Evolution of Speech and its Acquisition

Running title: Evolution of Speech

Bart de Boer

AI department

Rijksuniversiteit Groningen

Grote Kruisstraat 2/1

9712 TS Groningen

email: b.de.boer@ai.rug.nl

phone: + 31 50 363 6956

fax: + 31 50 363 6687

abstract

Much is known about the evolution speech. Fossil evidence points to modern adaptations for speech appearing between 1.5 million and 500 000 years ago. Studies of vocal behavior in apes show the ability to use combinatorial vocalizations in some species (but not chimpanzees) and some cultural influence on vocalizations, but little ability for vocal imitation. For modern speech, the comparison of many languages shows that speech can become extremely complex, but that it can also be astonishingly simple. Finally, the way in which infants acquire speech is becoming increasingly clear. We therefore know about the starting point of the evolution of speech, its end point and some steps in between.

Much less is known about the exact scenario and the dynamics of the evolution of the acquisition of speech. As it involves co-evolution between culturally transmitted sounds and genetic evolution, the dynamics are complex. Computer models are therefore ideal for studying these dynamics. This paper presents a computer model of the co-evolution between a repertoire of speech sounds and a population of learners that can either represent a lexicon holistically or combinatorially. It shows that cultural influences can change the dynamics of the transition between a population of holistic and combinatorial learners.

Keywords: Language Evolution, Phonology, Phonetics, Phonemic Structure, Co-evolution

1 Introduction

Speech is the most physical aspect of language and it is therefore an interesting topic when investigating the evolution of language. Adaptations for speech, such as our specialized vocal tract and adaptations for tongue- and breathing control have left traces in the fossil record. Vocalizations of other animals can be studied in detail, as can the sounds that are used in modern human languages. Acquisition of speech has also been studied in great detail. The capacities of production and perception of even the youngest infants have been investigated.

This paper provides an overview of what is known about the evolution of speech and its acquisition. The first part of the paper is devoted entirely to an introduction into what is known about human speech, its evolution, its relation to other animals' call systems and its acquisition. Although none of the individual facts is sufficient to give a clear view of how speech has evolved, the ensemble of the evidence makes it possible to get some idea of the important events in the evolution of speech. It can be said that we know the starting point of the evolution of speech (ape vocalizations) its end point (modern human speech and its acquisition) and some crucial steps in between (what has been found in the fossil record, and when this occurred).

There remain a large number of unknowns, however, and in order to make a more complete evolutionary scenario of the evolution of speech and its acquisition, it is necessary to understand more about the dynamics of the evolution of speech. It turns out that the evolution of (the acquisition of) speech is a co-evolution between the genetic evolution of adaptations to speech production and speech perception, genetic evolution of the brain so that it can learn speech (and possible evolution of instincts to *teach* speech to children, *i.e.* infant directed speech) and cultural evolution of the repertoire of speech sounds itself. The dynamics of such co-evolving systems are extremely complex, but have been investigated successfully in other areas of language using computer models. (see, e. g., Cangelosi & Parisi, 2002; Kirby, 2002 for overviews) .

The second part of the paper therefore presents a computer model for investigating the dynamics of the interaction between (cultural) evolution of a repertoire of speech sounds and (genetic) evolution of a population of learners. The topic the model investigates is the transition from holistic call systems (such as found in gorillas, chimpanzees and bonobos) to combinatorial (phonemic) systems of speech sounds. The model is extremely simplified. It is more comparable to models used in theoretical biology than to more traditional computational language evolution model. Its intention is to illustrate how such a model can be constructed and as an inspiration for further work. However, even in its simplest form it already shows how cultural evolution of a repertoire of speech sounds can influence the composition of a population of agents that have to learn this repertoire.

2 The Evolution of Speech

Before presenting data on the evolution of speech, it is perhaps interesting to consider the complexity of modern speech. This turns out to be very different for different languages. Systems of speech sounds can be extremely complex in some languages, while in other languages only very simple systems of speech sounds are used. The languages with the largest number of individual speech sounds are some of the Khoisan (Bushman) languages, with !Xóõ having the record of 186 different vowels and consonants (Traill, 1985). As for complexity of syllables, certain Caucasian languages appear to be the most complex (Catford, 1977). Georgian, for example has words such as *prtskvna* “to peel”. Although this complexity is impressive, there are also modern languages with very simple sound systems. Piraha, for example uses only 11 different speech sounds (Everett, 1982; Sheldon, 1974). Hawaiian appears to be the language with the fewest syllables, 162 according to Maddieson (1984). The conclusion must therefore be that although modern human speech can be extremely complex, such complex sound systems are not necessary for a modern language. In fact, complex linguistic communication can already be achieved with relatively few speech sounds. But what is known about the evolution towards human-like speech from ape-like vocalizations?

There are three sources of evidence for the evolution of speech: the fossil record, the vocalizations of other animals and the neuroanatomy of related species. Fitch (2000) gives an excellent review of recent papers on the first two topics. Hauser (1997, section 4.4) and Ploog (2002) give overviews on the comparative neuroanatomy of language, while Gannon et al. (1998) present interesting analogs between the human and chimpanzee brain. In order to understand the context of evolution of speech and its acquisition, it is interesting to consider this evidence in some more detail.

Speech itself does not fossilize, but there are anatomical adaptations that are related to speech. The best-known adaptation to speech is the shape of the human vocal tract. Because of the lowered larynx, and other modifications to the vocal tract, humans can produce a huge range of different speech sounds, but run the risk of choking on their food. The vocal tract itself, consisting mostly of soft tissue, does not fossilize well. There is one bone in the larynx (the *hyoid bone*) and this appears to have the same shape in Neanderthals (of approximately 60 000 years ago) as in modern humans (Arensburg et al., 1989). This is an indication that Neanderthals may have had a modern vocal tract, but more evidence is needed to draw firm conclusions.

Fortunately, there are other sources of evidence for the presence or absence of speech. As increased control of tongue movements and of breathing is needed for speech, it is to be expected that the nerves innervating the tongue and the diaphragm are larger in humans than in other species. This is indeed what is found when comparing humans and apes. As nerves pass through canals in the skull or the vertebrae, one can then measure the sizes of the relevant holes in fossils and check whether they are more human-like or ape-like. Kay, Cartmil and Balow (1998) have investigated the canal through which the nerve innervating the tongue passes (the *hypoglossal canal*). They found that it is human-like in early *Homo sapiens* and in Neanderthals, while it is ape-like in *Australopithecus* and *Homo habilis*. Based on their fossils they propose that “human-like vocal abilities” already existed about 400 000 years ago.

Adaptations related to breathing control were investigated by MacLarnon and Hewitt (1999). The nerves that control the diaphragm pass through the *thoracic vertebral canal*. It turns out that this canal is human-like in early *Homo sapiens* and Neanderthals, while in *Homo ergaster* and the Australopithecines it is not. MacLarnon and Hewitt conclude that adaptations for speech arose between 1.6 million and 100 000 years ago.

A final piece of evidence comes from the structure of the outer and middle ear. On the basis of fossils of roughly 350 000 years old, Martínez and colleagues (2004) have calculated the acoustic frequency sensitivity of *Homo heidelbergensis* (an ancestor of Neanderthals). They found that it was comparable to that of humans and not to that of apes. As humans are especially sensitive to the frequencies used in speech and apes are not, this is an indications that *Homo heidelbergensis* had adaptations to speech.

Although none of the above facts is conclusive in itself, the evidence does appear to converge to the conclusion that early *Homo sapiens* and Neanderthals had adaptations to speech, while australopithecines and early *Homo erectus* did not. This gives a time window for the evolution of modern adaptations for speech of about one million years, between about 1.5 million and 500 000 years ago. This corresponds roughly to about 50 000–100 000 generations.

Evolution did not have to start from scratch. Chimpanzees and other apes use vocal communication but this vocal communication is different from human speech in a number of ways. First of all, ape vocalizations appear to be mostly innate. There appears to be a little bit of learned flexibility in vocalization of chimpanzees (Crockford, Herbinger, Vigilant, & Boesch, 2004), but Fitch (2000) observes that they have great difficulties with real vocal imitation. Humans, on the other hand, imitate effortlessly, and imitation is essential for learning language.

Also, chimpanzee vocalizations do not appear to have combinatorial structure. While gibbon songs do have combinatorial structure, (Mitani & Marler, 1989) gibbons are more distant evolutionary. This must therefore be an example of convergent evolution. It does illustrate, however, that combinatorial structure is evolutionary adaptive for larger repertoires of calls and that ape brains are in principle capable of processing combinatorial calls.

There are a number of differences between modern humans and the last common ancestor of humans and chimpanzees with respect to speech. There have been changes in anatomy of the vocal tract and of the outer and middle ear, and there have been changes that have to do with control of breathing and control of the tongue. Although as (Fitch, 2000) suggests these adaptations might originally not have been for speech, they allowed for a better controlled and larger repertoire of speech sounds. There have also been cognitive changes, and these have to do with the ability to acquire and use a repertoire of combinatorial speech sounds.

3 Modern Acquisition of Speech

Given the complexity of the speech signal, infants learn to speak remarkably fast and remarkably accurately. While still in the womb, infants already learn the intonation patterns that are used by their mother, and might even learn more about the language used around them. This they can do because the vibrations of the mother’s vocal chords and even other sounds can be heard in the womb. After birth, development of perception and production of speech proceeds rapidly.

Production begins with non-speechlike sounds (de Boysson-Bardies, 1999; Vihman, 1996) such as crying, puffs, raspberries etc. Between two and five months, babies explore the possibilities of their vocal apparatus, while gradually getting more and more control over their vocalizations. However, these do not yet resemble the syllables found in adult language. At the same time, their perception changes in a way that makes it more tuned towards their native language. Whereas babies at first are able to hear all distinctions between speech sounds that are made in human languages, after a few months they are only able to hear the distinctions between sounds that are relevant in their native language. In a sense, they learn the categories that are used in their language. This happens at about six months of age for vowels (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992) and between six and twelve months for consonants (Werker & Tees, 1984).

Perception of speech is usually ahead of production. Production, however continues to develop rapidly as well. At around seven months on average, children start to babble. This means that they produce sounds that are very much like syllables in adult language. Often they are produced by simple oscillatory movement of the jaw and lips, without much movement of other articulators. Thus babbling sounds like the repetition of simple syllables: /bababa/. But more subtle control is learnt as well. At about eight months, the babbling of babies is already influenced by their native language. When presented with babbling of babies with different native languages, adults are quite good at picking out the baby of their own language (de Boysson-Bardies, Sagart, & Durand, 1984). At about ten months, babies start to vary their babbling more, and actively explore different possible variations.

After this, the first real words are uttered, and language learning starts to focus on vocabulary and grammar. Speech still continues to develop. More difficult speech sounds and combinations of speech sounds in particular can be mastered quite late, until the age of six or seven years in some cases.

The general conclusion is that children acquire the basics of speech amazingly quickly. They must therefore have adaptations for learning speech. One of these capacities must be the ability and the drive to imitate vocalizations, as Fitch (2000) suggest. Another capacity must be the ability to find complex (statistical) structure in the speech sounds they perceive. Other capacities may include innate sensitivities for different distinctions in speech sounds, or innate preferences for certain sounds. However, one must also take into account what the situation is in which children learn to speak.

If children were to base the acquisition of speech on adult speech alone, their task would indeed be formidable. Adult speech, especially the kind of speech used in informal settings, tends to be very fast, and because of its speed, articulations tend to be reduced. Reconstructing the sound system of a language from this informal speech is very difficult. It turns out, though, that parents tend to use a special register of speech to their children that is slower, more clearly articulated and that has more exaggerated intonation than adult-to-adult speech. This speech is called infant-directed speech or *motherese*, (although it is not exclusively used by mothers, but also by other caretakers) and should not be confused with the meaningless utterances that are sometimes used to soothe babies or to attract their attention. Infant-directed speech consists of meaningful, correct, but simple linguistic utterances and is generally used in direct face-to-face interactions with the infant. It turns out that children prefer to listen to infant-directed speech (Fernald, 1985), and that it occurs almost universally cross-culturally (e. g., Ferguson, 1964). Computer studies have indicated that it is indeed more learnable than adult-to-adult speech (de Boer & Kuhl, 2001, 2003). It is therefore plausible that not only the acquisition of speech has evolved, but also the instincts of caretakers interacting with infants. As the infant’s linguistic competence grows, caretakers increase the complexity of their utterances, therefore aiding the infant in a kind of bootstrapping procedure.

4 The Role of Computer Models

As can be concluded from the above-mentioned data, a lot is known about the evolution of speech. It is known how long the evolution can have taken, when adaptations for modern speech were in place, what the complexity of modern human speech is, but also that modern language is possible with relatively simple sound systems. We know approximately in what stages children acquire speech and what kind of input at which moments is needed. We also know that apes are unable to acquire speech and are quite bad at imitation, although they are capable of reasonably complex vocalizations. Other, less closely related animals are capable of complex learned vocalizations, and although there can be no direct evolutionary link, these tell us about possible ways in which vocalizations can evolve to become more complex. There is also a plausible way in which movements of the jaw that were already used for eating and breathing were exapted for the use in speech (MacNeilage & Davis, 2000). Finally, from mathematical and information-theoretical arguments it is known that combinatorial (phonemically coded) signaling systems are more efficient than holistically coded systems (Nowak & Krakauer, 1999; Plotkin & Nowak, 2000). A lot less is known about how these different pieces of the puzzle fit together.

The difficulty with making theories about the evolution of speech is twofold. First of all, evolution is a historical science, but we are unable to replay history. The second difficulty with speech is that it does not leave traces in the fossil record and that there are therefore only indirect indications. These factors cause a number of problems. As random factors play a role in any historical process, there will always be multiple possible scenarios. Random factors also make it more difficult to understand which aspects of speech are coincidence and which aspects are really the result of functional pressures. It is perhaps unavoidable that there will always be uncertainty about the evolution of speech, as information does get lost over time, and we will never be able to recover it all. However, with the aid of computer models it is possible to replay a simplified version of the evolution of speech, thus testing different possible scenarios and influences.

Speech (and language) are very complex systems. Speech consists of complex signals, speech is learned and used by complex brains, and it is used in the setting of a complex population of language users. This makes it very difficult to say what theories of the evolution of speech really predict. Because there are many complex feedback loops, it is quite possible that something unexpected will happen. Computer models can be useful in handling this complexity. Theories and scenarios can be implemented, and run for a large number of iterations (or generations). It then becomes clear immediately what the predictions of a given theory are, and it can be investigated whether this corresponds with the available data. Also, the implementation of a theory in the form of a computer model forces the researcher to make all the assumptions explicit. Thus computer models are a useful tool in investigating the evolution of speech.

There are many ways in which questions concerning the evolution of speech can be addressed with computer models. Many of these questions have already received attention from computer modelers. One of the questions that has been investigated most intensively is the influence of functional and cultural pressures on the evolution of speech. Functional pressures, such as acoustic distinctiveness, articulatory ease and learnability influence what utterances are selected for use in language. This can then influence the evolution of speech and its acquisition. Utterances that are easy to produce and easy to distinguish will be preferred, and perhaps evolution will adapt the language users to these preferences. From another perspective, if certain properties of modern speech can be explained as the result of functional pressures, it becomes unnecessary to explain them as the result of speech-specific adaptations of the brain. Understanding the functional influence on the evolution of speech is complex, because functional pressures often work in different directions and because there is also a pressure to conform to the population. This pressure to conform causes sub-optimal systems to remain stable, as making them more optimal would cause them to become unusable with the other members of the population.

A reasonable body of work exists in which computer models have been used to investigate functional pressures on speech. Probably the first effort was made by Liljencrants and Lindblom (1972) in which the effect of acoustic distinctiveness on vowel systems was investigated. A similar model was used to investigate consonant-vowel syllables (Lindblom, MacNeilage, & Studdert-Kennedy, 1984). Neither model took into account that language is used in a population, however. The first work that did take this into account by using an agent-based model was done by Glotin, Berrah and others (Berrah, 1998; Berrah, Glotin, Laboissière, Bessière, & Boë, 1996; Glotin, 1995) and was extended upon by de Boer (de Boer, 1997, 2000, 2001). This work showed that the universals of vowel systems could be explained as the result of functional pressures in a population. Innate features and preferences turned out to be unnecessary and hence need not have evolved. Oudeyer later showed that relatively simple interactions and neural mechanism could already give these results (e. g., Oudeyer, 2001). Models based on this work have been used to investigate the emergence of linguistic diversity (Livingstone & Fyfe, 1999) and evolution of tone, for example (Ke, Ogura, & Wang, 2003). Genetic algorithms have also been used to investigate the role of functional pressures on the evolution of speech, *e.g.* (Redford, Chen, & Miikkulainen, 2001). This last model has been used to test what factors are important for getting systems of speech sounds that are found in human language. Testing what happens when one changes the influence of different factors on the evolution of speech is an interesting and rather underexplored area of research.

The role of functional and cultural pressures is not the only one that can be addressed by computer models, but other questions have received notably less attention. One very interesting question is the emergence of compositional structure from holistic utterances. The more recent work of Oudeyer (e. g., Oudeyer, 2002, this issue) is on this topic, but it turns out to be hard to make good models of complex utterances. Another interesting question is whether parental behavior and infant language acquisition can have co-evolved. Work on this topic is presented in (de Boer, 2003). A hard-to-model, but crucial question is how the vocal tract, control of articulation and control of breathing on the one hand and speech on the other hand can have co-evolved.

Perhaps disappointingly, very little or no modeling research was carried out into the evolution of acquisition. There are models of speech acquisition, for example neural models (e. g., Guenther & Gjaja, 1996) but these are not about evolution. The lack of evolutionary models of acquisition is perhaps understandable, as they would involve the combination of an evolutionary system, a learning system and a population-based model. Also, the way in which acquisition of speech works in children is still not very well known, so the model would have to deal with many uncertainties, and it would be hard to evaluate its results with respect to human acquisition of speech.

5 Towards a Model of Combinatorial Acquisition

The model that will be presented in this paper is about the evolution of combinatorial (phonemic) acquisition of speech. As gorillas, chimpanzees and bonobos appear to use sounds in a holistic way, it is safe to assume that our common ancestor did so, too. As the use of a combinatorial sound system is crucial for extending the lexicon, it is interesting to investigate how its acquisition could have evolved. The model presented here is about the interaction of genetic and cultural influences on this process.

In order to make a computer model for investigating the evolution of combinatorial acquisition, agents must be made that can learn a system of speech sounds as either a set of holistically coded utterances, or as a set of combinatorially coded utterances. The crucial difference between such systems is how the speech sounds are eventually stored. Therefore, a definition of holistic storage and combinatorial storage is needed. In this paper, a holistic system is considered a system in which there is only one level of storage. For all utterances in the lexicon, all gestures needed to perform them are stored exactly. In a combinatorial system, there are two levels of storage. The level of the lexicon and the level of the building blocks of the lexicon (these could be phonemes, syllables, gestures etc.). At the level of the building blocks, gestures are stored exactly, just as in the holistic system. At the level of the lexicon, however, words are stored as sequences of building blocks. The two systems are illustrated in figure 1.

Figure 1 about here.

It is clear that if the number of building blocks is much smaller than the number of possible gestures, compression of the lexicon is possible. Also, a combinatorial system makes it easier to produce new utterances, by recombining building blocks, although producing new utterances is not impossible in a holistic system. In humans a combination of holistic and combinatorial storage is most likely used. Infants probably learn their first words as holistic gestures, and analyze these into building blocks only later. In adult language, too, there are some utterances that have communicative function and are learned, but that fall outside the standard phonology (and syntax) of the language. Examples of these are utterances such as “pssst” to get someone’s attention, “pffff” to express exasperation and “tsk tsk” to express disapproval. Such utterances are probably stored holistically.

In a model of the evolution of acquisition of combinatorial sound systems, fitness of the agents is determined by a number of factors. The amount of storage required to store a given system of sounds is one such factor. Another factor is the communicative success that can be achieved. A third factor is the amount of effort that is needed for learning and using the sound system. A final, but hard-to-quantify factor is the extra cognitive machinery that is necessary for using a combinatorial system. The research here will focus on the necessary storage exclusively.

The space *s* necessary for storing a lexicon *L* is as follows:

$$s = \begin{cases} \sum_{\forall w_i \in L} \alpha \cdot l_i & \text{,for a holistic system} \\ \alpha \cdot n + \sum_{\forall w_i \in L} \beta \cdot l_i + \gamma & \text{,for a phonemic system} \end{cases} \tag{1}$$

where α is the number of bits needed to store all information about a certain gesture, l_i is the number of gestures in word w_i, *n* is the number of articulatory gestures (“phonemes”) used by a phonemically coding agent, β is the number of bits needed to specify a phoneme in a word and γ the overhead needed for a phonemically coded system. From these formulas it is clear that if there are only words consisting of a single gesture in the lexicon, holistic coding is more efficient. If there are many words re-using the same gestures, phonemic coding is more efficient. The exact point where the switch occurs depends on the parameter values. Exact values cannot be determined, but it must be true that $\alpha \gg \beta$ (specifying a gesture exactly is more complex than referring to it), and that $\beta \propto \log_2 n$ (the more phonemes there are, the more bits are needed to distinguish them, and this number is proportional to the logarithm of *n*). Another consideration is that not all articulatory gestures that can be specified can be used as different phonemes. To preserve a margin of error and acoustic distinctiveness, some open space must remain between different gestures. Therefore there are fewer possible phonemes than possible gestures. In an equation: $n_{\max} \ll 2^{\alpha}$, where $n_{\max}$ is the maximal number of distinguishable phonemes and 2^{α} is the maximal number of speech sounds that can be specified articulatorily. From these considerations it follows that a lexicon consisting of a large number of words will always need to contain words that consist of multiple gestures, and therefore phonemic coding will be more efficient. A small lexicon, on the other hand, can consist of single-gesture words, and therefore holistic coding will be more efficient. Somewhere in the middle, a transition must occur. Such a transition can combine interesting genetic and cultural effects: holistic learners will prefer a system with many different gestures and short words, while phonemic learners prefer a system with long words and few different gestures. Through self-organization, the sound system in the population will tend to adapt to the preferences of the majority of the population. This will have effects on the dynamics in a system where both the learned system of speech sounds and the learners themselves change.

A computer model has been implemented to make a first investigation of the dynamics of a system in which a population of language learners evolves "genetically" and a system of speech sounds evolves "culturally". Of course the system is simplified enormously with respect to reality, but this has been done in order to create a point of reference from which other simulations can start and to get a model that shows the dynamics in as pure a way as possible. There are two kinds of agents in the model: holistic learners and phonemic learners. The population could have been modeled as a set of one-bit genomes, but it has been decided to just model the fraction of agents that learn holistically, p_h and the fraction that learns phonemically, p_p. Although this way of modeling a population is unusual in artificial life, it is an old tradition in theoretical biology (e. g., Maynard Smith, 1974). As there are no other agent types, it follows that $p_h + p_p = 1$. Both types have a fitness f_h and f_p, and these are used to calculate the fractions in the next generation. There is also the possibility μ of one type of agent mutating into another type of agent (set to 0.1 in the simulations presented here). The equations are as follows:

$$\begin{aligned} p_{h,t+1} &\leftarrow \nu\left(f_{h,t} \cdot p_{h,t} + \mu \cdot f_{p,t} \cdot p_{p,t} \right) \\ p_{p,t+1} &\leftarrow \nu\left(f_{p,t} \cdot p_{p,t} + \mu \cdot f_{h,t} \cdot p_{h,t} \right) \end{aligned} \qquad (2)$$

where ν is a factor that causes the sum of $p_{h,t+1}$ and $p_{p,t+1}$ to be one.

Figure 2 about here.

The fitness is determined by the number of bits needed to store a repertoire of speech sounds. This repertoire of speech sounds is shared by the whole "population". It consists of a list of words that each consist of one or more symbols. These symbols represent the basic gestures, and phonemic learners use them as their phonemes. Equation (1) is used for calculating the number of bits needed to store a repertoire. An example of a repertoire and the number of bits needed to store it is given in figure 2. The values of the parameters used here (and in the rest of the simulations) are as follows: number of bits for a gesture (α) = 10, penalty for using a phonemic system (γ) = 30 bits, number of bits used per phoneme (β) = $\log_2 n$.

Note that in the example, a more compact alternative holistic system using the same basic gestures would need only 110 bits (by changing *ao* into *o* and *aea* into *ea,* for example). A phonemically coded system with the same number of words could also be substantially more efficient, by for example using two one-phoneme, four two-phoneme and two three-phoneme words, using only two different phonemes. This would result in a total of 66 bits storage. From these considerations, it is clear that for lexicons with the same number of words, holistic and phonemic learners will prefer very different words.

Given the number of bits s_p and s_h for phonemic and holistic learners, respectively the fitnesses are calculated as follows:

$$f_p = 1 - \frac{s_p}{s_p + s_h}$$
$$f_h = 1 - \frac{s_h}{s_p + s_h} \qquad (3)$$

Each generation the lexicon can be modified. A new word can be added and agents can modify it to suit their preferences. New words are added by finding the shortest word that is not already present, and adding this. Adding is done holistically, in the sense that new articulatory gestures can be added. The maximum number of different gestures (n_{max}) is 16 in the simulation presented here. As words can sometimes be removed from the lexicon the added word is sometimes shorter than the longest word in the lexicon. Words are added with a probability of 10% per generation.

The lexicon can also be modified to better suit either holistic agents or phonemic agents. For holistic agents (who dislike long words) the longest word is removed from the lexicon and replaced with an unused shorter word, if possible. This can introduce new articulatory gestures as a side effect. Phonemic agents first find the phoneme that is least often used, and then find the word in which it occurs most frequently. They then try to replace this word with the shortest word that is build up of the phonemes already present in the lexicon. This can cause phonemes to disappear from the lexicon and average word length to increase. The two processes thus assert conflicting pressures on the lexicon.

In the simulation presented here replacement is performed two times per generation if culture is used, and no replacement is performed if there is no culture. Replacement is considered to model the effect of culture, as agents modify the system of speech sounds they transmit to the next generation to suit their learning preferences. This is something that would not occur if the calls were purely innate. For each replacement either a holistic or a phonemic agent is selected with probabilities that are proportional to their abundance in the population. Thus, if there are many holistic agents, the lexicon is pushed towards holism. If there are many phonemic agents, it is pushed towards phonemic coding.

Figure 3 about here.

The population is initialized with 50% holistic and 50% phonemic agents. The lexicon is initialized with a single word. The result of running the model without and with cultural influences is shown in figure 3. As can be seen from these figures, the proportion of holistic agents rapidly rises in the beginning. Apparently these are fitter than phonemic agents. The population then remains almost exclusively holistic for a while (the minority of phonemic agents remains present because of mutation). After a certain critical number of words is reached, a transition from a holistic majority to a phonemic majority takes place. There appears to be little difference in average behavior between systems with culture and without culture. This is due to artifacts of averaging, however. As can be observed in the graph, the error bars in the graph for populations with culture are much larger. This is caused by the fact that the transition happens much more quickly for systems with culture. This is illustrated in figure 4 that shows typical runs for systems without and with culture. A comparison of the time needed for the fraction of holistic agents to change from above 60% to below 40% confirms this. The time is 52.3 generations (with standard deviation $\sigma = 14.6$) for the population without culture and 12.1 generations (σ = 5.6) for the population with culture.

There appears to be no significant difference for the time at which the drop takes place; this happens after 191 (σ = 42.6) and 194 (σ = 49.7), respectively. The size of the lexicon at which it occurs is 22.4 (σ = 0.52) for populations without culture and 21.4 (σ = 1.07) for systems with culture. Although this is a significant difference, it is probably caused by the fact that populations with culture go through the transition faster than populations without culture. The transition probably starts when the number of words exceeds the number of possible articulatory gestures (in these simulations, that was fixed at 16). Even systems that have been optimized for holistic learners then become more efficient for phonemic learners.

Figure 4 about here.

6 Discussion and Conclusion

The simulation that has been described in the previous section has shown that cultural evolution in conjunction with genetic evolution changes the dynamics of a population of language users. It was shown that a population can change much more quickly from holistic to phonemic language use if there is co-evolution of the culturally transmitted system of speech sounds with the genetically evolving language learners. As even in this very much simplified model cultural evolution influences the dynamics of the emergence of a system of speech sounds, it is clear that the evolution of speech (and language) cannot be studied as either purely genetic or purely cultural evolution. Both mechanisms must be taken into account. It also confirms that phonemically coded systems win over holistically coded systems, at least as far as storage is concerned. They win, even though at first there is a cultural evolution towards systems that are more learnable for holistic learners.

The model that was used is admittedly simplistic, and only a limited number of experiments were performed. The model was made so simple in order to create the most basic model that shows some kind of interaction between the evolution of the acquisition of complex speech and the (cultural) evolution of the speech sounds themselves. There are many ways in which other experiments can be done: the influence of the different parameters can be investigated, it could be investigated at which point the transition from holistic to phonemic coding takes place exactly, different variants on the optimization procedures and the addition of new words could be tried out and many other small variants.

More interesting, however, is to strive for more realism in the simulation. As the model is about the evolution of acquisition, it is important to try to model a population of agents that really acquire the system of speech sounds. The acquisition mechanism should have parameters that make it exploit phonemic structure to a higher or lower degree, and these parameters should be able to evolve. The time it takes to acquire a repertoire of speech sounds and the accuracy with which this happens could be taken into account in the fitness function. A next step could then be to get rid of the global sound systems, and have them be emergent in the population, just as the sound systems are emergent in the population in (de Boer, 2000, 2001). Also, more realistic constraints on production and perception could be added. Such a model would already be quite realistic, but also have much more complicated and hard-to-understand dynamics. The model as presented in this paper gives a basis for understanding such dynamics.

Another important research topic would be finding independent evidence with which to compare the results of such a computer simulation. This can be evidence from language acquisition, evidence from the fossil record, or evidence from animal call systems. Perhaps the study of call systems from closely related primate species, for example the different species of gibbon, (Geissmann, 2002), can provide insight into the circumstances under which a holistic call system can change into a phonemic/combinatorial call system.

The understanding of the dynamics of the evolution of speech is a crucial piece of the puzzle that is missing in the evidence that we have from animal studies and from the fossil record. The amount of such evidence is already impressive and still increasing, as has been shown by the overview presented in this paper. The interaction between the evolution of a cultural repertoire of speech sounds, the neural adaptations for acquiring it and the physical adaptations for producing and perceiving cause the evolution of speech to have very complex dynamics. Computer models can provide insights in this dynamics. The computer model presented here is a first attempt to provide insight in the interaction between cultural and genetic evolution, and it is hoped that it can be used as an inspiration for further research.

Acknowledgement

I would like to thank Gert Kootstra for discussion that helped to shape the model presented in this paper.

References

Arensburg, B., Tillier, A. M., Vandermeersch, B., Duday, H., Schepartz, L. A., & Rak, Y. (1989). A middle palaeolithic human hyoid bone. *Nature, 338*(6218), 758–760.

Berrah, A.-R. (1998). *Évolution artificielle d'une société d'agents de parole: Un modèle pour l'émergence du code phonétique*.Grenoble: Thèse de l'Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives.

Berrah, A.-R., Glotin, H., Laboissière, R., Bessière, P., & Boë, L.-J. (1996). From form to formation of phonetic structures: An evolutionary computing perspective. In T. Fogarty & G. Venturini (Eds.), *Icml '96 workshop on evolutionary computing and machine learning, bari* (pp. 23–29).

Cangelosi, A., & Parisi, D. (Eds.). (2002). *Simulating the evolution of language*.Berlin: Springer Verlag.

Catford, J. C. (1977). Mountain of tongues: The languages of the caucasus. *Annual Review of Anthropology, 6*, 283–314.

Crockford, C., Herbinger, I., Vigilant, L., & Boesch, C. (2004). Wild chimpanzees produce group-specific calls: A case for vocal learning? *Ethology, 110*, 221–243.

de Boer, B. (1997). Generating vowel systems in a population of agents. In P. Husbands & I. Harvey (Eds.), *Fourth european conference on artificial life.* (pp. 503–510). Cambridge (MA): MIT Press.

de Boer, B. (2000). Self organization in vowel systems. *Journal of Phonetics, 28*(4), 441–465.

de Boer, B. (2001). *The origins of vowel systems*.Oxford: Oxford University Press.

de Boer, B. (2003). Conditions for stable vowel systems in a population. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim & J. Ziegler (Eds.), *Advances in artificial life, lecture notes in computer science 2801* (pp. 415–424). Berlin: Springer.

de Boer, B., & Kuhl, P. (2001). Infant-directed vowels are easier to learn for a computer model. *Journal of the Acoustical Society of America, 110*(5, pt. 2), 2703.

de Boer, B., & Kuhl, P. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online, 4*(4), 129–134.

de Boysson-Bardies, B. (1999). *How language comes to children*.Cambridge, MA: MIT Press.

de Boysson-Bardies, B., Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11*, 1–15.

Everett, D. L. (1982). Phonetic rarities in piraha. *Journal of the International Phonetic Association, 12*(2), 94–96.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist, 66*(6, part 2), 103–114.

Fernald, A. (1985). Four month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8*, 181–195.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in cognitve science, 4*(7), 258–267.

Gannon, P. J., Holloway, R. L., Broadfield, D. C., & Braun, A. R. (1998). Asymmetry of chimpanzee planum temporale:Humanlike pattern of wernicke's brain language area homolog. *Science, 279*, 220–222.

Geissmann. (2002). Duet-splitting and the evolution of gibbon songs. *Biological Reviews of the Cambridge Philosophical Society, 77,* 57–76.

Glotin, H. (1995). *La vie artificielle d'une société de robots parlants: Émergence et changement du code phonétique*.Grenoble: DEA sciences cognitives-Institut National Polytechnique de Grenoble.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100,* 1111–1121.

Hauser, M. D. (1997). *The evolution of communication*.Cambridge, MA: MIT Press.

Kay, R. F., Cartmill, M., & Balow, M. (1998). The hypoglossal canal and the origin of human vocal behavior. *Proceedings of the National Academy of Sciences, 95*, 5417–5419.

Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Optimization models of sound systems using genetic algorithms. *Computational Linguistics, 29*(1), 1–18.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life, 8*(2), 185–215.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*, 606–608.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulatons of vowel quality systems. *Language, 48*, 839–862.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of language universals. In M. Butterworth, B. Comrie & Ö. Dahl (Eds.), *Explanations for language universals* (pp. 181–203). Berlin: Walter de Gruyter & Co.

Livingstone, D., & Fyfe, C. (1999). Modelling the evolution of linguistic diversity. In D. Floreano, J.-D. Nicoud & F. Mondada (Eds.), *Advances in artificial life, lecture notes in artificial intelligence* (Vol. Volume 1674, pp. 704–708). Berlin: Springer.

MacLarnon, A., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology, 109*(3), 341–343.

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science, 288*, 527–531.

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

Martínez, I., Rosa, M., Arsuaga, J.-L., Jarabo, P., Quam, R., Lorenzo, C., et al. (2004). Auditory capacities in middle pleistocene humans from the sierra de atapuerca in spain. *Proceedings of the National Academy of Sciences, 101*(27), 9976–9981.

Maynard Smith, J. (1974). *Models in ecology*.Cambridge: Cambridge University Press.

Mitani, J. C., & Marler, P. (1989). A phonological analysis of male gibbon singing behavior. *Behaviour, 109,* 20–45.

Nowak, M. A., & Krakauer, D. (1999). The evolution of language. *Proceedings of the National Academy of Sciences, 96,* 8028–8033.

Oudeyer, P.-y. (2001). Coupled neural maps for the origins of vowel systems. In G. Dorffner & K. H. Bischof (Eds.), *Proceedings of the international conference* *on artificial neural networks, lecture notes in computer science 2130* (pp. 1171–1176). Berlin: Springer Verlag.

Oudeyer, P.-y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In J. Hallam (Ed.), *Proceedings of the international conference on the simulation of adaptive behavior (sab)* (pp. 406–416). Edinburgh: MIT Press.

Ploog, D. (2002). The neural basis of vocalization. In T. J. Crow (Ed.), *The speciation of modern homo sapiens* (pp. 121–135). Oxford: Oxford University Press.

Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology, 147–159*.

Redford, M. A., Chen, C. C., & Miikkulainen, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech*, *44*, 27–56.

Sheldon, S. N. (1974). Some morphophonemic and tone rules in mura-pirahã. *International Journal of American Linguistics, 40*, 279–282.

Traill, A. (1985). *Phonetic and phonological studies of!Xóõ bushman*.Hamburg: Helmut Buske Verlag.

Vihman, M. M. (1996). *Phonological development: The origins of language in the child*.Cambridge MA: Blackwell.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Figure captions

Figure 1: Diagrams of holistic storage and combinatorial storage.

Figure 2: Example of a lexicon and the number of bits needed for holistic and phonemic storage. For parameter values, see the text.

Figure 3: Comparison of average behavior (over 10 runs) of a population without culture (left graph) and a population with culture (right graph). As the fractions of phonemic and holistic fractions are symmetrical, error bars showing standard deviation are only shown for the holistic fraction.

Figure 4: Typical runs from a system without culture (left graph) and a system with culture (right graph). Note the faster transition in the graph with culture.

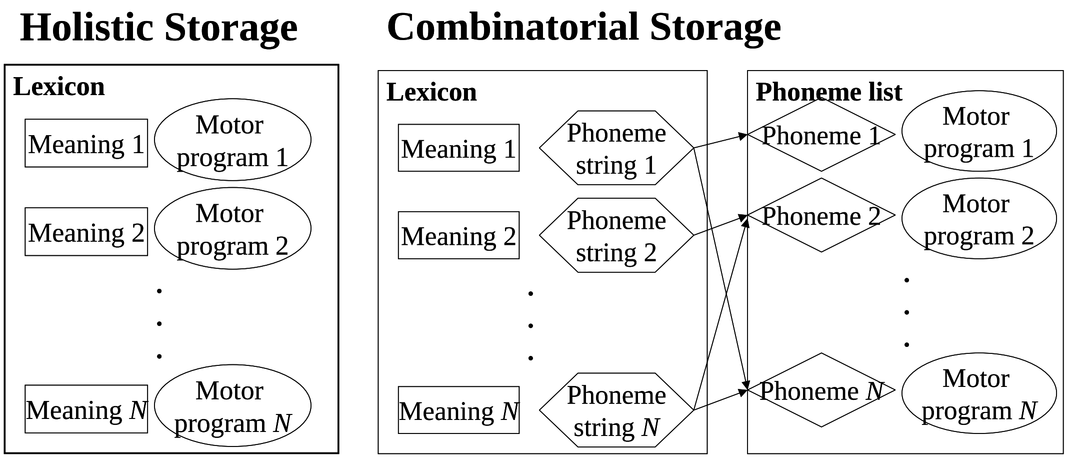
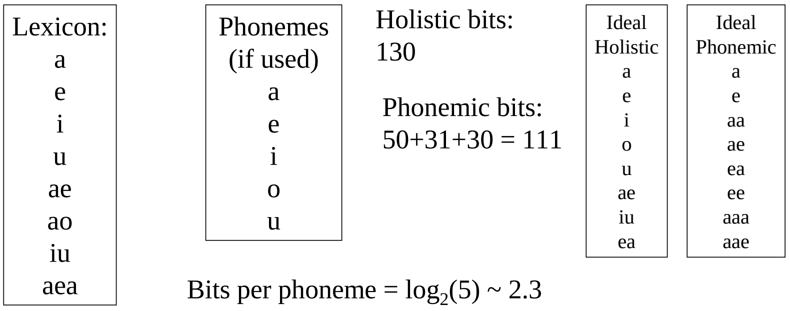

Holistic Storage
Combinatorial Storage
Lexicon
Meaning 1
Motor program 1
Meaning 2
Motor program 2
Meaning *N*
Motor program *N*
Lexicon
Meaning 1
Phoneme string 1
Meaning 2
Phoneme string 2
Meaning *N*
Phoneme string *N*
Phoneme list
Phoneme 1
Motor program 1
Phoneme 2
Motor program 2
Phoneme *N*
Motor program *N*

Figure 1



Lexicon:
a
e
i
u
ae
ao
iu
aea

Phonemes
(if used)
a
e
i
o
u

Holistic bits:
130

Phonemic bits:
50+31+30 = 111

Bits per phoneme = $\log_2(5) \sim 2.3$

Ideal
Holistic
a
e
i
o
u
ae
iu
ea

Ideal
Phonemic
a
e
aa
ae
ea
ee
aaa
aae

Figure 2

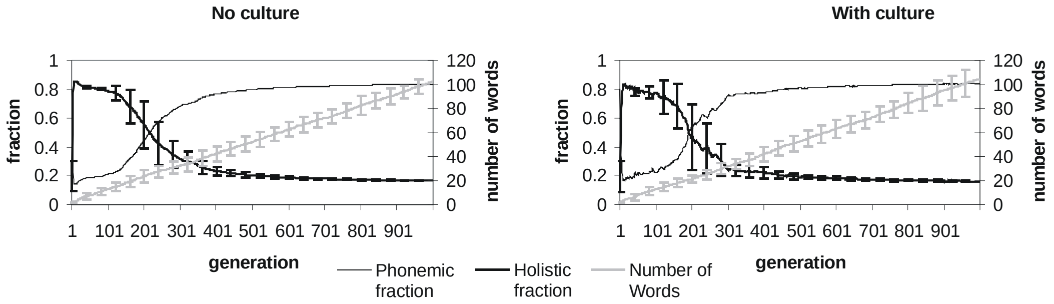

No culture
With culture
fraction
number of words
1
0.8
0.6
0.4
0.2
0
120
100
80
60
40
20
0
1 101 201 301 401 501 601 701 801 901
generation
Phonemic fraction
Holistic fraction
Number of Words

Figure 3

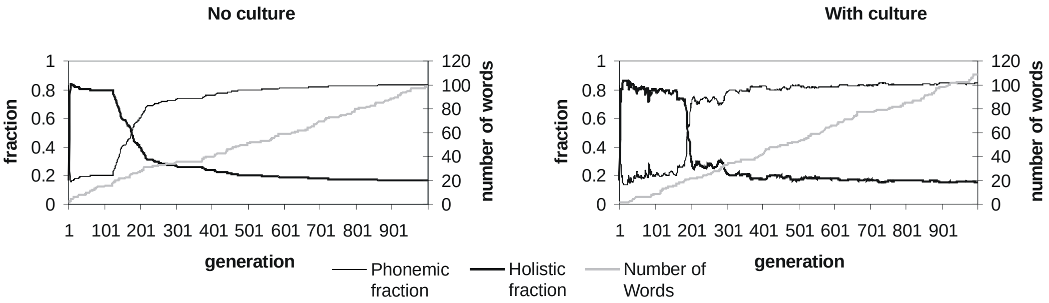

No culture
With culture
fraction
number of words
generation
1
0.8
0.6
0.4
0.2
0
120
100
80
60
40
20
0
1 101 201 301 401 501 601 701 801 901
Phonemic fraction
Holistic fraction
Number of Words

Figure 4

Bart de Boer has studied computer science at Leiden university and graduated in 1994 on the topic of artificial intelligence. He did his PhD at the artificial intelligence laboratory of the "Vrije Universiteit Brussel" under professor Luc Steels, where he first worked on learning robot systems and then on the evolution of language. He has worked as a postdoc at the University of Washington in Seattle with professor Patricia Kuhl on modeling infant speech acquisition. He is now assistant professor in cognitive robotics at the artificial intelligence department of the Rijksuniversiteit Groningen.