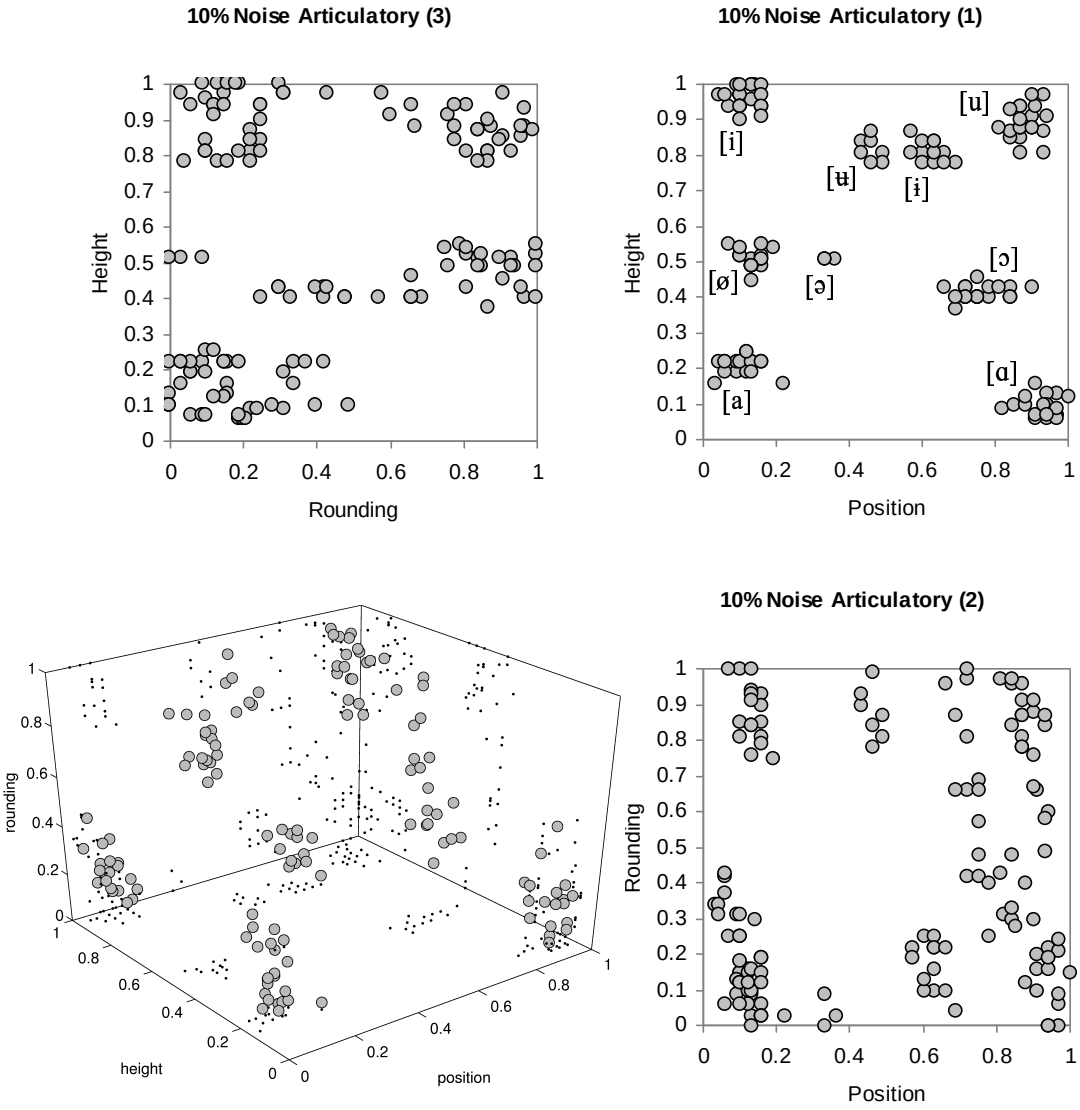


Figure 4.18: Articulatory representation of 10% noise system.

vowel is rounded (with for three agents the schwa [ə] as allophone). The high central vowel is rounded for some agents and unrounded for others, without any intermediate values. However, if one looks at the acoustic realisation of this vowel (shown in the right part of figure 4.19) one finds only one compact cluster. Apparently the acoustic realisations of the two different articulations are almost the same. This was to be expected, because the acoustic space is two-dimensional and the articulatory space is three-dimensional. But this phenomenon is not limited to the artificial system. Humans, too are able to produce the same sound with different articulatory settings (see e.g. Lindblom 1972, Maeda 1989).

The articulatory view of the emerging vowel systems shows that, even though articulatory constraints hardly play a role in the success of the imitation games, compact clusters do form in the articulatory space. Also, the articulatory parameters of the vowels in the agents' repertoires are comparable to the ones of the vowels in human systems, with, for example, rounded back vowels and unrounded front vowels (although this is not always the case). But it can be concluded that next to the

distribution in acoustic space, also the distribution in articulatory space of the emerging vowel systems is natural.

4.5 Conclusion

The results show that realistic systems emerge for a large range of different parameter settings. Realistic systems are systems that look like the vowel systems of human languages. These have a discrete number of non-overlapping vowels (if one compensates for the influence of context) that can usually be described by a number of distinctive features, such as high/low, front/back and rounded/unrounded. Furthermore, the vowels of a human system are typically dispersed through the available acoustic space. A typical vowel system of a human language (French) was shown in figure 4.2. The similarities between this figure and the figures of the vowel systems that emerged from the imitation games are striking. In chapter 6 the similarities will be explored further.

The similarity between the emerging vowel systems and the natural vowel systems is also illustrated by a number of measures. These are the success, the size and the energy of the vowel systems. Success is the ratio between the number of successful imitations and the total number of imitation games played. Although it is true that the performance of randomly generated vowel systems can already be quite high (over 50%, see appendix B and C) the success of the emerging vowel systems is usually over 90%, which is way better than random performance. The energy is a measure of how dispersed the vowel systems are. A minimal energy indicates maximal dispersion. Natural vowel systems are usually dispersed, so that their energy is near-minimal. The vowel systems that emerge from the simulations have energies that are much nearer to the minimal value than to the random value. This indicates that the systems are dispersed and therefore natural. The size of the vowel systems depends on a number of acoustic and articulatory parameters. The “easier” it is for agents to imitate sounds they heard (because of less noise and more accurate control of the articulators) the higher the number of vowels in their vowel systems. However, this was an emergent result of the settings of these parameters. Nowhere was it determined explicitly how high the number of vowels in the system would be.

Just like natural vowel systems, the systems that emerge can also be described by distinctive features, even though these do not play any role at all in their development. For example, in figure 4.19 tentative distinctive features are illustrated for two systems that emerged for the default parameter settings, the left one with ψ_{ac} = 20% and the other one for ψ_{ac} = 10% after 25 000 imitation games (the articulatory representations of these systems was given in figures 4.17 and 4.18, respectively). These features are just a convenient way of describing these systems, and are not “real” in the sense that they are part of the agents’ cognitive system. Nevertheless, the same features that are needed to describe human vowel

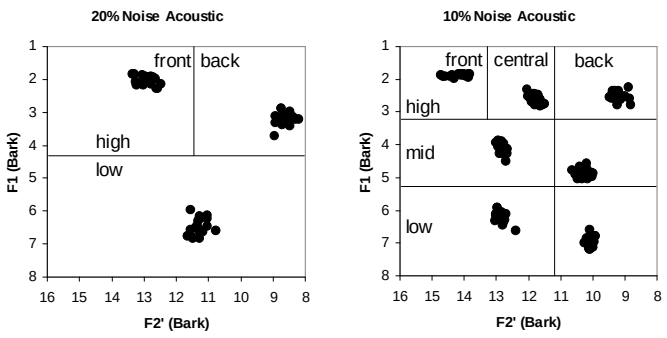


Figure 4.19: Distinctive features in emergent vowel systems.

systems can be used to describe the artificial systems.

The imitation games result in natural looking vowel systems, but not for all settings of parameters. However, the simulation is quite robust. It is not the case that realistic results are only obtained for certain very specific parameter values. As has been shown in this chapter, all parameters can be varied to a considerable extent and still realistic systems will emerge. The acoustic noise ψ_{ac} can be varied from 5% to 20-25%. This will influence the number of vowels in the emerging vowel systems, but this number will be in the range possible for human languages. The articulatory noise ψ_{art} could be varied between 0% and 5-15% (depending on the settings of the other parameters) without corrupting the ability of agents to imitate each other. The parameter λ which determines the relative influence on perceptual distance of F_1 and F_2' could be varied from 0.2 to 0.5 while still resulting in realistic vowel systems. The parameter ε, the step size with which agents would improve their vowel prototypes in reaction to the language games could be varied between 0.01 and 0.05 without disrupting the imitation games. The population size can be varied from 10 to at least 100 agents for realistic vowel systems. Smaller populations do achieve successful imitation, but their vowel systems are not stable, and somewhat smaller than the vowel systems of larger populations. The probability p_i with which new vowels were added could be varied from 0.001 to 0.05 without changing the performance of the imitation games. The only change is that higher probabilities of addition result in faster development of the vowel systems.

So, although it certainly is not true that the simulation produced realistic vowel systems for all parameter settings, it nevertheless produced realistic vowel systems for considerable ranges of the parameters. In fact, varying these parameters within the ranges that resulted in realistic systems, generally produced the different types of vowel systems that can be found in human languages. For certain settings of parameters, vertical two- or three vowel systems emerged, for other values symmetrical systems with five or more vowels emerged. Some settings resulted in systems with only peripheral vowels, while other settings resulted in systems with many central vowels.

One issue that should be addressed is the number of 5000 imitation games that was played in most of the experiments. In figure 4.16 it could be seen that for the default settings of the parameters the number of vowels in the agents' repertoires continued to grow after 5000 imitation games until approximately 25 000 imitation games. It is therefore possible that a number of phenomena have been missed because the vowel systems were not completely developed, yet. However, the limit of 5000 imitation games was universal for all experiments presented here. The experiments were meant to compare the results of different settings of parameters. As the conditions for all experiments were the same, the comparison should be fair. For the cases were the number of imitation games could play a role, such as in the experiments with different population sizes and different rates of addition of new vowels, the number of imitation games *was* varied.

The question that should be answered by the results presented in this chapter is whether the emerging vowel systems are really the result of the interactions between the agents or whether they are rather the result of the process of random vowel insertion. In Berrah's work (Berrah 1998) a population of agents was also used, but in fact, the structure of the resulting vowel systems was mostly determined by the repelling forces between the (fixed number of) vowels in each individual agent's vowel system. This was illustrated by the fact that a population of one agent

resulted in the same kinds of vowel systems as populations of more agents. Could a similar phenomenon also play a role here?

In fact, the random insertion does play an important role in the simulations presented here. Yet its role is not so much in determining the structure of the vowel systems, but rather in causing the number of vowels of the agents to increase. The interactions between the agents determine the structure of the emerging vowel systems and cause the rapid spread of new vowel prototypes within the population. The minor role of random insertion is illustrated by the vowel systems that emerge after 5000 imitation games, with, for example insertion probability $p_i = 0.002$ in a population of twenty agents (shown as the second frame from the left in figure 4.15). In this case ten random insertions have taken place. However, four (on average 4.41) distributed vowel prototypes have formed in all agents, and the prototypes of all agents already form compact clusters. This indicates that another process than random insertion must be responsible for most of the vowel prototypes in the agents. This other process, of course, is the imitation game. But the question remains to what extent the imitation games are responsible for the structure of the emerging systems. This question can be answered partly by the results of the experiments with the changing population size. In these it was found that vowel systems in smaller populations were less stable than vowel systems in larger populations, even though the number of insertions was *smaller* in the former than in the latter. This indicates that the interactions between the agents are responsible for the stability of the structures that emerge.

The random insertions provide possible positions for new vowel prototypes. The imitation games, together with the constraints on production and perception cause that certain vowels will be favoured over others. The other agents in the population will copy these vowels more easily, causing them to spread faster than less favoured vowels. Whether a vowel is favoured or not does not just depend on its absolute position but also on the positions of the other vowels that are already present in the system. Whether the imitation games themselves cause the prototypes to spread more evenly over the available acoustic space is not clear. But it is conceivable that random shifts of the vowel clusters can be selected for in the same way as the vowel prototypes themselves. If a certain shift makes it easier for agents to imitate a vowel, the other agents will rather adopt its position after the shift than its position before the shift. However, the experiments did not produce statistically significant data to settle this issue.

In any case, the results presented here show that the interactions between the agents play an important role in determining the vowel systems that emerge in the population of agents. None of the agents has central control over the emerging vowel system and the qualitative properties of the emerging vowel systems are not sensitively dependent on the history of the population. It can therefore be said that the vowel systems truly are the result of self-organising emergence under constraints of perception and production in the population of agents. The emerging vowel systems show remarkable similarities with human vowel systems, under different settings of parameters and even though the models of perception and production are rather crude. Therefore it is reasonable to believe that self-organisation also plays a role in the emergence of human vowel systems

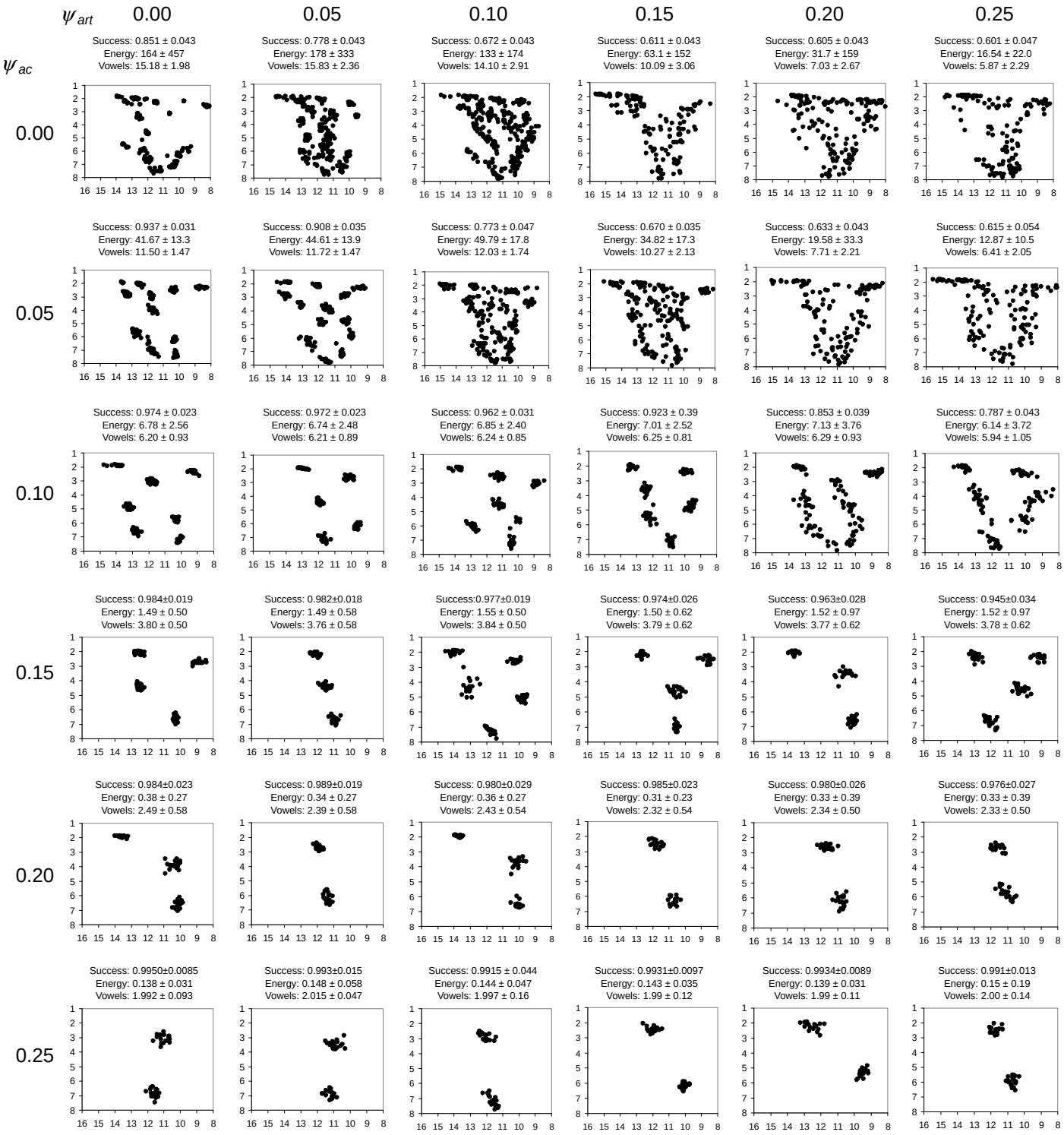


Figure 4.20: Results of changing articulatory and acoustic noise.

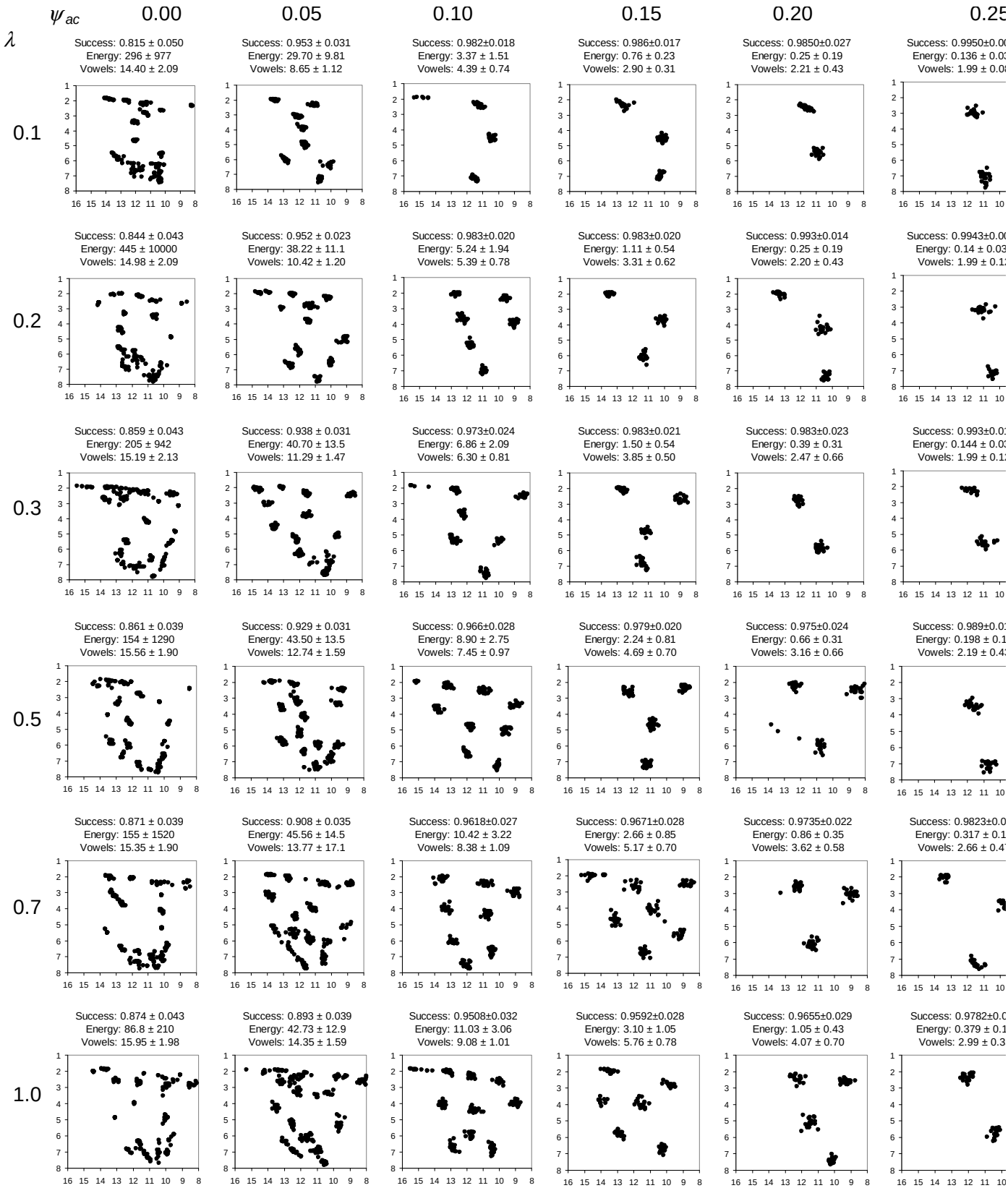


Figure 4.21: Results of changing λ and acoustic noise.

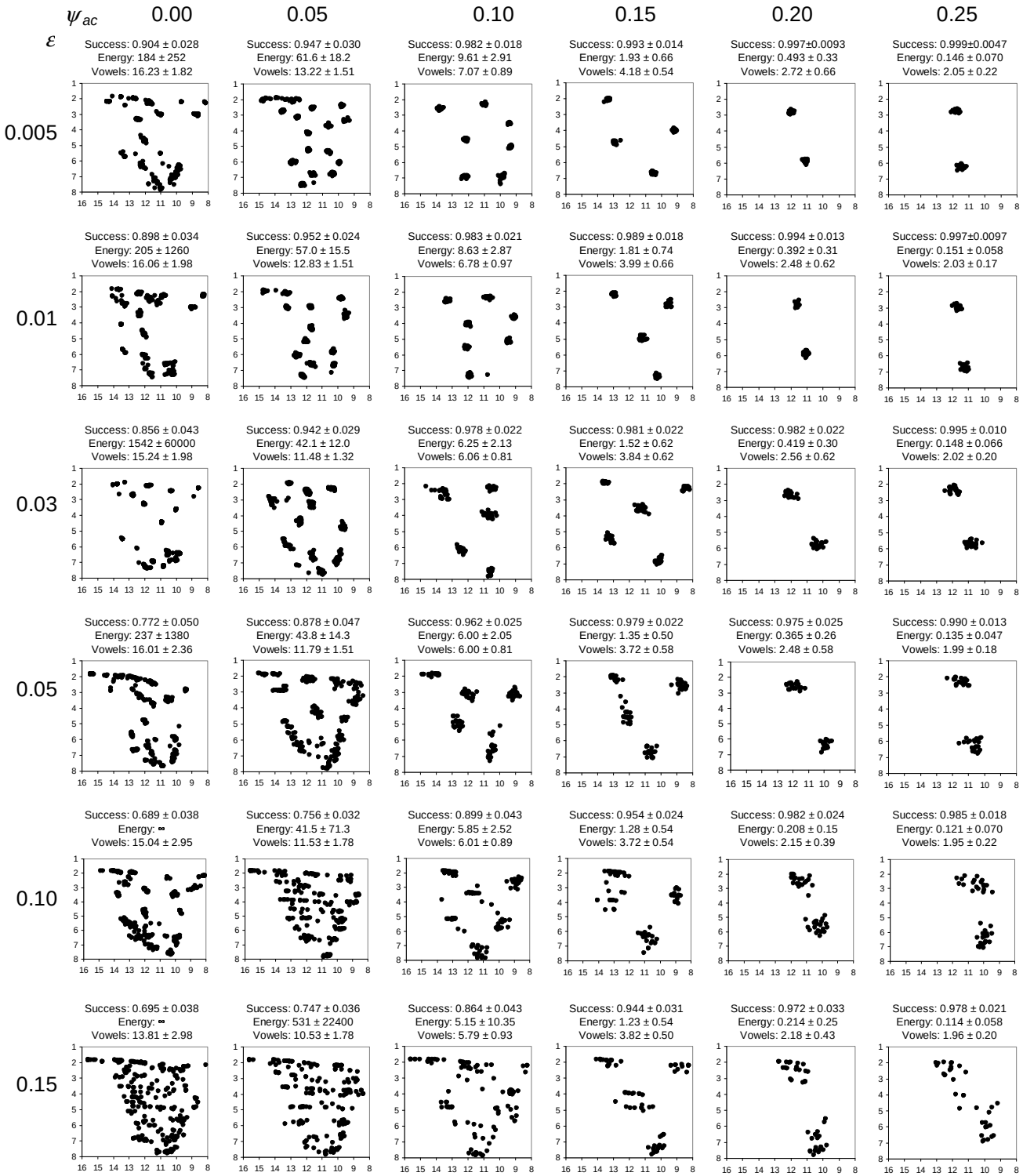


Figure 4.22: Results of changing step size and acoustic noise.

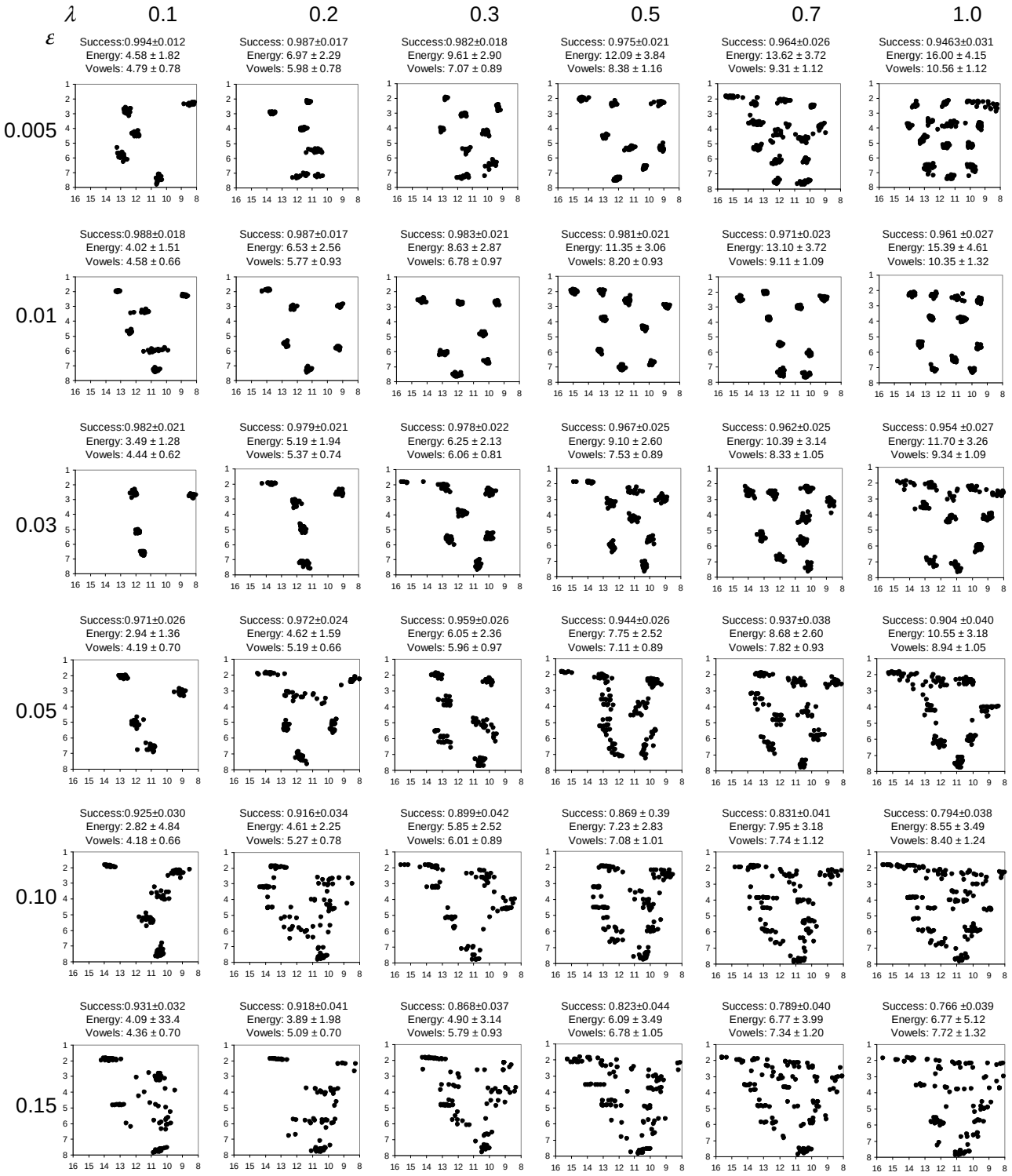


Figure 4.23: Results of changing step size and λ.

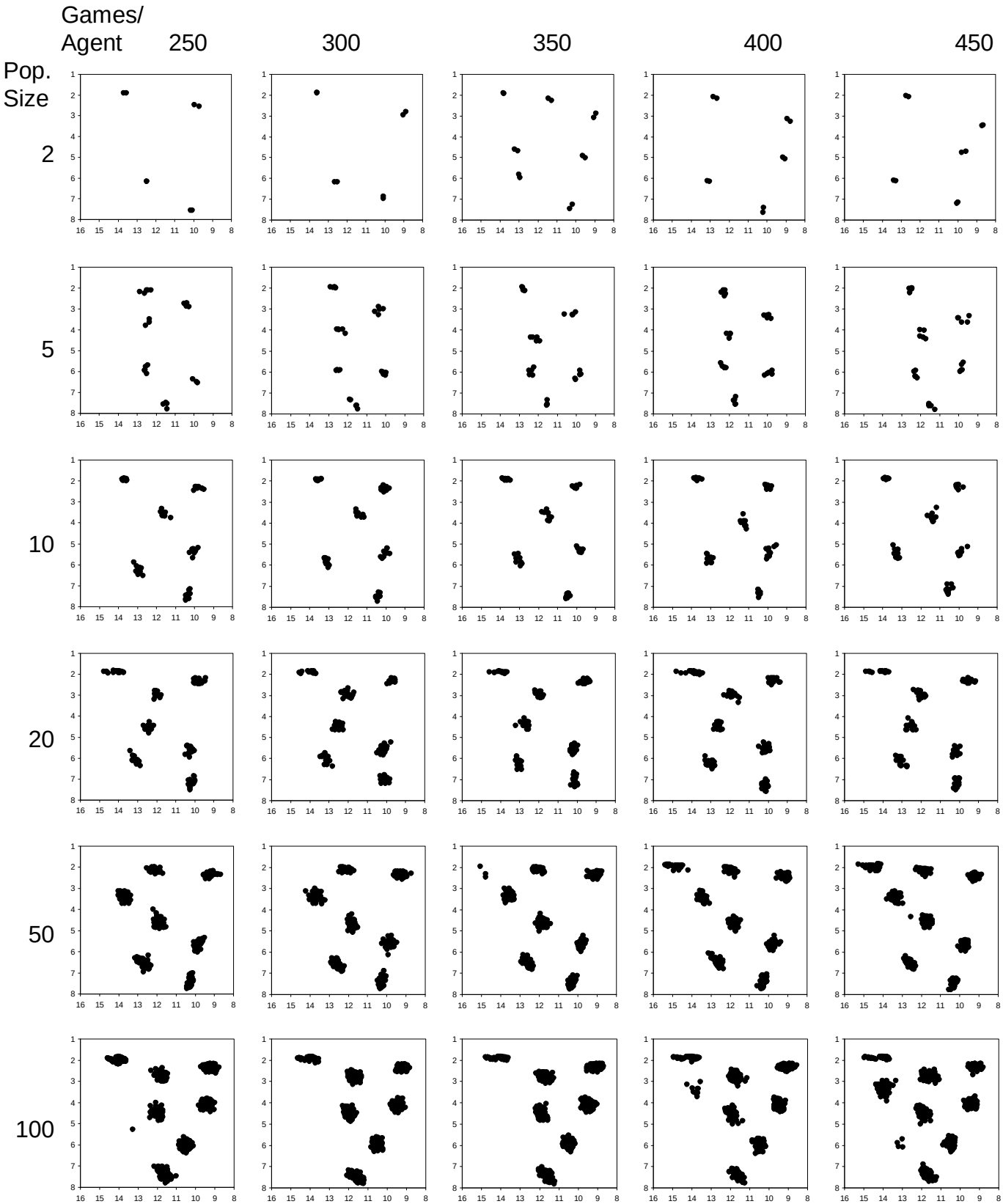


Figure 4.24: Evolution over time of vowel systems in populations of different sizes.

5. Qualitative Changes of the Simulation

Quantitative changes of the simulation, consisting of simple variations of parameters, have been investigated in the previous chapter. However, in the design of the simulation a number of more or less arbitrary decisions have been made. These were either made in order to simplify and speed up the implementation, or because it was not clear which of a number of possible alternatives was the most realistic. In this chapter these decisions will be reconsidered, now that the behaviour of the basic simulation has been investigated and its behaviour is understood in some detail. A number of experiments with alternative, qualitatively different simulations will be presented.

Of course there are many different possible variations, so only a subset of these can be tried out. Three variations on the simulation have been chosen, because they are found to be interesting in some respect. The first variation that will be investigated is a simulation with a variable population. In real language communities the population of speakers is not fixed. Speakers can enter (by birth or by immigration) or leave (through death or emigration) the population. The fact that language is an open system with respect to the things that can be expressed as well as to the population of language users is an essential part of Steels' theories (Steels 1997b, Steels 1998b, Steels & Kaplan 1998). It has already been shown in the previous chapter that the vowel simulation is an open system regarding the number of sounds of the agents, but it is also necessary to show that it is an open system regarding the population of speakers.

The second variation that will be discussed is a variation on the implementation of the imitation game. An unrealistic aspect of this imitation game, which was already mentioned in chapter 3 was the non-verbal feedback that the agents give each other at the end of each imitation game. It is generally accepted that humans do not learn language by being constantly told whether what they said was correct or not. Although the non-verbal feedback that humans receive does not need to be so explicit as the non-verbal feedback of the agents, it still is interesting to know the behaviour of a model that does not depend on it.

The last variation that will be discussed is a variation on the perception system of the agents. The signals transmitted between the agents in the simulations presented so far consisted of the frequencies of the first four formants of the vowel signals. The distance between two signals was calculated using a weighted non-linear distance function (Mantakas *et al.* 1986, Boë *et al.* 1995, Schwartz *et al.* 1997). The question might arise to what extent this is realistic. Therefore the model of perception is changed in order to work with real signals, so that the model can even be tested with a human participating in the imitation games.

5.1 Variable Populations

According to Steels' theory of language as an adaptive system (Steels 1996, Steels 1997b, Steels 1998b) a language should be an open system. A language should be able to accommodate linguistic innovation, but it should also continue to exist despite changes of the population. It has already been shown that the simulation presented in the previous chapter is able to generate successful vowel systems from scratch. Also, agents can introduce new vowels in their vowel repertoires. If these vowels can easily be distinguished from the other already existing vowels, all agents in the population will quickly adopt them. So it is clear that the "language" (consist-

ing of the vowel systems) of the agents is an open system regarding changes in the "language" itself. However, it remains to be shown whether it is also an open system with respect to changes of the population. In research within the same theoretical framework, but directed towards different aspects of language, such as lexicon and semantics, simulations have shown that changes in population (within certain limits) do not influence the "language" very much (Steels & Kaplan 1998, Kaplan 1998).

5.1.1 Definition of measures and parameters of population change.

In order to investigate this, changes in the population should be made possible. This means that it must be possible to add new agents and to remove old ones. There are several ways to do this, but the way that is simplest, most realistic and least likely to introduce artefacts is doing it stochastically. Two probabilities are introduced: the probability p_b of a new agent being "born" into the population and the probability p_d of an agent to die. Using a stochastic scheme for changing the population is less likely to introduce artefacts than any scheme that is based on regular replacement of the population (for example replacing agents after a fixed number of time steps). Regularities are always arbitrary and might interfere in unknown ways with the imitation games. Stochastic replacement is also most realistic, because in human populations, too, birth and death are stochastic phenomena that cannot be predicted. However one *can* make a good prediction of the average rate of birth and death. These average rates are determined by the two probabilities.

Birth and death of agents take place with probabilities p_b and p_d every imitation game. A new agent with an empty vowel repertoire is added with probability p_b, while both the imitator as well as the initiator can be taken out of the population with probability p_d. This means that if p_d and p_b are equal and non-zero, the population size will effectively decrease because on average two times as many agents will be removed from the population as will be added. For a stable population size on average, p_b should therefore be two times p_d.

A number of derived measures that describe the changes of the population can now be defined. The excess of birth over death, or growth of the population g_p is defined as:

5.1) $$g_p = p_b - 2p_d .$$

If two times the probability of agents dying is bigger than the probability of agents being born, g_p will be negative and the population size will decrease on average. The expected change of the number of agents after N_{games} imitation games is: $g_p \cdot N_{games}$.

The age a of an agent is defined as the number of imitation games in which it has participated. Because an agent has a fixed chance of dying in every imitation game in which it participates its life expectancy l is the reciprocal of the probability of dying, in other words:

5.2) $$l = 1/p_d .$$

Of course this does not mean that all agents disappear out of the population after l imitation games in which they participated.

It is also useful to know how fast agents are being replaced in the population. The replacement rate, or flux f_{pop} is the number of agents that changes in every imitation game. It can be calculated as follows:

5.3) $$f_{pop} = p_b + 2p_d$$

which is just the sum of the average number of agents that dies and gets born in every imitation game.

The half-life time τ of a population of agents is the number of imitation games that has to be played before half of the agents in the original population has died. Calculating the half-life time of a population of agents in terms of the number of games per agent is easy. Every time an agent participates in a game, it has a chance of $1-p_d$ to die. The expected value of the number of agents in the population after all agents have participated in i imitation games per agent is:

5.4) $$\langle N_i \rangle = N_0 (1 - p_d)^i$$

where N_i is the number of agents after i games per agent and N_0 is the original number of agents. In order to calculate the half-life time, one has to find the i_τ for which the expected value of $N_{i\tau}$ is half N_0:

5.5) $$\frac{\langle N_{i\tau} \rangle}{N_0} = (1 - p_d)^{i_\tau} = 0.5$$

which solves to:

5.6) $$i_\tau = \frac{\ln(0.5)}{\ln(1 - p_d)}.$$

However, it is much more interesting to know the absolute number of imitation games necessary so that half the original population has died out. This is quite hard to do in the general case of a changing population. But most of the experiments will be done with a population that is stable on average, in other words $g_p = 0$. In this case the number of games per agent i is:

5.7) $$i = 2 \cdot \frac{N_{games}}{N_{pop}}$$

where N_{pop} is the number of agents in the population and N_{games} is the total number of games played. Together with equation 5.6 and by taking $N_{games} = \tau$ this gives:

5.8) $$\tau = \frac{N_{pop}}{2} \cdot \frac{\ln(0.5)}{\ln(1 - p_d)}.$$

Knowing the half-life of a population of agents is useful, because with the law of exponential decay it can be calculated how many games are needed until for example 5% of the original population is left over—in this case slightly more than four times τ games.

5.1.2 Maintaining a vowel system.

The first experiment that will be done checks whether a vowel system that has emerged in a population that does *not* change can be maintained in a population that *does* change. The parameters of the simulation are set to the same default values as in the previous chapter (see appendix A for these default values). The population size is 50 agents and the acoustic noise ψ_{ac} is set to 10%. The population size is larger than in the experiments in the previous chapter, because the death and birth of new agents make the population size fluctuate. Fluctuations in a population of 20

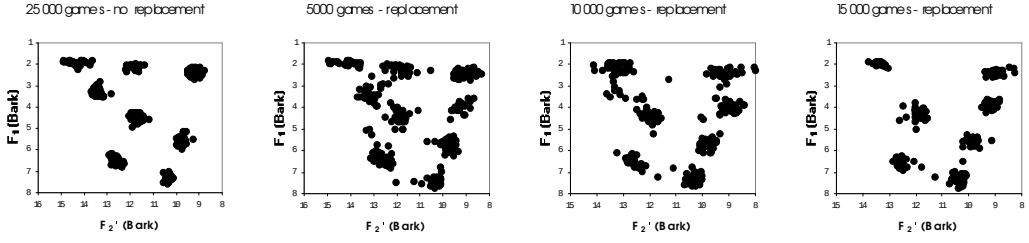


Figure 5.1: Vowel systems of imitation games with population replacement.

agents might easily make the population too small, whereas fluctuations in a population of 50 agents have a much smaller relative influence. The results are shown in figure 5.1. First the simulation is run for 25 000 imitation games in order to generate a vowel system that is almost fully developed. Then the probability of adding new agents p_b is set to 0.01 and the probability of agents being removed p_d is set to 0.005. Thus the growth of the population g_p is 0, the flux f_{pop} is 0.02, the life expectancy l of an agent is 200 games and the half-life of the population τ is 3457 imitation games. Figure 5.1 shows snapshots of the vowel systems of all the agents in the population with intervals of 5000 games. In the rightmost frame, 15 000 imitation games have been played since the first frame, meaning that only 5% of the agents in the original population are expected to be present. The population size is not completely stable. The original population consisted of 50 agents. In the second frame this has become 63, in the third 58 and in the last 42. This figure can be compared with figure 4.24, where the second row from the bottom shows the evolution of a vowel system of a stable population with intervals of 2500 games. It is clear that the stable population results in more stable vowel systems. On the other hand, the vowel system in figure 5.1 does not change completely randomly either. Apparently, the vowel systems that emerge from the stable population are too crowded for the changing population. A number of vowel clusters merge, there is a brief period of chaos and then a new stable system emerges. Still, it remains true that this vowel system is less stable and successful than the vowel system in a population that does not change.

A number of variations of this experiment can be imagined. Obviously, the parameters of the simulation can be changed in the same way as those of the simulation with the unchanging population were changed in the previous chapter. But more interesting experiments can be done by changing the parameters p_b and p_d that are unique to this simulation. The results of the previous experiment already suggested that the vowel systems of changing populations converge to configurations with fewer vowel clusters than in populations that do not change. In figure 5.2 the long-term behaviour of vowel systems in populations that change with different rates is explored. The frame to the left shows the vowel system with which the populations started, which is in fact the same system as with which the simulation in figure 5.1 started. The second frame shows the vowel systems of a population with p_b = 0.1 and p_d = 0.05 after 2500 imitation games. In this case τ = 338, so that the less than 1% of the original population is expected to be present. The third frame shows the vowel systems of a population with p_b = 0.01 and p_d = 0.005 (effectively the same settings as in figure 5.1) after 25 000 imitation games. The last frame shows a system with p_b = 0.001 and p_d = 0.0005 (τ = 34 649) after 250 000 imitation games. The population sizes have not remained completely constant. The population of the first system has increased to 59, of the second system to 81 and the popula-

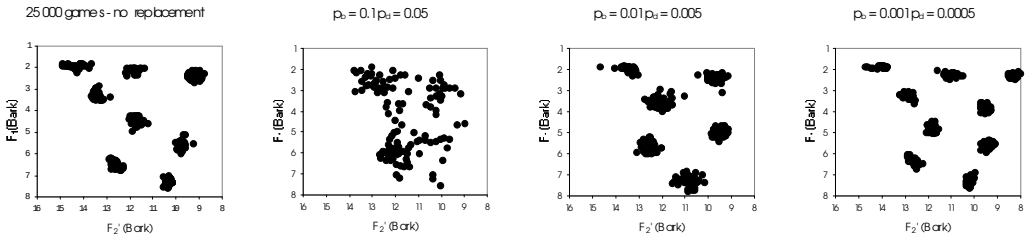


Figure 5.2: Vowel systems after complete population replacement.

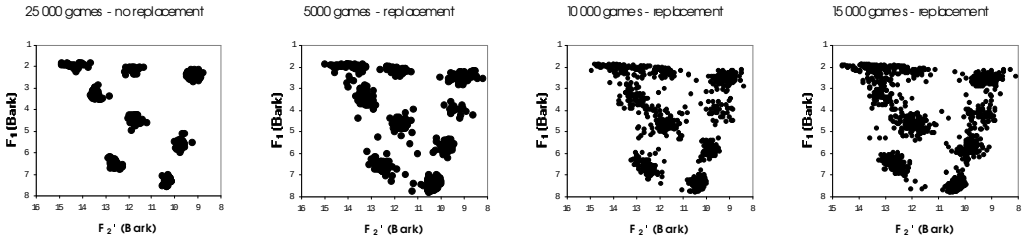


Figure 5.3: Evolution of vowel system in population with only births.

tion of the third system has decreased to 43. It is obvious that although the populations have undergone approximately equally many replacements, the system with the highest replacement rate becomes the most chaotic, while the system with the lower replacement rate remains most stable. This is to be expected, as new agents in populations with lower replacement rates have more time to get used to the vowel systems of the other agents. Apparently agents in a population with a replacement rate f_{pop} of 0.2 only have time to learn two or three vowels (the number of clusters in the plot of the population's vowel systems). Agents in a population with f_{pop} = 0.02 have time to learn six vowels and agents in a population with f_{pop} = 0.002 have time to learn the maximum number of vowels that is stable in an unchanging population with the same parameter settings.

Table 5.1 shows the average success values, energies and inventory sizes over 100 runs of the simulation for these three parameter settings, all starting with the vowel system that is given in the leftmost frame of figure 5.2. They are given with their standard deviations. It should always be kept in mind that the distributions are not normal, especially not of the energy. When this table is compared with the results of the previous chapter, it becomes clear that vowel systems in a changing population have less success than vowel systems in an unchanging population. It is also clear from this table that the final number of vowels is significantly (using the Kolmogorov-Smirnov test) lower for higher population fluxes than for lower ones.

	f_{pop} = 0.2 1500 games	f_{pop} = 0.02 15 000 games	f_{pop} = 0.002 150 000 games
Success:	0.8117 ± 0.032	0.7675 ± 0.035	0.8375 ± 0.043
Energy:	0.40 ± 0.79	2.84 ± 2.42	7.00 ± 3.34
Size:	1.88 ± 1.03	4.55 ± 1.67	7.02 ± 1.34

Table 5.1: Statistics of changing populations.

5.1.3 The sources of disturbance.

New agents that enter the population and that do not yet know the vowel system are the main source of disturbance. But it is not clear whether it is also necessary to remove the original agents from the population to really create disorder in the system. This question is answered in figure 5.3. In this figure the simulation is again started with the same population as the previous experiments and with the same parameters. The probability of adding new agents p_b is set to 0.01, but no agents will be removed from the population (i.e. p_d = 0). This means there is a growth rate g_p of 0.01. The first frame shows the original vowel system. Subsequent frames show the vowel systems after 5000, 10 000 and 15 000 imitation games. The number of imitation games played is the same as for the systems shown in figure 5.1. The population sizes are 50, 112, 183 and 235, respectively. It can be seen in this figure that the position and the number of the vowel clusters remain approximately the same,

but that a lot of new vowel prototypes have been formed in between, thus disrupting successful imitation.

Apparently vowel systems are disrupted by two processes: the rapid insertion of new agents that do not have time to learn the vowel system, and the removal of agents that have already learnt the vowel system and that could function as a target for younger agents. Obviously, if older agents are not removed from the population, something like a target for younger agents stays in the population for longer. However, if the number of new agents increases too rapidly, the number of interactions between old and young agents becomes insignificant compared to the number of interactions between young agents amongst themselves. Also, the older agents will adapt their vowels towards the vowel prototypes of the younger agents so that gradually the original positions of their vowels will also be lost. The vowel systems of a population in which no agents die are therefore not more stable than those of a population in which agents do die.

5.1.4 Emergence of a vowel system.

Stable vowel systems can be maintained in populations that change, even though the number of vowel prototypes might be smaller than in populations that do not change. The question now arises whether a vowel system could also emerge from scratch in a population that changes. It was shown in the previous chapter that vowel systems did emerge from scratch in the case of a stable population. In the data presented so far in this chapter, it has been shown that the same mechanism that was responsible for the emergence of the vowel systems in the unchanging populations could also be used by new agents to learn an existing vowel system. If it is shown that the same mechanism can be used for making a vowel system emerge in a changing population, this is strong support for Steels' hypothesis (Steels 1997b, Steels 1998b) that the same mechanisms that are responsible for learning language could be responsible for the emergence of a new language.

Figure 5.4 shows the results of such an experiment. In this figure, which can be compared to figure 4.1, the emergence of a coherent and realistic vowel system can be observed. The frames show the vowel system of the population of 50 agents after 1000, 2000, 5000, 10 000 and 20 000 imitation games from left to right. The probability of new agents p_b was set to 0.01 and the probability of agents being removed p_d was set to 0.005. The half life time τ of the population is then 3457 games, implying that after 20 000 imitation games on average only one agent of the original population is expected to remain. The number of agents in the populations shown in the frames was 59, 62, 67, 66 and 64, respectively. Two differences between figure 5.4 and figure 4.1 catch the eye. The first is that fewer and more spread out vowel prototypes emerge. This is in line with the findings of the experiments presented above. The second difference is that the emergence of the vowel system goes much

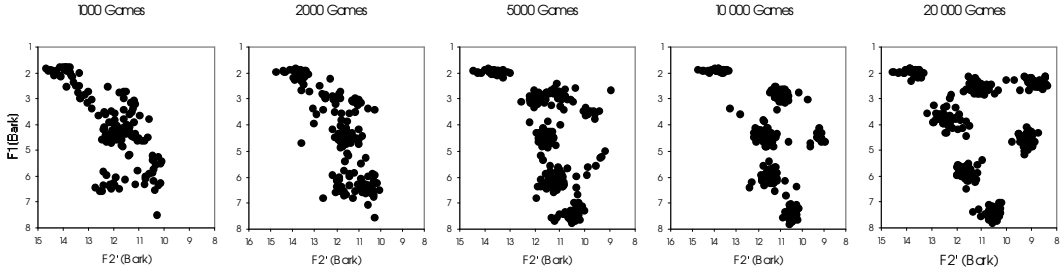


Figure 5.4: Emergence of a vowel system in a changing population.

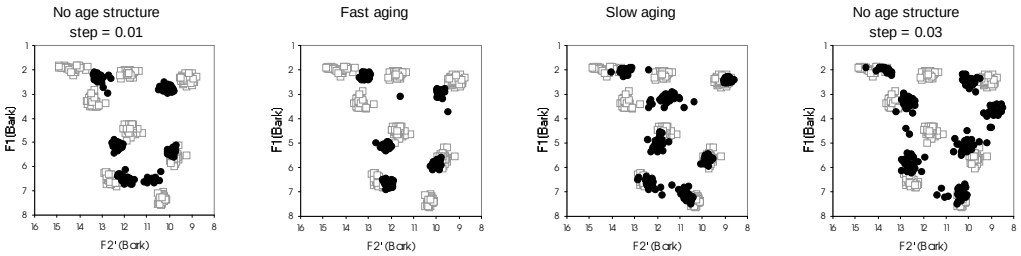


Figure 5.5: Influence of age structure on transfer of vowel systems.

slower than in the case of the unchanging population. Apparently the flux of agents that enter and leave the population makes it harder for a common vowel system to be accepted by all agents.

These results show that the imitation game truly is an open system, both with respect to the language itself and with respect to the population of speakers. Even though both the language and the population of speakers are highly dynamic, a stable (and realistic) system of sounds for successful imitation emerges.

5.1.5 Age structure.

It is a pity, however that in populations that change, not so many vowel clusters can be maintained as in systems in an unchanging population. Decreasing noise levels or decreasing the step size ε with which vowel prototypes are improved, would increase the number of vowel prototypes, but in unchanging populations with the same parameter settings, the number of vowel prototypes would be higher still. There is one case, however, where changing populations can maintain more vowel prototypes than unchanging ones. This is the case where the number of times an agent can improve a vowel prototype (the number of *practice steps*) is limited (see the section on step size and figure 4.14 in the previous chapter). It was found that if the number of practice steps was limited, there would be an intermediate step size ε where the number of vowels would be highest. In the case of the changing population, an age structure can be introduced, so that young agents have a large step size, so they can imitate new vowels relatively quickly, while older agents have a smaller step size, so they provide a stable target for the younger agents.

An example of the result of an experiment with such a model is shown in figure 5.5. In this figure the vowel systems of four populations are shown. Each of these populations consisted of fifty agents initially and was initialised with the same vowel system. The initial vowel system is shown as open squares in the plots. All populations were run for 15 000 imitation games, with different parameter settings, and the resulting vowel systems are plotted as black circles. All parameters were set to their default values, acoustic noise ψ_{ac} set to 10%, the probability of new agents p_b set to 0.01 and the probability of removing old agents p_d set to 0.005. The maximum number of practice steps for all agents was limited to 10. The leftmost and rightmost frames show the results of populations where there was no age structure. In the left one, a population is shown which had a practice step size ε of 0.01. As can be seen from the frame, the number of vowel clusters after 15 000 games is much lower than in the original system. Clearly, the younger agents have not been able to learn the older agents' vowels. In the right one, the practice step size was set to 0.03. Here the number of clusters has remained about the same, but they have become bigger and slightly more diffuse.

The two middle frames show systems that result from agents that were in between the agents from the leftmost and rightmost frames. These agents changed their practice step size from 0.03 to 0.01 over their lifetime, using the following mechanism:

5.9) $$\varepsilon_t = \varepsilon_{t-1} + \alpha_{aging}(\varepsilon_\infty - \varepsilon_{t-1})$$

where ε_t is the practice step size at time t, ε_∞ is the final step size of old agents (set to 0.01 in these experiments) and α_{aging} is the speed with which the agents age. The population in the middle left and the population in the middle right differ in their values for α_{aging}. The left one has $\alpha_{aging} = 0.1$, causing agents to age quickly and the right one has $\alpha_{aging} = 0.01$, causing agents to age slowly.

Statistics of simulations with parameter settings corresponding to these figures are presented in table 5.2. The columns have the same sequence in the table as the corresponding frames in figure 5.5. The table shows the averages of success, energy, size and similarity over 100 runs of each of the simulations, consisting of 15 000 imitation games, every time starting with the same population of 50 agents that was used as a starting point in all previous experiments. Also shown are the standard deviations. As always, these should not be taken too seriously, because the distribution of the different measures is not quite normal. The *similarity* is a new measure, and illustrates the difference between the present system and the system with which the agents started. It is calculated by playing 1000 imitation games (without updates of the vowel inventories) with a random agent of the original population as initiator and a random agent from the present population as imitator and 1000 imitation games with the roles reversed. The average success over these 2000 games is then taken as the similarity.

Population:	$\varepsilon_0 = 0.01$ $\varepsilon_\infty = 0.01$ $\alpha_{aging} = 0$	$\varepsilon_0 = 0.03$ $\varepsilon_\infty = 0.01$ $\alpha_{aging} = 0.1$	$\varepsilon_0 = 0.03$ $\varepsilon_\infty = 0.01$ $\alpha_{aging} = 0.01$	$\varepsilon_0 = 0.03$ $\varepsilon_\infty = 0.03$ $\alpha_{aging} = 0$
Success:	0.8531 ± 0.040	0.7701 ± 0.035	0.7930 ± 0.041	0.8041 ± 0.041
Energy:	4.04 ± 0.59	5.55 ± 1.10	5.10 ± 0.95	3.83 ± 0.72
Size:	4.77 ± 0.40	5.66 ± 0.63	5.61 ± 0.58	5.14 ± 0.50
Similarity:	0.7347 ± 0.023	0.8231 ± 0.028	0.8235 ± 0.032	0.7884 ± 0.032

Table 5.2: Statistics of populations with and without age structure.

The most interesting measures for assessing how well the original vowel system has been preserved are the size and the similarity measures. The original system contained eight vowels, so none of the populations preserves it completely, as they all end up with (on average) less vowels. The sizes of vowel systems of the two populations with age structure are significantly greater than those of the populations without age structure according to the Kolmogorov-Smirnov test. This test does not find a significant difference in inventory size *between* the two populations with age structure. It does find a significant (at the 1-% level) between the energies of the systems with the slowly ageing population having lowest energy on average. Comparing the similarities of the populations with and without age structure, it is found that the ones with age structure have significantly higher similarity (at the 1-% level) than the ones without age structure. It can therefore be concluded that populations with age structure preserve the vowel systems better when the number of practice steps is limited.

Many more experiments can be conceived with dynamic populations. However, this would go too far for this thesis. The experiments shown in this section are just

an illustration of the possibilities of a fast population-based simulation of linguistic phenomena.

5.2 No Non-Verbal Feedback

The imitation games presented so far have depended on non-verbal feedback for determining the success or the failure of an imitation game. This implies that agents should be able to determine, using non-linguistic means, whether the signal they produced as an imitation of the other agent's signal resulted in successful imitation. There are good points and bad points to this approach. The good point is that it provides a coupling between the linguistic knowledge of the agent and the outside world. Phonemes are defined as speech sounds that can distinguish meaning. In order to learn them, a coupling must be made between a signal and objects or situations in the world. But in order to do this, the language learner must necessarily make use of non-linguistic information in the beginning in order to boot-strap the learning process. This is what is being modelled in an admittedly simplistic way by the non-verbal feedback in the imitation game.

An additional advantage of this dependence on non-verbal feedback is that agents learn neither more nor fewer distinctions than are necessary for the sound system they are learning. The distinctions an agent has to make depend only on the distinctions that are being used by the other agents in the population. This is similar to the way humans learn language. Very early in language learning, the ability to hear distinctions that are not linguistically relevant is lost. Nevertheless, distinctions that are linguistically relevant, even though they can be extremely subtle, are still accurately perceived (Vihman 1996, ch. 3 and 4).

However, the non-verbal feedback is a bit of an embarrassment in the way it is implemented. The model could be criticised as being unrealistic because humans do not learn language by being corrected or rewarded for every sound they make. Children are hardly ever corrected for the mistakes in pronunciation they make. Either they are not understood at all, or it is clear from the context what it is they were trying to say, and the communication is successful, even though their pronunciation might have been way off. Although human children that learn a language are able to perceive in subtle and complex ways that communication might not proceed optimally, there also seems to be an internal drive to imitate the sounds of their parents and peers as closely as possible. As has been mentioned in chapter 2, pronunciation is learnt more accurately than is strictly speaking necessary for successful communication. Children do not only learn the distinctive features that define the sound system of their language, but also all the phonetic detail that defines their dialect.

For this reason, it is interesting to investigate what happens if agents do not rely on non-verbal feedback, but rather on an internal judgement of whether their imitation was close enough to the sound they just heard. In this way they effectively calculate their own feedback, and become truly autonomous language learners. This might be an exaggeration of the situation in human language learning, because children obviously depend on a number of non-linguistic cues for learning a language. But new insights might be gleaned from this exaggeration, which could later be used for improving the model of the imitation game.

The imitation game without non-verbal feedback was implemented in exactly the same way as the original imitation game, described in chapter 3 and summarised in table 3.4. However, instead of using non-verbal communication for transmitting feedback, the imitating agent determines its own feedback. It does this by calculat-

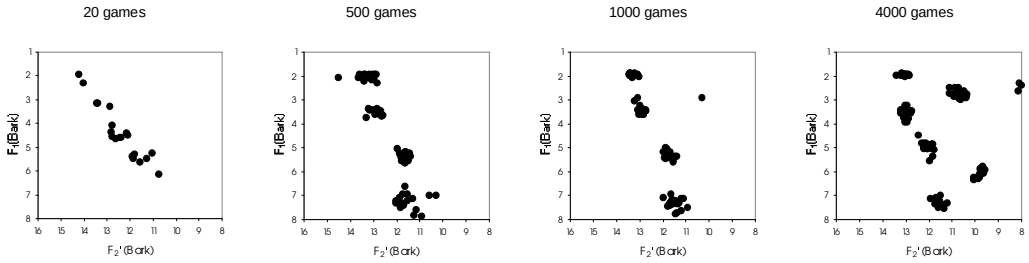


Figure 5.6: Emergence of vowel system without non-verbal feedback.

ing the acoustic distance between the original signal and the one it produced as an imitation. If this distance falls below a threshold distance D_θ the imitation is considered to be successful, otherwise it is a failure. The threshold distance D_θ is a parameter of the simulation.

5.2.1 Emergence of a system without non-verbal feedback.

The emergence of a vowel system from a population playing these imitation games is illustrated in figure 5.6. It can be compared with the emergence of a vowel system in a population that does use non-verbal feedback, which was illustrated in figure 4.1. The threshold distance D_θ was set to one Bark. The population size was 20 agents and the acoustic noise ψ_{ac} was 10%. All other parameters were set to their default values. After 20 imitation games no structure is visible, yet. The main factor operating is insertion of randomly created vowel prototypes and the direct imitation of these. After 500 imitation games, a number of more or less randomly distributed vowel clusters has formed. There are more vowel clusters than in the model with non-verbal feedback, because an imitation game can result in the formation of a new vowel prototype even if both participating agents only have one vowel prototype. In the game with non-verbal feedback, imitation was always successful in this situation, because no confusion was possible. In the imitation game without non-verbal feedback it can result in a failure (and thus the possible addition of a new vowel prototype) if the distance between the two agents' prototypes is larger than the threshold distance.

After 1000 imitation games, the situation has not changed very much. The vowel clusters have become somewhat more compact, but their number has remained the same. After 4000 imitation games, extra vowel prototypes have been added. These seem to be spread over the available acoustic space in the same dispersed way as the prototypes in the vowel systems that emerged from the imitation games with non-verbal feedback. Furthermore, just as in the games with non-verbal feedback, new vowel clusters keep on appearing until the available acoustic space is approximately optimally filled. This stage is reached after some 25 000 imitation games.

5.2.2 Variations on the distance threshold.

The vowel systems that emerge after 25 000 imitation games for different settings of the distance threshold parameter D_θ are shown in figure 5.7. The values that are shown are, from left to right 0.5, 0.75, 1 and 2. All other parameters of the simulations were the same as for the previous figure. It can be observed that for the smallest value of D_θ no coherent vowel system emerges. There can be two causes for this: either the acoustic noise level is so high that the agents try to learn multiple prototypes at a given location while there is in fact only one prototype that is shifted by noise. The agents try to discriminate with higher accuracy then the accuracy with

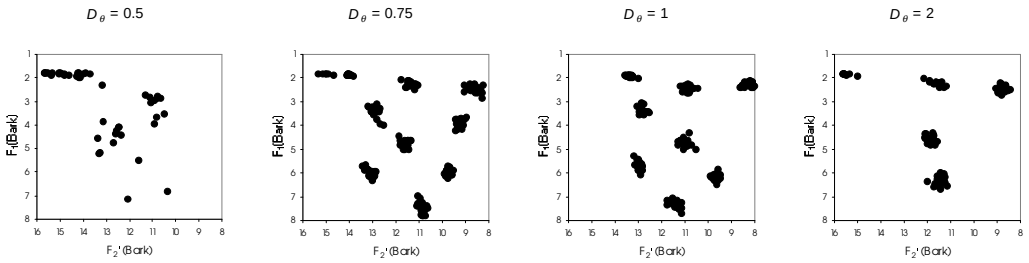


Figure 5.7: Limit systems of imitation games without non-verbal feedback.

which they can produce vowel prototypes. The other cause could be that the step size with which agents improve their vowel prototypes is bigger than the resolution with which they discriminate. They would then never be able to successfully imitate the sounds they hear. Experiments with changing both parameters show that it is rather the step size that determines the success of the emergence then the acoustic noise, as is illustrated in figure 5.8. In this figure the same parameter settings are used as in the first frame of figure 5.7. In the left frame the acoustic noise is made lower (ψ_{ac} = 5%) while in the right frame the step size ε is set to 0.015 (half of its default value). It can be seen that a vowel system does emerge in the right frame, but not in the left, indicating that step size plays a more important role than acoustic noise.

	Success	Energy	Size
D_θ = 0.5	0.4012 ± 0.14	2.74 ± 17.9	2.88 ± 1.47
D_θ = 0.75	0.9288 ± 0.30	13.33 ± 3.14	8.51 ± 0.81
D_θ = 1	0.9711 ± 0.020	11.46 ± 2.67	7.90 ± 0.78
D_θ = 2	0.9960 ± 0.011	1.66 ± 3.14	3.67 ± 0.66

Table 5.3: Measures of systems without non-verbal feedback.

Once D_θ exceeds a certain value, such as in the second frame of figure 5.7, a large number of vowel clusters emerges. The bigger D_θ gets, the smaller the number of vowel clusters gets, as can be seen in the third and fourth frame in the figure. This is because imitation games are considered to be unsuccessful for vowel prototypes that lie farther and farther apart. The measures of average success, energy and number of vowel prototypes, together with their standard deviations are presented in table 5.3. These numbers confirm the impression of figure 5.7.

5.2.3 Implications of not using non-verbal feedback.

In the simulation with no non-verbal feedback, the threshold distance D_θ is thus an extra factor in determining the number of vowel clusters in the vowel systems that emerge. It is mainly related to perception, and it is more of an internal parameter of the agents than the acoustic noise, which is rather something that exists outside the agents. In that sense the threshold distance is a more realistic parameter. It is now commonly accepted that

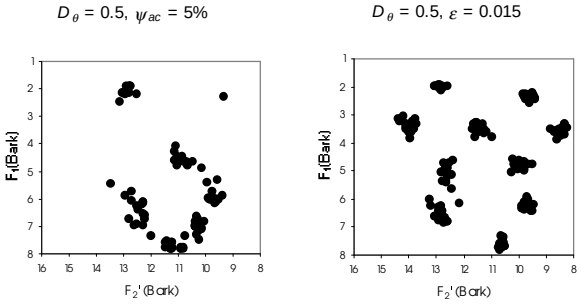


Figure 5.8: Influence of noise and step size on performance.

sound systems of human languages are not influenced by external factors, although in the past this was believed to be the case (see e.g. Rousseau 1986, on climatic factors, and Jespersen 1968, ch. XIV, § 3 and ch. XXI, § 4 on social factors). It is therefore preferable to have only parameters that are internal to the language user in any model of emergence of human speech. The distance threshold is such a parameter, but the acoustic noise is not.

Unfortunately, fixing the distance threshold to a given value limits the number of vowel prototypes that will eventually arise to a fixed and limited range. If D_θ is very high, the number of vowels will be small, and if D_θ is low, the number of vowels will be high. In the case of human language, the number of vowels a person can learn is obviously not fixed at birth. Languages have different numbers of vowels and the sound system of one language can evolve from one number of vowels to a different number. It would therefore be interesting if the agents' distance thresholds could converge towards a value that agrees with the number of vowels in the sound system they are learning. Agents would initially depend on non-verbal feedback mechanisms to determine the size of D_θ. When they get older they do not depend solely on non-verbal feedback anymore, but also on feedback they derive from their own distance threshold. In this way the role of the not-so-realistic non-verbal feedback can be reduced to a great extent. It can be imagined that some populations would evolve towards high distance thresholds, and thus to small numbers of vowels, while other populations would evolve towards low distance thresholds and large numbers of vowels. Another possible mechanism for determining D_θ in a more sophisticated model is to derive it through interactions with other parts of languages, such as lexicon and semantics.

The implementation of such a complicated simulation falls beyond the scope of this PhD. thesis. For reasons pointed out in chapter 7 it was deemed to be more interesting to work on implementing the imitation game with more complex utterances than to elaborate upon the simulation with vowels.

5.3 Realistic Signals

In the previous two sections the structure of the population and the behaviour of the agents in reaction to the imitation games have been changed. However, their perception and production of speech signals have remained the same. In this section, the agents will be changed so that they produce and perceive more realistic signals. This is done as an exercise for studying the feasibility of working with consonants (which require more complex signal processing) and as an opportunity to test whether the imitation game would work with human input as well.

In the imitation game as described so far the agents produce signals consisting of the frequencies of the first four formants. These formant patterns represent the properties of the acoustic filter that is formed by the shape of the agent's articulator, which in turn is determined by the articulatory parameters. In humans, this filter would be excited by the vibrating vocal chords in order to produce an audible sound. A listener derives information about the position of the articulators from this signal. Until now the generation and analysis of this acoustic signal was not modelled. The production of a signal from the formant frequencies and the reconstruction of the formant frequencies from a speech signal were assumed to be signal processing problems, rather than fundamental aspects of the imitation game. In order to speed up the simulation and make it more tractable, it was decided to work only with the formant frequencies. Now it will be investigated if this decision was justified.

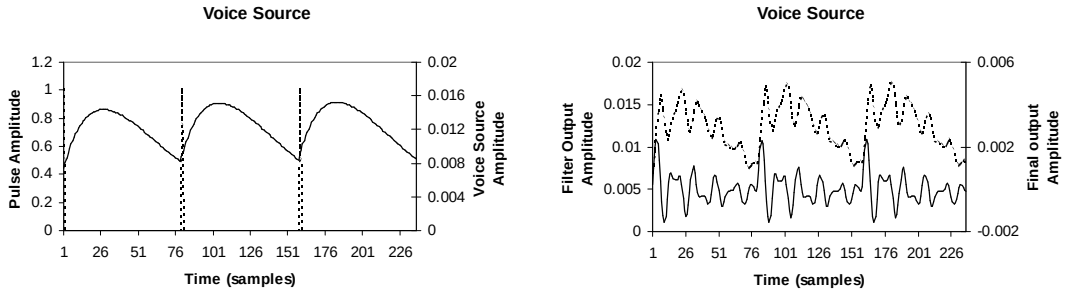


Figure 5.9: Pulse train (dashed, left) voice source (solid, left) filter output (dashed, right) and final output (solid, right) of the vowel synthesiser.

5.3.1 Generating a realistic signal.

Generating a signal from the four formant frequencies is relatively easy. It will be described in general terms in this chapter. A more detailed description can be found in appendix D. The formant frequencies were defined as the peaks in the spectrum of the signal. These peaks can be considered as the centre frequencies of a corresponding number of bandpass filters. The vocal tract can then be modelled as four consecutive bandpass filters whose centre frequencies are the four formant frequencies. In order to make a human-like sound, these bandpass filters have to be excited by a signal that is a good approximation of the vibration of the vocal chords. The vocal chord vibration can be modelled by low-pass filtering a train of pulses. Finally, the signal that comes out of the bandpass filters is high-pass filtered in order to stress the high frequencies. This gives quite realistic results, as illustrated in figure 5.9. In the left part of this figure, the original pulse train input and the artificial vocal chord vibration obtained after filtering these pulses are shown. In the right part of the figure, the output of the bandpass filters and the final high-pass filtered output for the low, front, unrounded vowel [a] (formant frequencies 742, 1266, 2330 and 3457 Hertz) are shown.

5.3.2 Perceiving a realistic signal

The perception, on the other hand, is more complicated and the technical details are discussed in appendix D. The main problem that was encountered, especially when working with real signals produced by a human speaker, was that not always all formant frequencies could be recovered. Sometimes formant peaks were so small that they could no longer be detected. This was especially the case with high back vowels, of which the second formant would disappear. For example, in the case of [u], which was assumed to have formant frequencies of 276, 740, 2177 and 3506 Hertz, the second formant would not be detected, and the perceived pattern became 276, 2177, 3506 and 4065 Hertz. This is actually quite close to the pattern of the [i] with formants at 252, 2202, 3242 and 3938 Hertz. This means that these two were confused, which would never happen in human perception. The performance of

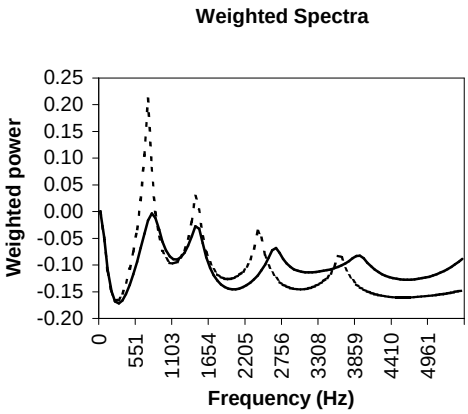


Figure 5.10: Example of weighted spectrum comparison.

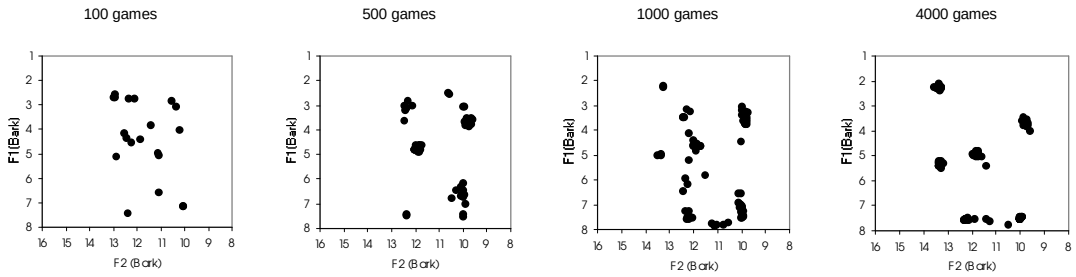


Figure 5.11: Emergence of a vowel system based on realistic signals.

direct comparison of formant frequencies is therefore unacceptable. It was decided to compare the whole spectrum of the signal, instead of only the formant frequencies.

A smoothed spectrum was calculated using linear predictive analysis (for details see appendix D). The spectra were then normalised for the strength of the signal, so that two vowels with the same quality, but with different volumes would have the same spectrum. Finally the spectrum was weighted in such a way that each octave contributed with equal weight to the comparison. The distance between two signals could now be calculated as the surface of the absolute value of the difference between the two spectra. An example of a comparison between the signal of a real [a] (solid line) and a reference signal (dotted line) is given in figure 5.10.

5.3.3 Realistic signal results.

Using this distance function, the agents can now play the unmodified imitation game, as it was described in tables 3.4, 3.6 and 3.5 in chapter 3. The results of such an imitation game are given in figure 5.11. In this figure acoustic representations of the vowel systems as they emerge in a population of twenty agents are given. The numbers of imitation games are chosen to be comparable to the number of games in figure 4.1 in chapter 4. Because of the way the agent populations were logged, it was not possible to show the system after only 20 games, so the first frame shows the system after 100 games. The vowels are shown in the familiar first formant-effective second formant plot, in order to make comparison with the results of the other simulations easier, but it should be noted that these formant values do not play a direct role in the comparison of vowels, as they did in all other systems that were shown so far. It can be seen that a vowel system consisting of a number of compact clusters does emerge in the case that realistic signals are used. After the first hundred games the vowel prototypes are still randomly dispersed through the available acoustic space. After 500 games, clusters start to form. It is remarkable that at this stage, the number of clusters is higher than in the corresponding frame in figure 4.1. After 1000 games, the situation is about the same, although the prototypes seem to be slightly more dispersed through the acoustic space. After 4000 games, compact clusters have formed that are evenly distributed through the available acoustic space, although it is not yet covered optimally. Also, one of the vowel clusters (the lowest one) is not yet very compact. Apparently the system is not yet fully developed. A more fully developed system (after 25 000 imitation games) is shown in figure 5.12. Although eight clusters are visible in this plot, the actual average number of vowels per agent is six. Not all the low clusters are shared by all agents and the high central cluster is only shared by two agents. The system is less spread out in the effective second formant dimension than systems that emerged from the experiments using simplified signals. However, it should be kept in mind that although

the signals that are used are realistic, the perception model is not, and was just developed to give satisfactory results, and was not directly based on a model of human perception.

5.3.4 Learning human vowel systems and modifying the imitation game

An interesting possibility of the imitation game with realistic signals is that it can be used to learn real human vowel systems. Although no serious experiments for testing this option have been performed, an implementation has been made that is able to learn and imitate vowels produced by a human speaker. An example of a vowel system that was learnt from one speaker is given in figure 5.13. The learning task consisted of learning a vowel system of five vowels ([i], [e], [a], [o] and [u]) from a single male speaker (the author). As can be seen in the figure, a vowel system was indeed learnt. This vowel system corresponds in four of the five vowels to the target vowel system. Only the mid back rounded vowel could not be learnt properly. This is probably due to the limitations of the perception model (described in appendix D).

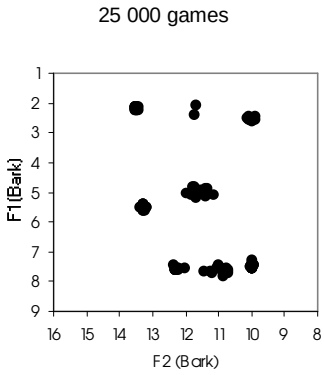


Figure 5.12: System based on realistic signals after 25 000 games.

This vowel system was not obtained with the original imitation game, though. It was found that this imitation game does not work with a human participant. The original imitation game assumes that roles are randomly assigned with equal probability and that the vowel prototype that is first used by the initiator is randomly chosen as well. These assumptions are not justified when interacting with a human. A human speaker, when confronted with an agent that is not competent in the vowel system that has to be taught, will usually assume the role of initiator more often than that of imitator. Also, the human speaker will generally start with giving examples of all vowels in the repertoire, rather than choosing all of the vowels in the repertoire with equal probability. Also, the human speaker will concentrate on the vowels that are imitated badly, and repeat these more often. This behaviour of the human speaker causes the first prototype that the agent forms to immediately get a low success/use ratio. This ratio will quickly drop below the threshold above which new prototypes are added in a failed imitation game (in the original game new prototypes were only added if the used prototype had proven to be successful previously). This will block the addition of new prototypes, and make it impossible for the agent to learn the human's vowel system.

The rules of the imitation game were therefore modified on the basis of two assumptions. The first assumption was that the human agent has perfect knowledge of the vowel system, and the second assumption is that the human always gives honest evaluations of the agent's performance. The rules were changed so that an agent would always add a vowel if an imitation game failed when it had the role of imitator. Fur-

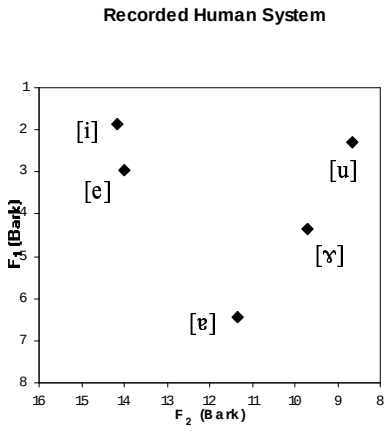


Figure 5.13: Vowel system learnt from a human speaker.

thermore, the agent would now be able to modify its vowel inventory when it had the role of initiator. In the original imitation game the initiator did not modify its vowel inventory in reaction to the outcome of the imitation game. Whenever an imitation game played with a human failed, the imitator removed the prototype that was used, because it could not be imitated properly by the human. Other processes, such as the merging of vowel prototypes that are too close together, were not changed.

These modifications allowed the agents to learn a simple human vowel system successfully. However, the fact that they added and removed vowels without taking into account their previous success, made the agents' vowel systems quite vulnerable. For a more robust system, more conservative ways of updating the agents' vowel inventories should be used. However, the experiment was only meant to show that it is possible *in principle* to learn a human vowel system.

The experiments with realistic signals have shown that the imitation game works with realistic signals and that it can even be used to learn human vowel systems. They also have shown that the way in which the agents update their vowel inventories in reaction to the imitation games should also depend on the way in which the roles of the agents are assigned, on the way in which the vowel prototypes are chosen and on which of the two agents has the most reliable knowledge of the sound system to be learnt.

5.4 Conclusion

In this chapter, a number of qualitative variations on the theme of the emergent vowel system have been presented. The first variation was the introduction of change in the population of agents. New, empty agents could be added to the population and old agents could be removed. It was shown that in this situation existing vowel systems could be preserved and new ones could emerge, even over periods of time that were so long that no agents of the original population were left in the final population. The emerging vowel systems, just as the ones that emerged from the simulation without age structure, show the same characteristics as the vowel systems found in human languages. The accuracy with which systems were preserved and the number of vowel prototypes in the final systems and in the systems that emerged depended on the speed with which agents were replaced in the population. It was also shown that under certain circumstances, giving agents of different ages different behaviours made for better preservation of the vowel systems.

The second variation on the simulation was to change the rules of the imitation game so that giving non-verbal feedback was no longer necessary. As non-verbal feedback does not seem to occur very often in human language learning, it was thought to be interesting to design a simulation that works without it. In the model without non-verbal feedback, the agents decided for themselves whether their imitation was good enough or not, based on a distance threshold. If an agent found that the acoustic distance between its imitation and the original signal was less than a given threshold, the imitation was considered successful, while if it fell above the threshold, the imitation was considered a failure. With this modification, coherent and realistic vowel systems also emerged. The disadvantage of this model is that the number of vowel prototypes an agent is able to learn is now pre-determined by the distance threshold, while in reality children can learn any vowel system. An agent model that can adapt its distance threshold on the basis of non-verbal feedback, and that then proceeds with learning the vowel system without non-verbal feedback was

proposed, but not implemented, because this would lead too far from the main goal of this chapter.

The third and last variation of the basic model was to change the production and perception of sounds by the agents so that they could work with realistic signals. The production was modified so that agents could produce a real signal based on the formant frequencies they calculated using the formant synthesiser that was used in all the other experiments. The perception was based on the calculation of the distance between the smoothed and weighted spectra of the two signals to be compared. It was shown that in a population of agents that produced and perceived vowels in this way, a coherent and realistic system also emerged. Although the production and perception were supposedly more realistic, the resulting vowel systems themselves were not necessarily more realistic. This was probably related to the fact that the perception model was quite crude and *ad hoc*. An interesting extra experiment that was done with this model was to learn vowels from a human speaker. A preliminary experiment showed that this was possible within limits. A remarkable finding was that due to the different behaviour and expectations of a human player in the imitation games, the reactions of the agents to the game had to be changed. Apparently the reaction to the imitation game should depend on the behaviour and the knowledge of the other agents.

The relative success of all these different variations in generating realistic vowel systems show that the occurrence of self-organisation is not dependent on the exact details of the implementation of the agents, nor on the exact dynamics of the population of agents. In the previous chapter it was shown that one single variant of the imitation game was not very sensitive to the settings of its parameters. In this chapter it has been shown that the outcome of the simulations are not very sensitive to details of implementation. Both the rules of the imitation game and the production and perception of the agents could be changed without qualitatively changing the outcome of the simulations. Of course, there are limits to the kinds of variations of the imitation game that still result in realistic systems (or, for that matter, in coherent systems at all). This was found in the experiment with a human speaker, who followed rules that were incompatible with the way the artificial agents learnt.

Nevertheless, the relative insensitivity to the exact details of implementation supports the claim that self-organisation plays a role in the emergence of universals of human sound systems as well. One could criticise the research presented in the previous chapter by saying that it is not sufficiently realistic, or that the self-organisation might only be the result of the idiosyncratic behaviour of a single implementation. The results presented in this chapter have shown that self-organised emergence of coherent and realistic vowel systems takes place in all the variations of the imitation game that have been investigated. Apparently self-organisation takes place independent of implementation details. This makes it more likely that self-organisation also plays an important role in human vowel systems, even though the implementation of production, perception and the way vowels are learnt is quite different in humans than in the experiments presented here.

6. Parallels with Human Vowel systems

Perhaps the most important question about the results of the previous two chapters is to what extent they agree with what is known about human vowel systems. In this chapter an answer to this question is sought. In order to understand the relation between the emerged systems and real human vowel systems some knowledge about the universals and the typology of human vowel systems is required. Universals were already discussed in a general way in chapter 2. In the first part of this chapter a more detailed and concrete overview of typology and universals will be presented. Also, it is discussed what predictions a theory that claims to model the emergence of vowel systems should make.

In the second part of the chapter, the different vowel systems that emerge from the simulation are compared with data on human vowel systems, and it is verified whether the frequency with which different systems emerge is comparable to the frequency with which similar systems are found in human languages.

In this chapter a lot of use is made of phonetic symbols and vowel positions, such as front, mid, central etc. Tables with International Phonetic Alphabet (IPA) symbols and with phonetic terms can be found in appendix H.

6.1 Human Vowel System Universals and Typology

It was already noted in chapter 2 that human sound systems, and more specifically human vowel systems, show a number of remarkable regularities. Humans are able to distinguish a huge number of different vowel sounds *in principle*. According to Ladefoged and Maddieson (1996) there are languages that make five distinctions in the height of vowels, languages that make three distinctions in their position and languages that make three distinctions in lip rounding. This would make for a total of at least 45 possible basic vowel qualities. However, any human language only uses a very limited subset of these. Vallée, (1994) who investigated the $UPSID_{317}$ found that the maximum number of different vowel qualities that are used in any language in the sample is 15 in Norwegian (Vanvik 1972). There are languages that have more vowel phonemes, but these will use other processes, such as length, nasalisation and pharyngealisation, not quality in order to distinguish vowels. Furthermore, the small subsets of the possible vowels that languages use are not chosen at random (see Crothers 1978, section 4.5 for a discussion of randomness with relation to five vowel systems). Some vowels appear more often than others do and vowel systems tend to be quite symmetrical. Typologies of possible human vowel systems have been based on these observations.

6.1.1 The basis of typologies of human vowel systems.

But before embarking on a description of the proposed typologies of human vowel systems, it needs to be made clear what it is exactly that these typologies are based on. They are based on phonetic descriptions of the vowel phonemes of languages. Phonemes are by definition minimal units of sound that can make a difference in meaning. However, it is quite possible that two speech sounds that are different (but close) phonetically do not make any distinction in meaning. These sounds are then called allophones of a phoneme. This happens for example through the influence exerted by neighbouring sounds. A description of the phonemes of a language necessarily abstracts away from this allophonic variation. If one wants to make a description of a language this is not a problem. On the other hand, if one wants to classify languages based on which phonetic signals are used for realising their vowel

phonemes it does become a problem. A choice needs to be made as to which phonetic realisation is representative of the phoneme. Usually the most frequent allophone of a phoneme is taken to be the representative one. These representative allophones can then serve as a basis for a typology of possible vowel systems. Some researchers have even considered vowel systems with phonetically different elements as belonging to the same category (e.g. Crothers 1978, who analyses [i], [a], [u] and [i], [a], [o] as belonging to the same type).

It will be assumed here that this is a valid methodology. However, it should be kept in mind that a typology and classification of vowel systems is based in the first place on abstract phonemes. The actual observed signals in a language can be quite a bit messier than would be expected from the typological classification of the language. A case in point is the vowel system of English. In figure 6.1 it is given as: [i], [ɪ], [ɛ], [æ], [ɑ], [ɔ], [ʊ], [u], [ɜ] and [ʌ]. This seems like a reasonably symmetrical ten vowel system with two central vowels. But if one looks at the reality of the figure, one sees that the actual clusters[1] (which are based on data from many different speakers) cover quite a considerable area of the acoustic space, meaning that the vowels could have been labelled differently as well. Also there seems to be overlap between the different vowel clusters, indicating that it is not always possible to say to which phoneme a given signal would have to be mapped (but probably this overlap disappears if higher formants are also taken into account).

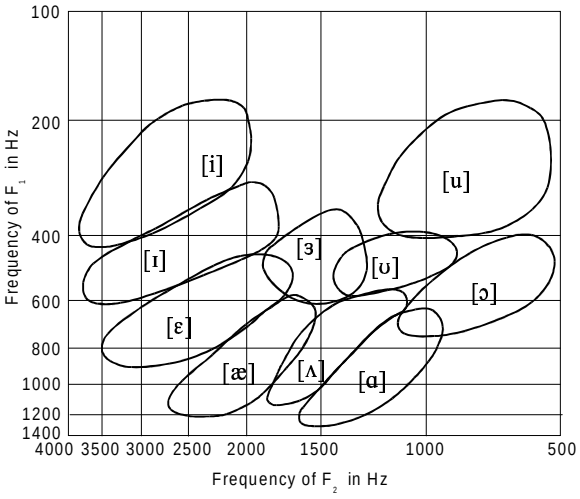


Figure 6.1: Vowels of English, adapted from Peterson & Barney 1952 through Rabiner & Schafer 1978.

The bottom line of this is that one should not always expect the actual observed sounds of a human language to follow a given typology. Typologies are based on data that is to at least some extent abstract and idealised. Vowel systems that emerge from a simulation should therefore not be expected to follow the typology exactly. They should rather be expected to follow it in a general way but not to consist of exactly the vowels predicted by the typology.

6.1.2 Classification and typology of human vowel systems.

Having made this warning, it is now time to introduce the classification of vowel systems as it has been developed by a large number of researchers. Already soon after the development of the idea of the phoneme, Trubezkoy (1929) already attempted to classify the vowel systems of the world's languages. This classification has been elaborated upon by others (Hockett 1955, Sedlak 1969), and has been used to con-

[1] This figure is based on figure 3.4 in Rabiner and Schafer (1978) which in turn is based on Peterson and Barney (1952). The axes have been changed in order to make the figure more comparable with the figures of the artificial systems in this thesis. Also, only the outlines of the clusters have been retained, and the individual data points have been removed.

struct a typology of vowel systems together with a number of universals (Crothers 1978, Vallée 1994, Ladefoged & Maddieson 1996, Schwartz *et al.* 1997.) From the seventies on explanations of the universal tendencies have been investigated with computer models that are based on functional criteria (see e.g. Liljencrants & Lindblom 1972, Schwartz *et al.* 1997b) and which have been discussed briefly in chapter 2.

In this chapter, mainly Crothers' (1978) typology will be followed. This typology is based on the Stanford Phonology Archive (Vihman 1976), the predecessor of UPSID, which consisted of 209 languages, and is therefore reasonably representative. An advantage of the typology for using it for the research presented here is that it classifies the vowels in acoustic space. This agrees with the way similarities between vowels are evaluated in the simulations presented in the previous chapters. It also ignores other articulatory parameters that might be used for distinguishing vowels, such as length, nasalisation and pharyngealisation. These parameters can not be used by the agents in the simulations, so they should not be used in evaluating the realism of the emerging systems, either. More recent work on the typology of vowel systems (Vallee 1994, Schwartz *et al.* 1997a) does take into account these distinctions and is therefore less applicable to the work at hand.

Crothers' typology is based on acoustic distinctions in the F_1-F_2 space. As rounding and tongue position both have the effect of changing the second formant, they are considered as one parameter, rather than two. This allows Crothers to lump together most central vowels, without taking into account whether the acoustic signals are produced through lip rounding or through centralising the tongue position. The vowel systems /i, e, a, o, u, ɨ/ and /i, e, a, o, u, y/ would thus be analysed as belonging to the same type. Whether this is sound practice when classifying actual human languages is questionable. Schwartz *et al.* (1997a) for example do make distinctions between the different central vowels. However, the acoustic representation of the vowels in the simulations does not make a distinction between different central vowels, so Crothers' (1978) typology is quite suited for comparing human languages with the outcomes of the simulations.

In general, Crothers seems to be more interested in the relation between the different positions of the vowel phonemes, than in their absolute positions. For example, he classifies vowel systems /i, a, u/, /i, a, ɯ/, /i, a, o/ etc. as the same three vowel system /i, a, u/. This is sound as long as one is interested in *classifying* vowel systems obtained from either a description of a language or from a computer simulation. This is what will be done in the next section. However, when lumping vowels together like this, one should be very careful about making inferences in the other direction, such as: "There are no languages without [i], [a] and [u]."

A third simplification that Crothers makes of the vowel systems in his sample is in the way he handles other articulatory parameters besides height, position and lip roundin. If other parameters are used, Crothers counts vowels that have different settings for this parameter, and that are very close (but not always equal) in quality as representing only one vowel quality in the system. For example, he analyses the vowel system of German, consisting of /ɪ, ɛ, ʏ, œ, ɐ, ʊ, ɔ, iː, eː, yː, øː, aː, uː, oː/ as a symmetrical seven vowel system with two central vowels. Again, this might not be the best approach for dealing with human languages (Schwartz *et al.* (1997a) analyse German as a system with 16 different vowel qualities, they also count /ə/ and /ɛː/as phonemes). However, although it is true that length (or other) distinctions are

often accompanied with quality distinctions, the length gives an extra cue for recognising the vowels. Vowel systems that do make length distinctions could therefore possibly be slightly more crowded than vowel systems in which length distinctions are not made. As the agents are not able to make length distinctions, it is probably not fair to try to fit their vowel systems in a typology that is based on possibly slightly more crowded vowel systems that *are* using length distinctions.

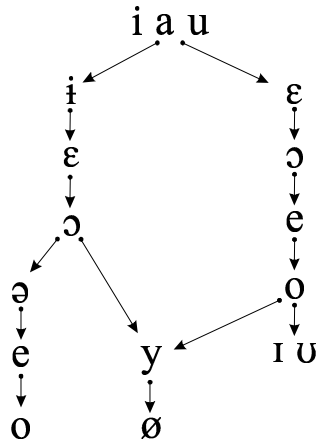


Figure 6.2: Vowel system hierarchy according to Crothers (1978).

The best way to illustrate Crothers' (1978) observations on vowel systems is with the diagram presented in figure 6.2. This diagram shows the sequence of vowels that differently sized vowel systems of the world's languages use. It should be interpreted as follows. If a vowel system has three vowels, it consists of /i, a, u/ or at least it as vowels that are near these three centres. If it has four vowels, it adds either /ɨ/ or /ɛ/. For systems with five and more vowels, one can just follow the arrows down, until one reaches the maximum size of eight or nine vowels. Although Crothers notes that there are exceptions to this hierarchy, the great majority of languages follow it. But, as has been mentioned above, Crothers allows considerable slack in the assignment of phonetic symbols to the phonemes of a language, so that his hierarchy says more about the relative positions of the vowels in the systems than about their actual precise phonetic value. However, Vallée (1994, p. 94) who uses a different sample of languages ($UPSID_{317}$) and a different methodology, comes to a rather similar hierarchy, although the central vowels' order of appearance is different.

Crothers also summarises his observations in a number of rules, of which the first twelve are quoted from (Crothers 1978, appendix I) below:

1. All languages have /i a u/.
2. All languages with four or more vowels have /ɨ/ or /ɛ/.
3. Languages with five or more vowels have /ɛ/. They generally also have /ɔ/.
4. Languages with six or more vowels have /ɔ/ and also either /ɨ/ or /e/, generally the former.
5. Languages with seven or more vowels have /e o/ or /ɨ ə/. (The types /ɨ ə/ may be represented by /ü ö/.)
6. Languages with eight or more vowels have /e/.
7. Languages with nine or more vowels generally have /o/.
8. A contrast between five basic vowel qualities is the norm for human language, and in general, the most common systems are those with close to this number of basic vowels.
9. The number of height distinctions in a system is typically equal to or greater than the number of backness distinctions.
10. Languages with two or more interior vowels always have a high one.
11. The number of vowels in a column of interior vowels cannot exceed the number in the front or back columns.

12. The number of height distinctions in front vowels is equal to or greater than the number in back vowels.

(end quotation). Rules 13, 14 and 15 have to do with vowel length and nasalisation, so they are not relevant for the present purposes. Note also that Crothers uses the American notation /ü ö/ for front rounded vowels, instead of the IPA notation /y ø/.

6.1.3 Conformation of emerged and real languages to the typology.

If the simulations presented in this thesis are realistic, the vowel systems that emerge should confirm to these rules, and to the hierarchy of figure 6.2. This will be investigated in the next section. However, it should be kept in mind that actual languages diverge from the typology, usually in detail, but sometimes completely. One could say that vowel systems of languages conform to the typology and the universals with a high probability. However they do not necessarily confirm. One should make a distinction between investigating which functional criteria play a role in determining the shape of vowel systems, and actually predicting the vowel systems that appear in the world's languages. Optimising an artificial vowel system according to criteria of acoustic dispersion, such as done for example by Liljencrants and Lindblom (1972) is good for testing whether acoustic distance plays a role in determining the shape of human vowel systems. As vowel systems that are (near-) optimal with respect to acoustic dispersion appear significantly more often than sub-optimal ones, it clearly plays a role. Such models are not complete, however, for actually predicting the vowel systems that do occur in human languages, as they will tend to produce only optimal systems, whereas non-optimal systems appear as well, albeit with lower frequency.

Any model that claims to predict vowel systems of languages should not only predict the vowel systems that are observed most often, but also, with a lower probability, the systems that appear less frequently. The frequency distribution of the predicted systems should conform to the frequency distribution of human vowel systems. Especially models that work with populations of agents, as opposed to models that simply optimise, should do this, because they do not only investigate which factors play a role in determining the shape of vowel systems, but actually how these factors are implemented as well. This is a weak point of the work of Berrah (1998) whose model only predicts the most frequently occurring systems.

The vowel systems that emerge from an agent simulation should therefore not only conform in their vowel inventories with human languages, but also in the frequency distributions of the different types of systems.

6.2 Relation between Emerged Systems and Real Systems

It will now be attempted to make a typology of the emergent vowel systems in the same way as for human systems. For this a number of simulation trials with different parameter settings have been run in order to generate a large number of stable artificial vowel systems. Each parameter setting resulted in vowel systems whose number of vowels is restricted to a limited range. These systems were classified in the same way as the systems of Crothers (1978) were classified. That is to say, more attention was paid to the relative arrangement of the vowel prototypes than to their exact phonetic values. It was then checked whether the types of vowel systems that emerged and their relative frequencies were comparable to Crothers' results. The parameter values were changed so that systems with different numbers of vowels emerged. The emerged systems were compared and classified and it was checked

whether the same hierarchy of appearance of vowels in systems of different sizes was found in the simulations as in real languages.

6.2.1 Three vowel systems

The first vowel systems that are shown were obtained from the vowel simulation with the standard parameter settings, (see appendix A) with a population of 20 agents and with acoustic noise ψ_{ac} set to 18%. The simulation was run 100 times. For each run 25 000 imitation games were played. From each of the resulting populations, the average number of vowels per agent was calculated. Then one of the agents that had a number of vowels that was equal to the average was selected, and its vowel system was classified. This was done on the basis of the number of front, back and central vowels, and on the basis whether the vowel system had one or two low vowels. This did not play a very important role in the trials with 18% noise, because only systems with three or four vowels emerged. There were 32 systems with on average three vowel prototypes, and 68 systems with on average four vowel prototypes. These numbers should not be compared with the frequencies of three- respectively four vowel systems in human languages, because through changing the noise parameter ψ_{ac}, the relative abundance of systems with any number of vowels can be assured.

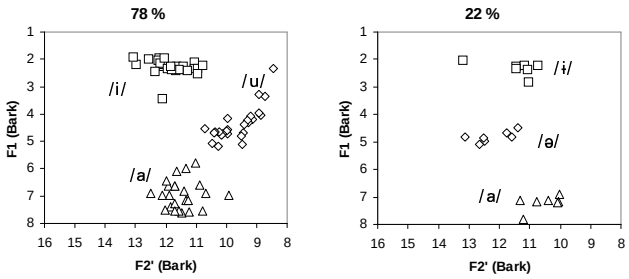


Figure 6.3: Classification of three vowel systems

The emerged three vowel systems are shown in figure 6.3. Again, the acoustic prototypes of the agents' vowels are shown in the acoustic space based on the first and effective second formant frequency in the logarithmic Bark scale. Although these plots look very much like the plots of vowel systems that have been shown so far, they actually show something quite different. The previous plots showed the vowel prototypes of all the agents in *one* population. The clusters in these plots corresponded with the vowel prototypes that were recognised by all agents in that population. In the plots shown here, the vowel systems of agents from *different* populations are plotted, classified per figure on the basis of the shape of each individual agent's vowel system. The individual agents' vowel systems are taken to be representative for the population from which they were taken. Whereas in the previous plots all agents shown had played imitation games with each other, in figure 6.3 (and other plots of classifications in this chapter) none of the agents have played imitation games with each other. The similarities between their vowel systems are therefore not due to their interactions, but to the fact that populations have been attracted towards similar vowel systems. The lack of interaction between the agents that are plotted also accounts for the larger size of the clusters and their greater degree of overlap.

The thirty-two vowel systems with three vowels can be classified into two types: one that is roughly triangular and one that is roughly vertical. The first type appears in 78% of the cases and the second type appears in 22% of the cases. This is not quite like in human languages. The first of Crother's (1978) universals says that all languages have /i a u/. However, here about one fifth of the emerging three vowel systems is vertical and although vertical systems do appear in human languages, (Choi 1991, Ladefoged & Maddieson 1996) they are quite rare. Also the triangular vowel systems have a mid back vowel [o] instead of a high back vowel [u].

Most probably this has to do with the position of the high front vowel, which appears to be consistently too far back, so that there is more distance between it and a mid back vowel than a high back vowel. This seems to be a problem with the synthesis- and perception functions that could also already be observed in chapter 4 in figures 4.8 and 4.9. It seems that although it is possible for the agents to produce and perceive completely high and front vowels, in practice it is almost impossible to learn this vowel, or to reach it through optimisations, as in figures 4.8 and 4.9.

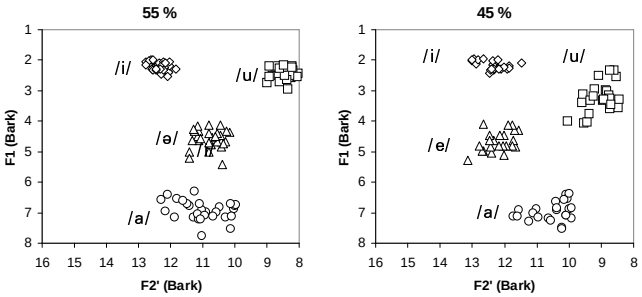


Figure 6.4: Classification of four vowel systems.

6.2.2 Four vowel systems.

The situation with four vowel systems is already much more realistic. The four vowel systems were not taken from the simulation with ψ_{ac} =18%, but from a simulation with ψ_{ac} =15%. The other parameters were set to the same values as in the previous experiment. From this run, 51 systems with four vowels and 49 systems with five vowels emerged. The classification of four vowel systems is shown in figure 6.4. In this figure it can be seen that all the four vowel systems that were found did contain /i a u/, thus conforming to Crothers' universal number one. About 55% of the systems contain a mid central vowel, and about 45% of the systems contain a mid front vowel. Although Crothers (1978) second universal says that all languages with four or more vowels have either /ɨ/ or /ɛ/, and the emerged four vowel systems with a central vowel rather have /ə/, they still confirm quite closely to Crothers' universals. The discrepancy is again probably due to the fact that the high front vowel is rather far to the back, so that there is not sufficient room for a high central vowel. If the results are compared with Schwartz *et al.* (1997a) it is found that in their data, the system without central vowels is much more abundant than the system with central vowels, whereas the simulation finds them in about equal proportions.

6.2.3 Five vowel systems.

The case of five vowel systems is shown in figure 6.5. The vowel systems were obtained from the same simulation with 15% acoustic noise as the four vowel systems. Here the resulting classification confirms nicely with human sound systems. From the 49 systems, 88% consisted of the symmetrical five-vowel system. Eight percent have a central vowel and two front vowels, while 4% have a central vowel and two back vowels. The most frequent type conforms to Crothers' first three universals.

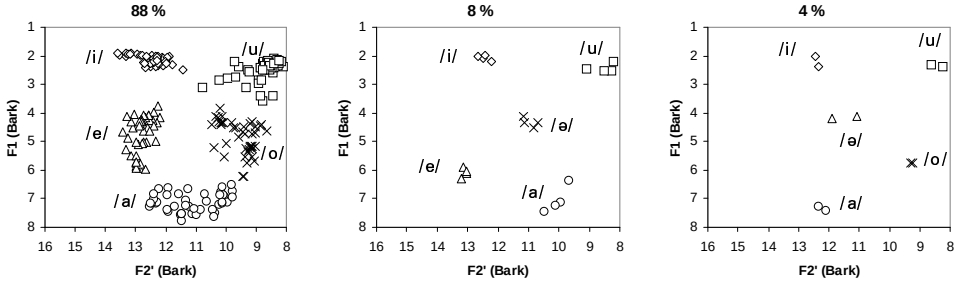


Figure 6.5: Classification of five vowel systems.

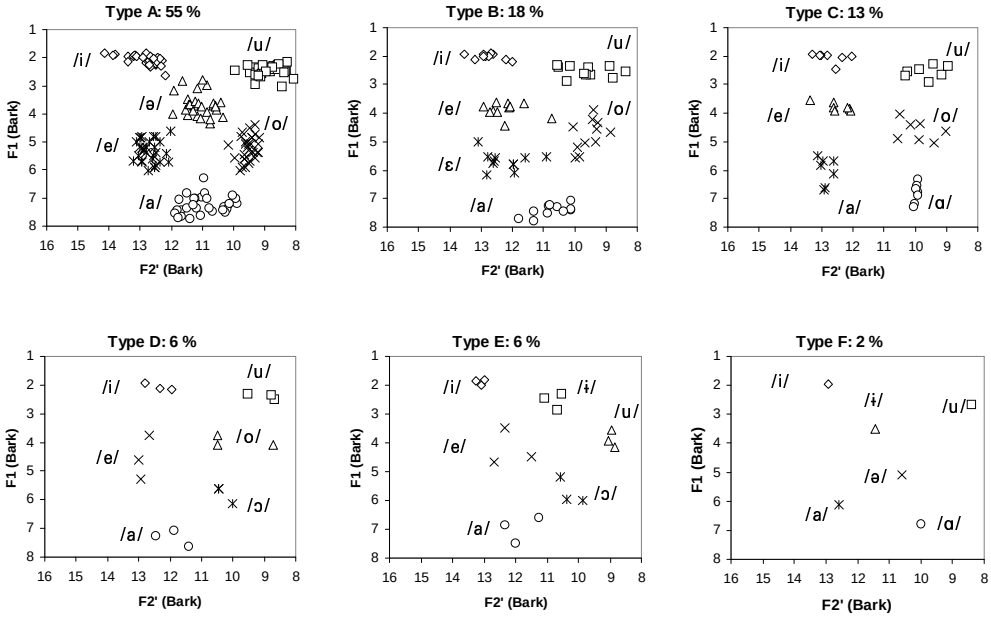


Figure 6.6: Classification of six vowel systems.

The type that occurs in 8% of the cases conforms to the first two universals, and the type that occurs in 4% of the cases only to the first universal. Schwartz *et al.* (1997a) found that the type leftmost in figure 6.5 occurs in 89% of the languages with five vowels in $UPSID_{317}$. Systems with one central vowel occur equally often with more front and more back vowels in their data (5% of the cases). The results of the simulation therefore conform very well with what is found in human languages for the case of five vowel systems.

6.2.4 Six vowel systems.

Systems with six vowels were obtained from a run with the acoustic noise parameter ψ_{ac} set to 12%. The rest of the parameters were exactly as in the previous experiments. From the hundred runs that were made, 54 resulted in vowel systems with six prototypes. The resulting classes of systems are shown in figure 6.6. There are more types in this figure than in the previous figure, because the bigger a system becomes, the more ways there are to distribute the vowel prototypes. For this reason the different types in this figure have been assigned letters in order to facilitate referring to them. Type A is the most frequent type, occurring in 55% of the cases. It consists of the symmetrical five vowel system with a more or less high central vowel. This system conforms to Crothers' first four universals, and so do type B and C (for a total of 86% of the systems). Types D and F lack both /e/ and /ɨ/, thus violating universal 4. Type E is quite similar to type A, except that the back vowels are lower. The systems also compare favourably with Schwartz *et al.*'s results. For the 60 six-vowel systems they found in $UPSID_{317}$, 68% were of type A and E, 20% were of type B, 5% were of type C and 7% were of type D. They did not encounter any systems of type F.

6.2.5 Seven vowel systems.

Seven vowel systems were obtained from simulations with the default parameter settings and ψ_{ac} =10%. From the 100 runs with this parameter setting, 25 resulted in systems with seven vowels. The resulting types of vowel systems are shown in figure 6.7. There are five types and again these have been assigned letters in order to facilitate reference. The types A and D conform to Crothers' universals 1 to 5, while types

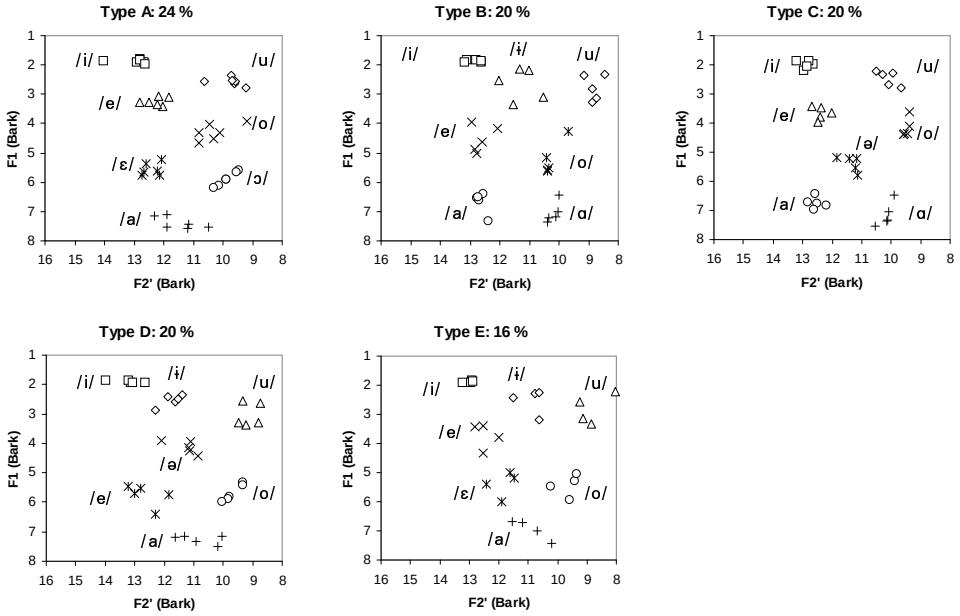


Figure 6.7: Classification of seven vowel systems.

B and E conform to his universals 1 to 4, but not to universal 5, which states that languages with seven or more vowels should have either /ɐ o/ or /ɨ ə/. Type C is truly anomalous by conforming only to universals 1 and 5, but not to numbers 2 to 4.

Schwartz *et al.*'s data contain 44 systems with seven vowels. Most of these systems, 52%, were of type A. Eighteen percent were of type E, while 27% were of type D. The remaining vowel system in their data does not fit any type of system that emerged from the simulations. The lack of systems in their data that fit type B is strange, because systems of this type do conform to four of the five "universals" for vowel systems (just as does type E, which did appear in the data quite frequently). A similar system with *six* vowels, but without the high central vowel (type C for six vowel systems) does appear in their data. It could be that the low front vowel [a] of type B systems is analysed as a low mid front vowel [ɛ], so that they are classified as type E. The lack of systems of type C for seven vowel systems is less surprising. This type of system does not conform at all well to Crothers' universals of vowel systems. It contains a low mid central vowel, whereas a mid or high central vowel would be expected. The relatively high frequency of this system can possibly be explained by the fact that the high front vowel seems to stay too much to the back in the simulations. Therefore there is less place for a high central vowel, and lower central vowels are preferred, just as in the case of the four vowel systems.

6.2.6 Eight vowel systems.

The next case are systems with eight vowels. From the simulation with ψ_{ac} =10%, 57 systems with eight vowels emerged. These systems are classified in figure 6.8. Again, the different types have been assigned letters. Types A, B and C conform to Crothers' universals 1 to 6. Type D does not conform to universal 6, type E does not conform to universal 5 and type F does not conform to universal 4 and 6. They do conform to all the other universals. In the data of Schwartz *et al.* that contained 19 languages with eight vowels, type A occurs in 42% of the cases, type B occurs in 16% and type C and F both in 5%, i.e. one case each. However, another four languages in Schwartz *et al*'s data seem to have three central vowels. In these systems the central

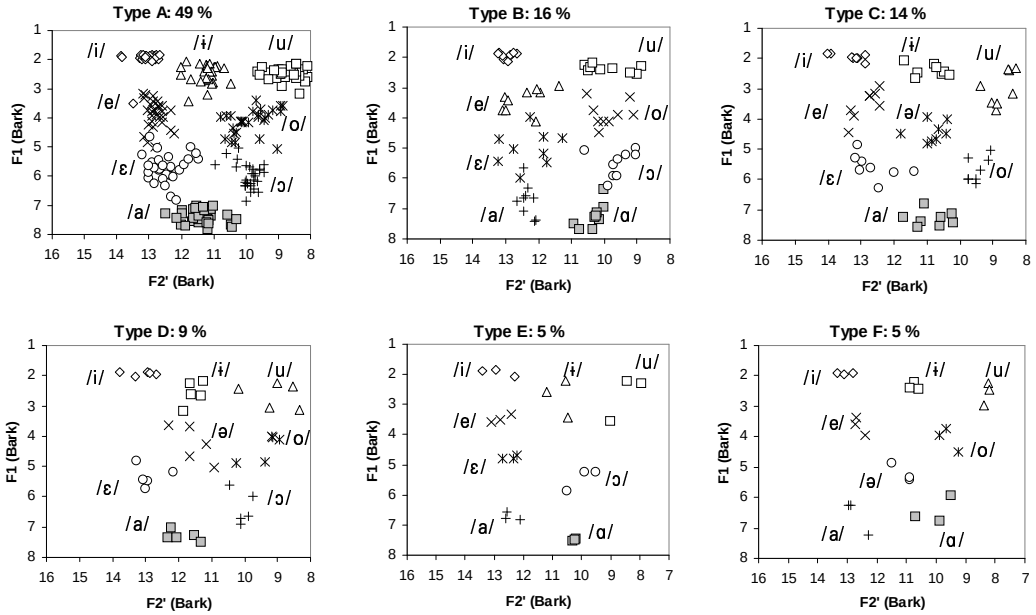


Figure 6.8: Classification of eight vowel systems.

vowels use two levels of height (corresponding to F_1 distinctions) and two levels of position/rounding (corresponding to F_2' distinctions). In the systems emerging from the simulations, only one degree in the F_2'-dimension seems to be used for central vowels. Again this could have to do with the fact that the high front vowel is usually too far to the back in the simulations.

6.2.7 Nine vowel systems.

The last vowel systems that have been analysed and classified were the nine vowel systems that emerged from the simulation with ψ_{ac} =10%. Of the hundred emerging vowel systems, 18 contained nine vowels. Representative agents with these vowel systems are shown in figure 6.9. In this figure, types A, B and D conform to Crothers' universals for nine vowel systems. Types E does not have /o/, so it does not conform to universal number 7. It does conform to all other universals, however. Schwartz *et al.* (1997a) present data on 24 nine-vowel systems. Of these, 29% were of type A, 4% (one system) is of type B, 17% were of either type C or F and 4% (one system) is of type D. However, the symmetrical nine vowel system without central vowels that accounts for 29% of the cases in their data, does not emerge in the simulations. Also, systems with three central vowels do not appear, just as in the case of the eight vowel system.

6.2.8 Crother's others.

There are other universals in Crothers' list that have not been checked for the emerged systems, yet. These are the universals 8–12 that are more or less independent of the number of vowels in the vowel systems. Universal number 8, which says that the preferred number of vowels in a human languages is five, can not be checked with the data that have been used so far. The number of vowels that emerges is dependent on the values of a number of parameters of the simulation (see the section on the parameters of the simulation in chapter 4). A number of values for these parameters have been chosen in the experiments presented so far in order to give interesting and realistic vowel systems with various numbers of vowels. It is therefore impossible to say anything about the preferred number of vowels that

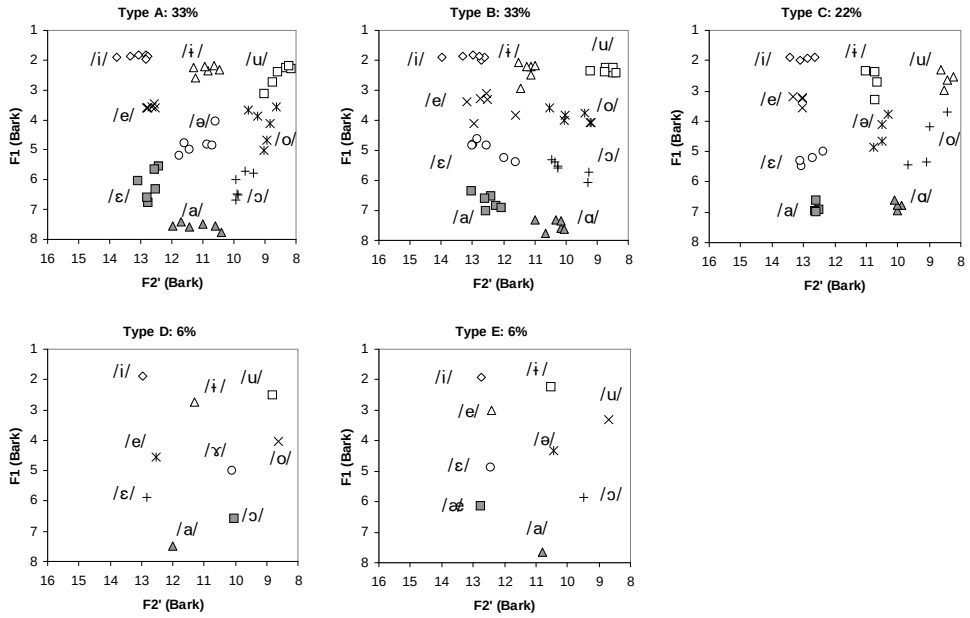


Figure 6.9: Classification of nine vowel systems.

emerges. Below an experiment will be presented that tries to check what the preferred number of vowels in the emerging vowel systems is.

Universal number 9, which says that the number of height distinctions is equal to or larger than the number of backness distinctions appears in the emerging vowel systems as well. None of the emerging vowel systems has more backness distinctions than height distinctions. The emerging systems also conform to universal number 10, which states that systems with two or more interior vowels always have a high one. The only exception to this universal is type F for systems with six vowels (figure 6.6), but this seems to be a rather anomalous system anyway, and only occurs one time (of the 54 systems with six vowels that emerged).

"The number of vowels in a column of interior vowels," as stated by Crothers' (1978) universal number 11, "cannot exceed the number in the front or back columns." This does not happen in the emerged systems, either, except, possibly in type F of the eight vowel systems (figure 6.8). However, this is also a rather anomalous system, occurring only in one of the 57 eight vowel systems, and it could even be argued that this system also has three vowels in the back column.

Crothers' last universal that is applicable to the emerged system, says that the number of height distinctions in the front column is equal to or greater than the number of height distinctions in the back column. This is not a strong universal, however. Schwartz *et al.* (1997a) found a number of exceptions. In the emerged systems, a number of exceptions against this universal also appear. If the systems with three vowels are not counted, (these will always have an equal number of back and front vowels) 254 vowel systems in total have been classified. Of these vowel systems there were 11 systems (4%) that had more back vowels than front vowels. This conforms nicely to Crothers' universal 12.

6.2.9 Preference for a certain number of vowel prototypes.

The last thing that remains to be investigated is whether the sizes of the systems tend towards five vowels. Unfortunately, this does not seem to be the case for the parameter setting used in the experiments presented above. The results of an experiment for investigating whether the emergent systems have a preference for a certain number of vowels shows (figure 6.10) that they prefer systems of four vowels. The data in the graph has been collected by running the simulation for many different values of the acoustic noise parameter ψ_{ac}. Furthermore, the standard parameter settings and a population of twenty agents were used. The simulations were run for 25 000 imitation games, exactly as in the experiments that resulted in the vowel systems that have been classified in this chapter. The values of ψ_{ac} that were used were the values 0.08 to 0.24 with intervals of 0.01 (17 values in total). These values were chosen so that below 0.08 only systems with more than nine vowels occurred, while above 0.24, only systems with less than three vowels occurred. It could therefore be that systems with two vowels occur much more frequently, because not all parameter settings for which they emerge have been tested. As they occur only very rarely in human systems, they have been ignored. For reference, the size distribution in Schwartz *et al.*'s (1997a) data is given as well in the upper part of the figure. The peak at five vowels is very clear.

For each of the values of ψ_{ac}, 100 runs of the simulation were done. It was then counted how many times each vowel system size occurred. The actual number of occurrences of the system sizes is shown on the right axis and with the solid line in figure 6.10. The left axis and the dashed line show a frequency that is weighted according to the relative size of the interval of ψ_{ac}. The difference between 0.23 and 0.24 for ψ_{ac} is relatively smaller than the difference between 0.08

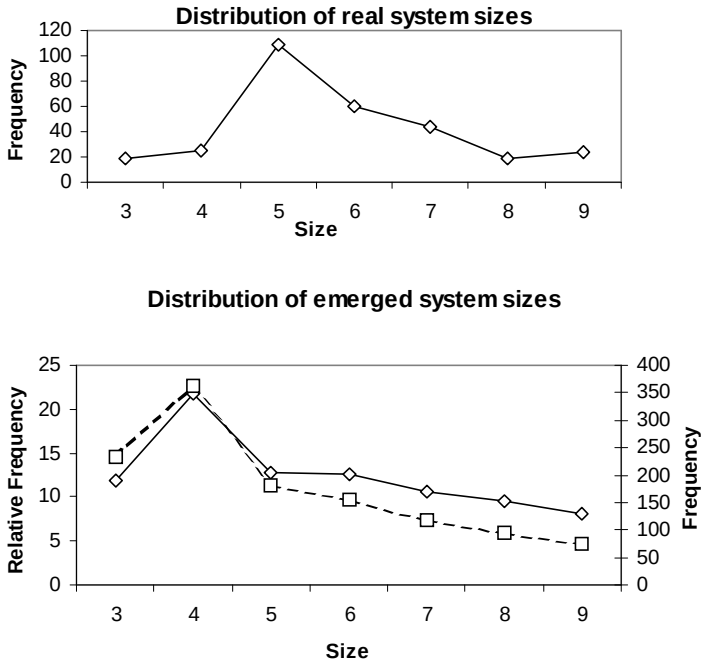


Figure 6.10: Distribution of vowel system sizes.

and 0.09. This is compensated for in the relative frequency by dividing all counts for that parameter setting by hundred times the value of ψ_{ac}. But in the qualitative appearance of the graph this makes no difference.

It can be seen that there is a large preference for systems of four vowels. Both systems of three and five vowels occur less frequently. Although this does not conform to the finding that systems of five vowels occur most frequently in human languages, it does show that the effect of a strong preference for a certain number of vowels also occurs in the emerged systems. Apparently systems of four vowels are the strongest attractors of the system for the largest range of values of ψ_{ac} for the specific parameter settings that were used in these experiments.

6.3 Conclusion

In this chapter it was shown that the vowel systems that emerge in populations of agents that play imitation games are quite realistic. Most of the systems that emerge show the same universal tendencies as the ones that Crothers (1978) found in human vowel systems. Also the frequencies with which different types of vowel systems appear are quite the same as the frequencies with which different types of human vowel systems appear. The systems (for given numbers of vowels) that appear most frequently in human languages also tend to emerge most frequently from the simulations. But other possible systems for given numbers of vowels appear, too, with frequencies that are comparable with the frequency with which they appear in human languages. Whereas previous simulations, based on optimisation of acoustic distinctiveness (Liljencrants & Lindblom 1972, Vallée 1994, Glotin 1995, Schwartz *et al.* 1997, Berrah 1998) have succeeded in producing the most frequent vowel systems, they generally did not predict alternative systems very well. This is probably because these simulations were actively optimising vowel systems, something that does not happen in human languages. The simulation investigated here only tries to develop successful imitation under human-like constraints of perception and production. Optimal and near-optimal vowel configurations can be considered attractors of the agents' vowel systems. The more optimal a system is, the stronger an attractor it is. This implies that the more optimal systems will emerge more frequently, but due to different histories and starting conditions less optimal systems will sometimes emerge as well.

Although these results are quite satisfying, they should be regarded with a little caution. For one thing, a number of vowel configurations were not predicted very well. Especially some types of systems with 3 and 4 vowels emerged that do not appear frequently in human languages, such as the vertical three-vowel system. It might be possible to explain this as a result of the difficulty of finding completely fronted high front vowels, which is due to non-linearities in the perception function. Also systems with more than two central vowels were not predicted. This could have to do with the fact that the acoustic representation does not easily distinguish between different types of central vowels, or with the fact that high front vowels are generally located too far back, so that there is no room for multiple distinctions in F_2'. Furthermore, the emerged systems were analysed in acoustic space, and the relative position of vowel prototypes in the systems was considered to be more important than their absolute positions. Crothers (1978) considered vowel systems in a similar way, but Schwartz *et al.* (1997a) considered more absolute positions of vowels. However, as has already been remarked in the discussion on human vowel system typology and in relation to figure 6.1, it is quite hard to determine the exact po-

sition (or even the exact number) of vowel phonemes in acoustic space for any human language. It is therefore not clear a priori how close the similarities between vowel systems that emerge from simulations and systems that are found in human languages should be, or, for that matter, what the theoretical value of such similarities as *are* found are.

Another problem with the results from the simulations is that the number of vowels that emerges seems to be determined by the acoustic noise level ψ_{ac} and a number of other parameters of the simulation (see the section on changing parameters in chapter 4). Most of the parameter settings in the system could probably be explained as resulting from properties of human perception and production of speech. But no single parameter setting reproduces the huge range of numbers of vowels (from two to approximately fifteen) that are observed in human languages. The simulations do produce systems with different numbers of vowels, but the range is quite limited. Of course, the fact that the number of vowels does not have to be predetermined, and is not fixed for a given parameter setting, is already a unique property of the simulations with respect to previous simulations. In all previous work the number of vowels had to be fixed beforehand. On the other hand, if one integrates over the possible values of ψ_{ac} that result in realistic numbers of vowels, one does find (see figure 6.10) that one vowel system size occurs more frequently than any other, just as one finds that in human vowel systems one size is preferred over all others. Unfortunately, in the simulations the preferred number of vowels is four, while in actual human systems the preferred number is five.

Even though the peak value is not the same as the peak value for human languages, it is interesting that the same phenomenon occurs. This is again an indication that self-organisation in a population is a good model for explaining the shape and size of vowel systems. The large range of possible sizes for human vowel systems can then only be explained if ψ_{ac} or a parameter with a similar function is learnt, depending on the number of vowels that is observed in the ambient language. This can perhaps be done in a way that is similar to what was suggested at the end of the section on simulations without non-verbal feedback in chapter 5.

Taking everything into account, it is justified to conclude that the emergent vowel simulations result in very realistic vowel systems. Although some discrepancies exist between the frequencies of occurrence of human vowel systems and the frequency of the different types of emerged systems, usually the predictions of the simulation were very accurate. Not only did it predict the most frequent systems very well, but also the systems that occur less frequently. Almost all the emerged systems conformed to the relevant universals found by Crothers (1978). Furthermore, by integrating over the acoustic noise parameter, it was found that one system size is preferred above all others, just as is the case in human vowel systems. The idea that self-organisation must be part of any explanation of the structure of human sound systems is sound.

7. On Complex Utterances

Of course, the results of the experiments with emerging vowel systems are impressive, but their importance for understanding the evolution of human languages should not be overestimated. No sound system of a human language contains only vowels, and even if it would, these vowels would not be uttered in isolation. In real human speech, sounds are uttered in a continuous stream, with all kinds of co-articulatory effects. As will be explained below, this has important repercussions for understanding the emergence of speech sounds. Also, even though some noise was added to the utterances of the agents, no systematic differences between the agents existed. On the other hand, humans produce sounds that differ systematically from each other. The length and shape of the vocal tract, the vocal cords as well as the dialect of the speaker all influence the actual sounds that are produced. The difference between the properties of vowels in isolation, spoken by a single speaker and vowels spoken by different speakers in the context of a word is clearly illustrated by the differences between figures 4.2 and 6.1. Humans seem to be able to cope with this perfectly. Linguistic variation is extremely hard to model, so one would like to abstract away from them. Unfortunately it is difficult to say whether these phenomena play a role in the learning and the emergence of systems of speech sounds or not. Abstracting away from them might perhaps not be justified.

One thing is clear, however. It is important to extend the methodology of studying the emergence of speech sounds with computer simulations to more complete models of speech. These models should ideally be able to work with speech signals that are more complex than isolated vowels, and ideally be as free of simplifications and abstractions as possible. In this chapter it will be explained why it is necessary to proceed towards more complex speech sounds, what has already been done and which techniques could be used for studying complex speech sounds.

7.1 Why Complex Utterances are Essential

Every human language has consonants (see e.g. Maddieson 1984). Although it would be possible to make a communication system that uses vowels only, this never happens in real languages. This has nothing to do with the fact that humans are capable of producing a much larger number of consonant sounds than vowel sounds. The largest number of different vowel sounds in a language (approximately 15 in Norwegian (Vanvik 1972)) is larger than the smallest total number of phonemes in a language (11 in both Rotokas (Firchow & Firchow 1969) en Murá-Pirahã (Everett 1982)). Apparently successful communication systems can be constructed with a small number of phonemes. And therefore, using only vowels would seem to be sufficient in principle.

But languages seem to organise speech in terms of units that are larger than the individual phonemes. These units usually have an onset, which often does not have high energy, they have a nucleus, which normally has high energy and sometimes they have a coda, again with low acoustic energy. These are obviously syllables. Consonants are the speech sounds that are usually found in the onset and the coda, whereas vowels are the sounds that are typically found in the nucleus. As the cyclic motion of the articulators that is a characteristic of syllables seems to be a very strong universal property of human speech, all languages tend to have consonants as well as vowels.

7.1.1 Universal tendencies of consonant systems.

Just as in the case of vowel systems, many universal tendencies are found in consonant systems. The first of these is one of frequency. Some consonants, such as [m], [p] and [k] (appearing in 94%, 83% and 89% of the languages in $UPSID_{451}$, respectively) are almost universally present. Others, such as [ɹ] (occurring in English) [χ] (occurring in Dutch) and [ʀ] (occurring in French) appear much less frequently (2%, 10% and 1% of the languages of $UPSID_{451}$, respectively). In fact Lindblom and Maddieson (1988) have found that consonants can be divided in basic, elaborated and complex articulations. Basic articulations are articulations that involve one action of the articulators, and only minimal displacement from their position at rest. All the really frequently occurring consonants are basic articulations. Linblom and Maddieson recognise 11 basic obstruents and 7 basic sonorants. Elaborate articulations, according to Lindblom and Maddieson (1988) "...are derived from a criterion of articulatory displacement: configurations representing departures from the near-rest position of lips, tongue-tip and tongue-body components of an articulatory model..." These include (among many others) creaky voiced, breathy voiced, ejective, prenasalised, aspirated, retroflex and uvular articulations. Complex articulations are combinations of elaborate articulations. It was found that languages with small phoneme inventories tend to use basic articulations, while languages with larger phoneme inventories use elaborated and complex articulations, next to approximately the maximum number of basic articulations. The conclusion that can be drawn from this is that in the case of consonant systems, not only acoustic distinctiveness, but also articulatory simplicity plays an important role.

But these are not the only universals that can be found in consonant systems. Just like vowel systems, consonant systems tend to be symmetrical. Vowel systems are symmetrical with respect to the acoustic characteristics and the places of articulation of its constituent vowels. Consonant systems are symmetrical with respect to the parsimonious use of articulations. If, for example a language makes a distinction between voiced and unvoiced stops, it tends to use this distinction at all places of articulation it uses. A system containing [p], [b], [t], [d], [k] and [g] is therefore much more likely than a system containing [b], [t] and [g] only. If different modes of voicing (breathy voice, aspiration etc.) are used, they tend to be used for all places of articulation as well. Also, if a language uses a certain place of articulation, it is likely to have a full set of consonants at this place of articulation. If a language uses retroflex articulations, for example, it is likely to have [ȼ] and [ɖ] as well as [ɳ] and possibly [ʂ]. Consonant systems are not always completely symmetric. Articulatory simplicity plays an important role in explaining the gaps in consonant systems. If a language makes the distinction between voiced stops and unvoiced stops, the most likely place of articulation not to have a voiced stop is the one most to the back, as there it is hardest to make clearly voiced stops. Also, a language that makes a distinction between voiced and voiceless sounds is more likely not to have voiced fricatives than not to have voiced plosives, as the former are much harder to make than the latter.

A final observation on consonants is that they can occur almost exclusively only in combination with other sounds. In fact, a lot of consonants, especially stops and affricates cannot appear in isolation at all. An important implication of this is that because of the nature of human articulators, sounds occurring in sequence will influence each other. Articulators cannot move from one position to the next instantaneously, causing phonemes to sound differently in each context in which they ap-

pear. This phenomenon is called co-articulation. Co-articulation is a very complicated process, some of which is language-dependent and some of which is language-independent. For example, the exact acoustic properties of a [t] are always dependent on the following vowel. For different vowels, the characteristic transitions of the formant frequencies will be different. This is independent of the language. However, the /l/ in English is pronounced much more velarised ("darker") at the end of a word than at the beginning of a word. This is a language dependent effect. The result of both these effects is similar. A consonant cannot be related to a single acoustic signal, and because consonants influence the neighbouring vowels as well, a vowel will also have a range of realisations rather than a single one.

7.1.2 Syllable structure.

The fact that consonants appear almost always in combination with other sounds implies that they cannot be studied in isolation, such as could be done for vowels. A simulation that models consonants should therefore model syllables. The problem is that there are a lot of different ways in which consonants and vowels can be combined into syllables. The most frequently occurring type of syllable (see e.g. Vennemann 1988) is the syllable consisting of a consonant (C) followed by a vowel (V). This type of syllable (called CV-syllable) occurs in every language of the world. Syllables that consist of only a vowel also appear in all languages of the world. Syllables that end in consonants are rarer and so are syllables that contain clusters of consonants. It is perhaps logical that consonant clusters are rarer than single consonants, but it is not directly obvious why syllables that start with a consonant occur much more frequently in the languages of the world than syllables that end in a consonant. Also, the possible combinations of consonants and vowels into syllables are quite restricted. They are governed by the so-called sonority hierarchy. This hierarchy states that some sounds are more sonorous than others are. Vowels are the most sonorous sounds, semivowels and liquids are less sonorous, nasals are even less sonorous, and voiceless fricatives and stops are least sonorous. Vennemann's (1988) proposal of a sonority hierarchy (which he calls consonantal strength) is illustrated in figure 7.1. The preferred shape of syllables can be "explained" by the fact that the sonority of segments increases towards the nucleus of the syllable. A syllable like /plis/ is therefore much more likely than one like /lpis/.

Unfortunately this is not a real explanation. The sonority of a phoneme is something that is very hard to define on other than subjective grounds, and quite often it is determined on the basis of the position in the syllable in which the sound usually occurs, thus creating a circularity in the explanation. In fact, the sonority hierarchy is often not more than a convenient way of notation for the preferred order in which segments tend to occur in syllables. Independent criteria for defining sonority, such as acoustic energy and opening of the jaw are better, but there is not always a direct relation. The best independent explanation based on functional criteria is the one by (MacNeilage, *to appear*) where it is said that the

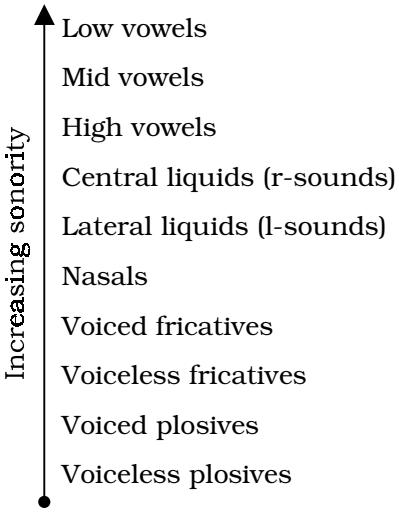


Figure 7.1: Sonority hierarchy (adapted from Vennemann 1988.)

form of syllables is determined by the preferred cyclic motion of the jaw. The jaw tends to open and close in a cyclic manner, thus causing syllables, and the sonority hierarchy.

7.1.3 Sound change and complex utterances.

In short, there are a number of interesting universals in the way in which consonant inventories are built up and in the way that consonants (and vowels) are combined in actual utterances in human languages. But these are not the only reasons for modelling more complex utterances. Another important reason for modelling complex utterances is for studying the way in which sound systems change. As has been said in the chapters on vowel systems, the vowel systems that emerge might be realistic, but the way in which they change is not. One way in which the vowel systems change, the slight movement of position of the prototypes is probably realistic. However, vowel systems of real human languages do not change by addition of random new vowels, the other way in which the systems in the simulation can change.

A much more important way in which human sound systems change is through the influence of neighbouring sounds on each other. Through co-articulation, as has been described above, phonemes can take over properties of neighbouring[1] phonemes, as long as this does not change it into another phoneme. The resulting sound is a conditional allophone of the phoneme, because it only occurs in the specific environment, whereas the original phoneme does not occur in that environment. As no minimal pairs of words can be formed, the sound is not a phoneme. Through other changes in the language, however, the conditioning environment might be lost, but the variation in articulation might be retained, because children keep on imitating their parents more closely than strictly necessary. The allophone then turns into a phoneme, because its occurrence is no longer predicted by its environment. A classical example is the appearance of nasalised vowels in a language. Originally, the language contains words that end in a nasal stop. Nasal consonants tend to influence the preceding vowels by nasalising them. As long as the nasal stops are present, the nasalisation of vowels is completely predictable. However, if the nasal stops disappear, the words with nasal vowels will start to contrast with non-nasal vowels, and the nasal vowels will get phonemic status. Many other examples of sound change caused by co-articulatory influence and subsequent loss of context have been found.

In other words, if one wants to model realistic sound change, one has to model complex utterances. There is another reason why it is necessary to model complex utterances in order to get realistic sound change. The types of sound changes described above can only take place in a population if variants of a certain phoneme can coexist. Therefore it must be possible to disambiguate and recognise the changed sounds. If one of the agents in a population starts to pronounce one of its sounds differently, the other agents must be able to recognise this as a variant of an existing sound and not as an entirely new sound. In order to be able to adapt their own sounds they must also be able to recognise of which sound it might possibly be a variation. This can only be done using the context in which the sound appears. In human language, the possible form of a word can be derived from the phonetic, but also from the syntactic, semantic and pragmatic context in which it appears. In the case of an agent simulation the only available context is the phonetic context. There-

[1] In fact, the influence does not have to come from only the nearest neighbours. Influences can extend over intervening sounds.

fore complex utterances are necessary to enable sound change. This also implies that at least two levels of representation are necessary. There must be a representation of individual sounds, (or short sequences of sound) which could be called the phoneme level and a representation of the contexts in which the sounds can appear and that could be called the word-level.

7.1.4 Lindblom's CV-experiment and other computer models of complex utterances.

The first to investigate complex utterances with a computer model were probably Lindblom *et al.* (1984). They investigated the possibility of explaining phonemic coding through optimisation of a number of articulatory and acoustic constraints. Human speech does not use completely different signals for different words and morphemes (this would be a *holistic* coding). Instead, it seems that words and morphemes are built up of smaller acoustic and articulatory units called phonemes (hence the term *phonemic* coding). Lindblom *et al.* investigated a model that could produce a large range of signals and that had to build up a repertoire of maximally distinct and easy-to-produce signals. The signals consisted of an onset (a stop consonant) and a nucleus (a vowel). The possible onsets were [b, d̪, d, ɖ, ɟ, g, ɢ]. Some of these were considered to be easier to articulate than others. The possible nuclei were 19 vowels that were evenly spread over the vowel space, making for a total of 133 signals. The nuclei were characterised acoustically by their formant frequencies. The onsets were characterised by their locus pattern, that is the formant frequencies from which the formants start to move to their values for the vowel. Acoustic effort was determined by the distance over which the articulators had to move, and by the inherent difficulty of producing the onset. Acoustic distance was determined by the total difference between the formant trajectories of two syllables. They found that by optimising these constraints a system of signals emerged that consistently used the same onsets and nuclei for the possible syllables, instead of using a random subset of the possible onsets and nuclei. The syllable system could therefore be analysed as phonemically coded, instead of holistically coded. It was also found that the onsets and nuclei that were used most often were the ones that occur most often in human language.

Although these results indicated that functional criteria could be used for explaining the structure of consonant systems, a possible criticism of this work is that it is too simplistic and that it takes too much for granted. Lindblom *et al.*'s (1984) simulation only allowed for a limited number of possible onsets, of which it had to specified beforehand whether they were hard to articulate or not. The number of possible nuclei was larger, but nevertheless discrete and finite. The modelling of the syllables did not take into account a number of other phenomena, such as coarticulation and timing. It also already assumed that syllables were built up along the lines of a consonant followed by a syllable, and could therefore not explain why this seems to be the preferred form for syllables in human language. Of course, this was not the objective of their research, but it remains an important question to answer. Another limitation of their work from the point of view of the research presented here is that it consisted of purely creating an optimised system of syllables. As has been mentioned in the section on vowels, humans do not actively optimise their systems of speech sounds. The optimisation should therefore be explained through dynamics in the population. This was not present in their model.

Carré & Mody (1997) have extended Carré's distinctive region model to predict places of articulation of consonants. Starting from a uniform acoustic tube, they try to determine the signals that cause the most distinct acoustic signals (formant tran-

sitions) for the least effort. In this way they find a number of candidate places of articulation that correspond to the most frequent places of articulation for consonants. A disadvantage of their model is that it predicts a large number of possible places of articulation (eight) and possible combinations of these, but it lacks a clear criterion for preferring the exact places of articulation that are found in human languages. It is also not able to predict the preferred sets of consonants that occur in human languages. The contribution of their model is to show that the sets of consonants that occur most frequently in human languages are maximally distinctive in their signals.

Recently Plaut & Kello (to appear) have constructed a sophisticated model for learning of phonology using neural networks. They are not interested in explaining the structure of speech, though, but rather in modelling the cognitive process of language acquisition in children. Their articulatory model is quite detailed, but not based on an actual simulation of the vocal tract, and therefore probably already biased towards the sounds that are found in human languages (although they do not provide sufficient detail in their paper for assessing this.) They also assume that words are already coded as phonemes and they train their model with words that are already split up into phonemes. Although they do not use their model for studying the explanation of the properties of human speech, it does seem to be one of the most detailed models of phonological acquisition so far.

Considering the problem from a different point of view, Redford *et al.* (1998, *to appear*) have built a computer model for explaining the kinds of syllable structures that appear in human languages on the basis of the interactions between the individual phonemes. Their approach uses a genetic algorithm, in which a population of candidate strings (the "words" of the language) are selected on the basis of a number of functional criteria and then mutated, crossed and multiplied in order to form the next generation of strings. After a number of generations, strings appear that have a structure that is reminiscent of the structure of human syllables. A number of criticisms apply to their work. First of all the strings are purely symbolic. No real articulation takes place and time does not play a role. With each symbol in the strings a number of features is associated, and these are used to determine the quality of the string. Secondly, the model works with genetic evolution, something which human languages clearly do not do. However, one could consider this genetic evolution as an abstract model of human cultural evolution. The most important problem with their work is that the constraints that are found to govern the structure of human syllables are already implemented as selection criteria. The model therefore does not explain the structure of syllables; it only shows that if the criteria are present, evolution (be it genetic or cultural) would be sufficient to produce syllables that conform to the constraints.

7.2 The Consonant-Vowel System

The first simulation that was build within the framework of Steels' ideas (Steels 1995, 1996, 1997b, 1998b) for working with more complex utterances was based on Lindblom *et al.*'s (1984) work on consonant-vowel syllables. It was decided to base the first experiments on this work, because it was well described, because it had results that could be easily represented and verified and because it could be implemented in a reasonably efficient way.

7.2.1 Production of CV-syllables.

As has been mentioned above, the possible onsets of the syllables consisted of the seven consonants: [b, d̪, d, ɖ, ɟ, g, ɢ]. The possible nuclei of the syllables consisted of the 19 vowels: [i, y, ɨ, ʉ, ɯ, u, e, ø, ə, ɤ, o, ɛ, œ, ʌ, ɔ, æ, ɶ, ɑ, a]. Linblom *et al.* (1984) are not very clear about the exact acoustic values of these signals. Therefore, a new interpretation of their acoustic values needed to be made. This was relatively straightforward in the case of the vowels. The synthesiser that was used in the vowel experiments described in this thesis was based on data points that were taken from Vallée (1994) pp. 162–164. All the formant values that were used in the consonant-vowel simulation were also taken from this table, except for the vowel [œ] which does not occur in Vallée's table. For this vowel a measurement of the formant frequencies of the author saying this vowel was used, which resulted in formant frequencies 600Hz, 1600Hz, 2500Hz and 3500Hz. The data for the consonants were taken from (Fant 1973; Stevens & Blumstein 1975; Cooper *et al.* 1976) and from measurements done by the author, which are described in appendix E. Lindblom *et al.* (1984) make the simplifying assumptions that a plosive consonant is uniquely characterised by a formant pattern (its *locus*) and that this locus pattern remains the same, independent of the following vowel. Both these assumptions are not strictly true. Plosive consonants are also characterised by the frequency pattern of the burst of noise that follows their release as well as by the timing of the burst and the onset of voicing. Nor is the locus independent of the following vowel. However, reasonably realistic sounds can still be generated if the two assumptions are accepted. In any case, Lindblom *et al.*'s (1984) implementation was followed as closely as possible, so these simplifying assumptions also had to be made in the simulation presented here. The formant patterns that were used in the simulation are given in table 7.1.

	[ɢ]	[g]	[ɟ]	[ɖ]	[d]	[d̪]	[b]
F_1	150	150	150	150	150	150	150
F_2	1150	1650	2200	1750	1700	2050	1060
F_3	3000	1950	2500	2250	2700	3080	2270
F_4	3600	2920	3600	2920	3300	3600	3080

Table 7.1: Locus patterns for consonants.

7.2.2 Perception of CV-syllables.

The distance between two syllables was calculated as the weighted sum of the distance between their onsets (plosive consonants) and their nuclei (vowels). The distance between the vowels was calculated using the same distance measure as the one used in the vowel experiments. This same distance measure was also used to calculate the distance between the consonants. The total distance was then calculated as follows:

$$7.1) \qquad D_{syl}(C_1V_1, C_2V_2) = \frac{D(C_1, C_2) + 0.5 \cdot (V_1, V_2)}{1.5}$$

where D_{syl} is the distance between the syllables. C_1 and V_1, respectively C_2 and V_2 are the consonants of the first, respectively the second syllabe. The factor 0.5 is a weighting factor for the relative influence of consonants and vowels, whose value was chosen to result in approximately equal influence of vowels and consonants on the syllable distance. Because all the first formants of consonants are equal, their distance is smaller on average than the distance between vowels.

The rules of the imitation game played in the population of agents using consonant-vowel syllables are equal to the ones given in tables 3.4, 3.6 and 3.5, except for one important difference. In the case of the vowel imitation game, new vowel prototypes that were added or updated in reaction to an imitation game, were improved on the basis of the same acoustic distance function that was used to determine which of the vowel prototypes corresponded to a signal that was perceived. In the consonant-vowel system a different function from the one given in equation 7.1 was used for improving CV-prototypes. This was done in order to model the articulatory effort involved in producing complex syllables, something that Lindblom *et al.* (1984) did explicitly in their optimisation function. The syllables were improved on the basis of a function that calculates the quality of any given combination of a consonant and a vowel with respect to a given acoustic signal:

7.2) $$Q_{syl}(CV, A) = \frac{D_{syl}(CV, A) + 0.2 \cdot |pos(C) - pos(V)| + 0.8 \cdot extreme(C)}{2}$$

where $Q_{syl}(CV,A)$ is the quality of syllable *CV* with respect to the acoustic signal *A*, $D_{syl}(CV,A)$ is the distance of syllable *CV* to acoustic signal *A* (which consists of an observed locus pattern and an observed vowel formant pattern) *pos*(*C*) and *pos*(*V*) are the articulatory positions for consonant *C* and vowel *V*, respectively and *extreme* is a function that is 1 if consonant *C* is extreme, and 0 otherwise. Again, the factors 0.2, 0.8 and 2 have been chosen to appropriately weigh the different terms of the function in order to get realistic results.

The term that calculates the difference in position between the vowel and the consonant models the articulatory effort needed to produce the syllable. The positions of the vowels are as usual, and the positions of the consonants are 1 for [ɢ] and [g], 0.5 for [ɖ], [d] and [d̪] and 0 for [ɟ] and [b]. The term that calculates the extremeness of utterances is used in order to disfavour consonants that involve extreme displacement from the rest positions of the articulators. It is 1 for [ɢ], [ɖ] and [ɟ] (defined by Lindblom *et al.* (1984) as extreme articulations), and 0 for all other consonants.

Another difference between the vowel simulation and the CV-syllable simulation was the way in which acoustic noise was added to the consonants and the vowels. In order to make it necessary for the systems to become robust, the formant patterns of consonants and vowels were shifted randomly. The amount with which formant patterns were shifted in the CV-syllable simulation was taken from a normal distribution, so that small shifts occurred more frequently than large shifts. In the vowel simulations, formants were shifted by values from the uniform distribution, so that extreme shifts had an equal probability of occurring as small shifts. The shift of CV-syllable formant patterns was calculated using the following formula:

7.3) $$F_i' = F_i(1 + S(0, \psi_{ac}))$$

where F_i is the unshifted formant frequency (for i=1..4) F_i' is the shifted formant frequency, ψ_{ac} is the noise percentage and $S(0, \psi_{ac})$ is a random variable taken from the normal distribution[2] with average 0 and standard deviation ψ_{ac}. A normal distribu-

[2] Normally distributed random numbers were calculated from uniformly distributed random numbers with the following formula: $S(\mu,\sigma) = \sigma + \left(\sum_{i=1}^{6} U(0,1) - 3\right) + \mu$

tion was chosen, because this seemed to be more realistic. In fact, it might be a good idea to do future experiments with the vowel system using normally distributed noise as well. Note that due to the way the shifts are calculated, a similar value for ψ_{ac} can result in much larger shifts in this method than in the method used in the vowel simulations.

7.2.3 Results of the CV-syllable simulations.

The results of the CV-syllable simulations have been somewhat confusing. It appears that the outcome of the simulations is very much dependent on the settings of the parameters that determine the way in which the distance between syllables is calculated, the way in which the quality of syllables with respect to a given acoustic signal is calculated and the way in which syllables are merged. Some of the parameter settings would result in systems where only a few consonants were used, while almost all vowels were used. Other settings resulted in systems where all consonants were used, but only a few vowels and some parameter settings resulted in systems were all possible combinations of vowels and consonants were used. Only very few parameter settings caused systems using a true subset of both the vowels and the consonants to arise. In still fewer of these systems, consonants and vowels appeared in different combinations, resulting in truly "phonemically coded" syllables. Nevertheless, phonemically coded systems did appear.

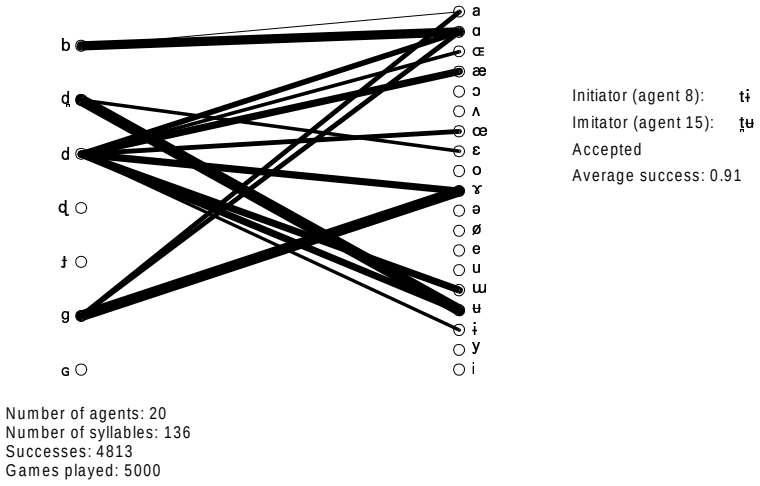


Figure 7.2: CV-imitation game simulation.

A further difficulty with the CV-simulations is that their results are harder to represent graphically. An example of a screen shot of the CV-simulation is given in figure 7.2. The points to the left represent the available consonants and the points to the right represent the available vowels. A line linking a consonant and a vowel means that at least one of the agents has the syllable with the consonant as onset and the vowel as core. The thickness of the line is a measure of how many agents share this syllable. Other information that is shown in the screen shot includes the number of gents, the total number of syllables in the system, the number of games played, the number of successful games, and the participants, the used syllables and the outcome of the game that was just played.

However, this is not a very clear way of representing the coherence in the population of agents. It is easier to evaluate the syllable inventories in a table. This is done in table 7.2, which shows the syllable inventories of a population of 20 agents after 10 000 imitation games. The noise level ψ_{ac} in this simulation was 5% (which should not be directly compared to 5% noise in the vowel imitation games). Each column contains the inventory of one single agent, while each row contains

---

where μ and σ are the mean and standard deviation and $U(0,1)$ is a random number taken from the uniform distribution between 0 and 1.

syllables that were considered good imitations of each other. The syllables in one row do not have to be exactly equal. This is illustrated many times in the table. The rows of the table are sorted so syllables that start with the same consonant are grouped together.

A phonemic analysis of the resulting syllables shows that all "ordinary" consonants are used in the syllables, while all "extreme" consonants are avoided. It could be argued that there is no pair of syllables with the same vowel that start with [ɡ] and [ɖ̟], but it would seem strange to analyse these two consonants as one phoneme. There seem to be at least five vowels that occur in minimal contrastive pairs, following the consonant [d], although not all agents share the syllable ending in [y]. The vowels phonemes in the agents' systems could then be analysed as: /œ/=[œ] and [ɶ], /ɑ/=[ɑ] and [a], /ɨ/ = [i], [ʉ], [ɯ] and [i], /ɛ/=[ɛ], [ə] and [æ] and /ɤ/ = [ɤ] and [y]. This analysis is somewhat fanciful and would result in a rather unrealistic vowel system. Nevertheless, it shows that certain vowel phonemes occur in different syllables.

	agents →																		
repertoire																		[gœ]	[gœ]
↓	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]	[gɑ]
	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]		[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]	[gɤ]
	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]	[gi]
	[bə]		[bə]	[bə]	[bə]	[bə]		[bə]		[bə]	[bə]	[bə]	[bə]	[bə]	[bə]	[bə]	[bə]	[bə]	[bə]
	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]	[ba]
	[dœ]	[dɶ]	[dœ]	[dœ]	[dɶ]	[dœ]	[dœ]	[dœ]	[dɶ]	[dɶ]	[dœ]	[dœ]	[dœ]	[dœ]	[dœ]	[dɶ]	[dœ]	[dœ]	[dœ]
	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]	[dɑ]
	[dæ]		[dæ]	[dæ]		[dæ]	[dæ]	[dæ]			[dæ]	[dæ]		[dæ]	[dæ]		[dæ]	[dæ]		[dæ]
	[dʉ]	[dɨ]	[dɯ]	[dɯ]	[dʉ]	[dɨ]	[dɨ]	[dɨ]	[dɨ]	[dʉ]	[dʉ]	[dɨ]	[dʉ]	[dɨ]	[dɨ]	[dɨ]	[dʉ]	[dʉ]	[dɨ]	[dɨ]
			[dy]					[dy]	[dy]	[dɤ]		[dy]			[dy]	[dy]			[dy]	[dy]
	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]	[ɖ̟ʉ]
	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]	[ɖ̟ɛ]

Table 7.2: Emerged CV-syllable repertoire.

7.2.4 Interpretation of the CV-syllable results.

What do these results mean? It seems that for the right parameter settings, a phonemically encoded system of syllables can emerge. However, the phonemes that emerge do not form a realistic system. It is quite rare for human languages to contain both a dental plosive and an alveolar plosive (in $UPSID_{451}$ only 31 languages, or 7% of the sample contains both types) because these consonants are very similar. The fact that no extreme consonants are selected is because these are assigned a lower quality than non-extreme consonants. Also the combination of vowel phonemes /ɨ/, /ɛ/, /œ/, /ɑ/ and /ɤ/ in the system is unrealistic.

Unlike the simulation by Lindblom *et al.* (1984) the agent-based model was not able to predict the most frequently occurring consonants and vowels from generating a set of good syllables. This can have several reasons. First of all, Lindblom *et al.*'s

model could directly select the best syllable (based on criteria of acoustic distinctiveness and articulatory ease) to be added to an expanding repertoire of syllables. In the agent-based simulation new syllables could only be added at random, while pressures towards distinctiveness and simple articulation had to be implemented by constraints on production and perception. The problem is that the recognition of phonemes is done on purely acoustic criteria, while the improvement of the prototypes is based on a compromise between acoustic similarity and articulatory ease. This can result in conflicting changes to the syllable repertoire, making emergence of coherence in the population impossible. This is the reason why the parameters have to be tuned much more carefully than in the vowel simulations. Of course, there is still some room for different parameter settings and it is conceivable that for certain settings of parameters, or for certain modifications to the perception and production of the speech sounds, more realistic systems would emerge. However, it would be much more desirable to have a system that is either robust with respect to different parameter settings or that tunes its own parameters.

A second reason for the lack of realism is probably the crude modelling of consonants and their interaction with the vowels. It is not clear what values for the formant loci of the consonants Lindblom *et al.* (1984) used in their simulations, nor is it completely clear how they calculated the distances between different syllables. Also, the data for vowels and consonants comes from different sources. The properties of vowels as well as consonants depend on the properties of an individual's vocal tract. If one mixes consonants from one speaker with vowels from another speaker, one could get unrealistic distances between the different syllables. The distances between the different vowels, consonants and syllables are of crucial importance to the structure of the syllable systems that will emerge. If they are unrealistic, the resulting syllable system and its constituent phonemes will be unrealistic.

The realism of the system could be improved by careful tuning of parameters, the addition of self-tuning mechanisms, the use of better data for generating vowels and consonants and more realistic perception- and production mechanisms. However, the question arises whether this is a good way to improve the understanding of the role of self-organisation in the emergence of speech sounds.

The experiments with the vowel sounds have already shown convincingly that a system of realistic (near-optimal) speech sounds can emerge under the right constraints of perception and production. The simulation of CV-syllables only changes the repertoire of possible signals and the articulatory and acoustic constraints that play a role. Lindblom *et al.* (1984) have already shown that a system of CV-syllables that optimises both acoustic distinctiveness and articulatory ease will have phonemically coded syllables, i.e. syllables that reuse the same onsets and codas. A similar result was obtained in the preliminary experiment described above, except that the onsets and codas that were found did not form a realistic sound system. The main contribution of the CV-syllable simulations would be to show how a near-optimal system of sounds can emerge under two conflicting pressures (acoustic and articulatory) instead of under only one pressure (acoustic) as in the vowel system. Also, it seems to be necessary for the emergence of realistic repertoires of syllables that agents can judge for themselves how accurate their imitations should be, i.e. how much articulatory effort should be invested in imitating the syllables. This is reminiscent of what was said in chapter 5 in the conclusion of the vowel simulations that did not make use of non-verbal feedback. Both these topics are interesting from the point of view of self-organisation in a population.

However, the disadvantages of the CV-simulation make it less interesting from a linguistic point of view. The first disadvantage, that of the rather artificial and arbitrary way of representing the acoustic signals of the syllables has already been discussed above. The second disadvantage is, of course, the fact that the consonants and vowels as well as the way in which they are combined into syllables have already been put into the model. One of the more interesting questions of the emergence of phonology, i.e. why is speech coded in syllables, or why are there consonants at all, is therefore not addressed. In order to address these questions, a more refined simulation is necessary. It should be free of linguistic assumptions and should directly model the function and dynamics of human articulation, hearing and learning op speech. Of course this is very ambitious, but in the next section a first effort to implement such a simulation, as well as the requirements it should fulfil will be described.

7.3 Towards a More Refined Simulation

This section describes an effort to build a simulation that models the perception and production of more complex utterances. The model consists of an articulatory and a perceptual part. The articulatory part consists of an articulatory synthesiser whose articulators move dynamically. The perceptual model consists of a number of feature detectors that can extract a large number of features from the speech signals it perceives. Unfortunately, no successful imitation games were played with this model. Why therefore bother with describing the simulation here at all? First of all, a number of design decisions had to be made in order to build it. These design decisions can be of relevance to people trying to elaborate on the research described in this thesis. Secondly, a number of experiments with learning to recognise and categorise speech signals have been performed, whose results will be presented in appendix F. Finally, building the model elicited a number of interesting research questions regarding the nature of speech signals and the emergence and learning of systems of complex speech signals. These questions will be presented and discussed here, even though no easy answers can be provided, and even though they are partly based on speculation and not on solid results.

The core of the simulation is formed by Mermelstein's articulatory model (Mermelstein 1973, Rubin *et al.* 1981). This is a geometrical model of a midsagittal cross-section of the human vocal tract. It models the position of the lips, the tongue, the palate, the velum and the pharynx and allows the calculation of the cross-sectional area of a slice of the vocal tract at any position from the vocal cords to the lips. The different articulators in the model can be moved in a continuous way, (within certain bounds) making realistic simulations of articulator movements possible. The model has nine degrees of freedom in total. The technical details of the model can be found in appendix F.

Mermelstein's articulatory model is rather straightforward and well-defined. No important design decisions had to be made in order to implement it. However, incorporating the model in an imitation game simulation entails interfacing it in two different ways. The positions of the articulators and the ensuing area function have to be used for producing an actual sound signal and the articulators have to be moved in a realistic way. The first task is relatively simple and has been well-researched (see e.g. Rabiner and Schafer Ch. 3 and references therein). The most important design decisions that have to be made here are in the trade-off between realism and speed of calculation. The details of how to go from a vocal tract outline in Mermel-

stein's model to a sound signal are given in appendix F. The second task is much more problematic, and involves a number of design decisions that have to be based on our quite incomplete knowledge of how human articulations are performed.

7.3.1 Movement of articulators.

Human speech organs do not move from one position to the next in one discrete step. Because of the finite speech with which they move they have to occupy intermediate positions. Because of their mass, they cannot change speed discontinuously. This will cause neighbouring sounds to merge into each other. A design decision about which model of articulator movement will be adopted has to be made. Also, realistic measurements representing the mass and the stiffness of the different articulators should be incorporated in the model. Further problems arise in the co-ordination of the different degrees of freedom of the articulatory model. The same speech signal can usually be produced with different settings of the articulatory parameters. A number of criteria will have to be defined on the basis of which certain parameter settings can be selected and others discarded. Research into modelling speech gestures (see e.g. Kaburagi and Honda, 1996) has shown that a minimisation of energy (in the form of articulator movement and the forces needed to produce this movement) results in realistic gestures. Other research has focused on modelling movements of articulators as a dynamical system (e.g. Kelso *et al.* 1986, Saltzman 1995). The many degrees of freedom in the vocal tract are seen in this approach as being controlled by a dynamical system with much fewer degrees of freedom, because the parameters of the vocal tract are *linked*, both in a physical as in a cognitive sense.

A related co-ordination problem is that articulatory goals are usually expressed in terms of constrictions of given degree (sufficient for say, a plosive or a fricative) at certain locations (say, palatal or alveolar) of the vocal tract. These have the most direct influence on the acoustic signal that is produced. But they can not be achieved in a straightforward way with the parameters of the articulatory model. The resulting vocal tract shape is dependent on all the parameters of the model at once, so a change of one articulatory parameter can result in quite different effects depending on the values of the other parameters. It is even possible to generate impossible vocal tract shapes, where the two walls of the model cross each other. A different set of parameters for controlling the shape of the vocal tract, from which the actual values of the parameters of Mermelstein's model can then be calculated, should therefore be adopted. This has already been done by the people at Haskins Laboratories who have developed an articulatory model of Mermelstein's synthesiser (see e.g. Saltzman 1986, 1995; Saltzman & Munhall 1989).

A good model of the kinematics of the vocal tract should take as input a number of articulatory positions, preferably specified in terms of place and degree of the desired constrictions of the vocal tract, and should output a realistic trajectory of the parameters of the articulatory synthesiser. This model will be called the *kinematics model*. The current implementation (for details, see appendix F) does not yet incorporate all of this, but is nevertheless able to produce dynamical movements of the articulators based on a sequence of (partly specified) articulator goals. One problem with the implementation as it is now, is that it works directly with the articulatory parameters, and not with parameters describing places and degrees of constriction. Why this is a disadvantage will be discussed below in the paragraphs on perception. The next problem of production is to determine what kind of commands will be given to the kinematics model.

7.3.2 Representation and learning of sounds.

This is the point where the (mostly) physical model of articulation and articulator movements meets the cognitive model of how sounds are represented and stored in the brain. The problem is that phonemes are never uttered in isolation. The smallest units that are uttered are words. Of course it is possible to store every word as a set of articulatory commands. However, there is abundant evidence, for example from speech errors or mispronunciations of words from an unfamiliar language, that indicates that words are analysed as consisting of smaller articulatory units that are stored separately. These units are usually considered to be phonemes. Alternatively they could be the possible onsets, nuclei and codas of syllables, which could consist of multiple phonemes. The difficulty is that it is not always possible to make the distinction between phonemes and clusters of phonemes on purely phonetic grounds. A sound that might be analysed in a certain language as a sequence of a plosive and a fricative might in another language be analysed as an affricate. It would therefore appear that both a level of representation[3] of individual sounds (in all their articulatory detail) as well as a level of representation for words (built up of the basic sounds) are necessary. Probably intermediate levels of representation, for example for clusters of phonemes or for possible syllables will also be necessary. For the individual sounds, the articulatory goals and the timing information needed for producing the sound should be stored. As the idea of the imitation game is that agents in the population start out with an empty repertoire of speech sounds. It therefore also depends on the learning algorithm that is used what representations the agents learn of the sounds they hear.

The learning model should model what is known about the way children acquire the sound system of their language (e.g. Vihman 1996) as closely as possible. Children start with a babbling phase, in which no recognisable imitations of words from the ambient language are produced. In this phase infants probably learn how their articulatory apparatus can be used to produce different kinds of sounds. Although their speech production is quite limited, they are nevertheless already able to learn a lot about the sound pattern of the ambient language passively. This indicates that production and perception are quite separate from each other. When children first start producing words, they produce quite imperfect imitations. Consonant clusters are reduced, complex sounds are reduced to simpler or already known ones and unstressed syllables are often reduced or dropped (see e.g. Vihman 1996 ch. 9). The children probably first imitate words holistically, i.e. without analysing them in terms of smaller units of sound, and only later, when their vocabulary starts to expand, learn a more phonemic representation. At all stages the passive knowledge of the language remains greater than the active capacity. Children may be able to hear the difference between two words, but may not be able to produce the same differ-

[3] The reference to *representation* of sounds and words is not meant to imply that sounds and words are actually represented in the human brain as identifiable symbolic entities. Neurophysiological research has shown that the brain works as a distributed neural network, and that it is probably not possible to find the exact locations where sounds or words are represented. For the sake of modelling, however, it is considered necessary to abstract away from the distributed neural representations towards a more tractable symbolic representation. It is assumed implicitly that this is possible. Whether this assumption is justified is of course open to debate.

ence. Gradually the child's command of the language improves until it is nearly perfect at five years of age.

The learning system of the agents should capture at least some of this complexity, and it is quite likely that a relatively simple learning mechanism would be able to do that. Agents start out with no knowledge of the sound system that is used around them. As far as they are aware, the sound system could either be phonemically or holistically coded (and it probably will be holistically coded in the first stages of a population trying to develop a sound system from scratch). The best they can do is to imitate the words as holistic units. But because the agents have a limited memory capacity, there will come a moment when it is more efficient to find recurring parts of the words in their vocabulary, store these separately and store words in terms of references to these parts. Different parts could be marked for the different contexts in which they occur, thus dividing them in different categories (for example: possible onsets, possible codas and possible nuclei). This opens the possibility for different treatments of different categories of sounds and thus for different phonological and phonotactical categories. This learning model closely follows Steels (1998a) ideas on the learning of syntax and syntactical categories.

As the agents try to minimise articulatory effort as well as storage, they will try to imitate the words they hear in a way that is as minimal as possible. Therefore consonant clusters and complex consonants will be reduced, syllable structure will be simplified and unstressed syllables will be removed. When the agent's vocabulary has expanded sufficiently, it will find out that it becomes necessary to make more complex articulations, and finds that it is easier to store these complex articulations because it already has coded its vocabulary (partly) in a phonemic way. Thus the kinds of speech reductions that children make seem to be explainable by a learning algorithm that minimises storage and articulatory effort. Unfortunately this learning algorithm has not been implemented, yet.

7.3.3 Main obstacles.

The main reason that the learning algorithm has not been implemented is that first two other big problems need to be solved. The first and simplest problem is that of recognising speech signals and analysing them in terms of their constituent sounds. The second problem is that of finding articulatory actions for imitating a (more or less) unknown sound the agents hears. This is also called the *inverse mapping problem*. Going from an articulatory representation to an acoustic signal is relatively straightforward: the agent just executes the articulatory moves and the sound is produced. Going from an acoustic signal to an articulatory representation is much more difficult.

The first problem that has to be tackled is which features of the speech signal to use for its analysis. Using the raw samples seems to be impractical. First of all the amount of data to be processed would be quite high. Secondly, there is lots of high frequency variation and noise in a speech signal, whereas the properties (noise bursts, silence, formant transitions) of the signal that humans use for recognition are of much lower frequency. Finally, human hearing mainly works by analysing the speech signal in its frequency spectrum. This spectrum is then analysed further by the brain. It therefore seems logical to extract from the speech signal those features of which it is known that they play a role in human recognition. The features that are extracted so far are: the strength of the signal, the prominence of the vocal cord vibration in the signal, (that is: the periodicity of the speech signal) the pitch of the vocal cord vibration if present and the properties of the first five formants of the

speech signal. The formant properties that are extracted are their centre frequency, their bandwidth and their strength. Details of the signal analysis procedures can be found in appendix F.

The problem of recognising acoustic signals is well known from speech processing. It will probably have to be solved by learning the possible acoustic realisations for every phoneme (or every pair of phonemes) either by storing all of them separately or by training a neural network (see e.g. Elman & Zipser 1988; Elman 1990) or a hidden Markov model (see e.g. Russel & Norvig 1995, pp. 762–767). As most speech sounds are not static through time, a learning method must be used that can cope with time series neural networks as well as hidden Markov Models can do this, but techniques for directly comparing time series also exist (Rosenstein & Cohen 1998). Solving this problem should be simpler than solving the problem of general human speech recognition for several reasons. In the population of agents there will be no variation between vocal tracts of speakers. There will be no difference between allegro and careful speech. Agents will always pronounce their words as carefully as possible (although initially they will try to learn the simplest possible forms). Also, as the agents themselves develop the vocabulary of the population, it can never become more difficult than what the agents can handle (but it is probably true that in order to make it sufficiently human-like, the agents will have to be able to handle rather difficult sounds). Finally the range of sounds that agents can produce will be much more limited than the range that humans can produce.

The problem of the inverse mapping is more complex for the agents than simple recognition. In the beginning, when the agents' repertoires are still empty, the only way in which they can reliably imitate the other agents is by finding an articulatory gesture that produces a sound that is a reasonably accurate reproduction of the sound they heard. In the vowel imitation game, agents could solve this problem by generating a large number of vowel signals and improving these by a hill-climbing heuristic. This was feasible, because vowels could be generated so quickly. The articulation of consonants, on the other hand, is so complex and time consuming that it is not feasible to use the hill-climbing heuristic. Therefore a direct mapping between acoustic signals and articulatory actions is needed in order to enable the agents to make an initial imitation and to improve the sounds they already know. The agents have to learn this mapping by talking to themselves in a process that is rather reminiscent of the babbling of babies. It is therefore suspected that babbling serves, at least in part, for learning the inverse mapping in children. This seems to be a rather neglected topic in child phonology acquisition. Vihman (1996) who is otherwise very concise does not say much on the topic of the acquisition of the mapping from acoustic signals to articulator actions. It seems as if this is assumed to be innate.

Unfortunately, it has not been possible to learn a satisfactory inverse mapping so far. Attempts to learn an inverse mapping are discussed in appendix F. The problem is made extra complex because of the large amount of input and output involved, the fact that there are many dependencies on time and articulatory context and the fact that, because of slight differences in timing, two time series taken from two signals that sound the same can still be quite different. Furthermore, the direct use of the articulatory parameters rather than more convenient tract constriction parameters makes the learning problem more difficult. The articulatory parameters do not always have a direct link with the acoustic signal. Different values of one articulatory parameter can cause the same acoustic signal and identical values can

cause different acoustic signals depending on the settings of other articulatory parameters. With the tract constriction parameters the mapping would be much more direct. Clearly further research is needed to solve the inverse mapping problem to an extent that enables agents to make initial imitations of sufficient quality.

The experiments with more realistic signals are in a sense a return to the over-optimistic first attempt at implementing the imitation game that was described in the first part of chapter 3. Although the simulation described here is based on a much profounder knowledge of speech synthesis, speech processing and the dynamics of the interactions between agents, a fully functional simulation that works with complex signals still remains a bit of a chimera. A lot of effort notwithstanding, no single realistic imitation game has still been played. Nevertheless, it is worth continuing in this direction. The complexity of the problem is enormous: it combines dynamical systems, articulatory synthesis, learning of time series and speech processing. However, building the simulation should be able to shed light on a number of fundamental questions on how speech is organised, learned and how it emerged.

7.4 Conclusion

In this chapter two different approaches to more implementing the imitation game with more complex speech signals were illustrated. The first approach was based on an extension of the vowel simulation by adding consonants and articulatory constraints. It was shown that in this way systems of coherent, phonemically coded CV-syllables could be formed. Unfortunately, the syllable inventories were not quite realistic and unlike the vowel imitation games, the success of the imitation games was rather sensitive to the settings of a large number of parameters. Still, the experiments show that even under two conflicting optimisation criteria, successful imitation can emerge.

The second model that was described is a new attempt towards a simulation that can work with real speech signals with all the complexity and flexibility of human speech. So far, no experiments have been performed with this system, but in trying to build it, a number of important lessons about the implementation and learning of speech sounds in the agents have been learnt. It appears that a description in terms of tract parameters is better than a description in terms of direct articulatory parameters. It also appears that one should fix on the learning of only phonemes or syllables. Rather, multiple emerging levels of complexity will be necessary. Finally, it appears that learning the mapping from acoustic signals to articulatory movements is an important, and in research into infant learning of speech rather neglected topic. Further research is necessary to see whether all these ideas work out.

The two main ways of extending the work that present themselves at the moment are the investigation of the simultaneous satisfaction of multiple conflicting criteria through self-organisation and the implementation of an agent that can produce, perceive and learn more complex signals. The experiments with the CV-syllables have shown that it is necessary to minimise articulatory effort and at the same time maximise acoustic distinctiveness. In order for the agents to converge to a realistic solution without being too sensitive to parameter settings, they must be able to "tune" these parameter settings themselves. In the experiment that was presented, fixed weights were assigned to articulatory effort and acoustic distinctiveness. In a more realistic system, the agents should be able to determine these weights dynamically, based on the language they are trying to learn. If the language

makes use of many subtle distinctions, it is clear that acoustic distinctiveness must be assigned more weight than the articulatory effort. If the language only makes use of fewer coarse distinctions, articulatory effort must be assigned more weight than acoustic distinctiveness. This would add an extra layer of complexity to the agents' behaviour with respect to their behaviour in the simulations so far. The self-tuning of parameters should probably be based on the non-verbal feedback that agents receive, as was already suggested at the end of the section on the vowel simulations without non-verbal feedback in chapter 5.

As for more complex utterances, these are a very interesting, but very difficult extension to the model. If the really interesting aspects of the emergence and evolution of sound systems of languages are to be investigated, it is absolutely essential that the model be extended to utterances that involve multiple moving articulators and sequences of sounds. The structure of human sound systems, as well as their development seems to depend on the interactions between the sounds as much as on the properties of the individual sounds. Also, the way in which sounds are combined into larger utterances is subject to rules and regularities that are reminiscent of a rudimentary form of syntax. It is likely that insights in the functioning of syntax can be learnt from a computer model that learns the phonology of a language.

There is a snag, of course. The implementation of a simulation for producing, perceiving and learning complex utterances is extremely difficult. It involves building an articulatory synthesiser, a system that makes the articulators move in a realistic way, a system that extracts the right features from the acoustic signal, a system that is able to estimate articulatory movements that are necessary to create a certain acoustic signal and, last but not least, a system that is able to learn the sounds it hears and to find patterns in them that can be used to store them more effectively, eventually resulting in the discovery of phonological and phonotactic rules.

In implementing this model, it is essential that it be as realistic as possible. If too many simplifications or arbitrary implementation decisions are made, the results of the model can not be compared with what is known about human languages anymore. This would make the whole exercise rather futile. Of course, certain simplifications will always have to be made, but they should be made very carefully, and should be taken into account when comparing the results of the model with observations of human languages.

The task ahead might seem daunting. However, a lot of work on implementing the model has already been done and a lot of work on human production, perception and learning of speech sounds already exists. This work can be used in building the model. Building and testing the model will hopefully result in valuable new insights in how humans produce, perceive and learn sounds. Running the model in imitation games will then result in new insights in how speech sounds, and therefore language, emerged and changed over time.

8. Conclusion

This thesis investigated whether the emergence of vowel systems can be explained as the result of interactions in a population of agents that learn and use vowel sounds. It was shown with computer simulations that this was indeed possible. Moreover, it was shown that the universal properties of human vowel systems can be predicted accurately from these simulations. The frequently found vowel systems can be considered as attractors of the dynamic system that is formed by the articulatory and perceptual constraints of the agents and by the rules of the imitation game. Self-organisation ensured that the resulting systems were coherent. It appears that these mechanisms must also be taken into account in the explanation of human vowel systems, and probably for other human speech sounds as well.

8.1 Summary

The vowel systems of human languages show remarkable cross-linguistic regularities (Hocket 1955; Crothers 1978; Schwartz *et al.* 1997a). As the total number of vowels that humans can possibly produce is quite impressive, (Ladefoged & Maddieson 1996, ch. 9) this cannot be due to chance. It has been assumed quite often that these regularities can be explained by innate distinctive features (especially since Jakobson & Halle 1956; Chomsky & Halle 1968) and by assuming that some of the features and some of their values are more *marked* than others. Marked features occur less often than unmarked ones, so the corresponding speech sounds will also be rarer. There are a number of problems with the theory of distinctive features, but the most important for predicting human vowel systems is that the markedness values are derived from observing the occurrence of certain speech sounds in human languages. These markedness values are then used to explain the very observations they are based on thus rendering the explanation circular and invalid.

More promising are independent functional explanations, based on such criteria as acoustic distinctiveness and articulatory ease (e.g. Liljencrants & Lindblom 1972; Stevens 1972; Carré 1994). The computer simulations by Liljencrants and Lindblom (1972) were the first to show that by maximising acoustic distance between the vowels in a vowel system, the most frequent vowel systems can be predicted quite well. Further improvements of this method (see e.g. Schwartz *et al.* 1997b) have confirmed these results. However, even though it turns out that human vowel systems are in general optimised for acoustic distinctiveness (at least for up to approximately nine vowels) it is not quite clear who is doing the optimisation. Certainly children, when learning a language, do not optimise their vowel systems. The answer must therefore be sought in the interactions in a population of language users over a longer period.

This fits in well with the work of Steels (1995, 1997b, 1998a, 1998b) and others (see e.g. the papers in Hurford *et al.* 1998, part III, and Kirby & Hurford, 1997; Kirby *to appear;* Hurford *to appear*) on modelling the origins of language using computer simulations of populations. This kind of work is part of the emerging field of artificial life (see e.g. Langton, 1989, Langton *et al.* 1990). In this work, language is seen as a complex system that is as much an emergent phenomenon of a population as it is individual knowledge of the speakers. In Steels' view, language must not be seen in terms of ideal speaker-hearer interactions or in terms of an idealised competence, but rather in terms of all the different, imperfect versions of the language that exist in the individual speakers' minds. Through repeated interactions between

speakers and through a pressure to remain understandable, coherence in the language will be maintained. Because speakers all have a different version of the language, that what is spoken in the population will be in constant change and variation. Change and variation are therefore seen as inherent properties of language, which are essential for understanding its dynamics and the way it originates. As the proposed models are so complex that all their implications cannot be predicted by hand, they have to be simulated on computers. Therefore the central focus in this thesis was on computer simulations.

The computer simulations were based on a population of agents. These agents are small, independent computer programs that can produce, perceive and learn vowels in a human-like way. The production was based on a simple articulatory vowel synthesiser that produces the first four formant frequencies of vowels whose articulation is specified by the three main vowel parameters. The perception was based on a weighted Euclidean distance in a two-dimensional space with the signals' first formant values as one dimension and their effective second formant values as the second dimension. This effective second formant is calculated with a formula that is based on work by Mantakas *et al.* (1986). The agents can store a number of vowel prototypes as well as the number of times these prototypes were used and the number of times they were successfully used. Initially the agents' vowel repertoires are empty. They develop a repertoire through interactions with other agents.

The interactions between the agents consisted of so-called imitation games. In an imitation game, two agents are picked at random from the population. One of these agents (the *initiator*) picks a random vowel from its repertoire, and produces its acoustic signal, which is deformed by a certain amount of noise. The other agent (the *imitator*) listens to this signal, finds the vowel in its repertoire that is closest to this signal and considers this vowel as the one that it recognised. It then produces the acoustic signal (with noise added) of the vowel (which might be quite different from the signal produced by the first agent). The first agent then analyses the signal in terms of its vowels, and checks whether the closest vowel it finds is the same as the one it originally produced. If this is the case, the imitation game is considered to be successful, if not it is a failure.

Depending on the outcome of the imitation game, the agents update their repertoire of vowels. Vowels with a success/use ratio that is too low are discarded. Whenever an imitation game was successful, the vowel prototype that was used by the imitator is shifted closer. When it was a failure, and the success/use ratio of the imitator's vowel was low, it is shifted closer to the perceived signal. If the success/use ratio was high, however, a new vowel prototype is added that is close to the perceived signal. Vowel prototypes that come too close together are merged. Also, in order to put a pressure on the agents to increase their number of vowels and in order to get the imitation games started, new, random vowel prototypes are added with low probability.

It is found that these imitation games, when iterated in the same population for a sufficiently large number of games, result in the emergence of realistic and successful vowel repertoires. The number of vowel prototypes per agent turns out to be equal for all agents in the population and the corresponding vowels of the agents are located in approximately the same place in acoustic as well as articulatory space. Moreover, the emerging vowel systems correspond to the vowel systems that are most frequently found in human languages. It can be shown with a number of quantitative measures (the energy of the vowel systems, the average success of imi-

tation and the average number of vowels) that the emerging vowel systems are much better than random, and quite realistic, because the vowels are distributed in a near-optimal way.

Of course, the exact shape of the emerging vowel systems and the number of vowel prototypes per agent depend on the settings of a number of parameters. It can be shown, however, that the simulation is relatively insensitive to the exact values of the parameters. There is a substantial range of parameter settings that result in realistic systems. Tentative independent evidence (Lindblom & Lubker 1985) could even be found for the relative weight of the first formant versus the effective second formant, one of the important parameters that determines the shape of the resulting vowel systems, that its value should be approximately 0.3 (which actually turned out to result in the most realistic vowel systems, see also Vallée 1994). The main parameter that was used to change the shape of the vowel systems is the acoustic noise. This parameter determines the size and number of the resulting clusters of vowel prototypes. It is not exactly clear yet what a plausible correlate in human languages for this noise parameter would be, but it is probably related to a desired level of accuracy of imitation.

By making further variations on the agents and the rules of the imitation game, it was shown that the outcome of the imitation game is not very sensitive to qualitative changes either. If another model of production and perception was used, the emerging vowel systems remained approximately the same. The alternative model was based on processing of real signals, rather than on processing of formant patterns only. The fact that it could work with real signals made it possible to try the imitation games for learning the vowel system of a human. This worked only partially, (some vowels turned out to be quite difficult to distinguish from each other) and because of the way in which humans interact with computers, the rules of the imitation games had to be changed.

Another variation of the imitation game that was investigated did not use the non-verbal feedback that agents need in order to find out whether the imitation game was successful or not. Instead of the non-verbal feedback, the agents used an internal distance measure for determining whether their imitations were close enough to the original signal. If the distance was larger than a certain threshold, the imitation was considered to be unsuccessful. This also resulted in the emergence of realistic vowel systems. However, the number of vowel prototypes in the emerging vowel systems was determined (within a narrow range) by the distance threshold. The agents could therefore not learn vowel systems with arbitrary numbers of vowels. In order to learn vowel systems with arbitrary numbers of vowel prototypes, agents must get at least some non-verbal feedback. How this relates with the way in which children acquire the sound system of their language (see e.g. Vihman 1996) is not quite clear, yet. Nevertheless, these results show that even when the rules of the imitation game are changed, realistic and successful vowel systems still emerge.

The most important variation on the basic imitation game that was tried out was to introduce change in the population of speakers. In real human populations, speakers of a language may die and infants without any knowledge of the language may be born without changing the language very much. This was simulated by allowing random insertion and removal of agents. It was found that if the flux of the population was not too great, vowel systems could be maintained and new vowel systems could emerge. It was even found that under certain circumstances, it was beneficial that the agents have an age structure. If the amount of articulatory effort

that could be invested in imitating other agents' sounds was limited, it was found that preservation of vowel systems was better if old agent changed their vowel repertoire slower than young agents. These experiments have also shown that the system is truly open in the way human languages are open: both with respect to what can be expressed (the number and position of the vowel prototypes) and with respect to the speakers of the language themselves.

The final test of the vowel system was a detailed comparison with the vowel systems that are found in human languages. It was found that the vowel systems that emerge from the simulations do not just predict the most frequent vowel systems in human languages very well, such as previous work (e.g. Liljencrants & Lindblom 1972; Schwartz *et al.* 1997a) already did, but that they also predicted the less frequently occurring types. Only for small numbers of vowels (three and four) and for large systems with more than two central vowels important discrepancies were found. This might have to do with the fact that high front vowels are not perceived very realistically in the models that were used, so that the agents seem to prefer more centralised high front vowels.

Finally, a new attempt was made to implement the imitation game for complex utterances. The first experiment that was conducted was with CV-syllables. This experiment was based on an optimisation experiment by Lindblom *et al.* (1984). It was shown that for the right parameter settings phonemically coded CV-syllables would emerge. However, this outcome was very sensitive to parameter settings and the constituent vowels and consonants were not very realistic. Also, the implementation of the syllables was not quite realistic. Furthermore, it would be interesting if the model could predict the occurrence of CV-syllables rather than to have to assume it as a given. The most important conclusion of these experiments was that it is possible to optimise two competing criteria, acoustic distinctiveness and articulatory ease, although this is harder than optimising only one criterion (acoustic distinctiveness) as in the case of the vowel systems. It was decided to build an agent that was able to produce the full complexity of human sounds, but unfortunately, linking perception and production turned out to be too hard, so that experiments could not be done.

8.2 Which Aims have been Achieved?

Most of the aims that were stated in the introduction of this thesis have been achieved. Aims a) and b) stated that it had to be shown that coherent vowel systems can emerge from scratch in a population of agents and that these systems have to be realistic. In chapter 4 it was shown that such systems do indeed appear, that they are better than random and that they are near optimal, if optimal is defined as being maximally dispersed in perceptual space. In chapter 6 the emerging vowel systems were compared with real human vowel systems and it was found that the similarities are striking. Therefore it can be concluded that coherent and realistic vowel systems can indeed emerge from scratch in a population of agents. Moreover, it appears that the interactions in the population are essential for the emergence of vowel systems. Without interactions between the agents, no vowel systems at all would emerge and their composition would be much less regular than what was found.

Aims c) and d) have also been achieved. It has been shown in chapters 4 that the vowel systems emerge for largely different parameter settings, although they are not always realistic for all parameter settings. In chapter 5 it was shown that for different qualitative variation of the composition of the agents and of the rules of the

imitation game, coherent and realistic vowel systems still emerge. It can therefore be concluded that the emergence of coherent and realistic vowel systems is not an idiosyncrasy of one particular implementation. It appears that it is a phenomenon that will happen in a large class of implementions, as long as there are similar (human-like) constraints on perception and production and as long as there is a pressure to imitate each other as well as possible. New systems can be derived from scratch if there is also a pressure to increase the number of vowel prototypes for each agent.

The only aim that was not achieved was aim e), which stated that complex utterances with consonants and sequences of sounds should be investigated. A small experiment with consonant-vowel syllables was implemented, but the results were not quite satisfactory. It also turned out to be too hard to build a simulation that could work with all possible human speech sounds, because it turned out to be too hard to find a good mapping between perceived sounds and articulations. Nevertheless, important parts of an agent that can potentially work with complex utterances have been build and are described in chapter 7 and appendix F.

The result of the first experiments in chapter 4 showed that the systems that emerged are *coherent*. The experiments with quantitative and qualitative variations of the simulation in chapter 4 and 5 showed that the resulting systems are *robust*. In chapter 5 the experiments with changing populations also showed that the system is *open* with respect to the language users. Finally, the comparison with natural vowel systems in chapter 6 showed that the emerging vowel systems are *natural*.

In short, all aims that have to do with vowel systems have been achieved. This has mainly to do with the fact that production and perception of vowels is relatively well understood. Steels' (1995, 1997b, 1998b) theory of language as a complex dynamic system in which self-organisation plays an important role is supported in the realm of vowel systems with the findings described in this thesis.

8.3 Implications of the Results

What are the implications of these results for the understanding of human intelligence and for the understanding of the sound systems of human languages? Most importantly, the results show that complex, recurrent structures in vowel systems do not have to be caused by innate dispositions towards these structures, nor that they have to be caused by explicit optimisation of functional constraints by individual agents. Rather, these structures can be considered as the result of self-organisation in a population. The perception, production and learning of the individual agents together with their interactions form a complex dynamic system. The dynamics of this system cause certain configurations of vowels to be stronger attractors than others are. The system will tend to evolve towards the strongest attractors most often. Because of random variations in the dynamic system (on production of the vowels, on the way that agents and vowel prototypes are chosen and on the insertion of new vowel prototypes) the vowel systems will not always settle down in only one attractor, but they will keep on changing. Also, because the system is so complex, it is quite likely that it gets stuck in a local optimum. For this reason, the vowel systems that are found are not always the ones in which the dispersion of the vowel prototypes is maximal. The success of the vowel prototypes of an agent is not just determined by their distribution through acoustic space, but also by the extent in which they resemble the vowel prototypes of the other agents in the population. There is no advantage to be gained from optimising an individual vowel system if this makes it impossible to imitate the other agents in the population. Lindblom and

Liljencrants (1972) have already remarked that the communicative success of vowel segments does not depend on their absolute position in acoustic space, but on their position relative to the other vowels in the system. This study has shown that their success also depends on the positions of the vowels of the other members of the population.

The implication for the study of sound systems of human languages is therefore that for the explanation of universal tendencies of sound systems it is not always necessary to look for innate structures or individual behaviours. The explanation can also be found in the interactions between the agents and in the history of the sound system. This makes it necessary to view language as a changing, social phenomenon rather than (or next to) viewing it as an abstract capacity of an individual. In order to understand language, its history is equally important as the capacities and the knowledge of the individual speakers. This view might help to bridge the (apparently rather artificial) distinction between diachronic and synchronic linguistics.

The implications for the field of artificial intelligence are less direct. The work presented in this thesis is an example of work in the subfield of artificial life (see e.g. Langton, 1989, Langton *et al.* 1990) where life-like phenomena that are too complex to understand directly are simulated with computer models in order to gain a better understanding. Artificial intelligence and artificial life in this thesis have been used as a methodology rather than as ends in themselves. No new learning methods or models of perception and reasoning have been developed. However, the work has shown that computer simulations can be used successfully to understand cognitive phenomena. Together with the work of others on different aspects of language (De Jong 1998; Steels 1995, 1997a, 1997c, 1998b; Steels & Kaplan 1998; Steels & Vogt, Kirby & Hurford, 1997; Hurford *to appear*; Kirby 1998, *to appear*) this work provides support for the theory that the origins of language are in a large part cultural rather than biological. Of course, this leaves room for the Baldwin effect (Baldwin 1896) to internalise the capacity for language and make it genetic, but this has not been investigated. This in itself has repercussions for the understanding of the origins of human *intelligence*, which seems to be related to language to a large extent. In any case it must be concluded that the results of the research presented here indicate that for understanding (and modelling) human intelligence the interactions between individuals *have to* be taken into account.

8.4 What Remains to be Done?

There are three ways in which the model described in this thesis can (and should) be extended. The first way is to get rid (as much as possible) of the non-verbal feedback between the agents. The agents should have a way to evaluate their own performance without the need for the explicit non-verbal feedback. The outcome of the experiment that does not use non-verbal feedback, described in chapter 5, was that it is possible to have the agents develop coherent vowel systems without non-verbal feedback. However, the number of vowel prototypes they develop depends on the acoustic distance threshold they use to determine the accuracy of their imitation. This accuracy should be language dependent, of course. The value for the threshold could be estimated from some initial non-verbal feedback, after which the agents develop their vowel systems further without it (or relying on it only when necessary). Conversely, it might be possible to make an estimate of the distance measure by passive observation of the sounds of the ambient language. Apparently infants seem

to learn to recognise the vowels of their ambient language before they start actively using them (Grieser & Kuhl 1989; Vihman 1996). Both these methods could be implemented and their repercussions for the emergence of vowel systems should be investigated.

The second extension of the method is to introduce extra constraints. So far, the only constraint that has been investigated into any detail is the acoustic distinctiveness. This seems to be sufficient for predicting vowel systems. It was shown in the experiments with CV-syllables that coping with two conflicting constraints at the same time is much harder than coping with a single constraint. It should be investigated whether this was just and idiosyncrasy of the CV-syllables experiment or whether it is really much harder to cope with conflicting constraints. This can probably already be done in the framework of the vowel systems experiment. It is a tendency of vowel systems in natural languages that once the number of vowels in the system exceeds a certain value (about eight or nine) secondary articulations such as length or nasalisation will be used. If systems have this secondary articulation, it is usually a subset of the ordinary vowels that has the extra articulation and usually this subset consists of at least three vowels (Vallée 1994, Schwartz *et al.* 1997a). This is probably related to articulatory as well as cognitive (learning and generalisation) constraints. The vowel simulation can be extended to incorporate one or more secondary articulations with an associated cognitive and articulatory cost. It can then be investigated under which circumstances natural systems emerge.

The third way in which the model can be extended is to make more complex utterances possible. These utterances will be extended over time and will contain consonants as well as vowels. This would automatically introduce multiple constraints in the system, because for producing the utterances actual articulatory movements (with their associated costs) will have to be made. For this, a realistic model of articulation (which has already been partially built) as well as a realistic model of human perception (parts of which have also already been built). The way in which the stream of speech sounds will be broken up into its constituent syllables and phonemes and the way in which the possible combinations into words are learnt will be similar to the way in which syntax in sentences is learnt in the models of Steels (1998a). With such a model universal tendencies of consonant systems, phonotactics and historical sound change can (in principle) be investigated. However, implementing such a model is still very difficult.

8.5 Some Idle Speculation

It is tempting to speculate a bit about what the findings reported in this thesis mean for other aspects of the study of language. Of course, the results in this thesis are only on vowel systems and for a small part on CV-syllables, so there is no substantial proof here that self-organisation also plays a role in other parts of language, or even in sounds that are more complicated than simple vowels. However, related research has shown that in the realm of lexicon formation and innovation (Steels 1995) the formation and sharing of semantic distinctions (Steels 1997a; Steels & Vogt 1997) and possibly in the formation of syntax (Steels 1998a; Kirby & Hurford, 1997; Hurford *to appear*; Kirby 1998, *to appear*) the mechanisms of self-organisation and cultural evolution also play an important role. In more mainstream linguistics, too, there seems to be a tendency to move from overtly abstract and innate explanations to more functional and less abstract explanations of linguistic phenomena.

The research presented here has shown that universal tendencies in vowel systems can be explained through the simultaneous action of functional constraints and the way the agents in the population interact. It is very likely that this plays a role in the formation of human vowel systems as well, as the similarity between the obtained vowel systems and the vowel systems that are found in actual human systems was extremely high. As functional constraints and interactions play a role in all parts of language, this mechanism will probably work in other parts of language as well. Take the example of the occurrence of both voiced and unvoiced consonants at the same place of articulation. This can be explained through the interaction of functional constraints and historical processes (basically interactions between young and old speakers). If a language does not make use of a voiced/voiceless contrast at a given place of articulation, it is likely that this becomes conditioned by the environment, so that voiced consonants will occur between vowels and voiceless consonants will occur utterance initially and finally. However, if the conditioning environment is lost, for example through the loss of syllables, the language will now have contrasting voiced and voiceless consonants. This process is much more likely to occur than the reverse process (merging of voiced and voiceless segments) so it is expected that more languages will make voiced/voiceless distinctions at the same place of articulation than not. Note that this explanation does not make any assumption about how people learn or store language cognitively. The only thing that is necessary is that they imitate the other speakers of the language under functional constraints. Of course, there will remain phenomena that will have to be explained by cognitive processes, but these can probably also be expressed in terms of functional constraints: minimisation of storage, minimisation of exceptions (resulting in analogical change) loss of frequently used items etc.

As a final speculation, it can be remarked that the learning of phonology is in a way quite similar to the learning of syntax and morphology. A language user has to learn which sounds there are, but also in which combinations they can be used and therefore which classes of sounds there are. The language user also has to learn in which way sounds influence each other. This influence can be of neighbouring sounds on each other, but it can also work over longer stretches of sounds (nasal harmony or vowel harmony, for example). For learning syntax, similar distributions and influences have to be learnt. It is therefore quite possible that the mechanisms that are responsible for learning the sound system of a language are similar to the mechanisms that learn syntax. The study of the emergence of sound systems might therefore also shed light on the study of syntax.

8.6 Finally

With these idle speculations this thesis has come to an end. Its substantial contribution has been to show that universal tendencies of vowel systems can be explained through functional constraints and (self-organising) interactions in a population of agents. The learning mechanism of the agents is simple rote learning of vowel prototypes, where the necessary prototypes are determined through their success in imitation. In any case, no innate predisposition towards certain structures (features, markedness constraints, or rules) turns out to be necessary. Although this result is modest, the similarities between the emerged vowel systems and the vowel systems that are actually found in human languages are such that the method must be considered very promising. It seems that interactions in the population, both between agents that already know the sound system and between agents that are

learning the sound system are factors that have to be taken into account in explaining human sound systems and their origins. The research presented here is just the beginning. There are many ways in which it can be extended. It might be a very exciting new way of exploring the complexities of human speech sounds and language.

References

Allen, Jonathan, M. Sharon Hunnicutt & Dennis Klatt (1987) *From text to speech: The MITalk system*, Cambridge: Cambridge University Press.

Baldwin, J. Mark (1896) A new Factor in Evolution, *The American Naturalist* 30 (June 1896) pp. 441–451, 536–553. Reprinted in R.K. Belew and M. Mitchell (eds.) *Adaptive Individuals in Evolving Populations: Models and Algorithms*, SFI Studies in the Sciences of Complexity, Proc. Vol. XXVI, Addison Wesley, Reading, MA, 1996.

Batali, John (1998) Computational simulations of the emergence of grammar. In Hurford *et al.* 1998, pp. 405–426.

Berrah, Ahmed Réda (1998) *Évolution Artificielle d'une Société d'Agents de Parole: Un Modèle pour l'Émergence du Code Phonétique*, Thèse de l'Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives.

Berrah, Ahmed-Réda, Hervé Glotin, Rafael Laboissière, Pierre Bessière & Louis-Jean Boë (1996) From Form to Formation of Phonetic Structures: An evolutionary computing perspective. In Terry Fogarty & Gilles Venturini (eds.) *ICML '96 workshop on Evolutionary Computing and Machine Learning*, Bari 1996, pp. 23–29.

Boë, Louis-Jean, Jean-Luc Schwartz & Nathalie Vallée (1995), The Prediction of Vowel systems: perceptual Contrast and Stability. In Eric Keller (ed.) *Fundamentals of Speech Synthesis and Speech Recognition*, John Wiley, pp. 185–213

Boersma, Paul (1998) *Functional Phonology* The Hague: Holland Academic Graphics.

Browman, Catherine P. & Louis Goldstein (1995) Dynamics and Articulatory Phonology. In Robert F. Port & Timothy van Gelder (eds.) *Mind as Motion*, MIT Press, Cambridge Mass. pp. 175–194.

Carlson, R, B. Granström & G. Fant (1970) Some studies concerning perception of isolated vowels, *STL/QPSR* (Speech Transmission Laboratory Quarterly Progress and Status Report, Department of Speech Communication and Music Acoustics, KTH, Stockholm) 2/3 pp. 19–35.

Carré, René (1994) 'Speaker' and 'Speech' Characteristics: A Deductive Approach. *Phonetica* 51, pp. 7–16

Carré, René (1996) Prediction of vowel systems using a deductive approach. In *Proceedings of the ICSLP 96*, Philadelphia, pp. 434–437.

Carré, René & Maria Mody (1997), Prediction of Vowel and Consonant Place of Articulation. In John Coleman (ed.) *Computational Phonology, Proceedings of a Workshop Sponsored by the Association for Computational Lingistics*, Association of Computational Linguistics, pp. 26–32.

Carré, René & Mohamad Mrayati (1995), Vowel transitions, vowel systems, and the Distinctive Region Model. In C. Sorin *et al.* (eds.) *Levels in Speech Communication: Relations and Interactions*, Elevier pp. 73–89.

Choi, John D. (1991) Kabardian Vowels Revisited. *Journal of the International Phonetic Association* 21 pp. 4–12.

Chomsky, Noam (1965) *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Mass.

Chomsky, Noam (1972) *Language and Mind, enlarged edition*, New York: Harcourt Brace Jovanovich, Inc.

Chomsky, Noam (1975) *Reflections on Language*, New York: Pantheon Books.

Chomsky, Noam (1980) Rules and representations. *The behavioral and brain sciences* 3, pp. 1–21.

Chomsky, Noam & Morris Halle (1968) *The sound pattern of English*, MIT Press, Cambridge, Mass.

Comrie, Bernard (1981) *Language typology and linguistic universals*, Oxford: Blackwell

Cooper, Franklin S, Pierre C. Delattre, Alvin M. Liberman, John M. Borst & Louis J. Gerstman (1976), Some Experiments on the Perception of Synthetic Speech Sounds. In D.B Fry (ed.) *Acoustic Phonetics*, Cambridge University Press, pp. 258–283.

Crothers, John (1978) Typology and Universals of Vowel systems. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.) *Universals of Human Language*, Volume 2 Phonology, Stanford: Stanford University Press pp. 93–152.

Darwin, Charles (1859) *On the origin of species*, reprinted: Penguin Classics, 1985.

Dawkins, Richard (1976) *The selfish gene*. Oxford: Oxford university press

De Jong, Edwin D. (1998) *The development of a lexicon based on behavior*. In Han La Poutré & Jaap van den Herik (eds.) Proceedings Xth Netherlands/Belgium Conference on Artificial Intelligence, Amsterdam, 18–19 November 1998, pp. 27–36.

De Saussure, Ferdinand (1987*) Cours de linguistique générale, édition préparée par Tullio de Mauro*, Paris: Payot.

Dunbar, Robin (1996) *Grooming, gossip and the evolution of language*, London: Faber and Faber.

Elman, Jeffrey L. (1990) Finding Structure in Time. *Cognitive Science* 14 pp. 179–211.

Elman, Jeffrey L. & David Zipser (1988) Learning the hidden structure of speech. *Journal of the Acoustical Society of America* 83(4) pp. 1615–1626.

Everett, Daniel. L. (1982) Phonetic rarities in Piraha. *Journal of the International Phonetic Association* 12/2 pp. 94–96.

Fant, Gunnar (1973) *Speech sounds and Features*, Cambridge (MS): MIT Press.

Firchow, Iwin & Jacqueline Firchow (1969) An abbreviated phoneme inventory. *Anthropological Linguistics* 11, pp. 271–276.

Gasser, Michael (1998) (ed.) *The Grounding of Word Meaning: Data and Models, Papers from the 1998 Workshop*, Menlo Park (CA): AAAI Press.

Glotin, Hervé (1995) *La Vie Artificielle d'une société de robots parlants: émergence et changement du code phonétique.* DEA sciences cognitives-Institut National Polytechnique de Grenoble.

Glotin, Hervé & Rafael Laboissière (1996) Emergence du code phonétique dans une societe de robots parlants. *Actes de la Conférence de Rochebrune 1996 : du Collectif au social*, Ecole Nationale Supérieure des Télécommunications – Paris.

Goldberg, David E. (1998) *Genetic algorithms in search, optimization, and machine learning*, Reading (Mass.): Addison-Wesley.

Grieser, DiAnne & Patricia K. Kuhl (1989) Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology* 25, pp. 577–588.

Grimes, Barbara F. (ed.)(1996) *Ethnologue: Languages of the World, 13th edition*, SIL.

Handel, Stephen (1989) *Listening*, An introduction to the Perception of Auditory Events, Cambridge (MS) "A Bradford book" MIT Press.

Hashimoto, M. J. (1973) *The Hakka Dialect; A Linguistic Study of Phonology, Syntax and Lexicon*. Cambridge: Cambridge University Press.

Hasselbrink, Gustav (1965) *Alternative analyses of the phonemic system in Central South-Lappish*, The Hague: Bloomington.

Hauser, Marc D. (1997) *The evolution of communication*, Cambridge (MS) MIT Press.

Haykin, Simon (1994) *Neural Networks: A Comprehensive Foundation*, New York: MacMillan.

Hertz, John, Anders Krogh, & Richard G. Palmer (1991), *Introduction to the Theory of Neural Computation*, Addison Wesley.

Hockett, Charles F. (1955) *A manual of phonology*, Baltimore: Waverley Press.

Hopper, Paul J. & Elizabeth Closs Traugot (1993) *Grammaticalization*, Cambridge: Cambridge University Press.

Hurford, James R. (*to appear*) Social transmission favours linguistic generalization. In Chris Knight & Jim Hurford (eds.) *The evolution of language*, Cambridge: Cambridge University Press.

Hurford, James R,Michael Studdert-Kennedy & Chris Knight (eds.) (1998) *Approaches to the Evolution of Language* (selected papers from the 2nd International Conference on the evolution of Language, London, April 6–9 1998), Cambridge: Cambridge University Press

Ifeachor, Emmanuel C. & Barrie W. Jervis (1993) *Digital Signal Processing, A practical approach*, Workingham: Addison Wesley.

Jakobson, Roman & Morris Halle (1956) *Fundamentals of Language*, the Hague: Mouton & Co.

Jespersen, Otto (1968) *Language, its nature, development and origin*, London: Allen and Unwin.

Johnson, Mark H. (1997) *Developmental Cognitive Neuroscience*, Oxford: Blackwell.

Kaburagi, Tokihiko & Masaaki Honda (1996) A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes. *Journal of the Acoustical Society of America* 99 (5) pp. 3154–3170.

Kandel, Eric R, James H. Schwartz & Thomas M. Jessell, (1991) *Principles of Neuroscience, 3rd edition.* London: Prentice-Hall.

Kaplan, Frédéric (1998) Role de la simulation multi-agent pour comprendre l'origine et l'evolution du langage. In J-P Barthes, V. Chevrier & C. Brassac (eds.) *Systemes multi-agents: de l'interaction à la socialité* (JFIADSMA 1998), Hermes, Paris.

Kaplan, Frédéric, Luc Steels & Angus McIntyre (1998) An architecture for evolving robust shared communication systems in noisy environments in: *Proceedings of Sony Research Forum 98*, Tokyo.

Kegl, Judy (1994) The Nicaraguan sign Language Project: an overview. *Signpost* 7, volume 1, pp. 24–31.

Kegl, Judy & Iwata, Gayla A. (1989) Lenguaje de signos Nicaragüense: A pidgin sheds light on the "creole?" ASL. In R. Carlson, S. DeLancey, S. Gildea, D. Payne, A. Saxena (eds.) *Proceedings of the Fourth Meetings of the Pacific Linguistic Conference*, Eugen, Oregon: Department of Linguistics, University of Oregon. pp. 266–294.

Kelso, J. A. S, E.L. Saltzman & B. Tuller (1986) The dynamical perspective on speech production: data and theory. *Journal of Phonetics* 14, pp. 29–59.

Kirby, Simon (1998) Fitness and the selective adaptation of language. In Hurford *et al.* pp. 359–383

Kirby, Simon (*to appear*) Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In Chris knight & Jim Hurford (eds.) *The evolution of language* (selected papers from the 2nd International Conference on the evolution of Language, London, April 6–9 1998). Cambridge: Cambridge University Press.

Kirby, Simon & James R. Hurford (1997) Learning, Culture and Evolution in the Origin of Linguistic Constraints. In Phil Husbands & Inman Harvey (eds.) *Fourth European Conference on Artificial Life*, Cambridge (MS): MIT Press, pp. 493–502

Ladefoged, Peter (1981), *Preliminaries to Linguistic Phonetics*, Midway Reprint, The University of Chicago Press.

Ladefoged, Peter & Ian Maddieson (1996) *The Sounds of the World's Languages*, Oxford: Blackwell.

Lakoff, G. (1987) *Women, fire, and dangerous things: what categories reveal about the mind.* Chicago: Chicago University Press.

Langton, Christopher G. (ed.) (1989) *Artificial Life*, Addison Wesley.

Langton, Christopher G, Charles Taylor, J. Doyne Farmer & Steen Rasmussen (eds.)(1990) *Artificial Life II*, Addison Wesley.

Liberman, Alvin M, Pierre C. Delattre, Franklin S. Cooper & Louis J. Gerstman (1976) The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants. In D.B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press.

Liljencrants, L. & Björn Lindblom (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast. *Language* 48 pp. 839–862.

Lindblom, Björn (1972) Phonetics and the Description of Language. In André Rigault & René Charbonneau (eds.) *Proceedings of the seventh international congress on phonetic sciences*, The Hague: Mouton pp. 63–93.

Lindblom, Björn (1986) Phonetic universals in vowel systems. In: Ohala, J. J. & Jaeger, J. J. (eds.) *Experimental Phonology*, Orlando (FL): Academic Press, pp. 13–44.

Lindblom, Björn (*in press*) From second thoughts to first principles. In Pierrehumbert, J. & Broe, M. (eds.) *Papers from the fifth Laboratory Phonology Conference*, Cambridge University Press.

Lindblom, Björn & Olle Engstrand (1989) In what sense is speech quantal? In: *Journal of Phonetics* 17, pp. 107–121.

Lindblom, Björn & James Lubker (1985) The Speech Homunculus and a Problem of Phonetic Linguistics. In Vicoria A. Fromkin (ed.) *Phonetic Linguistics: essays in honor of Peter Ladefoged*, Orlando: Academic Press pp. 169–192.

Lindblom, Björn, Peter MacNeilage & Michael Studdert-Kennedy (1984) Self-organizing processes and the explanation of language universals. In Brian Butterworth, Bernard Comrie & Östen Dahl (eds.) *Explanations for language universals*, Walter de Gruyter & Co. pp. 181–203.

Lindblom, Björn & Ian Maddieson (1988), Phonetic Universals in Consonant Systems. In Larry M. Hyman & Charles N. Li (eds.) *Language, Speech and Mind*, pp. 62–78.

MacNeilage, Peter (*to appear*) The Frame/Content Theory of Evolution of Speech Production, *Behavioral and Brain Sciences*.

Maddieson, Ian (1984) *Patterns of sounds*, Cambridge University Press.

Maddieson, Ian & Kristin Precoda (1990) Updating UPSID. In *UCLA Working Papers in Phonetics* 74, pp. 104–111.

Maeda, Shinji (1989) Compensatrory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal Tract Shapes using an Articulatory Model. In Hardcastle, W.J. & Marchal, A. (eds.) *Speech Production and Speech Modelling*, Kluwer, pp. 131–149.

Mantakas, M, J.L. Schwartz & P. Escudier (1986) *Modèle de prédiction du 'deuxiéme formant effectif' F_2'—application à l'étude de la labialité des voyelles avant du français*. In Proceesings of the 15th journées d'étude sur la parole. Société Française d'Acoustique, pp. 157–161.

Mermelstein, P. (1973) Articulatory model for the study of speech production, *The Journal of the Acoustical Society of America*, 53(4) pp. 1070–1082.

Oppenheim, Alan V, Alan S. Willsky & Ian T. Young (1983) *Signals and Systems*, London, Prentice-Hall international.

Peterson, Gordon. E. & Harold L. Barney (1952) Control Methods Used in a Study of the Vowels, *The Journal of the Acoustical Society of America*, 24(2) pp. 175–184.

Pinker, Steven (1995) *The language instinct*, Penguin

Pinker, Steven & Bloom, P. (1990) Natural Language and Natural Selection. The *Behavioral and Brain Sciences* 13, pp. 707–784.

Plaut, David C. & Christopher T. Kello (*to appear*) The Emergence of Phonology from the Interplay of Speech Comprehension and Production: A Distributed Connectionist Approach. In B. MacWhinney (ed.) *The emergence of language*, Mahweh (NJ): Erlbaum.

Press, William H, Saul A. Teukolsky, William T. Vetterling & Brian P. Flannery (1992), *Numerical Recipies in C: The art of Scientific Computing* second edition, Cambridge University Press.

Pullum, Geoffrey K. & William A. Ladusaw (1996) *Phonetic Symbol Guide, second edition*, Chicago: The Chicago University Press.

Rabiner, Lawrence R. & Ronald W. Schafer (1978) Digital Processing of Speech Signals, Englewood Cliffs (NJ): Prentice-Hall, Inc.

Redford, Melissa Annette, Chun Chi Chen & Risto Miikkulainen (1998) Modeling the Emergence of Syllable Systems. In Morton Ann Gernsbacher & Sharon J. Derry (eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (COGSCI-98) pp. 882–886.

Redford, Melissa Annette, Chun Chi Chen & Risto Miikkulainen (*to appear*) Constrained Evolution of Syllable Systems. In Chris Knight & J. R. Hurford (eds.) *The Evolution of Language* (selected papers from the 2nd International Conference on the evolution of Language, London, April 6–9 1998), Cambridge: Cambridge University Press.

Rober-Ribes, J. (1995) *Modèles d'intégration audiovisuelle de signaux linguistiques*. Thèse de docteur de l'Institut National Polytechnique de Grenoble.

Rosenstein, Michael T. & Paul R. Cohen (1998) Concepts From Time Series. In *Proceedings of AAAI-98*, Fifteenth National Conference on Artificial Intelligence, Madison, Wisconsin July 26–30, 1998, Menlo Park (CA) AAAI Press/MIT Press pp. 739–745.

Rousseau, Jean-Jacques (1986) Essay on the origin of languages which treats of melody and musical imitation. In John H. Moran & Alexander Gode *On the origin of language*, pp. 1–74.

Rubin, Philip, Thomas Baer and Paul Mermelstein (1981) An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America* 70(2) pp. 321–328.

Russel, Stuart J. & Peter Norvig (1995) *Artificial Intelligence A Modern Approach*, Englewood Cliffs (NJ): Prentice-Hall.

Saltzman, Elliot L. (1986) Task dynamic coordination of the speech articulators. *Experimental brain research series* 15 pp. 129–144.

Saltzman, Elliot L. (1995) Dynamics and Coordinate Systems in Skilled Sensorimotor Activity. In Port, Robert F. & van Gelder, Timothy (eds.) *Mind as Motion*, MIT Press, Cambridge Mass. pp. 149–174.

Saltzman, Elliot L. & Kevin G. Munhall (1989) A Dynamical Approach to Patterning in Speech Production. *Ecological Psychology* 1(4) pp. 333–382.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry (1997a), Major trends in vowel system inventories. *Journal of Phonetics* 25, pp. 233–253

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry (1997b), The Dispersion-Focalization Theory of vowel systems. *Journal of Phonetics* 25, pp. 255-286.

Sedlak, P. (1969) Typological considerations of vowel quality systems. *Working Papers on Language Universals* 1, Stanford University pp. 1–40.

Seiden, W. (1960) Chamorro Phonemes. *Anthropological Linguistics* 2(4) pp. 6–33.

Senghas, Ann (1994) Nicaragua's Lessons for Language Acquisition. *Signpost* 7, volume 1, pp. 32–39.

Senghas, Richard J. & Judy Kegl (1994) Social considerations in the emergence of Idioma de Signos Nicaragüense (Nicaraguan Sign Language). *Signpost* 7, volume 1, pp. 40–46.

Sheldon, S. N. (1974) Some morphophonemic and tone rules in Mura-Pirahã. *International Journal of American Linguistics* 40 pp. 279–82.

Snyman, J. W. (1970) *An introduction to the !Xũ (!Kung) language*, Cape Town: Balkema.

Steels, Luc (1992) *PDL Reference manual*, Vrije Universiteit Brussel AI-memo 92-05.

Steels, Luc (1995) A Self-Organizing Spatial Vocabulary. *Artificial Life* 2(3), pp. 319–332.

Steels, Luc (1996) The Spontaneous Self-organization of an Adaptive Language. In S. Muggleton (ed.) *Machine Intelligence* 15.

Steels, Luc (1997a) Constructing and Sharing Perceptual Distinctions. In Maarten van Someren and G. Widmer (eds.) *Proceedings of the ECML*, Berlin: Springer Verlag.

Steels, Luc (1997b) The Synthetic Modelling of Language Origins, *Evolution of Communication* 1(1): pp. 1–34.

Steels, Luc (1997c) The Origins of Syntax in Visually Grounded Robotic Agents. In: Pollack, M. (ed.) *Proceedings of the IJCAI-97 Conference*. The AAAI Press, Morgan Kaufmann Pub. Los Angeles.

Steels, Luc (1998a) The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103(1–2) pp. 133–156.

Steels, Luc (1998b) Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation. In James R. Hurford, Michael Studdert-Kennedy & Chris Knight (eds.) *Approaches to the Evolution of Language*, Cambridge: Cambridge University Press pp. 384–404.

Steels, Luc & Frédéric Kaplan (1998) Spontaneous Lexicon Change. In: *Proceedings of COLING-ACL 1998, Montreal*, pp.1243–1249.

Steels, Luc & Paul Vogt (1997) Grounding adaptive language games in robotic agents. In Husbands, Phil & Harvey, Inman (eds.) *Proceedings of the Fourth European Conference on Artificial Life*, Cambridge (MS): MIT Press, pp. 474–482.

Stevens, Kenneth. N. (1972). The Quantal Nature of Speech: Evidence from articulatory-acoustic data. In: E. E. David, Jr. & P. B. Denes (Eds.) *Human communication: a unified view*. New York: McGraw-Hill pp. 51–66.

Stevens, Kenneth N. (1989) On the quantal nature of speech, *Journal of Phonetics* 17, 1, pp. 3–45

Stevens, Kenneth N. & Sheila E. Blumstein (1975) Quantal aspects of consonant production and perception, Journal of Phonetics 3, pp. 215–233.

Stiles, Joan & Donna Thal (1993) Linguistic and Spatial Cognitive Development Following Early Focal Brain Injury: Patterns of Deficit and Recovery. In Johnson, Mark Henry (ed.*) Brain development and cognition: a reader*, Oxford: Basil Blackwell, pp. 643–664.

Suzuki, Junji & Kunihiko Kaneko (1994) Imitation Games. *Physica D* 75 pp. 328–342.

Trubetzkoy, N. S. (1929) Zur allgemeinen Theorie der phonologischen Vokalsysteme. *Travaux du cercle linguistique de Prague* 7, pp. 39–67.

Turing, Alan M. (1950) Computing Machinery and Intelligence. In E. A. Feigenbaum (ed.) *Computers and Thought*, McGraw-Hill.

Vallée, Nathalie (1994) *Systèmes vocaliques: de la typologie aux prédictions*, Thèse préparée au sein de l'Institut de la Communication Parlée (Grenoble-URA C.N.R.S. no 368).

Vanvik, A (1972) A phonetic-phonemic analysis of Standard Eastern Norwegian. In *Norwegian Journal of Linguistics* 26, pp. 119–164.

Vennemann, Theo (1988) *Preference Laws for Syllable Structure*, Berlin: Mouton de Gruyter.

Vihman, M. (ed.) (1976) *A reference manual and user's guide for the Stanford Phonology Archive*, part I, Stanford University.

Vihman, Marilyn May (1996) *Phonological Development, The origins of language in the child*, Blackwell.

Wittgenstein, Ludwig (1967) *Philosophische Untersuchungen*, Frankfurt: Suhrkamp.

Appendix A: Symbols

This appendix provides a list with symbols, with a short definition of the symbol and the default value if they are a parameter of one of the simulations in this thesis. The symbols are ordered in alphabetical order, where Greek symbols are inserted in the place where they would appear in the Greek alphabet. The letter ψ, for example appears at the end of the list.

A Abstract symbol for an acoustic signal in an imitation game.

ac_v Acoustic representation of a vowel v.

ar_v Articulatory representation of a vowel v.

α_{aging} The speed with which the step size ε of ageing agents changes. Parameter of the changing population simulation. Should be between zero and one.

c Critical distance in the calculation of the effective second formant. Always set to 3.5 Bark.

D Distance between two acoustic signals.

D_θ Acoustic distance between a perceived signal and its imitation above which the imitation is considered to be a failure. Parameter of the vowel imitation simulation without non-verbal feedback.

E The energy of a vowel system.

ε Step size with which articulatory vowel prototypes are shifted to make them sound more like observed signals. Parameter of the vowel simulation. Normally 0.03.

F_1, F_2, F_3, F_4: The first four formant frequencies. Peaks in the frequency spectrum of a vowel.

$\not{F}_1$, $\not{F}_2$, $\not{F}_3$, $\not{F}_4$: The first four formant frequencies shifted to model noise.

F_2' Effective second formant. Weighted sum of the second, third and fourth actual formant frequencies. Corresponds to the frequency of the second formant in a signal with two formant peaks that best approximates a vowel to human subjects.

f Any frequency.

f_{rate} Sampling frequency of simulations that work with realistic signals. Set to $\frac{1}{4}^{th}$ of CD-frequency, i.e. 11025 Hz.

G_p The excess of birth over death or growth of the population. Measure of simulations with a changing population.

h Articulatory height of a vowel. In the vowel simulations its value will be in the range [0,1]. Zero means lowest (most open) and one means highest.

θ_c The threshold below which vowels are considered not to be successful enough and will be removed from an agent's vowel repertoire during cleanup. Parameter of the vowel simulations. Normally 0.7

θ_s The threshold above which vowels are considered successful enough so that new phonemes will have to be added in case of a failure of the imitation game. Parameter of the vowel simulations. Normally 0.5

θ_u The minimal number of times a phoneme will have to be used before it can be removed in a cleanup operation in the vowel simulations. Normally 5.

λ The weighting factor of the effective second formant frequency relative to the first formant frequency. Parameter of the vowel simulation. Normally 0.3

N The number of agents in the population. Normally 20.

N_{games} The number of games that has been played.

n The number of vowels in a vowel system.

p Articulatory position in the front-back dimension of a vowel. In the vowel simulations its value will be in the range [0,1]. Zero means most front, while one means most back.

p_b Probability of a new agent being added to the population in an imitation game. Parameter of the vowel simulation with changing population. See also p_d.

p_c The probability of cleaning up a vowel system. Parameter of the vowel simulation. Normally 0.1

p_d Probability per imitation game of removing the imitator and/or the initiator from the population. Should be half of p_b for a stable population size.

p_i The probability of inserting a random new vowel into an agent's vowel system. Parameter of the vowel simulation. Normally 0.01

r Articulatory rounding of a vowel. In the vowel simulations its value will be in the range [0,1]. Zero means least rounded, one means most rounded.

r_{ij} Distance between two vowels in a vowel system.

s_t Sample of a real signal at time t.

s_v Number of times an item has been successfully used in an imitation game.

t Any point in time.

τ Half-life time of a population of agents. The number of imitation games it takes before half of the agents in the original population have died out.

u_v Number of times an item has been used in an imitation game.

v Any vowel.

V The set of vowel prototypes of an agent.

ψ_{ac} Maximal percentage by which formant frequencies are shifted in order to model acoustic noise in the vowel simulations. No default value.

ψ_{art} Maximal percentage by which articulatory parameters are shifted in order to model inaccuracy of articulation. Parameter of the vowel simulation. Normally set to zero.

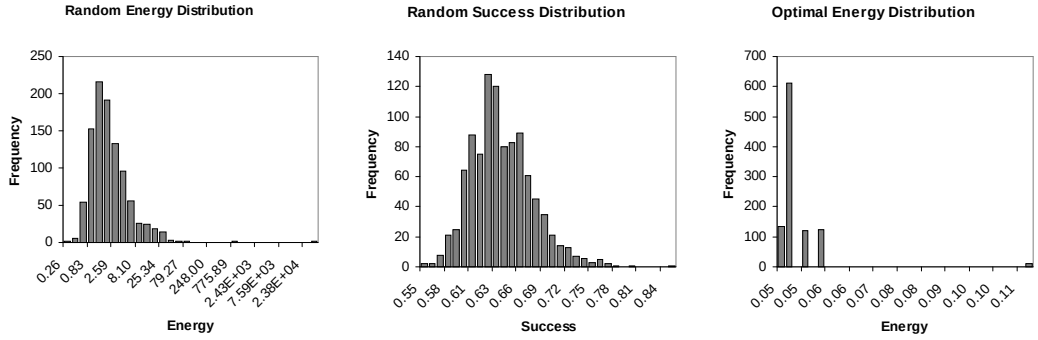
w_1, w_2 Weights by which the formant frequencies are weighted in the calculation of the effective second formant frequency in the vowel simulations.

Appendix B: Random and Optimal Vowel systems

In this appendix the energy and success of random systems as well as the energy of optimal systems with a given number of vowels from two to ten are given. The energy of both random and optimal systems was calculated using Liljencrants and Lindblom's (1972) formula (equation 2.1 in chapter 2). The success value of the random systems was calculated by randomly initialising a population of size twenty, and then using each vowel of each agent in an imitation game with each other agent. The success was then calculated as the ratio between the total number of successful games and the total number of games.

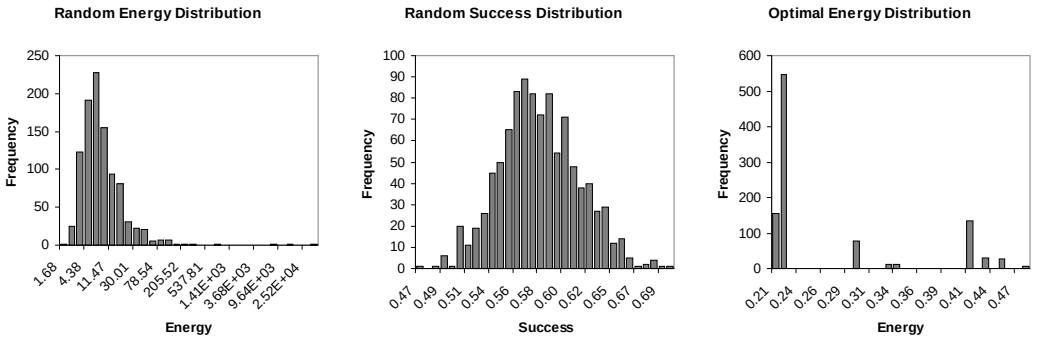
Optimal systems were obtained by optimising the energy function for vowel systems with a given number of random vowels in the way described in the analysis section of chapter 4. Because of the random initialisation and the presence of local minima, this optimisation procedure usually resulted in a number of different near-optimal systems, so that there are a number of peaks in the energy distribution.

B.1 Two Vowels



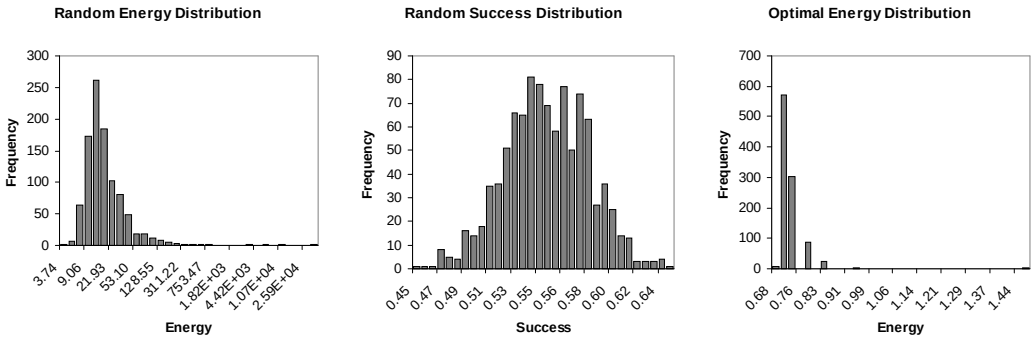
The random energy has mean 38.22, standard deviation 1099 and median 1.40. The random success has mean 0.637, standard deviation 0.038 and median 0.632. The optimal energy has mean 0.0485, standard deviation 0.0071 and median 0.0462.

B.2 Three Vowels



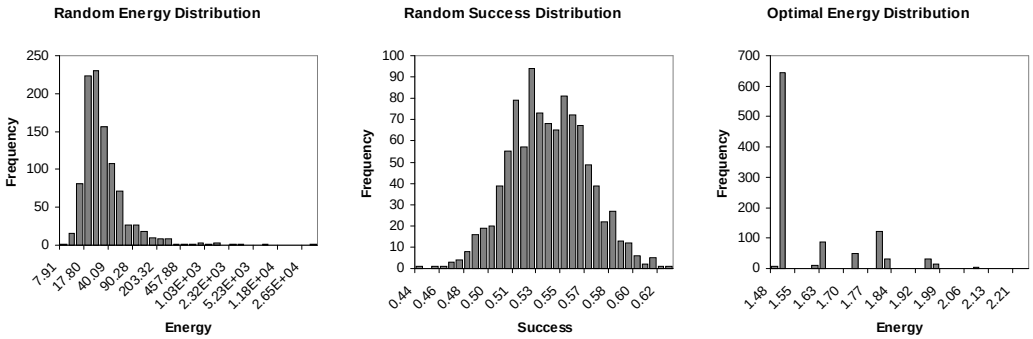
The random energy has mean 60.15, standard deviation 1154 and median 5.47. The random success has mean 0.575, standard deviation 0.036 and median 0.574. The optimal energy has mean 0.262, standard deviation 0.083 and median 0.212.

B.3 Four Vowels



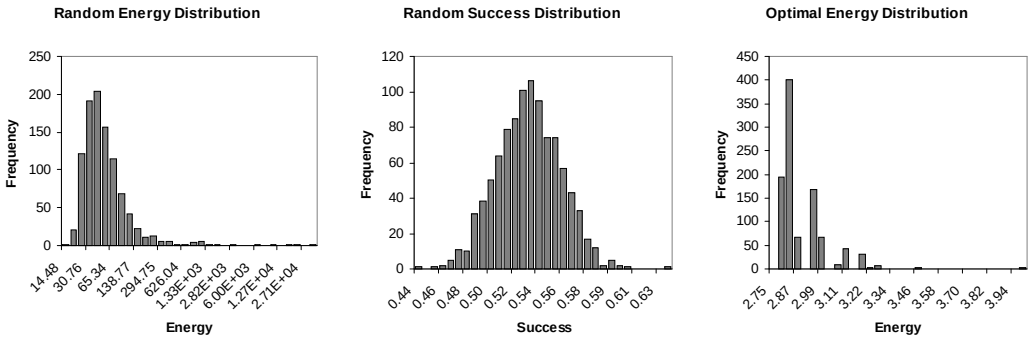
The random energy has mean 74.26, standard deviation 1158 and median 12.08. The random success has mean 0.549, standard deviation 0.031 and median 0.549. The optimal energy has mean 0.706, standard deviation 0.058 and median 0.680.

B.4 Five Vowels



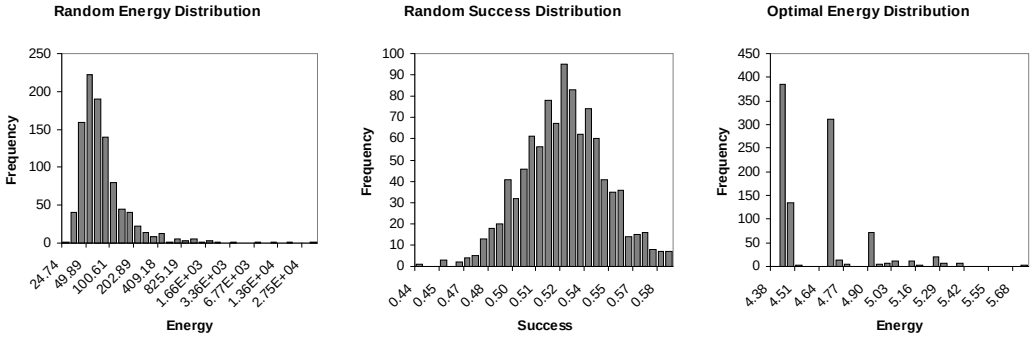
The random energy has mean 85.35, standard deviation 1121 and median 21.79. The random success has mean 0.535, standard deviation 0.028 and median 0.534. The optimal energy has mean 1.576, standard deviation 0.146 and median 1.482.

B.5 Six Vowels



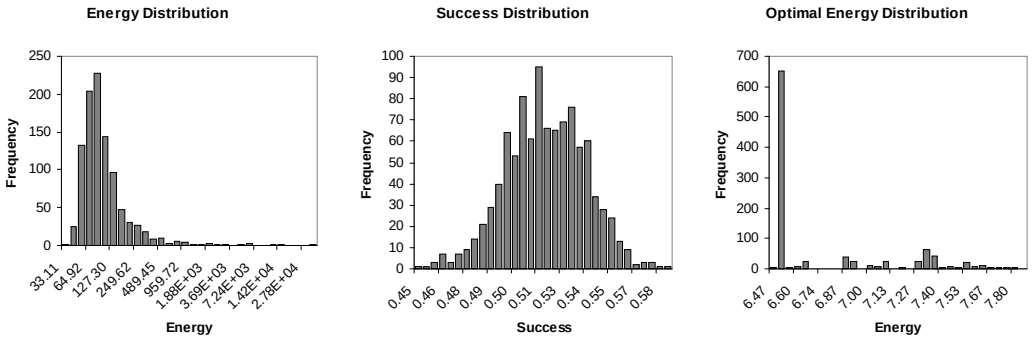
The random energy has mean 147.5, standard deviation 1404 and median 37.81. The random success has mean 0.526, standard deviation 0.026 and median 0.525. The optimal energy has mean 2.87, standard deviation 0.13 and median 2.80.

B.6 Seven Vowels



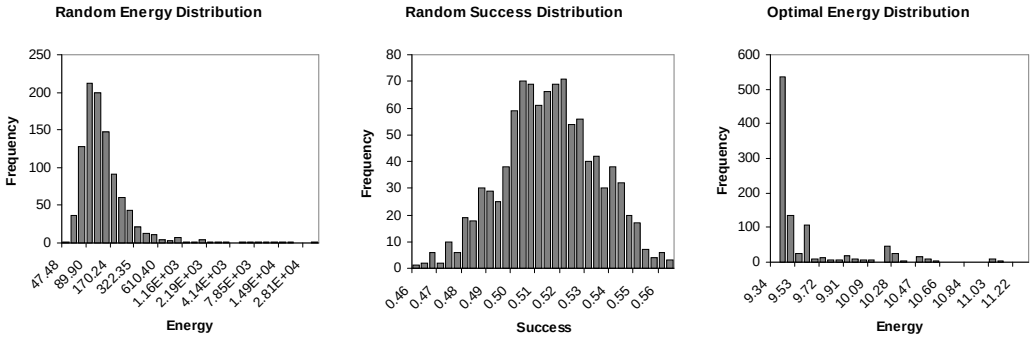
The random energy has mean 156.86, standard deviation 1263 and median 54.68. The random success has mean 0.521, standard deviation 0.024 and median 0.521. The optimal energy has mean 4.58, standard deviation 0.23 and median 4.46.

B.7 Eight Vowels



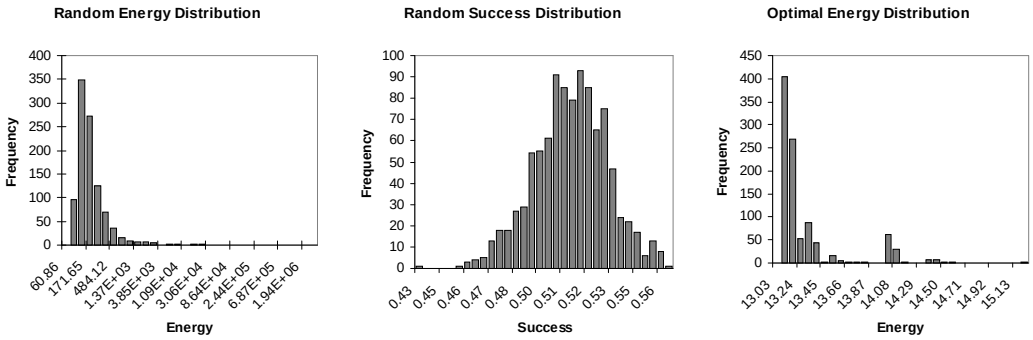
The random energy has mean 191.20, standard deviation 1258 and median 73.51. The random success has mean 0.515, standard deviation 0.021 and median 0.515. The optimal energy has mean 6.72, standard deviation 0.37 and median 6.48.

B.8 Nine Vowels



The random energy has mean 258.5, standard deviation 1421 and median 101.3. The random success has mean 0.513, standard deviation 0.020 and median 0.513. The optimal energy has mean 9.56, standard deviation 0.35 and median 9.40.

B.9 Ten Vowels



The random energy has mean 3043, standard deviation 86 600 and median 127.20. The random success has mean 0.511, standard deviation 0.019 and median 0.512. The optimal energy has mean 13.25, standard deviation 0.34 and median 13.12.

B.10 Trends

A number of trends can be observed in these data. The energy of both random and optimal vowel systems increases with the number of vowels. This was to be expected, as the number of terms in equation 2.1 increases with the square of the number of vowels in the system. Also, the distance between the vowels decreases with the number of vowels and therefore the terms that are added in equation 2.1 increase. The increasing energy for random as well as optimal systems is plotted in figure B.1. As the energy distribution is extremely skewed, the median rather than the average of the energy is plotted.

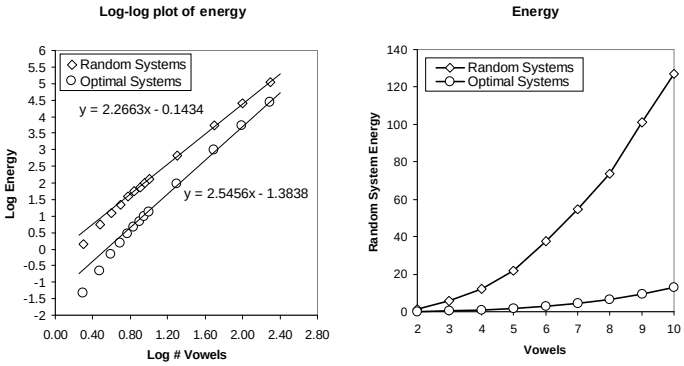


Figure B.1: Log-log plot and linear plot of energy.

In the right part of this figure the medians of the energies of the random and optimal systems of two to ten vowels are shown in a linear plot. It can be seen that random system energy is always much higher than optimal system energy, but it is not quite clear what the asymptotic behaviour of the energy is. For this the left part of the graph is more interesting. Here the scales are logarithmic. Shown are the data points for systems of two to ten vowels as well as for systems of twenty, fifty, hundred and two hundred vowels. The data points for systems up to a hundred vowels are the medians of one thousand runs. The data points for two hundred vowels are the median of 522 and 320 for random and optimal systems, respectively, because of limitations on computational resources. Systems with more than twenty vowels have no linguistic significance, as the maximum number of different vowel qualities in any language is near twenty. However, they have been calculated in order to investigate the asymptotic behaviour of the energy. As can be seen, the growth of both the random and the optimal energy converges towards straight lines in the log-log plot. The equations for these lines are shown in the graph. Although they are not quite equal for both systems, yet it is safe to assume

that they will converge to the same value, as optimal systems will always have lower energy than random systems. The limit of the slope of the asymptote of the growth curves will probably be two, indicating that the energy will finally grow with the square of the vowel system size. It should also be noted that the energy of the optimal systems becomes less reliable for larger systems, because the optimisation worked with a fixed step size of 0.03. The more vowels there are in a system, the closer they are together this step size should become smaller for larger systems. The calculated optimal energy will therefore be higher than the actual optimal energy.

Although this asymptotic growth is interesting from a theoretical point of view, it is not very relevant to the performance of the agent simulations, as systems with more than twenty vowels emerge only very rarely. For the relevant range of vowel system sizes the only things that can be said are 1) that the energy increases much faster than linear for both the optimal and the random systems, 2) that the relative growth of the energy of optimal systems is slightly faster than that of random systems and 3) that there is no easy formula for estimating the energy of optimal or random systems. The energy of the systems found should therefore always be compared with the values presented in this appendix.

The asymptotic behaviour of the success for random systems is much more relevant for the vowel systems that emerge from the simulations. As can be seen in figure B.2 (showing random success and error bars for the 1% reliability interval based on 1000 runs) the random success converges quite rapidly towards a value of 50% for higher numbers of vowels. The random success rate for lower numbers of vowels is even higher than this. This means that the performance which could be expected if the agents were doing nothing at all is already quite "good". In the case of discrete signals, one would expect a random performance of $1/n$ for n signals. However, due to the continuous nature of the distance function that is used to distinguish signals form each other and due to the rules of the imitation game the random performance of the imitation game is much higher. A mathematical analysis of imitation games with randomly initialised agents is presented in appendix C.

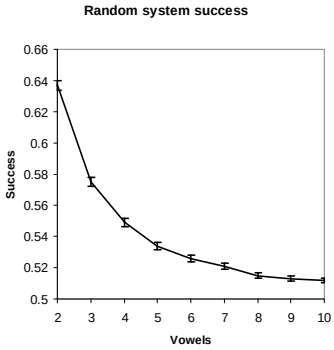


Figure B.2: Success of random systems.

Appendix C: Analysis of Random Vowel Imitation

In chapter 4 it was found that the average success/use ratio (hence the success rate) for populations of agents with randomly initialised vowel systems tends towards 50% for large number of vowels. This seems rather high as a naïve analysis of random imitation with n elements would suggest a success rate of $1/n$ which would tend to zero for large n. Indeed, if one considers random imitation to be the process in which the imitating agent randomly selects an element from its repertoire of n elements in order to imitate the other agent's signal, one would find a $1/n$ chance of correct imitation. However, this is not what happens in the imitation games.

In an imitation game the imitator does not choose a random element from its repertoire, but chooses the element that is closest to the generated signal from n randomly generated prototypes. Imitation games played in this way will tend towards a success of 50% for large numbers of vowels, as will be shown by the tentative mathematical analysis in this appendix. Making a mathematical analysis of the imitation game in this way is useful, because it provides more insight into the precise dynamics of the imitation games and is a first step of describing and analysing the imitation games more thoroughly.

The distances between the vowel prototypes of agents can be considered as a matrix of distances R:

C.1) $$R = \begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nm} \end{pmatrix}$$

Where m is the number of vowels of the imitator and n is the number of vowels from the initiator. The individual distances between vowel i of the initiator of the imitation game and vowel j of the imitator is r_{ij}. For agents with randomly generated vowel systems this is a matrix of random numbers, although the numbers are not independent from each other, because the distances are Euclidean. This means that if vowel v_1 is close to vowel v_2 and vowel v_2 is close to vowel v_3, vowel v_1 will more likely than not also be close to vowel v_3.

However, numbers within a row or within a column *are* independent from each other. This is because the vowels within one agent have been chosen independently from each other. This does not mean that they are distributed uniformly. Even though the articulator positions of random vowels have been chosen uniformly distributed between zero and one, the mapping from articulator positions to distances is non-linear, so the distance distribution is not expected to be uniform. However, this does not matter.

In an imitation game, the initiator randomly chooses a vowel from its repertoire, say vowel i. The imitator than chooses the vowel j, so that the distance r_{ij} is minimal, in short:

C.2) $$\left\{ j \,\middle|\, (1 \le j \le m) \wedge \left(\neg\exists k : (1 \le k \le m) \wedge r_{ik} < r_{ij}\right)\right\}$$

The initiator in turn chooses the vowel l, so that the distance r_{lj} is minimal, in short:

C.3) $$\left\{ l \,\middle|\, (1 \le l \le m) \wedge \left(\neg\exists k : (1 \le k \le m) \wedge r_{kj} < r_{lj}\right)\right\}$$

The imitation game is successful if l=i. This process might seem complicated, but it is in fact quite simple. It consists of finding the minimum of row i and of the minimum of column j, where j is determined by the first minimum. If the two minima are the same, the imitation game is successful. As all the numbers can be considered independent random numbers, the probability of success of an imitation game boils

down to the probability that the minimum of a row of m random numbers is smaller than the minimum of another row of n–1 random numbers, taken from the same distribution.

First, it will be assumed that the numbers are taken from the uniform distribution between zero and one. This can be assumed, because if the distribution of the numbers is not uniform (but still continuous) a monotonously increasing function f: R→ R can always be found that maps the numbers in such a way that their distribution does become uniform between zero and one. As this mapping is monotonously increasing, the sequence of the numbers will not be changed. Any proof about the sequence of the uniformly distributed numbers will therefore also be applicable to the original numbers.

The function $B_N(x)$ gives the probability that x is smaller than the smallest number in a row of N numbers taken from the uniform distribution between zero and one. For N=1 this probability is 1–x, in short:

C.4) $$B_1(x) = 1 - x,$$

because the probability that a number (the one number from the row of numbers) that is taken from the uniform distribution is bigger than x is equal to the surface under the distribution function between x and one. This is equal to 1–x for the uniform distribution.

With induction $B_N(x)$ can now be calculated from $B_{N-1}(x)$. The probability that x is smaller than the smallest number of a row of N numbers is equal to the probability that x is smaller than the smallest number of a row of N–1 numbers multiplied by the probability that the N^{th} number (taken from the uniform distribution between zero and one) is larger than x. This last probability, is equal to 1–x as has been shown above. Thus: $B_N(x) = (1-x)B_{N-1}(x)$, which, using equation C.4, solves to:

C.5) $$B_N(x) = (1-x)^N$$

The probability that the lowest number of a row of m numbers is also lower than the lowest number of a row of n–1 numbers is equal to the probability that x is lower than the lowest number of n–1 numbers under the condition that x is also lower than the lowest number of m–1 numbers. This conditional probability can be calculated as follows:

C.6) $$p(a|b) = \frac{p(a \wedge b)}{p(b)}$$

where a is the probability that x is smaller than the smallest number of n–1 numbers and b is the probability that x is smaller than the smallest number of m–1 numbers.

The probability $p(a \wedge b)$ can be calculated by integrating over all values of x the product of the probability that x is smaller than the smallest number of a row of n–1 numbers and the probability that x is smaller than the smallest number of a row of m–1 numbers. But these probabilities can be calculated using equation C.5, resulting in the following equation:

C.7) $$p(a|b) = \int_0^1 B_{n-1}(x) B_{m-1}(x) dx = \int_0^1 (1-x)^{n-1} (1-x)^{m-1} dx$$

which solves to $1/(m+n-1)$.

The probability p(b) can be calculated in a similar way, by integrating over all values of x the probability that x is smaller than the smallest number of a row of m–1 numbers:

C.8) $$p(b) = \int_0^1 B_{m-1}(x)dx = \int_0^1 (1-x)^{m-1} dx$$

which solves to $1/m$. Substituting both these values in equation C.6 gives the probability: $m/(m+n-1)$. If one assumes that m and n are equal, this would predict a success value of 0.60 for systems with three vowels and a success value of 0.55 for systems with six vowels. Both these values agree quite well with the values that were found in the random vowel experiments in chapter 4 and that are shown in figures 4.6 and 4.7, respectively. The values are not totally equal, because the numbers of vowels in the agents were not all equal.

Appendix D: Realistic Vowel Synthesis and Analysis

This appendix describes the technical details of the production and perception of real vowel signals, as used by the simulation described in the section on real signals in chapter 5. In the first section of this appendix, the production of the real vowels is described, and in the second section the perception, or rather the calculation of the distance between two realistic vowel signals. In this section the way in which a vowel is extracted from a real speech signal is also described.

D.1 Production

The production of vowels is based on the same formant frequencies that were calculated for the simulation with simplified signals. Figure 3.3 in chapter 3 shows how these formant frequencies are calculated from the articulatory positions. These formant frequencies are then used as the centre frequencies of four band-pass filters that will be excited with an artificial glottal pulse. The whole process of generating the realistic signal is illustrated in figure D.1 while the signals at the different stages are illustrated in figure 5.9.

The band-pass filters are implemented as infinite impulse response filters (the implementation details were taken from section 12.1.6 of (Allen *et al.* 1987) and the error in their description was corrected on the basis of the re-implementation of Dennis Klatt's formant synthesiser by Jon Iles and Nick Ing-Simmons, also Ifeachor & Jervis (1993) and Oppenheim *et al.* (1983) were used) on a time series. They work as follows:

D.1) $$o_t = a \cdot i_t + b \cdot o_{t-1} + c \cdot o_{t-2}$$

Where o_t is the output of the filter at time t, i_t is its input at time t, and a, b and c are coefficients that are calculated as follows:

D.2) $$c = -e^{-2\pi \cdot \frac{bw}{f_{rate}}}$$

$$b = 2e^{-\pi \cdot \frac{bw}{f_{rate}}} \cdot \cos\left(2\pi \frac{f_{centre}}{rate}\right)$$

$$a = 1 - b - c$$

Where bw is the bandwidth of the filter, f_{centre} is the centre frequency of the filter and f_{rate} is the rate with which the signal is sampled. This means that for working with realistic signals a sampling rate should be fixed and a bandwidth should be specified for each formant. The sampling rate was chosen to be ¼th of Compact Disc frequency, or 11025 Hertz. This sampling rate is high enough to capture the highest distinguishing formant frequencies of vowels, but low enough to make calculating the samples feasible.

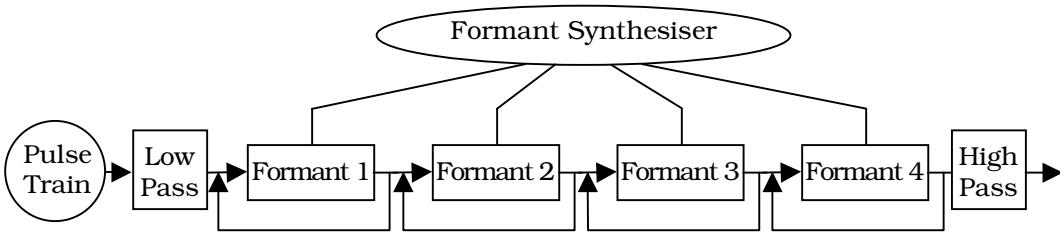


Figure D.1: Signal flow through realistic vowel synthesiser.

The bandwidth for formant frequencies is more problematic. Normally, the shape of the vocal tract does not only change the centre frequencies of the formants, but also their bandwidths. One could imagine calculating these bandwidths with a similar interpolation function as the one that is used to calculate the centre frequencies. Unfortunately, the bandwidths change much less nicely with changing vocal tract shape than the centre frequencies do. However, it was found that the value of the bandwidths did not really influence the quality of the sound that was produced very much. It was therefore decided to give all formants a bandwidth of 70 Hertz.

The filters for all formants were put in series. This means that the filter for the second formant takes as its input the output of the first formant filter, while the third formant filter takes as its input the output of the second formant filter, etc. The input of the first formant was an artificially generated glottal pulse. This glottal pulse was generated by low-pass filtering an impulse train. An impulse train is a signal that is one only at certain times, usually with a fixed frequency, and zero otherwise. The low pass filter was in fact a band-pass filter, identical to the one described in equation D.1, but with a centre frequency of zero Hertz and a bandwidth of 100 Hertz. In order to make the glottal pulse more realistic, the frequency was not completely constant. The distance d_p between two consecutive pulses was calculated as follows:

D.3) $$d_p = \frac{f_{rate}}{f_{voice}}\left(0.99 + U\left(-0.02, 0.02\right)\right)$$

Where f_{voice} is the voicing frequency, f_{rate} is the sampling rate and U (-0.02, 0.02) is a random number taken from the uniform distribution between –0.02 and 0.02. In order to create a realistic intonation contour, the voicing frequency can be decreased over the duration of the vowel, but this was not used in the results presented in chapter 5.

Finally, the signal that resulted from applying the formant frequency filters to the artificial glottal pulse was high pass filtered in order to emphasise the higher formant frequencies, and in order to model the radiation effects of the lips. Although for the most realistic results the characteristics of this filtering should depend on the lip rounding, it was found that a very good result was already obtained by just calculating the derivative of the signal as follows:

D.4) $$o_t = i_t - i_{t-1}$$

Where o_t is the output of the filter at time t, and i_t is its input (in this case the output of the fourth formant filter).

D.2 Perception

The perception of realistic vowels is based on comparing smoothed and weighted spectra. First a smoothed spectrum of a perceived signal is calculated, and the energies of the different frequencies in this signal are weighted. Then the absolute value of the difference of this smoothed and weighted spectrum and the reference spectra that are stored with the vowel prototypes are calculated. The signal with the smallest difference is considered to be recognised.

The smoothed spectrum is calculated on the basis of a *linear predictive coding analysis* (see e.g. Rabiner & Schafer 1978). The technique of linear predictive coding tries to code a speech signal as the coefficients needed to predict a sample as the linear combination of a number of the previous samples. This technique works well

with periodic signals such as the vowels. The basic linear prediction formula is as follows:

D.5) $$\bar{s}_t = \sum_{k=1}^{n} \alpha_k \cdot s_{t-k}$$

Where $\bar{s}_t$ is the predicted sample at time t, s_{t-k} are the observed samples at t-k, n is the number of samples on which the prediction is based and the α_k are the linear prediction coefficients.

The coefficients for a given sample of speech can be estimated in a number of different ways. In the realistic vowel system, the coefficients were calculated with the linear predictive coding routines in Press *et al.* (1992) section 13.6. This implementation estimates 14 coefficients based on a sample of length 512. They are then used to calculate a smoothed power spectrum of the original sample.

Calculation of the power spectrum is based on the maximum entropy method described in Press *et al.* (1992), section 13.7. In short, this method relies on the observation that if one considers equation D.5 as a linear filter, its frequency response for frequency f, $H(f)$ is:

D.6) $$H(f) = \frac{1}{1 - \sum_{k=1}^{n} \left(\alpha_k \cdot e^{2\pi i f k / f_{rate}} \right)}$$

The power of the spectrum at frequency f is then equal to $|H(f)|^2$.

Due to the fact that a number of coefficients is used that is small compared to the number of samples, the resulting power spectrum is smoothed. In the realistic vowel system, the power of the spectrum is calculated for 100 frequencies that are evenly space between zero and half the sampling frequency f_{rate}. These power estimates are called P_1–P_{100}. However, these power spectrum values can not yet be used for comparing two signals. First of all, the spectra have to be normalised, so that they are always zero for frequency zero. Also, the relative influence of the higher frequencies is too high. Therefore weighted, normalised power estimates W_i are calculated from the original power estimates P_i in the following way:

D.7) $$W_i = \frac{P_i - P_0}{i + 1}.$$

This weighting procedure results in all octaves contributing approximately equally to the total power of the spectrum. In the original power estimate, the higher octaves were more influential, because equal distances in Hertz separated the frequencies whose power was calculated.

The distance $D(a,b)$ between the weighted power spectra of two signals a and b can now be calculated as follows:

D.8) $$D(a,b) = \sum_{k=1}^{100} |W_{i,a} - W_{i,b}|$$

Where $W_{i,a}$ and $W_{i,b}$ are the i^{th} weighted power estimates of signals a and b, respectively.

A last thing that needs to be explained is the way in which the vowel is located in an utterance by a human speaker. The human utterance will be more complicated and longer than the artificial utterances generated by the agents. It is therefore essential to locate the best part of the signal for analysis. It is assumed that the vowel will be clearest at the core of the syllable, and that the core of the syllable is

the place where its volume is highest. The volume is calculated as the running average over the power of the signal in the following way:

D.9) $$R_t = \gamma R_{t-1} + (1-\gamma)s_t^2$$

Here R_t is the running average of the volume at time t, s_t is the sample at time t and γ is a factor between zero and one that determines over how much time the running average is calculated. The closer γ is to one, the longer the period. In the realistic vowel system it was given the value 0.995. The point where R_t reaches its maximum is taken to be the start of a sample of length 512 that is then used for analysis. This process is illustrated in figure D.2.

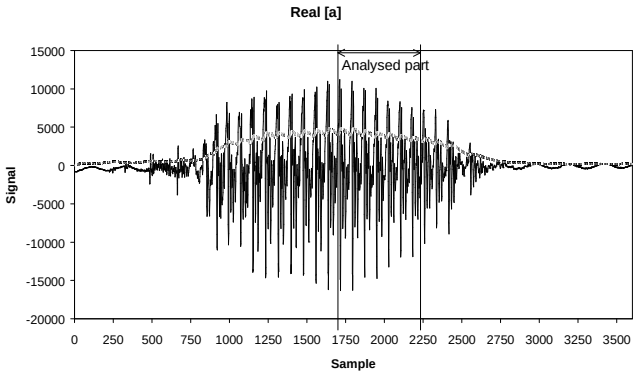


Figure D.2: Finding the core of a vowel.

Appendix E: Consonant Data

For the experiments on CV-syllables, described in chapter 7, data on plosive consonants were needed. Although a lot of measurements of perception and production of plosives (Fant 1973; Stevens & Blumstein 1975; Cooper *et al.* 1976) have been made, most of this data is limited to a few places of articulation. Most abundant data is found on bilabial, alveolar/dental and velar plosives that occur in well-studied European languages such as English, French and Swedish. However, the aim of the experiments in this thesis is to find the reason why some places of articulation are preferred over others. Consequently, it is essential not to constrain the agents beforehand to a limited set of places of articulation. A number of measurements of the acoustic properties of plosive consonants have therefore been made. The selected places of articulation were: uvular, velar, palatal, retroflex, alveolar, dental, linguo-labial, labio-dental and bilabial. The author produced the consonants before [i], [a] and [u]. A spectrogram was made of each syllable and the frequencies of the starting points of the formant transitions as well as the frequency of the most energetic part of the burst were measured. In order to verify the measurements, the consonants were re-synthesised, and the measurements were refined if the result was not satisfactory. The measurements are presented below. For the experiments, a selection was made. The measurements were compared (and if necessary corrected) with data from the literature. Only the loci were used, and the effect of the following vowel was ignored (in order to keep in line with Lindblom et al's (1984) experiment).

	[ɢ]	[g]	[ɟ]	[ɖ]	[d]	[d̪]	[d̻]	[b̪]	[b]	[i]
Burst:	1600	2000	3000	2500	2500	3000	2400	2100	1600	
F1:	300	300	300	250	300	250	250	200	200	310
F2:	1700	2100	2100	2000	2100	1500	1600	1600	1600	2300
F3:	2400	2700	3000	2400	2700	2200	2300	2300	2300	2800
F4:	3800	3400	3700	3000	3500	3500	3000	3400	3200	3600

Table E.1: Consonants before **[i]**.

	[ɢ]	[g]	[ɟ]	[ɖ]	[d]	[d̪]	[d̻]	[b̪]	[b]	[a]
Burst:	1330	1800	2800	2500	3200	3800	2500	2800	1750[1]	
F1:	500	340	280	310	370	310	400	400	400	700
F2:	1200	1800	2000	1800	1800	1800	1600	1500	1500	1450
F3:	3000	2100	2100	2100	2600	2700	2600	2500	2400	2700
F4:	3500	3200	3300	3100	3600	3500	3300	3900	3500	3800

Table E.2: Consonants before **[a]**

	[ɢ]	[g]	[ɟ]	[ɖ]	[d]	[d̪]	[d̻]	[b̪]	[b]	[u]
Burst:	900	900	2300-1700	1500	2500	2500	1500	2500	2000	
F1:	350	250	250	350	400	400	400	400	400	250
F2:	900	900	1700	1200	1500	1900	1600	900	900	900
F3:	2800	1800	1900	2100	1700	2200	2200	2400	2400	2400
F4:	3700	3600	3500	2800	3700	3700	3200	3000	3300	3300

Table E.3: Consonants before **[u]**.

[1] Actually, it appears as if [pa] is best perceived without a burst.

Appendix F: Details of the Complex Utterance Model

In chapter 7 the importance of implementing the imitation games with more complex utterances than vowels was pointed out. It was also pointed out that a model that is sufficiently sophisticated to model complex utterances is very hard to implement. A part of the necessary software has already been implemented as part of the research for this thesis. In order to make it easier to replicate or extend the results, the details of the implementation will be presented. Parts of the implementation are based on the literature, and parts are based on new ideas. References to the original sources will be made at the appropriate places.

F.1 Production

The core of the production of realistic sounds is Mermelstein's (1973) articulatory synthesiser. This model has also been described (but not in so much detail) in (Rubin *et al.* 1981). Other models that can be used as articulatory synthesisers also exist, most notably Maeda's (1989) model, but has not been described in sufficient detail in the open literature to re-implement it. Also, Maeda's model is based on a principal components analysis of a large number of vocal tract shapes. The degrees of freedom are the first few principal components that were extracted from the measurements. Although these principal components correspond to some extent with the different articulators, they do not have a direct geometrical interpretation. Mermelstein's model, on the other hand, is a geometrical model. Its degrees of freedom can be interpreted directly as movements of the different articulators. Boersma (1998) implemented a very detailed and original articulatory synthesiser based on Mermelstein's model. However, his synthesiser is computationally too complex and was discovered too late by the author to be used in this thesis.

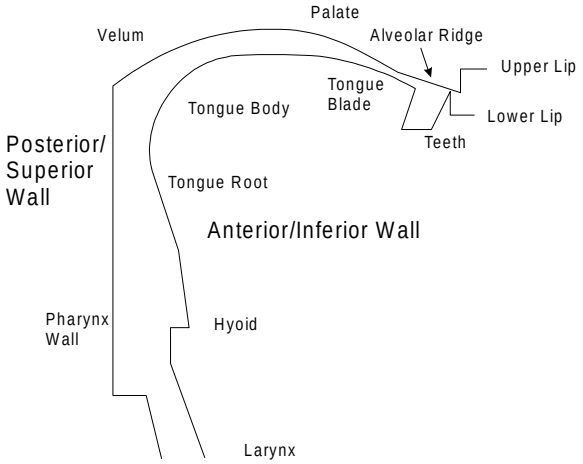


Figure F.1: Mermelstein's (1973) model.

F.1.1 Calculating the shape of the vocal tract

The Mermelstein model is illustrated in figure F.1. It models a midsagittal cross section of the vocal tract. It consists of an anterior (towards the face) and a posterior (towards the back of the head) wall. The posterior wall is mostly static, except for the upper lip, the soft palate and the velic opening. The anterior wall is highly flexible, as its shape is influenced by the action of the lower lip, the tongue, the front wall of the pharynx and the jaw.

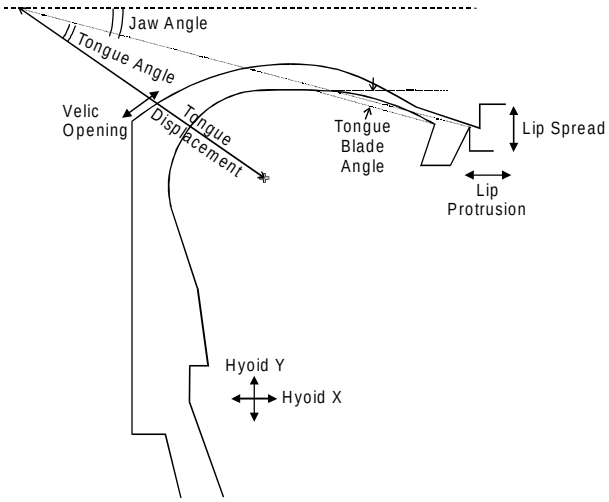


Figure F.2: Control parameters.

The shape of the walls of the model can be influenced by a number of parameters, also called its control parameters or its degrees of freedom. These are illustrated in figure F.2. They are jaw angle, tongue angle, tongue displacement, tongue blade angle, hyoid vertical displacement, hyoid horizontal displacement, lip spread, lip protrusion and velic opening. Following Mermelstein's notation, tongue angle is written as ϑ_c, tongue displacement as s_c, tongue blade angle as ϑ_t, lip spread as h_l, lip protrusion as p_l (see also figure F.4). For velic opening and for the position of the hyoid there is no separate notation in Mermelstein (1973). He uses the point V, which is the position of the uvula for calculating the velic opening and the point H to represent the position of the hyoid. In this description, V_p will be used for the control parameters of the velum and h_x and h_y will be used for the control parameters of the hyoid. Note that in figure F.4, the angles and lengths seem to indicate the whole length and whole angle. This is not the case; they only indicate the controls on the total lengths and angles. In the formulas given below, the control parameter symbols will be used as variations (preferably centred around zero) on the lengths and angles indicated. An offset is added to the control parameters to get the total lengths and angles. In this respect the use of the symbols differs from Mermelstein's (1973) use.

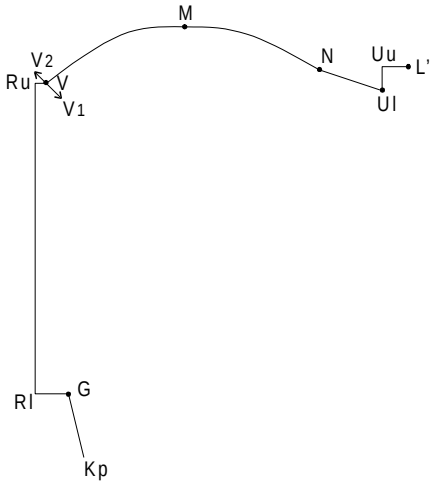


Figure F.3: Posterior/superior wall of Mermelstein's model.

All co-ordinates of the model are relative to the point F[1], which corresponds to the turning point of the jaw. The position of the tongue, lower teeth and lower lip depend on the position of the jaw. The movement of the jaw is rotational, so it is convenient to have the centre of rotation of the jaw at the origin of the co-ordinate system. For extra convenience, the movement of the centre of the tongue, which can in fact move freely in two dimensions, is also linked to the origin. The shape of the posterior/superior wall, illustrated in figure F.3 does not depend on the position of the jaw, and will therefore be described first.

The dimensions and shape of the both walls have been taken from the text and figures of Mermelstein's (1973) paper. Most of the information needed to reconstruct the model could be found in the text, but some of it had to be measured from the figures. For the reasons of the different ways of calculating the positions of the points, the reader is referred to the original article. The point Kp is the position of the larynx on the posterior wall of the vocal tract. Its position is related to the position of the hyoid, as follows:

F.1) $\quad \mathrm{Kp} = \left(5 + h_x/2\,,\, 10.3 + h_y\right)$

where the first term in parenthesis is the x co-ordinate and the second term is the y co-ordinate, relative to point *F*.

[1] In naming points in the model, Mermelstein's (1973) original paper will be followed as closely as possible.

Point G is fixed at (4.7, –9), point Rl is fixed at (4, –9), the *x* co-ordinate of point Ru is equal to that of Rl, but its *y* co-ordinate is equal to that of point V. Their positions are calculated as follows:

$$\begin{aligned} \text{Ru} &= \left(4, -2.6 + 0.6V_p\right) \\ \text{V} &= \left(4 + 0.4V_p, -2.6 - 0.6V_p\right) \end{aligned} \qquad \text{F.2)}$$

The control parameter V_p has the range [0,1]. Point V can thus change from (4, -2.6) to (4.4, –3.2). Note that the ranges have been determined by experimenting with the model, and are not to be found in Mermelstein (1973).

The point M is at position (7.2, –1.4). The line between V and M is a circle segment that goes through these two points and has a horizontal tangent in M. Point N has position (10.2, –2.27) in fact, point N is located on two thirds of the way on the straight line from M to Ul. The position of point Ul is (11.2, –2.7). The *y* co-ordinate of Uu is related to the *y* co-ordinate of the upper lip, L'. They are calculated as follows:

$$\begin{aligned} \text{Uu} &= \left(11.2, -10.3 + h_l\right) \\ \text{L'} &= \left(11.2 + p_l, -10.3 + h_l\right) \end{aligned} \qquad \text{F.3)}$$

where p_l is in the range [0,1] and h_l is in the range [-1,1].

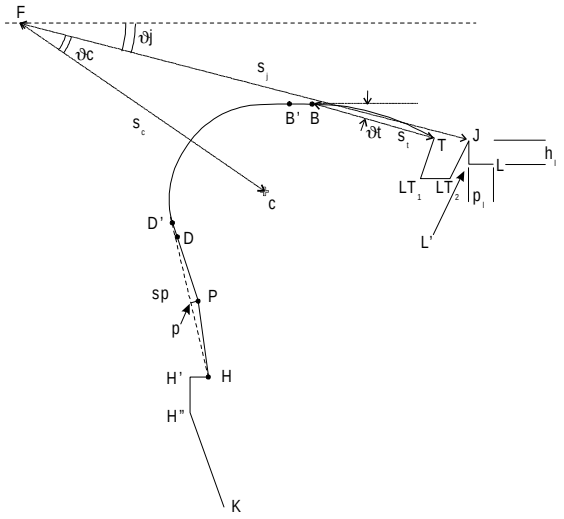


Figure F.4: Anterior/inferior wall of Mermelstein's model.

The calculation of the shape of the anterior/inferior wall of the vocal tract is slightly more complicated, because it depends in part on the angular movements of the jaw and the tongue. The lower parts of the vocal tract, however, depend only on the movement of the hyoid. The model is illustrated in figure F.4. The reader is again referred to Mermelstein (1973) for the science behind the model. Here it will only be described how the different points are calculated.

The position of points K, H, H' and H" depend only on the movements of the hyoid. They are calculated as follows:

$$\begin{aligned} \text{K} &= \left(5.9 + h_x/2, -10.3 + h_y\right) \\ \text{H} &= \left(5.6 + h_x, -7.6 + h_y\right) \\ \text{H}' &= \left(5.2 + h_x, -7.6 + h_y\right) \\ \text{H}'' &= \left(5.2 + h_x, -8.3 + h_y\right) \end{aligned} \qquad \text{F.4)}$$

The position of the points LT1, LT2, J, L' and L only depends on the position of the jaw and the lips. They can be calculates as follows:

$$\begin{aligned} \text{J} &= \left(11.3\cos\left(-0.237 - \vartheta_j\right), 11.3\sin\left(-0.237 - \vartheta_j\right)\right) \\ \text{LT}_1 &= \left(\text{J}x - 1.0, \text{J}y - 0.8\right) \\ \text{LT}_2 &= \left(\text{J}x - 0.4, \text{J}y - 0.8\right) \\ \text{L}' &= \left(\text{J}x, \text{J}y - h_l\right) \\ \text{L} &= \left(\text{J}x + p_l, \text{J}y - h_l\right) \end{aligned} \qquad \text{F.5)}$$

where J*x* and J*y* are the *x* and *y* co-ordinate of J. The angle ϑ_j should fall in the range [-0.2, 0.2].

The rest of the points depend on the movement of the tongue. The tongue is modelled as a circle of a constant radius of 2.0cm. The position of the centre of the tongue c is calculated as follows:

F.6) $$c = \left((7.8 + s_c)\cos(-0.527 - \vartheta_j - \vartheta_t), (7.8 + s_c)\sin(-0.527 - \vartheta_j - \vartheta_t)\right)$$

where s_c should fall in the range [-1,1] and ϑ_t should fall in the range [-0.2, 0.2].

The rest of the points are defined relative to the tongue body. Point D' is the point furthest to the back on the tongue body where the line HD' is tangent to the tongue body. Line pP is the line halfway the line HD', perpendicular to HD'. The length of this line is dependent on the length of line HD' as follows:

F.7) $$|pP| = 0.57(3.48 - |HD'|).$$

The anterior pharynx wall is now formed by the straight line from H to P and the straight line from P to D, where D is the point furthest on the back of the tongue body where the line PD is tangent to the tongue body.

The position of the tongue tip and tongue blade depend on the movements of the jaw, the tongue body and the tongue blade. The point B is the point about which the tongue blade rotates, B' is the point where the tongue blade starts and T is the tongue tip. They are calculated as follows:

$$B = \left(Cx + 2.0\cos(1.493 + \vartheta_j), Cy + 2.0\sin(1.493 + \vartheta_j)\right)$$

F.8) $$B' = \left(Cx + 3.4\cos(1.493 + \vartheta_j + \vartheta_c - \vartheta_t), Cy + 3.4\sin(1.493 + \vartheta_j + \vartheta_c - \vartheta_t)\right)$$

$$T = \left(Bx + 3.4\cos(-0.237 + \vartheta_j + \vartheta_c - \vartheta_t), By + 3.4\sin(-0.237 + \vartheta_j + \vartheta_c - \vartheta_t)\right)$$

Where Cx and Cy are the x and y co-ordinate of the centre of the tongue t, and Bx and By are the x and y co-ordinate of point B. The range of ϑ_t is [-0.2, 0.2]. Mermelstein's (1973) paper presents a method for calculating a default value for this control parameter using the following formula:

F.9) $$\vartheta_t = 0.004(s_c - 0.8)$$

The tongue blade outline itself is determined by a curve which is tangent to the tongue body at B' and goes through point T. Mermelstein (1973) says that this curve is "a radial coordinate about the tongue-body center C which varies as the square of the angular difference with respect to the starting point B' on the tongue body". The formula is reconstructed here as follows:

F.10) $$r = \frac{\alpha^2}{(\angle B'cT)^2}(|cT| - 2.0) + 2.0$$

Where r is the necessary length at angle α, $\angle$B'cT is the angle between lines B'c and cT and |cT| is the length of line cT. The value of 2.0 is the radius of the circle, equal to the length of line B'c.

F.1.2 Calculating the Areas

The next problem is how and where to calculate the areas of the vocal tract. A regular grid, shown in figure F.5 defines the places where the cross sectional areas have to be determined. Mermelstein (1973) uses a grid that is fixed with respect to the co-ordinate system, but in the implementation used for this thesis, the horizontal part of the grid was taken relative to the larynx, in order to prevent the lowest parts of

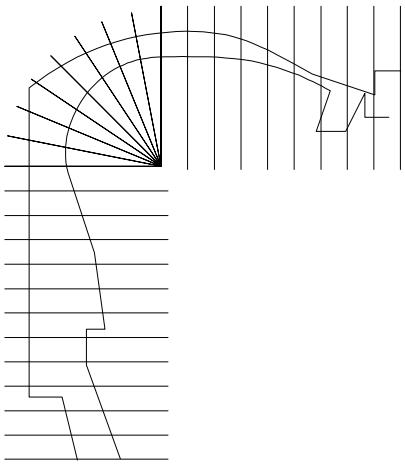


Figure F.5: The measured cross sections.

the grid from not intersecting anything when the larynx moved up. The number of sections in which the tract was divided was taken constant in the implementation in this thesis. As the length of the vocal tract varies slightly, this means that the length of the sections also varies somewhat. This was not taken into account in the implementation, but it did not seem to influence the realism of the model very much.

In the implementation used for this thesis, the length, and subsequently the area of the sections was used directly. Mermelstein (1973) uses a more complicated and more realistic model, in which he connects the midpoints of the intersections, and compensates the length and area of the resulting tube for the angle that these connections make with the intersections. Unfortunately this results in tubes with different lengths, which makes it impossible to use a lossless tube model for modelling the vocal tract. Mermelstein's (1973) solution to this problem is insufficiently clear to be reproduced. Therefore it was decided to use the lengths of the cross sections directly in the implementation used for this thesis.

Mermelstein (1973) does not use a fixed point of termination for the vocal tract. In his model the termination plane of the vocal tract is at the intersection of two lines, one drawn from the frontmost point of the upper lip with an angle of 45 degrees and one drawn from the frontmost point of the lower lip with an angle of –45 degrees. As this, too, would change the length of the tract, and thus either the number of tubes or the lengths of the tubes, it was decided to take the termination plane always at a fixed point.

The model calculates the area of 30 tubes. The first 29 depend on the shape of the vocal tract, the last one has a fixed area of 12, in order to make a good termination of the tube model. The first 12 tubes are based on horizontal intersections, then there are 8 tubes based on the radial intersections that make the 90 degree turn in the tract. The final 9 tubes are based on vertical intersections. The total length of the vocal tract is taken to be around 17.5 cm, so the length per tube is 6 mm. The lines on for the radial intersections are radiating from the point 3.4cm below the highest point on the palate, at co-ordinates (7.2, -4.8).

Mermelstein (1973) provides a number of formulas for calculating the area of the vocal tract in different locations that have been taken from the literature. The areas are calculated on the basis of the mid-sagittal cross sections. The area of the pharyngeal part of the tract (tubes number 1–14) is modelled as an ellipse with one axis the length of the cross section, and the other axis increasing linearly from 1.5 to 3 cm when moving upwards from the larynx. The area of the region of the velum (tubes 15–20) is calculated with:

F.11) $a = 2d^{1.5}$

Where a is the area and d is the length of the cross section. The area in the region of the hard palate (tubes 21–25) is calculated with:

F.12) $a = 1.6d^{1.5}$.

The area in the alveolar region (tubes 26–28) is calculated with:

F.13) $$a = \begin{cases} 1.5d, & \text{for } d < 0.5 \\ 0.75 + 3(d - 0.5) & \text{for } 0.5 \le d < 2 \\ 5.25 + 5(d - 2) & \text{for } d \ge 2 \end{cases}.$$

The area in the region of the lips (tube 29) is calculated as an ellipse with height the lip height hl and as width w the following:

F.14) $w = 2 + 1.5(\mathrm{L}'y - \mathrm{L}y - p_l)$

where L'y and Ly are the vertical position of the upper lip and lower lip, respectively (see above for the way to calculate these).

F.1.3 Making noise

The vocal tract can now be considered as a number of connected lossless cylindrical tubes. The acoustic properties of these tubes are quite simple and relatively easy to model. For a derivation of their properties as well as a comprehensive discussion of lossless tube models, see (Rabiner & Schafer, 1978, chapter 3). The propagation of sound waves in the frequency range of speech signals can be considered as planar waves. The tubes are considered lossless at the moment (this will be reconsidered later) meaning that the waves loose no energy while traversing the tube. The combination of these two properties makes it possible to assume that a wave that traverses the tube does not change in doing so. The time for a wave to traverse the tubes is equal to the length of the tube divided by the speed of sound. As the tubes are 6 mm long, the time to traverse them is 18µs, or 55 000 Hertz. When a wave goes from one tube to the next, part of it is reflected. The fraction of the wave that is reflected, R_i depends on the areas of the tubes A_i and A_{i+1} between which it travels:

F.15) $$R_i = \frac{A_{i+1} - A_i}{A_{i+1} + A_i}.$$

The value of R was taken to be one for the junction between the first tube and the glottis, effectively reflecting all sound back into the vocal tract. This is the simplest way to do this and amounts to assuming that there is no coupling between the glottis and the vocal tract. The value of R at the last tube was taken to be zero, effectively assuming that all sound is radiated. As the area of the last tube is fixed at twelve, this results in realistic sounds. If this last tube were not present, this would not result in a realistic signal. Because of the reflections, there will be waves travelling in both directions in the tubes. If at time t, there is a wave with energy $F_{i,t}$ travelling forward through tube i and a wave with energy $B_{i+1,t}$ travelling backwards through tube i+1, the waves at time t+1 can be calculated as follows:

F.16) $$\begin{aligned} F_{i+1,t+1} &= (1+R)F_{i,t} + R \cdot B_{i+1,t} \\ B_{i,t+1} &= (1-R)B_{i+1,t} - R \cdot F_{i,t} \end{aligned}.$$

However, there are two aspects that need to be taken into account in order to make the simulation more realistic. The first is that the human vocal tract is not lossless. In a lossless tube model, some waves bounce around too long, effectively resulting in a metallic quality to the produced sound. Therefore a damping factor was introduced. Reflected waves at every junction were multiplied with a damping factor slightly smaller than one, dependent of the area of the area through which the wave would be traversing:

F.17) $$\begin{aligned} F_{i,t} &\leftarrow \left(1 - \frac{0.0007}{\sqrt{A_i}}\right) F_{i,t} \\ B_{i,t} &\leftarrow \left(1 - \frac{0.0007}{\sqrt{A_i}}\right) B_{i,t} \end{aligned}.$$

The idea of this formula was taken from appendix B of (Rubin *et al.* 1981).

The second factor that needs to be taken into account is that airflow gets turbulent if there is a too powerful air stream through a too narrow opening, in other words if the Reynolds number gets higher than a threshold value. In the model used for the experiments, this is modelled simplistically with the following condition:

F.18) $$\frac{F_{i,t}}{A_i} < 300$$

If this condition is fulfilled, a random signal is added to the air stream. Right now this is implemented as multiplying the original air stream with a random number taken from a uniform distribution in the range [0,1]. This is perhaps a bit too crude and is one of the points that should be refined.

The input to the model is either a steady airflow (for example for generating voiceless fricative sounds) or a voicing signal that is generated in the same way as the voicing signal in the realistic vowel experiment, described in appendix D.

One important thing that is not yet modelled is the nasal tract. This is because a good model of the nasal tract has not yet been implemented. The crude modelling of a number of aspects notwithstanding, the signals produced by the synthesiser are quite acceptable. Sample output is presented in the section on experiments in this appendix.

F.1.4 Moving the articulators

Now that it has been described how a setting of the articulatory parameters can be converted in an acoustic signal, it is necessary to describe how the articulators can move. The movement of the articulators is implemented in a rather *ad hoc* way, and is therefore the least realistic aspect of the sound production part of the model. It can probably be improved by making use of the relevant literature (e.g. Browman & Goldstein 1995; Saltzman 1986; Saltzman & Munhall 1989; Saltzman 1995; Kaburagi & Honda 1996). At the moment, however, the model seems to be able to make satisfying sounds.

The dynamics of a single articulator are determined by its movement towards a certain articulatory goal. This movement is constrained by the fact that an articulator cannot change position or speed instantaneously. It has been inspired by the way dynamic behaviours are implemented on mobile robots in the AI-lab of the Vrije Universiteit Brussel (Steels 1992). The original way of changing quantities in the model described in the beginning of chapter 3, equation 3.2 and on robots was as follows:

F.19) $$p_t \leftarrow p_{t-1} + \alpha(g - p_{t-1})$$

Where p_t and p_{t-1} are the values of a position p at times t and t-1, g is the goal value and α is a constant between zero and one that determines the speed of change. Although this formula results in continuous movements, it can also result in discontinuous changes in speed. As articulators have mass and stiffness, this is not realistic. A slightly more complex (second order) formula has been used in the model described here in order to solve this problem:

F.20) $$\begin{aligned} v_t &\leftarrow v_{t-1} + 0.3 \cdot \alpha(\alpha(g - p_{t-1}) - v_{t-1}) \\ p_t &\leftarrow p_{t-1} + 0.3 \cdot v_t \end{aligned}$$

Where p, g and α have the same meaning as above. The difference is that the articulators now also have a speed v_t, whose change depends on the distance to the articulatory goal. The factor α is determined by the mass and the stiffness of the articulators in a rather crude way. For the different articulators in the model this factor was determined by trial and error. The value of α was set to 0.5 for all articulators, ex-

cept for the lips, were it was set to 0.9. The factor 0.3 was introduced for making the equations work most realistically. The difference in behaviour between these two formulas is illustrated in figure F.6. In this figure, the movement of an articulator that can change speed discontinuously (formula F.19) and the movement of an articulator that cannot change speed discontinuously are compared. The value of α was set to 0.2 for the first type of articulator and to 0.5 for the second type. It can be seen that the first type changes speed discontinuously at the moment the articulatory goal changes. The second type takes more time to accelerate. The problem with the second type is that it also takes more time to decelerate. This causes overshoot and oscillation around the goal value. This can probably be remedied by a better model of movement that is based on the movement of damped mass-spring systems.

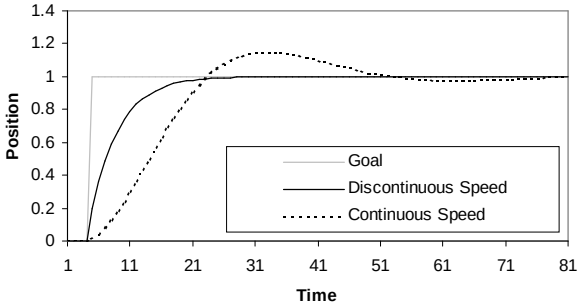


Figure F.6: Articulator movements with and without restrictions on speed change.

Problems sometimes occur when the two walls of the vocal tract touch each other and get blocked. This is detected if one of the tube areas becomes zero. In this case the solution is to stop all articulators that move towards the other wall, and to set their goal values to their current position. This allows articulators that move in the other direction to deblock the vocal tract, whilst preventing the two walls of the tract from intersecting each other. Note that this *can* result in a discontinuous change in speed (but not in position).

F.1.5 Co-ordinating the articulators

The last problem of producing sound is how to co-ordinate the different articulatory goals. In human speech co-articulation is very important. Therefore, articulations in sequence should influence each other. In the model presented here, co-articulation is implemented by making it possible that articulators that are not used for a certain articulatory gesture already start moving towards the target of the next articulatory goal. This is implemented by means of a *sequencer*.

An articulatory command to the sequencer consists of a number of articulatory goals for the different articulators in the model. A goal does not have to be specified for all articulators, though. Furthermore, a duration for which the articulatory goals will have to be pursued has to be specified. It is possible to send multiple articulatory commands to the sequencer simultaneously. Whenever the first articulatory command to be executed does not specify goals for all articulators, the sequencer looks whether in one of the subsequent commands a goal is specified for that particular articulator. It then already starts to move that articulator towards the articulatory goal specified. If no articulatory goals are specified for a given articulator, a default goal value (the rest value) is used. Articulatory goals are pursued until the duration that was specified for the articulatory command from which it was taken, has elapsed. The duration for a certain articulatory command only starts being counted after all previous articulatory commands have finished. This implies that, due to co-articulation, it is possible that all articulatory goals for a given command have already been reached even before its duration starts being counted.

The specification of the articulatory sequencer and of articulatory commands raises the question what articulatory commands actually mean in terms of linguistic concepts, such as phonemes and syllables. It could naively be expected that one articulatory command correspond with one phoneme. In the experiments described below, it was found that one articulatory command is often perceived as one CV-syllable. But before describing the experiments, the perception of the agents will be described.

F.2 Perception

Perception is based on a number of properties (features) that can be calculated from the speech signal. So far, the properties that have been calculated are the strength of the signal, whether the signal is voiced (periodic) or not, its fundamental frequency if it is voiced, and the frequency, bandwidth and strength of the first five formant frequencies. The different techniques that are used will now be discussed in turn, with references to relevant literature where more detailed descriptions and their mathematical derivation can be found.

Two standard techniques are used for analysing the signal. Fast Fourier transforms (FFT) are used for calculating the power, voicing and voicing frequency of the signal, while linear predictive coding analysis (LPC, see also appendix D) is used for estimating the formant frequencies of the signal. Although it is possible to do both analyses with both methods, the best routines that were found for voicing analysis used FFT and the best routines for formant analysis used LPC. Many of the ideas for both methods were found in Rabiner & Schafer (1978) and Press *et al.* (1992). Implementation was mostly based on Press *et al.* (1992).

The input of the agents is a continuous stream of speech. This stream is sampled at 11 025 Hertz—in fact, it is generated by the articulatory model at this frequency. In order to do an FFT or an LPC, a number of samples are needed. It was decided to analyse signals of 512 samples. This corresponds to a time of 46.4 ms. Analysing the signal in this way is sufficient to detect frequencies of minimally 21 Hertz and maximally 5512.5 Hertz. This is sufficient for analysis of speech. However, some events in speech take a shorter time than 46.4 ms. Therefore it was decided to do a signal analysis every 64 samples (5.8 ms). Every time the last 64 samples were thrown away and 64 new samples were added to the signal. Thus even very fast events in the speech signal could be detected.

F.2.1 Calculating power, voicing and voicing frequency with autocorrelation.

The power, voicing and voicing frequency were calculated on the basis of a autocorrelation analysis (Rabiner & Schafer 1978, ch. 4). The autocorrelation of a signal *S* is calculated as follows:

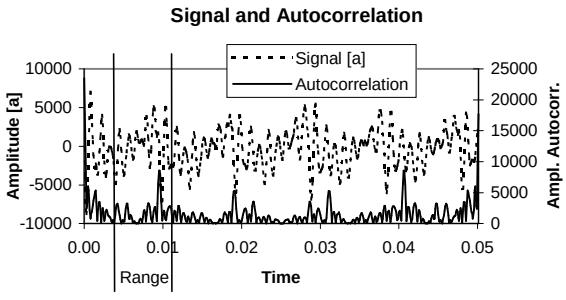


Figure F.7: The autocorrelation of a signal.

F.21) $$A = \mathrm{FFT}^{-1}\left(\left|\mathrm{FFT}(S)\right|\right)$$

where FFT is the fast Fourier transform and *A* is the autocorrelation of signal S. The autocorrelation *A* consists of a sequence of numbers as long as S, but it is symmetric (see figure F.7). It has a number of interesting properties. The total power of the

signal is the first element of the sequence. Voiced signals have strong peaks in the range that is indicated in the figure—in the implementation discussed here between the 44th and the 125th sample, corresponding to a minimal frequency of 88 Hertz and a maximal frequency of 251 Hertz. The position of the peak corresponds to the frequency of the voicing signal and its height relative to the total strength of the signal corresponds to the prominence of the voicing. In this implementation it was decided that if the ratio between the height of the secondary peak and the height of the first sample was higher than 0.25, the signal was voiced. There are a number of ways of improving the extraction of voicing information from a speech signal, such as filtering out the irrelevant frequencies, clipping the signal etc. For details, see Rabiner & Schafer (1978) chapter 4.

F.2.2 Extracting formants with linear predicitive coding.

Formants were extracted from the signal by linear predictive coding (LPC) analysis. As described in appendix D, LPC-analysis tries to predict the next sample in a signal as a linear sum of the previous samples. This was stated in equation D.5, which is reproduced below for reference:

$$\bar{s}_t = \sum_{k=1}^{n} \alpha_k \cdot s_{t-k}$$

where $\bar{s}_t$ is the predicted sample, $s_{t-1}...s_{t-n}$ are the n measured samples and $\alpha_1...\alpha_n$ are the coefficients. The value for n was chosen to be 12. This turned out to work well for calculating the first five formants. Estimating the coefficients is done by a technique described in Press *et al.* (1992), §13.6. The linear predictive coding sum from equation D.5 is in fact a recursive filter that is the closest possible approximation (given the number of coefficients) of the vocal tract transfer function. This can be used to calculate the formant frequencies of the vocal tract. These correspond to the places where the frequency response of the filter is maximal. The frequency response of the filter was given in equation D.6, which is related to the characteristic polynomial of the filter:

F.22) $$z^n - \sum_{k=1}^{n} \left(\alpha_k \cdot z^{n-k}\right) = 0 .$$

The frequency response can be derived from this function by substituting $e^{-\frac{2\pi i f}{f_{rate}}}$ for z, taking the reciprocal and multiplying by $e^{-\frac{2\pi i f}{f_{rate}} n}$, where f_{rate} is the sample rate of the signal. It is therefore clear that the roots of the characteristic polynomial are related to the maxima of the frequency response function. The polynomial has n roots that are either real numbers or complex conjugate pairs, as illustrated in figure F.8. The unit circle in this figure illustrated the projection of the real frequencies on the z-plane. The position of the (complex) roots can be expressed as a distance r from zero and an angle β with the positive real axis. The angle is directly related to the formant frequency F by:

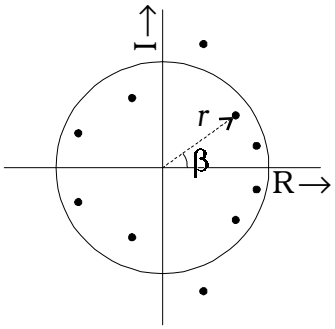


Figure F.8: Roots in z-plane.

F.23) $$F = \frac{\beta}{2\pi} f_{rate} .$$

Note that as all roots occur as either a real number or a complex conjugate pair, formant peaks occur either at frequency 0, half f_{rate} or in pairs that are mirrored

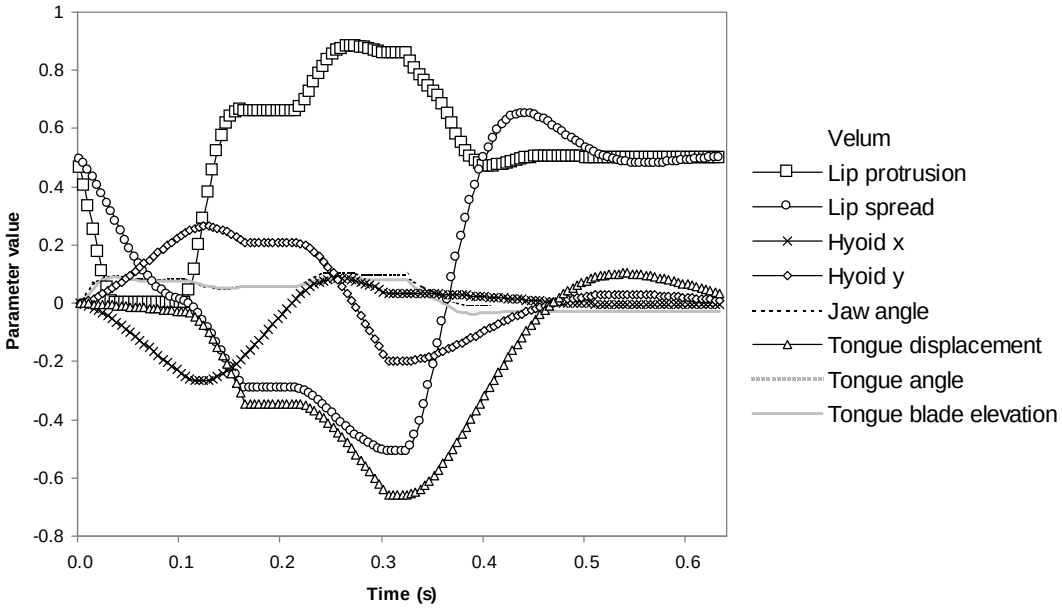


Figure F.9: Example articulator movement with three random commands.

around f_{rate}. This is exactly as expected, as a frequency spectrum of a real signal has to be symmetric around half the sampling frequency. The bandwidth of the formants can be estimated as follows:

F.24) $$B = \frac{f_{rate}}{\pi}|1 - r|$$

Where r is the distance between the complex root and zero. This formula is only accurate when r is sufficiently close to 1.

The power at the formant frequencies can be calculated by substituting them in equation D.6.

F.3 Experiments

Although the complex simulation was never used for playing imitation games, a number of experiments were performed to see how it would work and whether an inverse mapping between acoustic signals and articulatory gestures could be learnt. It turned out that the sounds the model could produce were reasonably realistic. Unfortunately, the inverse mapping problem turned out to be too complex to learn with the different methods that have been tried out.

F.3.1 A simple sound

The first experiment that was tried was whether the model was capable of making realistic sounds. For testing this, three random articulatory commands were sent to the sequencer. In all three commands, articulatory goals for all articulators were specified. The voice source was instructed to produce voicing throughout the speech signal. An example of articulator movements generated by the model is given in figure F.9. In this figure it can be seen that the articulators start from a rest position, move fluently towards three different target values and then return to their rest positions. The sequencer was instructed to pursue each articulatory goal for 0.1 seconds. It can also be seen in the figure that sometimes the articulators get blocked

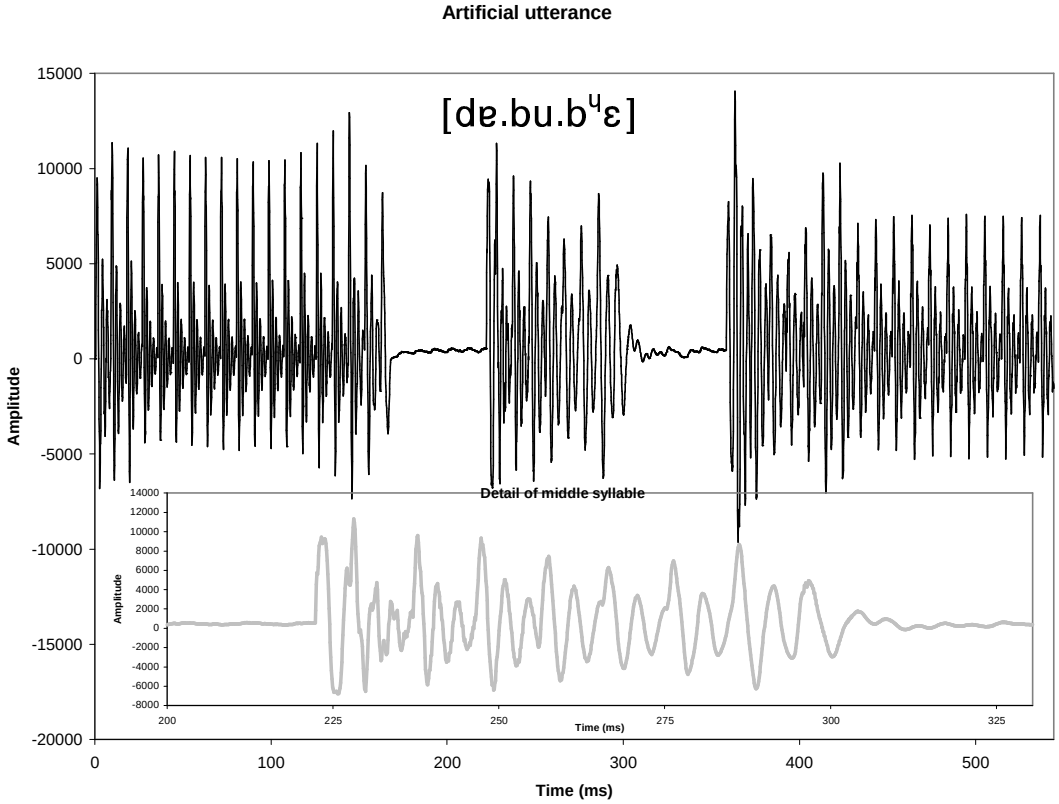


Figure F.10: Acoustic signal of artificial utterance.

and cannot move further towards their articulatory goals (for example at approximately 0.15 seconds). However, not all articulators get blocked simultaneously. For example, the hyoid x position does not get blocked in this example. Furthermore the same overshoot that was already seen in figure F.6 can be observed. Overshoot seems to be most present in the movements of the lips. This is to be expected as the lips are the articulators that are capable of the highest acceleration, and whose parameter range is greatest.

The acoustic signal that results from these articulator movements is presented in figure F.10. It can be seen that the signal does not simply consist of three phonemes in sequence, but rather of three syllables in sequences. Apparently the movement of the articulators from and towards the rest position also influences the sound that is produced. The utterance itself can be transcribed phonetically as [dɐ.bu.bᵘɛ] but in fact boundaries in the signal (except for the stops) are as hard to find as in human speech. This is illustrated with the middle part of the signal that is enlarged in the figure. Here one can see that the stop really merges fluently into the following vowel. To the human observer the signals that are produced sound like the babbling of an infant.

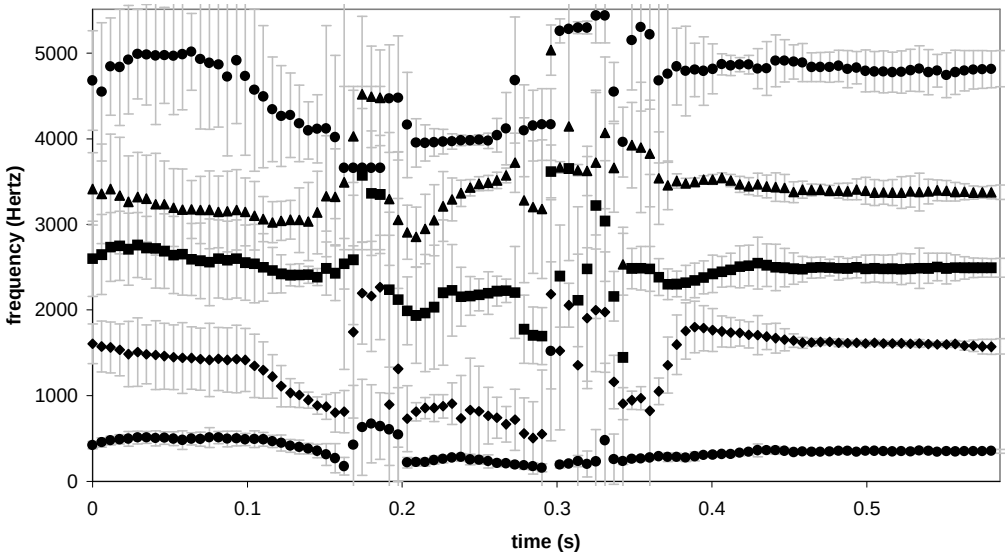


Figure F.13: Formant frequency and bandwidth of signal.

The voicing properties and (the logarithm of) the power of the signal are shown in figure F.11. It can be seen that power and voicing prominence drop enormously when the stops in the signal are articulated. It can also be seen that the pitch of the voicing drops over the utterance. This was because an (artificial) intonation contour was added to the signal in order to make them more realistic for human observers. Even though the signal was generated without noise, the properties fluctuate. This has to do with the fact that the window that is used to take the samples is not in any way synchronised to the periodic fluctuations in the speech signal. Sometimes there are huge peaks in the voicing frequency. This is because there are sometimes higher peaks than the one that actually corresponds to the frequency of voicing in the relevant range of the autocorrelation function. Why this happens is not clear. Apparently the estimation functions are quite sensitive to the exact alignment of data and the sample window.

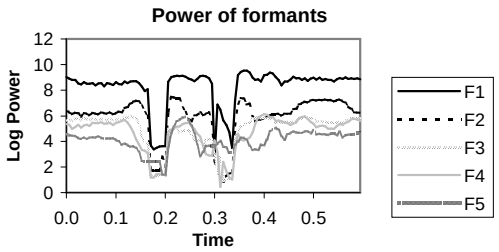


Figure F.12: Power of formants.

In figure F.13 the formant frequencies and their bandwidths (as error bars) are shown. It can be seen that the frequencies and bandwidths of the formants are well defined in the parts were the signal is clearly voiced. However, in the parts of the signal around the stops, the bandwidths and frequencies tend to be more confused. They do seem to be sufficiently well defined for recognising signals.

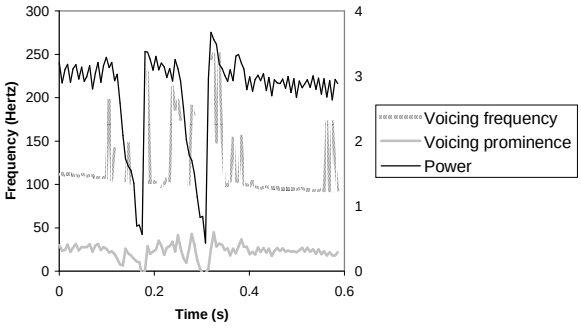


Figure F.11: Voicing and power of signal.

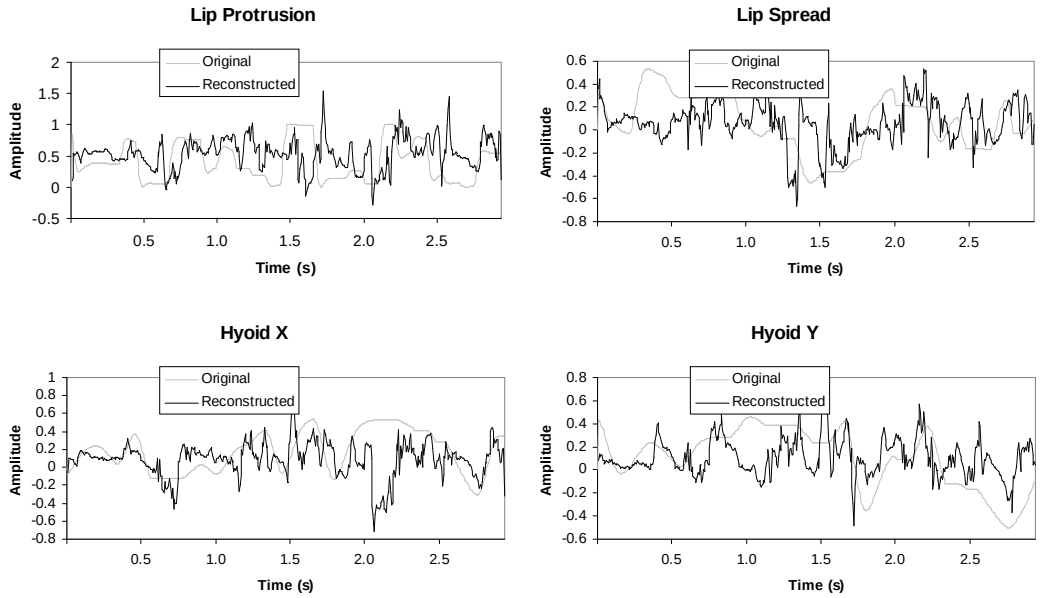


Figure F.14: Actual and reconstructed movements of lips and hyoid.

The power of the formant frequencies is shown in figure F.12. It can be seen that the power of the higher formants tends to be lower than the power of the lower formants. Their power also drops at the times when the vocal tract is blocked because of the stop articulations. Because their behaviour is so predictable, it seems that the power of the formant frequencies does not add much information.

F.3.2 Inverse mapping of a complex utterance

The second kind of experiment that has been done with the complex model was to see whether it was possible to learn the mapping between acoustic properties of the signal (such as the ones shown in figures F.11, F.13 and F.12) and the articulatory movements (such as the one shown in figure F.9) or the articulatory commands (for example, the goal values of figure F.9). A number of different learning algorithms were tried out, such as back propagation networks (see e.g. Hertz *et al.* 1991, Haykin 1994), Elman networks (Elman 1990), storage of prototypes and least mean squared error methods. None of these seemed to work. It is important that an agent be able to do an inverse mapping from a signal it perceives to an articulatory action in order to make a first approximation for imitation and in order to improve existing phonemes. Starting a hill-climbing procedure from a fixed point in articulatory space, as was done in the simulations with vowels, would not work because of the higher number of degrees of freedom and the time it takes to articulate a sound. As an example the attempt to use a least mean squared error method to learn the inverse mapping will be presented.

F.3.3 The Least Mean Squared Error method

The least mean squared error (LMS) method tries to find an optimal solution to the following problem:

F.25) $\vec{y} = A\vec{x}$

Where $\vec{y}$ is a given vector of length m, A is a given matrix of size n by m and $\vec{x}$ is a vector of length n that has to be calculated. If $m > n$, the problem is underdetermined, and the equation will not have an exact solution. It is therefore necessary to find a compromise vector $\vec{x}'$, which when multiplied with A will result in a vector

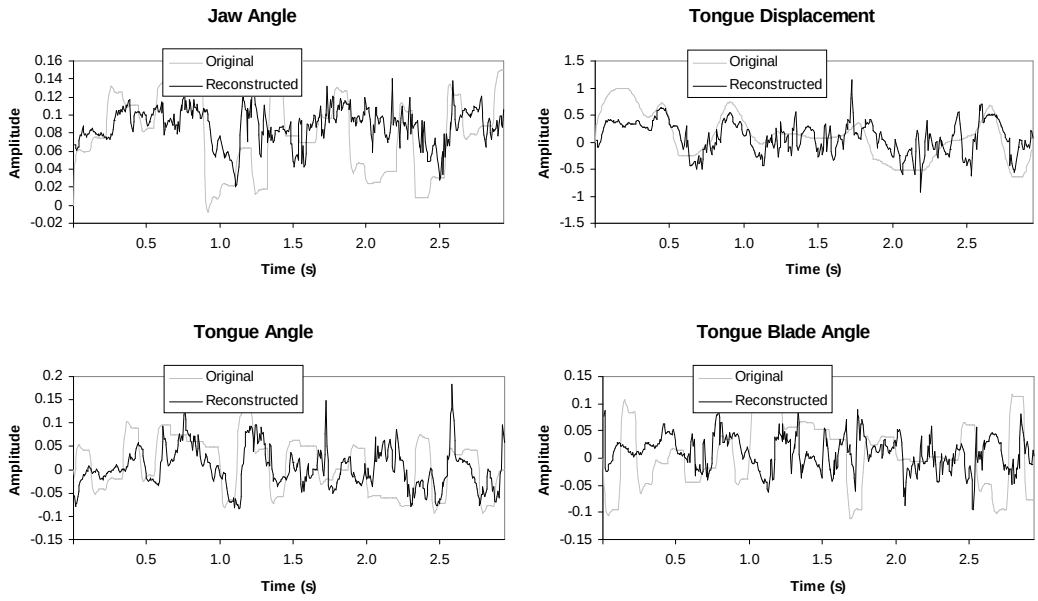


Figure F.15: Actual and reconstructed movements of jaw and tongue.

$\vec{y}'$. A good criterion for finding a compromise vector $\vec{x}'$ is to minimise the squared difference e between vectors $\vec{y}$ and $\vec{y}'$:

F.26) $$e = \sum_{i=1}^{m} \left(\vec{y}_i - \vec{y}'_i \right)^2 .$$

Where $\vec{y}_i$ and $\vec{y}'_i$ are the i^{th} element of $\vec{y}$ and $\vec{y}'$, respectively. This problem is well known, and the algorithm for solving it was adopted from Press *et al.* (1992) chapter 15.

The LMS method can be used for tackling the inverse mapping problem. In this problem, it is necessary to estimate articulator positions from acoustic data. The agent can create a set of example forward mappings by making random movements with its articulator and recording the articulator positions and the acoustic properties for a large number of time steps. It can now be tried to find a linear combination of acoustic properties that best predicts the articulator positions. In terms of equation F.25, A is the set of acoustic properties for all time steps, $\vec{y}$ is the articulator position to be predicted and $\vec{x}'$ is the set of linear coefficients to be found. The LMS problem is solved for every articulatory degree of freedom. The coefficients that are found can then be used to predict the articulator positions corresponding to new acoustic properties that are observed.

Unfortunately, this method does not work very well if all degrees of freedom of the articulatory model are used. Figures F.14 and F.15 show the original movements (in grey) and the reconstructed movements (in black) of the eight different articulatory degrees of freedom that were used. The ninth degree of freedom, the velum, was not used, because a nasal tract could not be modelled.

As can be seen in the figures, the reconstruction of the signals is not very good. Sometimes the original signal is followed quite closely, but usually there is not much agreement. Also, the reconstructed signal is very noisy, whereas the original signal is smooth. This is due to the high-frequency fluctuations in the acoustic properties of the signal, which could already be observed in figures F.11, F.13 and F.12. In order to give a numerical impression of the agreement between the original

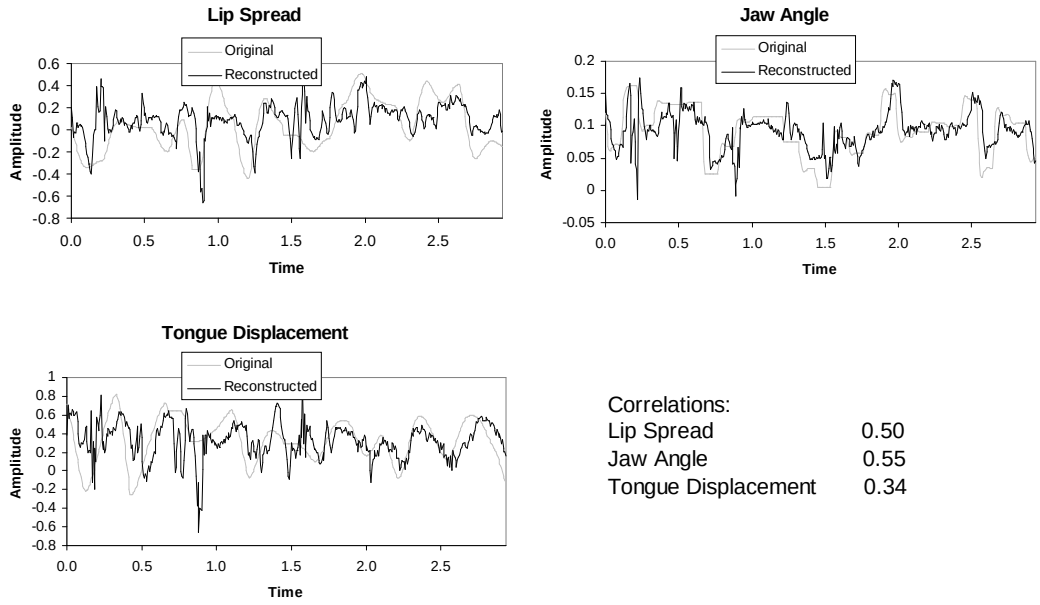


Figure F.16: Reconstruction of limited movements.

and the reconstructed movements, the correlations have been calculated and are given in table F.1.

Lip Protrusion	0.01	Jaw Angle	0.37
Lip Spread	0.31	Tongue Displacement	0.65
Hyoid X	0.13	Tongue Angle	0.35
Hyoid Y	0.32	Tongue Blade Angle	-0.15

Table F.1: Correlations of original and reconstructed movements.

Most of the correlations are rather bad. Apparently tongue displacement can be reconstructed the best. This was to be expected as this is directly related to the frequency of the second formant. Tongue angle and jaw angle are directly related to the first formant, so these are also reconstructed with some accuracy. Apparently hyoid vertical movement is also related with the acoustic signal to some extent. Hyoid horizontal movement, lip protrusion and tongue blade angle cannot be reconstructed to any reasonable degree.

F.3.4 Inverse mapping of a simple utterance.

It is somewhat easier to reconstruct a signal if only a few articulatory movements are allowed. If one moves only the articulators that are relevant for vowel production, tongue displacement, tongue angle and lip spread, reconstruction is slightly better. The results of this experiment and the correlation values are shown in figure F.16. Here the agreement is better, but there is still a lot of noise on the reconstructed signal. The correlations are not very impressive either.

Unfortunately, other learning methods suffer from the same problems. The experiments that were tried with neural networks, for example, resulted in the network learning to produce a constant output. The constant values were the rest positions of the articulators. Apparently this minimised the error so far that further improvement through learning became impossible. Possible solutions to the problem of learning the inverse mapping will be discussed in the conclusion of this appendix.

F.4 Conclusion

Both the production and perception models work satisfactorily. The main problem seems to be the mapping from acoustic signals to articulatory movements. When this mapping has been made, articulatory goals can be extracted from the speech stream and learning of the words and phonemes can proceed in ways similar to how multiple word utterances have been learned in Steels' framework (1998a). However, none of this has been implemented yet, because the first step of mapping acoustic signals onto articulations has turned out to be so difficult.

Of course there are still a lot of things that can be improved in articulation and perception. First of all, the articulatory model must be extended with a nasal passage. Nasals are very frequent in human languages and appear very early in infant speech. Investigating the emergence of sound systems without the ability of producing nasals therefore does not seem right. Recently a description of the nasal passage has been located in Boersma (1998). Also, the way in which the articulators move and the way in which blocking of articulators is handled can be improved a lot. Right now *ad hoc* solutions have been adopted. Better solutions can be found in the literature (see e.g. Browman & Goldstein 1995; Kaburagi & Honda 1996). Furthermore, the model should probably be controlled with parameters that are more directly related to the shape of the vocal tract, such as specifications of place and degree of constriction, rather than the direct manipulation of angles and distances used so far. This has also been done by the people at Haskins laboratories that developed the Mermelstein synthesiser in the first place (see Saltzman 1986; Saltzman & Munhall 1989; Saltzman 1995). This would probably make the task of finding the inverse mapping easier.

Perception is also still a problem that has not completely been solved. It is now possible to calculate a number of properties of the speech signal. One important thing that still needs to be implemented is the detection of noise in the speech signal, such as occurs in fricatives and the release of plosives. For recognition it is probably also necessary to be able to extract higher level features, such as where utterances begin, where the cores of the syllables occur and possibly the features of a larger stretch of the signal, such as intonation contours. Probably more research on how people perceive sound (see e.g. Handel 1989) should be taken into account.

It is always a pity to conclude that much still has to be done and that the goals that were aimed for were not achieved. However, the aim of this appendix was to provide researchers who want to continue this research with sufficient details and background to be able to do so without having to explore all the disparate fields of speech production, perception and machine learning. The work has just begun.

Appendix G: Languages Used

A number of languages have been used as examples in this thesis. Some of these languages are well known while others are relatively exotic. In this appendix all the languages that have been referred to in this thesis are listed alphabetically. Per language its genetic affiliation, its (original) geographic location, its number of native speakers and its vowel inventory are presented, as well as a reference to information on its phonology if it is a less well known language. Also a world map is presented with the locations of all the languages referred to in the thesis. Sources for this information, apart from the original sources mentioned at every language entry individually have been Vallée, (1994) the electronic version of $UPSID_{451}$ (Maddieson en Precoda 1990) and the online version of the Ethnologue (originally Grimes 1996) at: http://www.sil.org/ethnologue/ethnologue.html

G.1 Chamorro

Chamorro (Seiden 1960) is an Austronesian language of the Western-Malaya Polynesian group spoken by approximately 78 000 people in Guam and the Northern Mariana Islands. Its vowel system consists of [i], [‘e’], [æ], [ɑ], [‘o’] and [u].

G.2 Dutch

Dutch is an Indo-European language of the West-Germanic group spoken by approximately 20 000 000 people in the Netherlands, the North of Belgium and in a small corner in the Northwest of France. Its vowel system (with diphthongs) consists of [iː], [yː], [ɪ], [ɛ], [aː], [ɑ], [ɔ], [uː], [ə], [ɵ], [e^{i}], [$ø^{y}$], [o^{u}], [$æ^{i}$], [$ɑ^{u}$] and [$œ^{y}$] although there is considerable dialectal variation, especially in what is realised as a diphthong and what not.

G.3 English

English is an Indo-European language of the West-Germanic group spoken by approximately 322 000 000 people in the United Kingdom, the United States of America, Canada, Australia and ex-colonies of the UK. This is no doubt the best studied language in the world, so it should be not surprising that there is a huge controversy about the exact vowel system of English. Usually it is considered to be something like this: [iː], [ɪ], [ɛ], [æ], [ɑː], [ɒ], [ɔ], [ʊ], [uː], [ʌ], [$e^{ɪ}$], [$o^{ʊ}$], [$a^{ɪ}$], [$a^{ʊ}$] and [$ɒ^{ɪ}$] with considerable dialectal variation.

G.4 French

French is an Indo-European language of the Romance group spoken by approximately 72 000 000 people in France, Belgium, Switzerland and Canada and in France’s ex-colonies. French is also quite well researched, so also here there is some controversy about its vowel inventory. $UPSID_{451}$ gives: [i], [y], [e], [ø], [ɛ], [ɛː], [œ], [a], [ɒ], [ɔ], [o], [oː], [u] and nasalised vowels [ɛ̃], [œ̃], [ɒ̃] and [õ], although there is also dialectal variation.

G.5 German

German is an Indo-European language of the West-Germanic group, spoken by 98 000 000 people, mainly in Germany and Austria. Also used as a national lan-

guage in Switzerland, but the colloquial dialect is usually considered to be a different language. There are very many dialects, some of which are mutually incomprehensible. Depending on the dialect and the person who makes the analysis, the vowel inventory differs. UPSID states the following vowels for German: [ɪ], [ɛ], [ɐ], [ʏ], [‘ə’], [œ], [ʊ], [ɔ], [iː], [eː], [yː], [øː], [ɛː], [aː], [uː] and [oː].

G.6 Hakka

Hakka (Hashimoto 1973) is a Sino-Tibetan language of the Sinitic group. It is spoken by approximately 34 000 000 people, mainly in China’s Guangdong province. It’s vowel inventory consists of: [i], [ɛ], [æ], [a], [ɔ] and [u].

G.7 Kabardian

Kabardian (Choi 1991) is a Northwest-Caucasian language, spoken by approximately 647 000 people in the Caucasus, in Russia and Turkey. An interesting feature of the language is its small and “vertical” vowel inventory: [ɨ], [ə] and [a], although the actual phonetic realisation of these sounds is influenced to a large extent by the preceding consonants.

G.8 Murá-Pirahã

Murá-Pirahã or Pirahã (Everett 1982) is a South-American language of the Mura-group, spoken by approximately 250 people in the Northwest of Brazil. An interesting feature is its small phoneme inventory, (only 11 phonemes in total) although it does have two contrasting tones. Its vowel inventory is: [i], [a] and [o].

G.9 Norwegian

Norwegian (Vanvik 1972) is an Indo-European language of the North-Germanic group, spoken by approximately 4 000 000 people in Norway. There are two main varieties of the language, Bokmal and Nynorsk. The variant that is referred to in this thesis is standard Eastern Norwegian. According to UPSID_{451}, it contains the following vowels: [ɪ], [ʏ], [‘ɐ’], [æ], [œ], [a], [ɑ], [ɒ], [ʉ], [ʊ], [‘ə’], [iː], [yː], [eː], [øː], [æː], [‘oː’], [ʉː], [uː] (which Valleé (1994) and Schwartz *et al.* (1997a) analyse as 15 different vowel qualities) and diphthongs: [æ$^{\text{i}}$], [æ$^{\text{ʉ}}$], [a$^{\text{i}}$], [ɔ$^{\text{y}}$] and [ø$^{\text{y}}$].

G.10 Rotokas

Rotokas (Firchow & Firchow 1969) is an East-Papuan language spoken by approximately 4 000 people on Bougainville Island. An interesting feature is its extremely small phoneme inventory, totalling only 11 phonemes, although there is quite a lot of allophonic variation. Its vowel inventory is: [i], [‘ɐ’], [a], [‘o’] and [u].

G.11 Saami

The Saami language is spoken by the Saami (or Lapp) people of Northern Scandinavia and bordering parts of Russia. The variant that is used in this thesis is a Southern variant spoken mainly in the Västerbotten province of Sweden. Its vowel inventory is [i], [‘ɐ’], [a], [‘o’], [u] and [ɨ], although there is considerable allophonic variation due to neighbouring consonants.

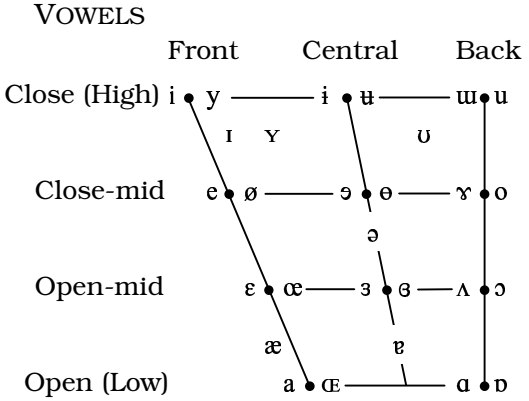
G.12 !X•

!Xũ (Snyman 1970) is a Khoisan language spoken by approximately 5 000 people in Angola and Namibia. An interesting feature is its huge phoneme inventory, consisting of 141 phonemes, many of which are clicks. Its vowel inventory is: [i], [iː], [ĩ], [ʻe’], [ʻeː’], [a], [aː], [ã], [aˠ], [aˠː], [ãˠ], [ãˠː], [ʻo’], [ʻoː’], [õ’], [õː’], [ʻoˠ’], [õˠ’], [ʻoˠː’], [õˠː’], [u], [ũ], [uː], [ũː] and diphthongs: [əi], [əĩ], [əu], [əũ], [aˠe], [aˠẽ], [aˠo], [aˠõ], [ae], [ao], [ɔˠi], [ɔˠĩ], [oi], [oĩ], [ia], [oe], [oa], [ui], [oãˠ], [oã] and [uĩ].

Appendix H: The International Phonetic Alphabet

The following tables contain the symbols that form the International Phonetic Alphabet. The first table contains the IPA-vowel symbols. The second table contains the consonants that are produced with pulmonic egressive air stream, (air flowing out from the lungs) ordered according to place (horizontal) and manner (vertical) of articulation. In each column, symbols representing voiceless sounds are at the left of the column, while symbols for voiced sounds are at the right. Shaded areas indicate articulations that are impossible. The next table contains other consonants, produced with other air stream mechanisms, such as clicks implosives and ejective consonants. The last table contains miscellaneous consonant symbols (mostly consonants with double articulations). For more in depth information on phonetics and phonetic symbols, see (Ladefoged 1981, Ladefoged & Maddieson 1996, Pullum & Ladusaw 1996).

VOWELS



PULMONIC EGRESSIVE CONSONANTS

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

CLICKS		VOICED IMPLOSIVES		EJECTIVES	
ʘ	Bilabial	ɓ	Bilabial	’	Such as:
ǀ	Dental	ɗ	Dental/alveolar	p’	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	t’	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	k’	Velar
ǁ	Alveolar lateral	ʛ	Uvular	ʃ’	Postalveolar fricative

OTHER SYMBOLS			
ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʃ͡ʃ	Simultaneous ʃ and x
ʜ	Voiceless epiglotal fricative	Affricates and double articulations can be represented by two symbols joined by a tie-bar if necessary:	
ʢ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive	k͡p	t͡s

IPA SUPRASEGMENTAL SYMBOLS			TONES AND WORD ACCENTS			
ˈ	Primary stress	ˌfoʊnəˈtɪʃən	Level		Contour	
ˌ	Secondary stress					
ː	Long	aː	a̋ or ˥	Extra high	ǎ or ˩˥	Rising
ˑ	Half-long	aˑ	á or ˦	High	â or ˥˩	Falling
˘	Extra short	ă	ā or ˧	Mid	ǎ or ˦˥	High rising
.	Syllable break	ɹi.ækt	à or ˨	Low	à or ˩˨	Low rising
|	Minor (foot) group		ȁ or ˩	Extra low	ȃ or ˦˩	Rising-falling etc.
‖	Major (intonation) group		↓	Downstep	↗	Global rise
‿	Linking (absence of a break)		↑	Upstep	↘	Global fall

IPA DIACRITICS			Diacritics may be placed above a symbol with a descender e.g. ŋ̊					
◌̥	Voiceless	n̥	d̥	◌̤	Breathy-voiced	b̤	a̤	◌̪	Dental	t̪	d̪
◌̬	Voiced	s̬	t̬	◌̰	Creaky-voiced	b̰	a̰	◌̺	Apical	t̺	d̺
◌ʰ	Aspirated	tʰ	dʰ	◌̼	Linguolabial	t̼	d̼	◌̻	Laminal	t̻	d̻
◌̹	More Rounded	ɔ̹		◌ʷ	Labialised	tʷ	dʷ	◌̃	Nasalised		ẽ
◌̜	Less Rounded	ɔ̜		◌ʲ	Palatalised	tʲ	dʲ	◌ⁿ	Nasal Release		dⁿ
◌̟	Advanced	u̟		◌ˠ	Velarised	tˠ	dˠ	◌ˡ	Lateral Release		dˡ
◌̠	Retracted	i̠		◌ˤ	Pharyngealised	tˤ	aˤ	◌̚	No Audible Release		d̚
◌̈	Centralised	ë		◌̴	Velarised or Pharyngealised	ɬ̴					
◌̽	Mid-centralised	e̽		◌̝	Raised	e̝	(ɻ̝ = voiced alveolar fricative)				
◌̩	Syllabic	ɹ̩		◌̞	Lowered	e̞	(β̞ = voiced bilabial approximant)				
◌̯	Non-syllabic	e̯		◌̘	Advanced Tongue Root	e̘					
◌˞	Rhoticity	ə˞		◌̙	Retracted Tongue Root	e̙					

Index