3

Evolving Sound Systems

Human languages use an amazing variety of subtly different speech sounds to convey meaning. With the exception of sign languages that are used and developed by communities of deaf people, all human languages use sound as the primary signal. The sounds, or more accurately the differences between sounds, that humans use for distinguishing meanings can be very subtle. Two different sounds that would be perceived as identical by a speaker of one language might make an important distinction in meaning in another. For example, in the Bahing language of East Nepal, the word /mərə/ means "monkey", while the word /mɯrɯ/ means "man". Speakers of neighboring and European languages alike are generally not able to perceive this distinction, an unlimited source of fun to the Bahing people.

In the UCLA Phonological Segment Inventory Database (UPSID), a database that now contains 451 languages (Maddieson 1984, Maddieson & Precoda 1990) 921 different speech sounds occur. The language with the largest inventory of speech sounds in the database is !Xũ (Snyman, 1970, 1975), a Khoisan language of Southwest Africa with 141 sounds, while the languages with the smallest inventories are Rotokas (Firchow & Firchow, 1969) a East-Papuan language and Mura-Pirahã (Sheldon 1974, Everett 1982) a South-American language, both with only 11 sounds. According to Maddieson (1984) usually languages have between 20 and 37 sounds in their repertoires. However, these repertoires are not chosen randomly. Some sounds occur much more often in the languages of the world than others. Lindblom and Maddieson (1988) have found that languages tend to use a set of basic articulations first. Such basic articulations are simple articulations that involve only one articulatory gesture and minimal displacements of the articulators. When the repertoire becomes larger, languages tend to use what Lindblom and Maddieson call 'elaborate' articulations, which involve larger displacements and simultaneous actions of multiple articulators. Finally, when a language's repertoire becomes even larger, 'complex' articulations will be used. These consist of combinations of the two previous types.

There are other patterns to be found in the sound repertoires as well. Examples of such patterns are symmetries. In consonant systems, for example, if a language has a voiced sound at a certain place of articulation, it is very likely to have a voiceless sound at the same place of articulation. Comparable symmetries are found in vowel systems.

Regularities are not just found in the repertoires of sound systems, but also in the way sounds are combined into words and syllables. It is possible to make a hierarchy of sounds with respect to whether they tend to occur close to or far from the nucleus of a syllable. This hierarchy is called the *sonority hierarchy* (Vennemann 1988). Some sounds, such as vowels, are very sonorous and tend to occur at the nucleus of a syllable, while others, such as plosive consonants (p, b, t etc.) are little sonorous and tend to occur at the periphery of a syllable. Whenever sounds occur in sequence, it turns out that they almost always increase in sonority towards the nucleus of a syllable. For this reason, "play" is a possible word in English, while "*lpay" is not.

Phenomena that occur in many languages are often called *universals*. Although the term universal implies validity for all languages, there are very few non-trivial phenomena that occur in all known human languages. For this reason the term universal is often used for phenomena that occur in a (large) majority of human languages. All parts of language: syntax, morphology, semantics, phonology, can have their own universals. This paper will concentrate on universals that have to do with sound systems.

Universals might be explained in different ways. The first possible explanation would be that all languages are historically related. Although there is still some controversy over the exact evolution of Homo sapiens, it is most likely that modern humans came from Africa some 200 000–300 000 years ago. Genetic diversity within the species Homo sapiens is so small that it is very likely that at one time in its early history the species must have consisted of only a few thousand individuals. It is not unlikely that all these individuals spoke dialects of the same language. However, given the speed with which languages change, and given the amount of time during which different groups of humans have been isolated from each other, it is highly unlikely that any trace of the original relationship between all human languages remains. Tentative reconstructions of "proto-world" (Ruhlen 1994) although enthusiastically embraced by the popular press, should be regarded with the utmost scepsis. Another reason why deep historical relations between human languages alone cannot explain universals is that there are also universals of language change (e.g. Labov 1994). Quite different languages seem to change along similar paths.

A second possible explanation is that language universals are a reflection of innate human capacities for language. Such an innate capacity does not only have to be in the form of a "universal grammar" as investigated by some researchers, but could also consist of more general cognitive mechanisms that are used for using and learning language. The innate capacity for language is also determined by physical and physiological factors, such as the shape of the vocal tract, accurate control over breathing and the way the ear processes sound. Innate factors obviously play a role in determining universals of human language. However, the problem of innate factors as explanation for language universals is that they themselves have to be explained as the result of evolution, or possibly as exaptations of pre-existing body parts and cognitive mechanisms.

This leads to the third possible explanation of language universals: that they are functional optimizations for communication over a noisy channel. Human language seems to be optimized for communication in a number of respects. The frequency

with which different vowels occur in human language can for example be explained by the optimization of acoustic distinctiveness. If one optimizes a system with a fixed number of vowels so that the average distance between them is maximized, systems that occur frequently in human languages tend to appear. Now such functional optimization could be a result of the interactions between the speakers, listeners and learners of a language or the result of an evolutionary process. Also, the preference for languages that are functionally optimal over languages that are not could, over a long period of time, influence the mechanisms that are used for learning language through a process that is called the Baldwin effect (Baldwin 1896).

Possibilities for Modeling

The role of innate properties versus the role of functional optimization and the way by which the different human adaptations to speech have evolved can be investigated with computer models. Traditionally, linguists prefer to solve theoretical disputes with linguistic data and physical, cognitive or philosophical arguments. However, language origins and evolution can hardly be investigated by looking at modern languages, and the complexity of theories of evolution of populations is such that their behavior cannot be predicted by simple philosophical argument. For this reason computer models are used more and more to test and create hypotheses. The study of speech has a long tradition of using computer and other electronic equipment. Due to the fact that speech works with objectively measurable and recordable signals, it can be manipulated relatively easily. From the nineteenfifities onwards important discoveries were made by manipulating recorded signals and synthesizing artificial ones. Another advantage of the fact that speech signals can be measured in a relatively objective way, is that predictions of models can be easily compared with observations of real language data.
Different aspects of the evolution of speech can be investigated with computer models. One can try to reconstruct the evolution of the human vocal tract, one can use computer simulations to find out what factors (such as articulatory ease, acoustic distinctiveness etc.) have played a role in evolution, but one can also use computer models to investigate how much of speech is learnt and how much of it is innate.

Different approaches to modelling speech

One interesting and important way in which computer models have been used to study the evolution of speech (and language indirectly) is by reconstructing the vocal tract of fossil hominids, most notable Neanderthals. These vocal tract models can then manipulated and excited with an artificially generated glottal pulse. By studying the resonances of the model, the range of possible vowel sounds that could be made by the hominid under study can be estimated. Although this technique comes closest to actually being able to listen to our hominid ancestors,

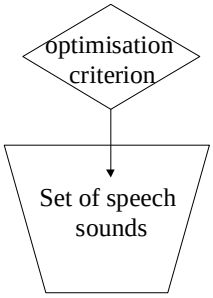

Figure 1: Schematic view of optimisation.

the technique is not quite uncontroversial, mostly because important parts of the vocal tract (tongue, pharynx, larynx) do not fossilize very well. Interesting and exciting as these results may be, they do not quite model the origins and evolution of speech (they only reconstruct one stage of the evolution from fossil data) so they fall somewhat outside the scope of this chapter.

Apart from such direct modeling techniques, roughly three computational paradigms have been used for investigating the evolution of speech. The first paradigm is that of straightforward optimization of sound systems on the basis of different criteria. The paradigm is illustrated in figure 1. The figures are added for illustration, but also to be able to compare the different paradigms at a glance.

The optimization criteria include factors such as acoustic distinctiveness, acoustic stability, articulatory ease or learnability. Through optimizing different (combinations of) criteria and checking whether the sound systems that are predicted conform to what is found in human languages, one can find out what criteria are important for the formation of human sound systems.

Optimization is probably the technique that is least controversial in its applications, as its dynamics are relatively simple: there is an optimization criterion and it results in sound systems that look like human sound systems or not. Discussion is possible on the implementation of the optimization criteria or on the interpretation of the sound systems that are found, but the optimization process itself is not controversial. The relative simplicity of optimization is also a disadvantage. It can only be applied to relatively simple problems. As soon as multiple optimization criteria interact, the optimization process becomes more difficult and decisions have to be made about which solutions to investigate. Also, the relative importance of the different criteria and the way they interact might be controversial. However, optimization is a good technique for checking which criteria play a role in determining the sound systems that are found in human languages. How these criteria have become important and how the optimization process takes place in human language use and learning can then be investigated with different techniques.

The second paradigm is that of genetic algorithms (GA's). The genetic algorithm is a technique that is based on the way evolution works in nature. The algorithm has a population of potential solutions, all of which are coded as artificial genes (usually in the form of bit strings). These genes are converted into possible solutions to the

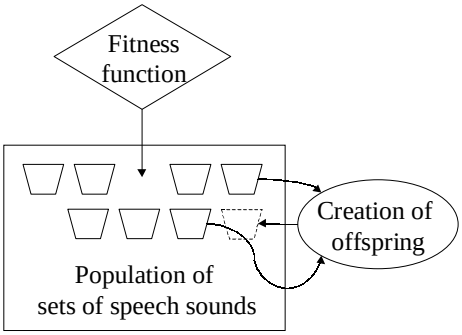

Figure 2: Schematic view of a genetic algorithm.

problem at hand (sound systems in the case of evolution of speech) and are evaluated with a fitness function. This fitness function is a function that gives a high value for good solutions and a low value for bad solutions. Just as in nature, solutions with a high fitness are allowed to create offspring, while bad solutions are removed from the population. The genes of the offspring are created by combining the genes of the parent solutions. Usually combination methods inspired by nature, such as mutation and crossover, are used. It is clear that for the proper functioning of a genetic algorithm the right fitness function as well as the right coding of solutions in genes are essential. The GA is illustrated in figure 2.

Basically, GA's also optimize on the basis of an optimization criterion (the fitness function), but they are much more flexible and robust than straightforward optimization algorithms. They can therefore be used to model more complex optimization problems and even problems in which the optimization criterion changes over time. Also, GA's work with a population of solutions, instead of with a single one. This is more realistic in the case of language, as language is typically used in a group of individuals rather than by a single individual. Finally, genetic algorithms are modeled after Darwinian evolution, and are as such ideally suited for modeling real evolution.

Their resemblance to real biological evolution is possibly the biggest advantage of genetic algorithms for research into the evolution of speech. But modelers who enthusiastically embraces genetic algorithms as their paradigm of choice should be aware that there are a large number of design decisions to be made in building a GA for investigating the evolution of speech. Decisions have to be made what to encode as genes and how to implement the fitness function. Also, it is very important to not confuse biological evolution of the human faculty for speech and cultural evolution of human languages. Although historical relations between languages and historical change of languages are often expressed in terms similar to those of biological evolution and although there are definite and valid similarities between the processes of biological evolution and language change, one should not confuse the two processes in one's model. The two processes are clearly distinct and operate on totally different time scales. They do influence each other, but this influence happens because the properties of a learned system (the

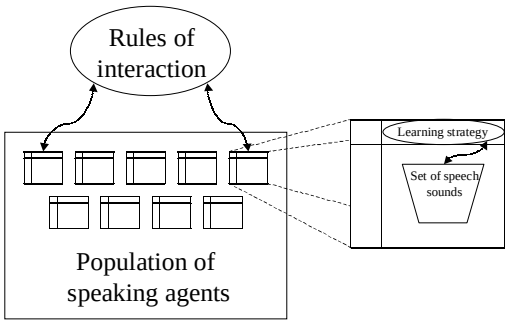

Figure 3: Schematic view of a language game.

language) influence the fitness of individuals that have to learn it, and is an interesting subject of investigation in itself.

The third paradigm is inspired by game theory and Wittgenstein's (1967) ideas on language games. Language games as a paradigm for modeling of evolution of language have first been used by Steels (1995, 1997) In this research the notion of a game is not very well defined, but language games have a number of properties in common. There usually is a population of agents that each have certain linguistic knowledge and that can interact with each other. The rules of the game determine how the interactions are structured and what information is exchanged. The agents can update their linguistic knowledge on the basis of the interactions they have taken part in. Usually all agents follow the same strategy for updating their knowledge. The language game paradigm is illustrated in figure 3.

Language games are a useful model of linguistic interactions between humans. The rules of the game and the strategy for updating an agent's knowledge can be varied to create different types of games for investigating different parts of language. Of course, one has to make simplifications while using language games. In real human language, different parts of language influence each other and interactions between language users can be highly complicated and dependent on extra-linguistic context. In this respect, the language game model is not different from other computational models of the study of language, but it is necessary to keep in mind what simplifications one has made and how these might influence the outcome of the games.

Strictly speaking, language games cannot be used to study the evolution of language, as the agents do not change over time. However, language games can be used to investigate to what extent properties of language can be explained as the result of interactions between agents and to investigate what must be programmed into the agent (i.e. what must be innate) so that it can learn a certain aspect of language. Such aspects as have to be pre-programmed will have to be explained by evolutionary models, such as genetic algorithms.

As both the genetic algorithm and the language game paradigm work with a population of agents, it is obvious that the two can be combined. However, not the sound systems would be coded into genes, but the properties of individual agents. In such a system it could be investigated, for example, how different learning

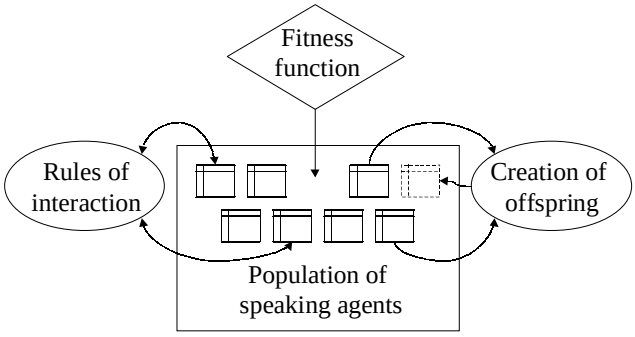

Figure 4: Combination of GA and language game.

techniques can evolve, or whether it is possible to reconstruct the evolution of the human vocal tract on the basis that it enables speakers to produce a wider range of possible speech sounds. The combination of the language game and the GA is illustrated in figure 4.

The combination of these two techniques makes it possible to investigate the interactions between biological evolution and cultural evolution, without running the risk of confusing timescales or genetically and culturally transmitted information, as mentioned above. Although the paradigm of language games with evolving agents is the one that comes closest to human reality, there are still a number of problems. All problems with respect to how agents are coded into genes, and how the fitness function is implemented also occur here, as well as the problems with respect to the simplification of interactions that were mentioned with the language games. Another important problem is that the combination of the two highly complex mechanisms might result in behavior that is hard to explain. It might not be possible anymore to determine which mechanism caused which part of the complete behavior, or to reconstruct how the system came up with the solution that was found.

Another problem with systems that work with populations of agents or sets of speech sounds and that have to simulate many iterated operations with these, is that their running time can become prohibitively long. For example, the most realistic speech synthesizers that exist take approximately 1000 times as long to calculate a speech signal than the actual duration of the signal. It is not possible to simulate a realistic number of interactions in a population of any size with such a model. It is therefore essential that the right simplifications be found and that reasonably realistic, but fast models of the speech phenomena under study be used. An important part of modeling the evolution of speech (and perhaps of any cognitive phenomenon) is therefore the trade-off between speed and realism.

Modeling different aspects of speech

Not only are there different possible approaches to the problem of modeling speech, there are also different aspects of speech that can be modeled. Here again, there is a trade-off between accuracy and speed. As speech sounds pronounced in sequence influence each other, and as this influence is of great importance to understand language change, it would be desirable to have a model that is as complete as possible. That is, a model that is able to produce a sequence of consonants and vowels as well as an intonation contour. However, there are a number of problems with modeling such complex utterances. The first problem, that was already mentioned in the previous section, has to do with the lack of speed of complex articulatory models. But this is not the only problem. Another problem is that actually very little is known about how sounds in sequence are produced, perceived and processed.

Linguists generally make descriptions of human languages in terms of *phonemes*, the sounds that are able to make distinctions in meaning. An example is the distinction between English /r/ and /l/ which have many minimal pairs (words that differ only in one sound, and that have different meanings) such as ‘rate’ and ‘late’. However, in a language such as Japanese, this particular distinction is not used, there are no minimal pairs with [r] and [l], and so in that language [r] and [l] are said to be *allophones* of one phoneme /l/.

Although phonemes have great descriptive value, it is not quite clear what their role is in storage and processing of speech sounds. It is quite possible that processing of speech is done on different levels of complexity, both on a level higher and lower than that of the phoneme. This is because when people pronounce words, they do not produce a string of nicely distinguishable phonemes. Instead, they produce a sequence of speech gestures that influence each other mutually, so that different phonemes overlap and become indistinguishable. Little is known about how this process works in articulation, and even less is known about how the speech signal is converted into strings of phonemes and words by the listener. Any model that works with complex utterances therefore has to make assumptions about how these processes work. But such assumptions reduce the realism that was sought by using more complex speech signals.

A final problem with modeling complex utterances with the computer is that inevitably time sequences have to be learnt. This is actually an area of machine learning that is very hard, and for which very few general purpose algorithms are available.

For the time being, all attempts at modeling have tried to tackle only a subset of the possible speech sounds and the possible speech universals. Successful models have been made of models and simple (abstract) syllables, while work is in progress on tone systems and intonation.

A short History of Modeling

Probably the first attempt at making a computer model to explain universals of speech sounds was made by Liljencrants & Lindblom (1972). This model performed an optimization of randomly initialized vowel systems with a fixed number of vowels. The optimization used a function that was based on the potential energy of repelling magnets or electrically charged particles with equal polarity (this potential energy is higher whenever such particles are closer together). By shifting the individual vowels in the system, this energy function was minimized. Liljencrants and Lindblom found that vowel systems that were optimized in this way showed remarkable similarities with vowel systems found in human languages, although there were some discrepancies. Later re-implementations of that used modified distance functions (e.g. Vallée 1994, Schwartz *et al*. 1997b) have succeeded in making progressively better approximations of human vowel systems.

Subsequently, Lindblom *et al*. (1984) have tried to use an optimizing model for explaining phonemic (that is combinatorial) coding of syllables. The syllables consisted of a simple consonant followed by a vowel. Although the systems that emerged were phonemically coded, their model has not had the success of the model for vowels, because there are many more parameters in it and it is much more difficult to replicate the results.

Only in the mid-nineties did work on explaining sound systems with computer models get a new impulse with systems that were based on populations of sound systems and agents. The first to make an agent-based implementation to investigate the emergence of vowel systems was Glotin (Glotin 1995; Glotin & Laboissière 1996; Berrah *et al*. 1996) of the Institut de Communication Parlée (ICP) in Grenoble, the same institute were Schwartz *et al*. (1997b) do their research. He made a model in which a population of talking agents tries to develop a shared repertoire of (a fixed number) vowels. His agents have both an acoustic and an articulatory representation of the vowels, and adapt their vowel systems on the basis of their interactions. The agents are also subject to a genetic algorithm, which is (according to Glotin, personal communication) not meant to be a model of actual biological evolution of the agents, but rather of the way sound systems are transferred from parents to children. This is a weak point of the research, as the influence of the genetic algorithm and the interactions between the agents are difficult to separate. Another problem with the model was that it was computationally too involved, and that therefore only few simulations with small populations and small numbers of vowels could be run. In a way, this work was ahead of the computing power of the time.

It has been at the basis of a number of subsequent research efforts, however. In the first place those of Berrah (1998) and myself (de Boer 1997, 2000; de Boer & Vogt 1999). Berrah's work was a direct continuation of Glotin's research. Berrah's model is a simplification of Glotin's model, in that the agents do no longer have an articulatory representation of the sounds they use, only an acoustic one. This reduces the computational load considerably and allows more experiments with larger populations and larger numbers of vowels to be run. Berrah extends Glotin's model by investigating what he calls the "Maximum Use of Available Features".

By allowing the agents to use an extra feature (which could be length, nasalization etc. in human languages, but which he models as an extra abstract dimension of the acoustic space) he shows that this is only used whenever the number of vowels in the agents' repertoires exceeds a certain threshold. His simulations also contain a genetic component, which makes it sometimes hard to tell when a particular phenomenon is due to interactions between the agents and when it is due to the actions of the genetic algorithm.

My own work has concentrated on predicting vowel systems from interactions in a population. The agents have both an articulatory as well as an acoustic representation of their vowels, but use a much simpler articulatory model than Glotin's model. Also, the agents do not evolve, although experiments have been done with changing populations (de Boer & Vogt 1999). They interact through language games (in this experiment called imitation games) only. It has been shown that vowel systems of human languages, and the relative frequencies with which they occur can be predicted quite well with this model.

More recently research has started to investigate syllable systems with genetic algorithms and population models relating in a similar way to the optimizing simulation used by Lindblom *et al*. (1984) as Glotin's, Berrah's and my own work relates to Liljencrants' and Lindblom's (1972) model. Redford *et al*. (1998, *to appear*) have made a model that is based on a genetic algorithm. The population consists of words, which in turn consist of a closed set of phonemes. Redford *et al*. use a number of rules that determine how hard it is to produce and perceive different combinations and sequences of phonemes. On the basis of this a fitness for all the words in the population is calculated and selection and recombination take place. They try out different combinations of rules and investigate which rules are most important to predict syllables that are like those found in human languages.

Other work on predicting properties of more complex utterances is underway, but still largely unpublished. Pierre-yves Oudeyer of the Sony computer science laboratory in Paris, France is working on predicting repertoires of syllables using more realistic signals. Emmanuelle Perrone of the Institut des Sciences de l'Homme is also working on predicting consonant-vowel syllables in the framework of imitation games. Eduardo Miranda of the Sony computer science laboratory in Paris, France is working on modeling intonation contours, while professor William Wang of the electronic engineering department of the City University of Hong Kong and co-workers Mieko Ogura and Jinyun Ke are working on modelling tone systems within the framework of genetic algorithms.

A case-study

In order to illustrate the ideas outlined above, a case-study will now be presented. As the work with which I am most familiar is my own, I will present my model of the emergence of vowel systems. At every point in the description I will discuss the design decisions that have been made. I will not present full details, as these can be found in the references (de Boer 1997, 1999; de Boer & Vogt 2000). Of course I do not mean to imply that my work is more interesting, or more typical

than the other work mentioned above. On the contrary, the fact that a genetic component is lacking in my system makes it somewhat different from most computational modeling of the origins of sound systems. However, the other work is best studied in the original sources. As a genetic component is a very important factor in modeling evolution and origins of language, I *will* discuss the possibilities of integrating my model with a genetic algorithm, although so far this has not been implemented.

Vowels were chosen as the subject of research for two reasons. First of all, they are the easiest speech sounds to model. Typically, a vowel signal is constant over time and both its articulatory and acoustic characteristics can be described by very few parameters: in my model three real numbers for articulation and four real numbers for the acoustic signal. Secondly, vowels are the speech signals for which most is known about their distribution over the languages of the world. This makes it relatively easy to compare results of simulations with what we know about real human languages. Easy and objective comparison with human language data makes simulations much more convincing for a linguistic audience.

It was decided to investigate change of vowel systems from a cultural perspective rather than from an evolutionary perspective, because vowel systems of human languages change over time, but continue to show the same near-universal characteristics. However, there are exceptional vowel systems that do not conform to the universals. Therefore, it would seem unlikely that a strong innate constraint determines their shape. Rather, as was proposed by Steels (1995) in the context of vocabulary, self-organization in a population might be the force that causes human vowel systems to show universal tendencies. Of course, genetic evolution has also played an important role in shaping the vocal tract, but this might then be considered as a process that is driven by cultural evolution.

Therefore, it was decided to leave out any genetic component in the first implementations of the model and rather to work with a population of agents playing language games. This also makes it easier to analyze the behavior of the system and to determine what phenomena are caused by which processes. Of course, genetic evolution of the agents can be introduced as well, and suggestions will be made as to where this could be done.

The agents that make up the population were designed to be as simple as possible while preserving the crucial characteristics necessary for investigating the characteristics of human vowel systems. They were equipped with a simple articulatory synthesizer that was based on measurements of vowel parameters taken from (Vallée 1994). This synthesiser takes as input the three articulatory parameters necessary to describe a simple vowel: position, height and rounding (Ladefoged and Maddieson, 1996) and outputs the first four formant frequencies. These represent the center frequencies of the four most important peaks of the vowel's acoustic spectrum. The agents' perception uses a distance function that is calculated in the space that has as dimensions the first and the so-called effective second formant. The effective second formant is the weighted sum of the three highest formants and represents the perceptual phenomenon that multiple peaks in the higher part of the spectrum can be replaced by one single peak and still be perceived as the same. The particular calculation used is adapted from (Mantakas *et al*. 1986).

The agents store vowels in terms of both acoustic and articulatory prototypes. There is a one-to-one association between the two types of prototypes. Prototypes are centers of categories. Whenever a signal is perceived, the distance to all the acoustic prototypes is calculated and the one that is closest is considered to be the one that is recognized. In the case of production, an articulatory prototype is chosen and the corresponding acoustic signal is produced, but noise is added to this by shifting the formant frequencies somewhat. During the process of learning a repertoire of vowels, prototypes can be added, deleted or shifted in order to match the vowels of other agents in the population more closely. For doing this, agents can only base themselves on the behavior of other agents; they cannot look at the other agents' vowel repertoires directly. Storing phonemes in terms of prototypes seems to be cognitively plausible. It has been observed that different types of speech signals are perceived in terms of prototypes (see e.g. Cooper *et al*. 1952; Frieda *et al*. 1999) and also that other linguistic and cognitive concepts are stored and processed in terms of prototypes as well (e.g. Lakoff 1987).

In a model of this kind, the interactions between the agents are as important as the architecture of the agents themselves. In human language, linguistic interactions do not just consist of an exchange of linguistic symbols. There is always a context, both in the form of a linguistic context and the situation in which the conversation is taking place. This situation has a physical aspect, i.e. the environment in which the conversation is taking place, but it also has a social aspect and a pragmatical aspect (and possibly other dimensions as well). All these aspects influence the linguistic exchange. It is clear that modeling a complete linguistic exchange is extremely difficult.

However, when one is only interested in the sounds of language, one can in principle ignore everything that has to do with meaning. Instead, one can use interactions that are based on imitation. In imitation, the same constraints on sound systems apply as in real linguistic interactions. For imitation to be successful, sounds have to be easily distinguishable, as well as easy to produce, just as they should be in a complete communication system. For this reason, the interactions between the agents in the system under study consisted of agents trying to imitate each other. In analogy with the term language game, these interactions will be called *imitation games*.

In an imitation game two agents are picked from the population at random. One of these agents is assigned the role of *initiator* of the imitation game, the other is assigned the role of *imitator*. Although the roles of the agents in an imitation game are not symmetrical, all agents in the population have equal probability to play both roles. Although it is the case that in human learning of sound systems the roles of infants and adults are not symmetrical, it was decided not to implement this in the model. First of all, it would have introduced more parameters and more arbitrary design decisions and secondly, the aim of the research was not so much to model the way sound systems are acquired, but to investigate whether universal tendencies of vowel systems can be explained as self-organization in a population of language users.

The initiator of the imitation game chooses a random vowel from its repertoire, and produces it, while adding a small amount of noise. The imitator perceives this sound, finds the acoustic prototype of the vowel from its repertoire that is closest to

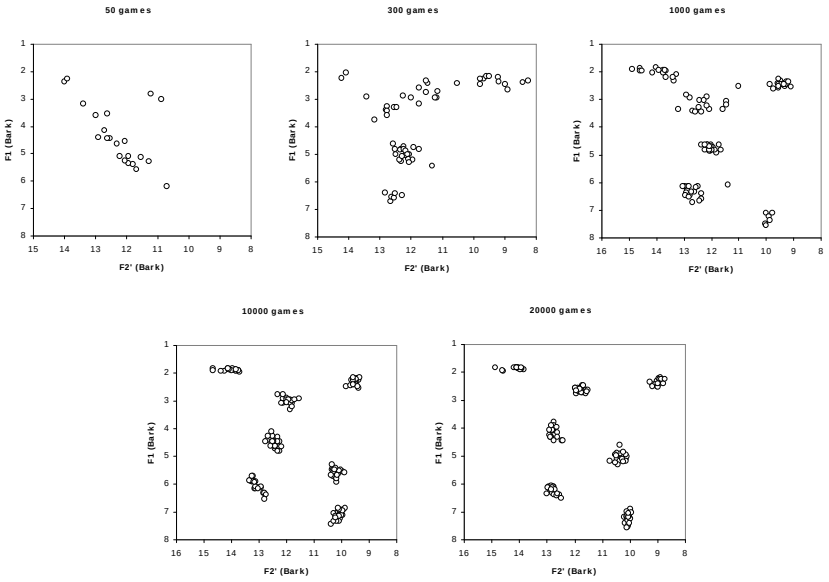

Figure 5: Emergence of a realistic vowel system.

it and produces the corresponding articulation, again adding noise. The initiator then perceives this signal, finds its closest vowel, and checks whether this is the same as the one it originally produced. If it is the same, it gives a "non-verbal feedback" to the imitator that the imitation was successful, while if it was not the same, it gives feedback that it was a failure. These steps include the main aspects of a linguistic utterance using a sound: production under constraints and with error, analysis in terms of a finite set of categories, and grounding of these categories outside the agent using non-linguistic cues. Although it is true that infants do not receive direct feedback about the quality of the sounds they produce, there must be a mechanism to provide a connection between meanings in the outside world and the sounds an infant perceives, otherwise an infant would not be able to learn which sounds in its language can distinguish meaning and which sounds can not. The feedback in the case of human infants learning language is probably derived from the extra-linguistic context in which the utterance takes place, or by the ability to achieve a goal with a given utterance or not.

In reaction to the feedback, and based on the success of the vowel in previous games, the imitator can shift the vowel it used or add a new vowel. Both agents also keep track of how many times the vowel was used and how many times it was used successfully. Also, both agents regularly throw away vowels that have been tested a few times and have been found to be unsuccessful most of the time, and merge vowels that are too close together. Finally, a random vowel can be inserted with low probability, in order to make sure that the agents' repertoires become as large as possible. The details of the way in which the agents update their repertoires will not be discussed here, but can be found in (de Boer 1997, 2000; de Boer & Vogt 2000).

The agents start out with an empty repertoire and are in principle able to produce all basic vowels. This means that the system is not biased towards any language in particular, and that the results of the simulations can therefore be assumed to say something about human language in general.

Running the simulations results in the emergence of realistic vowel systems. A representative example is given in figure 5. The figure consists of five frames, each representing a stage in the development of the vowel system. In each frame, the effective second formant and the first formant of all the acoustic prototypes of all the agents in the population are projected. The effective second formant is projected on the horizontal axis and the first formant is projected on the vertical axis. The usual directions of the axes are reversed, so that the vowels are projected in the way phoneticians usually project vowel systems, with [i] in the upper left corner, [u] in the upper right corner and [a] below. Note that not every point in the square can be reached by the agents' articulations. The available acoustic space is roughly triangular with the tip at the bottom of the graph.

The first frame shows the situation after 50 games. The agents start out empty, and as there has only been little time for agents to interact with each other, the most important process so far is random insertion of new vowels by agents that initiated an imitation game and direct imitation of these vowels by the agents that played the role of imitator in an imitation game. The vowels are therefore quite widely dispersed through the available acoustic space, but they do not cluster very much. During subsequent imitation games, the agents' vowels gradually move together. Also, due to the random insertion of new vowels, other clusters emerge, but not all agents have prototypes that correspond to all clusters. This situation is illustrated by the second frame of figure 5, taken after 300 games. When the interactions continue, the clusters tend to stabilize and contract, and become dispersed over the available acoustic space. This becomes apparent after about 1000 imitation games (frame 3) and is almost finished after 10,000 imitation games (frame 5). After 10,000 imitation games, the clusters have become compact, and the available acoustic space is almost completely covered. However, the dispersion of the clusters over the available space is perhaps not quite optimal, yet. The dispersion gradually becomes better, until it is quite natural after 20,000 imitation games (frame 6). The vowel system that emerges is natural, and could be found in a human language. It is not completely static, though. Vowel prototypes can move, so that the actual phonetic realizations of the vowels might change a little over time. Also, in rare cases, clusters may approach each other and be merged, or, if there is room, a new cluster might emerge.

Although a realistic vowel system emerges from the simulation illustrated in figure 5, this does not establish that the simulation always results in realistic vowel systems emerging. In order to investigate this, many runs of the system need to be done, and the results be compared with what is known about human languages. For one thing it is possible to define a measure of the dispersion of the vowels in the population of agents. It has been found that vowels in human languages tend to be dispersed more than in randomly created systems, and are actually quite close to being optimally dispersed (Liljencrants & Lindblom, 1972). It turns out that emerged systems, too, are almost optimally dispersed over the available acoustic space.

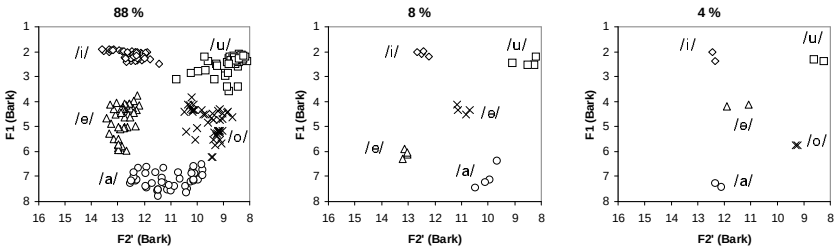

Figure 6: Classification of emerged five-vowel systems.

But it is also possible to compare emerged vowel systems with human ones directly. This can be done by running the simulation many times, then classifying the emerged vowel systems and comparing this classification with the classification one can make of human vowel systems. This is illustrated in figure 6. Here five-vowel systems that emerged from the simulation for one setting of the parameters are classified in three different types. The symmetrical type occurs in 88% of the cases, the type with more front vowels than back vowels (and one central vowel) occurs in 8% of the cases, while the type with more back than front vowels occurs in 4% of the cases. This compares very well with the percentages that Schwartz *et al*. (1997a). They have found 87% for the first type, 4% of the second type and 2% of the third type (these percentages do not add up to 100%, as they also found types that did not emerge from my simulations). Although the match between merged systems and real human language data is particularly good, excellent matches were also found for systems of six and seven vowels. For systems of four, eight and nine vowels, matches were good, but not as good. For three-vowel systems, the right types were predicted, but the so-called "vertical" three-vowel system, which is quite rare in human languages, occurred relatively frequently. However, the study has shown that the universal tendencies of human vowel systems can be explained as the result of self-organization under constraints of perception and production.

The model could be augmented with a genetic algorithm that works on the agents in the population in several ways. One way is to let the learning parameters of the agents change over time in a genetically determined way, and select for the agents that imitate the best. In this way, parameters that have to be tuned by hand in the present model could be set in a more objective way. Another way is to let the agent's production or perception evolve over time. Especially production would be interesting as it seems that the human vocal tract is specially adapted to language. One could imagine a population of agents that start with a uniform tube with only a few control parameters, which is evaluated on how well they can imitate each other and how many different sounds they can distinguish. It would be interesting to investigate whether a vocal tract that is similar to that found in humans evolves.

Conclusion and future work

It has been demonstrated by different researchers that the evolution of human speech sounds can be investigated successfully with computer models. Different aspects of speech, such as vowel systems, syllables, tone systems and intonation have been investigated, or are being investigated. The approaches taken have consisted of either pure optimization, the use of genetic algorithms, the use of a population of language-using agents or a combination of these. The most realistic would be a system consisting of a population of agents that learn speech from each other, but that are also subject to genetic evolution. However, such a system would have many parameters and many points on which a (more or less) arbitrary design decision would have to be made. Also, it might turn out to be difficult to analyze the behavior of such a system. For the time being most systems either concentrate on population dynamics or on evolution, but in the future the two will definitely have to be combined.

In future work, too, more complex utterances have to be tackled. So far vowels in isolation and simple consonant-vowel syllables have been the main subjects of investigation. But for more insight into language change and evolution, longer combinations of arbitrary sounds have to be studied. For this, more realistic and more computation intensive models will be needed. However, computing power available to the average researcher has increased so much in recent years that such models have now become computationally feasible. It will still be necessary, though, to find appropriate simplifications in order to make realistic, but tractable models.

Also, for the study of more complex sounds, machine learning algorithms are needed that are able to learn temporal sequences and that are able to extract patterns from such sequences. This is an area of research that is still very open in the machine learning community. An interesting aspect is that the ability to learn sequences and to find patterns in them is also a necessary prerequisite for learning syntax and grammar. Perhaps an interesting exchange of ideas and models between the investigation of the origins of syntax and the origins of speech is possible.

Speech is the aspect of language that is most concrete. It is therefore easiest to make an objective comparison between real linguistic data and the outcomes of a computer model in research into the evolution of speech. Also, paleontologic data can only tell us something about our ancestor’s capacity for speech, never about other aspects of language. Speech is therefore ideal for investigating and modeling the evolution of language. So far, we have only scratched the surface.

Acknowledgements

The work that is described in this chapter was performed at the artificial intelligence lab of the Vrije Universiteit Brussel in Brussels Belgium. It was written down at the Center for Mind, Brain and Learning at the University of Washington in Seattle. I would especially like to thank all the researchers whose ongoing and unpublished work is mentioned.

References

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.

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.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J. (1952), Some Experiments on the Perception of Synthetic Speech Sounds. *Journal of the Acoustical Society of America* **24** pp. 597–606. Reprinted in: D.B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press pp. 258–272

de Boer, B. (1997) Generating vowels in a population of agents. In P. Husbands & I. Harvey (eds.) *Proceedings of the Fourth European Conference on Artificial Life*, MIT Press, pp. 503-510

de Boer, B. (2000) Emergence of vowel systems through self-organisation *AI Communications* **13** (2000) pp. 27-39

de Boer, B. & Paul Vogt (1999) Emergence of Speech Sounds in Changing Populations In: Dario Floreano, Jean-Daniel Nicoud & Francesco Mondada (eds.) *Advances in Artifical Life, Lecture Notes in Artificial Intelligence* **1674**, Berlin: Springer Verlag, pp. 664-673

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

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

Frieda, E. M., Walley, A. C., Flege J. E. & Sloane, M. E. (1999) Adults' perception of native and nonnative vowels: Implications for the perceptual magnet effect. Perception & *Psychophysics* **61**(3), pp. 561—577

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.

Labov, William (1994) *Principles of linguistic change*, Oxford: Blackwell

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.

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

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*, Berlin: 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*, London: Routledge pp. 62–78.

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.

Mantakas, M, J.L. Schwartz & P. Escudier (1986) Modèle de prédiction du 'deuxiéme formant effectif' F2'—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.

Redford, Melissa A., Chun Chi Chen, and Risto Miikkulainen (1998). Modeling the Emergence of Syllable Systems. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society* (COGSCI-98, Madison, WI), 882-886, 1998. Hillsdale, NJ: Erlbaum.

Redford, Melissa A., Chun Chi Chen and Risto Miikulainen (*to appear*) Constrained Emergence of Universals and Variation in Syllable Systems, *Language and Speech*.

Ruhlen, Merritt (1994) *The origin of language: tracing the evolution of the mother tongue*, New York: Wiley.

Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée & Christian Abry (1997a), Major trends in vowel system inventories. *Journal of Phonetics* **25**(3), 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**(3), pp. 255-286.

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.

Snyman, J. W. (1975) *Zu/ohasi: fonologie & woordeboek*, Cape Town: Balkema

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

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

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).

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

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