Investigating the Emergence of Speech Sounds

Bart de Boer
Vrije Universiteit Brussel AI-Lab
Pleinlaan 2
1050, Brussels
Belgium
bartb@arti.vub.ac.be

Abstract

This paper presents a system that simulates the emergence of realistic vowel systems in a population of agents that try to imitate each other as well as possible. The agents start with no knowledge of the sound system at all. Although none of the agents has a global view of the language, and none of the agents does explicit optimization, a coherent vowel system emerges that happens to be optimal for acoustic distinctiveness.
The results presented here fit in and confirm the theory of Luc Steels [Steels 1995, 1997, 1998] that views languages as a complex dynamic system and the origins of language as the result of self-organization and cultural evolution.

1 Introduction

Language is considered to be important for the understanding of intelligence. Although animals are often quite capable of behavior that can be described as adaptive or intelligent, they are not capable, with the possible exception of the higher primates, of the more abstract intelligence (abstract reasoning, working with hierarchical structures, learning of arbitrary mappings) that is characteristic of humans. This more abstract kind of intelligence is of a symbolic nature, and therefore associated with language. Understanding the nature and the origin of language is therefore of crucial importance to the understanding of the nature and origin of human intelligence [Steels 1995, 1997, 1998].

1.1 The origins of language

Some scholars have assumed that the human faculty for language is innate and genetically determined in a very specific way [Chomsky 1980; Pinker & Bloom 1990]. It is obviously true that humans have a unique capability for learning and using language. If a bonobo chimpanzee (our evolutionary closest relative) is raised in the same (linguistic) environment as a human child, it will only learn a very rudimentary set of words, and no grammatical structure, whereas the human child will learn the full language. There are also a number of features of human anatomy (lowered larynx, very accurate control of breathing, accurate control of the tongue) that can only be explained as adaptations to language. However, it is questionable whether the human brain is really so specifically adapted to language that it contains a language organ and a set of "principles and parameters" [Chomsky 1980]. Although a couple of areas in the brain (most notably Broca's and Wernicke's area in the left hemisphere) do seem to be used for language processing in most humans, it is quite possible for other areas of the brain to take over their function. For example, children that are born with damage to these areas, or that receive the damage at a very early age, are still able to learn language very well [Johnson 1997]. Also, the neural pathways in the brain do not seem to be determined in sufficient detail genetically to explain something as specific as the proposed language organ.

It seems more likely that humans have a number of general capacities for learning and abstraction that enable them to learn language. How then did language emerge? Steels [1995, 1997, 1998] considers language the product of *cultural evolution.* Language, from his point of view is a distributed, complex and adaptive system. Important properties of language are that it is spoken in a population, where none of the speakers has perfect knowledge or central control. The language is not dependent on the individual speakers; they can enter and leave the population without changing the language. Also, new words and constructions can be adopted and spread in the language. From his point of view, language is not so much determined by an abstract individual grammar, but is rather an emergent phenomenon of a population of speakers. Whenever a group of humans is brought together, they will spontaneously develop a language. This has actually been observed in the emergence of pidgins and Creoles and in the emergence of sign languages in communities of deaf people [Senghas 1994].

In Steels' theory, humans developed a need to cooperate and communicate under pressure of environmental circumstances. The first communication systems were developed on the basis of the general intelligence of the speakers. Complexity in the language was increased through innovation under the (conflicting) selection pressures of ease of production and ease of understanding.

The first pressure tends to reduce the utterances, while the second one tends to expand them. Variation will be introduced either through speech errors and reductions or through conscious innovation by the speakers themselves. Reproduction of the language is ensured through learning and imitation. All elements for an evolutionary system are present: reproduction, variation and selection. Therefore the process is called cultural evolution. According to Steels, coherence of the language is maintained through self-organization in the population of language users. In this framework it is not the biological evolution that drives the development of language, but rather the development of language that drives the biological evolution through the Baldwin effect [Baldwin 1896].

1.2 The origin of speech sounds

Steels tries to test all of his theories using computer simulations. A number of aspects of language, such as lexicon formation and formation of meanings have already been modeled, both in computer simulations and on robots [Steels 1995; Steels & Vogt 1997, Steels & Kaplan 1998]. The work presented in this paper applies the theory of language as a complex adaptive system to the emergence of speech sounds and more specifically to the emergence of vowels.

Speech sounds are an ideal test case for the role of self-organization and cultural evolution in the emergence of language. Speech sounds are the most physical aspect of language. It is therefore easy to measure their properties and the properties of human speech production and perception. The constraints on a system that works with speech sounds are therefore much more explicit and less controversial than the constraints on a system that works with e.g. grammar. Earlier work [Liljencrants & Lindblom 1972] has shown that in the case of vowel systems, the constraints are mostly acoustic. At the same time, the kinds of sound systems that can appear in human languages are well researched (see e.g. [Lindblom & Maddieson 1988; Schwartz *et al.* 1997a] and references therein). It is therefore easy to verify whether the sound systems that are predicted by the simulation are realistic or not.

Humans can distinguish a large number of different vowels: phoneticians have found at least 44 different basic vowels in the world's languages and the number of different vowel qualities that humans can distinguish in one single language is at least 15 (in Norwegian). However, vowel systems of the worlds' languages do not use a random subset of these vowels. Almost all languages contain [i], [a] and [u] (they appear in 87%, 87% and 82% of the languages in the UPSHW database [Maddieson 1984]) many languages also contain [e] (65%) and [o] (69%). Other sounds are much rarer. Also, if a language contains a back, rounded vowel of a certain height, for example [o], it will usually also contain the front, unrounded vowel of the same height, [e]. In other words, vowel systems tend to be symmetric. Furthermore, the world's languages have a strong tendency towards systems with five vowels, which is neither the minimum, nor the maximum number of possible vowels. Of course, these are just tendencies, not universal rules. There are always languages that are exceptions.

It has already been known for some time [Liljencrants & Lindblom, 1972] that the symmetry of vowel systems, the abundance of certain vowels and the rarity of others can be explained as the result of optimizing acoustic distinctiveness. This has been shown with computer simulations. However, these simulations do not explain who is doing the optimization. No human language learner actively optimizes the sound system he or she learns. Instead, they try to imitate the sound system as accurately as possible. Until now, simulations of vowel systems were forced to explicitly implement the optimization, even in simulations that were based on populations of agents [Glotin 1995, Berrah 1998]. This paper will show that the optimization is an emergent result of self-organizing interactions in the population.

2 The System

The simulations are based on a population of agents that are each able to produce, perceive and learn realistic vowel sounds. For this purpose, they are equipped with a realistic vowel synthesizer, an associative memory for storing vowel prototypes and a model of vowel perception for calculating the distance between the vowel prototypes and the acoustic signals that the agents receive.

2.1 Production and Perception

The production module is an articulatory synthesizer that takes as input the three major vowel parameters and that produces as outputs the first four formant frequencies of the corresponding vowel. The major vowel parameters [Ladefoged & Maddieson 1996, ch. 9] are tongue height, tongue position and lip rounding. In the model the parameters are real numbers in the range [0,1]. For tongue position, 0 means most to the front, for tongue height 0 means lowest and for lip rounding 0 means least rounded. Thus the parameter setting (0,0,0) (in the order position, height, rounding) generates [a], (0, 1, 0) generates [i] and (1, 1, 1) generates [u]. The formant frequencies are defined as the peaks in the frequency spectrum of the vowel. The precise position of the peaks for different vowels depends on the speaker. The articulatory synthesizer that is used here is based on data from [Vallee 1994 pp. 162—164]. For [a] the formant values are (708, 1517, 2427, 3678), for [i] (252, 2202, 3242, 3938) and for [u] (276, 740, 2177, 3506). The mapping from articulatory to acoustic space is highly non-linear. In order to make the simulations more realistic and more interesting, noise is added to all four formant frequencies as follows:

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

where *F,* is the formant frequency without noise, *F,* is the formant frequency with noise and v, is a random value taken from the uniform distribution in the range $\left[\frac{-noise}{2},\frac{noise}{2}\right]$, where *noise* is the noise level of the simulation.

---

[1] UCLA Phonological Segment Inventory Database with 451 languages.

The perception of vowels is based on a comparison with a list of prototypes. Research into perception of linguistic signals has shown that humans perceive them in terms of prototypes. Therefore each agent maintains a list of vowel prototypes. Whenever it perceives a signal, it compares it with all its vowel prototypes and considers the closest prototype as the one that is recognized. The realism of the simulation depends on the distance function. It is based on work by [Mantakas *et al* 1986, Schwartz *et al* 1997b]. It calculates the distance between the acoustic signals of two vowels. This distance is a weighted Euclidean distance between two 2-dimensional vectors that consist of the first formant frequency *F*\ of the vowels and their effective second formant frequency F_2. The effective formant frequency is a non-linear weighted sum of the second to the fourth formant. The idea of the effective second formant stems from the way humans perceive formant patterns. Because of the higher bandwidth of human receptors of higher frequencies, peaks at higher frequencies tend to merge into each other and are perceived as one peak. It is calculated as follows:

$$2)\quad F_2' = \begin{cases} F_2, \text{if } F_3 - F_2 > c \\ \dfrac{(2-w_1)F_2 + w_1F_3}{2}, \text{if } F_3 - F_2 \le c \wedge F_4 - F_2 > c \\ \dfrac{w_2F_2 + (2-w_2)F_3}{2} - 1, \text{if } F_4 - F_2 \le c \wedge F_3 - F_2 < F_4 - F_3 \\ \dfrac{(2+w_2)F_3 - w_2F_4}{2} - 1, \text{if } F_4 - F_2 \le c \wedge F_3 - F_2 \ge F_4 - F_3 \end{cases}$$

where F_2, F_3 and F_4 are the formant frequencies expressed in Bark[2], *c* is a threshold distance, equal to 3.5 Bark, and w_1 and w_2 are weights, which in the original formulation are based on the strengths of the formants. As the articulatory model does not generate the strengths of the formants, they are considered to be proportional to the distance between the peaks, as follows:

$$3)\quad w_1 = \frac{c - (F_3 - F_2)}{c}$$

$$4)\quad w_2 = \frac{(F_4 - F_3) - (F_3 - F_2)}{F_4 - F_2}.$$

Finally, the distance *D* between signal *a* and signal *b* is calculated as follows:

$$5)\quad D = \sqrt{\left(F_1^a - F_1^b\right)^2 + \lambda\left(F_2^{a\prime} - F_2^{b\prime}\right)^2}$$

where *X* is a parameter of the system that determines how the effective second formant frequency should be weighted with respect to the first formant frequency. Investigation of the behavior of this function in prediction of vowel systems [Vailce 1994, Schwartz *et al* 1997b] as well as observations of human perception suggest a value of 0.3 for this parameter.

2.2The imitation game

The interactions between the agents are called imitation games. The intention of the interactions is to develop a coherent and realistic vowel system with which the agents can imitate each other as well as possible from scratch. For each imitation game, two agents are picked from the population at random. One of the agents is the *initiator* of the game, the other the *imitator.* The initiator picks a random vowel from its repertoire. If its repertoire is empty (as is the case at the beginning of the simulation) it adds a random vowel. It then produces the acoustic signal of that vowel. The other agent listens to this signal and finds its closest prototype. If its prototype list is empty, it finds a good imitation by talking and listening to itself, while improving the signal using a hill-climbing heuristic. It then produces the acoustic signal of the vowel it found. The initiator then listens to this signal and finds its closest prototype. If this is the same prototype as the one it used to initiate the game, the game is successful. If it is not the same, it is a failure. It communicates the success or the failure of the game using non-verbal feedback. Explicit non-verbal feedback is usually not given to children that learn language. However, they do get feedback on the quality of their communication through gesture, facial expression or the achievement (or lack thereof) of the communicative goal.

The imitator and the initiator react to the language game in a number of ways. Both update the use count of the vowels they produced. If the game was successful, they also update the success count. On average every ten imitation games, the agents throw away vowels that have been used at least 5 times and have a success/use ratio that is lower than 0.7. They also merge prototypes that are so close together in articulatory space that they will always be confused by the noise that is added.

The imitator also modifies its vowel inventory depending on the outcome of the imitation games. If the imitation game was successful, it shifts the vowel prototype it used closer to the signal it perceived in order to increase coherence. In the case the imitation game was a failure, this can have two reasons: cither the initiator has more prototypes than the imitator, causing confusion, or the imitator simply used a bad phoneme. If the success/use ratio of the used vowel is low, then it is considered to be bad, and it is shifted closer to the perceived signal in the hope that it will be improved. If its ratio is high, this means it was used successfully in earlier games, so the reason of the failure was probably confusion. Therefore, a new prototype is added that is a close imitation of the perceived signal, using the same hill-climbing procedure that was used to add first prototypes.

A last possible change of the agents' vowel inventories is random addition of a new vowel (with probability typically 0.01). This is done in order to put a pressure on the agents to increase their number of vowels. In humans this pressure could for example come from a need to express new meanings. Iterating the imitation game in a large enough population of agents results in the emergence of realistic vowel systems.

---

[2] A (partly) logarithmic frequency scale based on the properties of human perception. An equal interval in Bark corresponds to an equal perceptual distance.

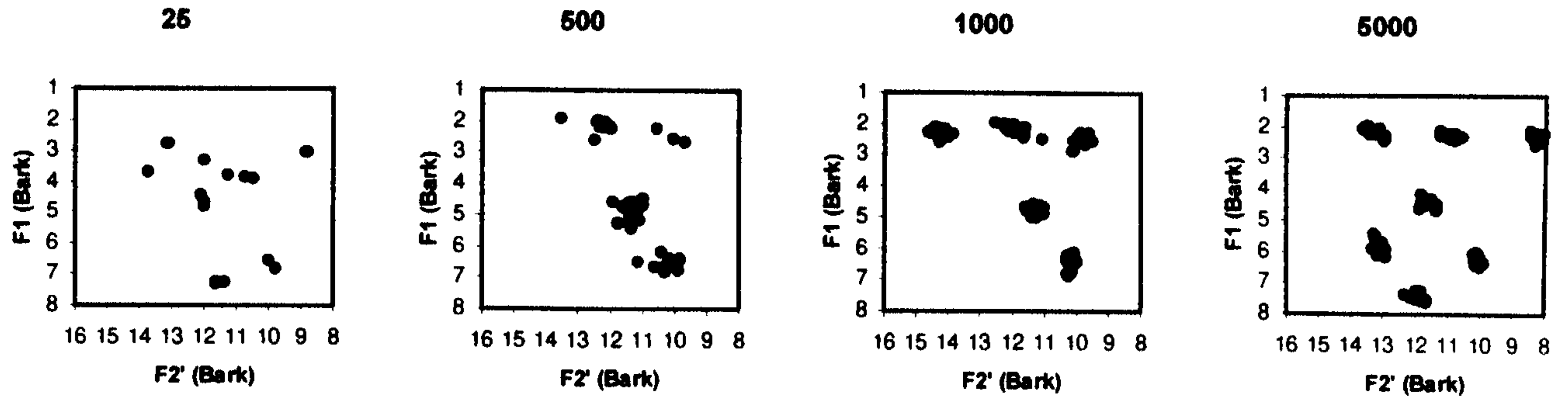

Figure 1: Emergence of a vowel system in a population of 20 agents with 10% noise.

3 The Results

The first result, which is shown in figure 1, is the emergence of a vowel system in a population of twenty agents and a noise level of 10%. In this figure, all vowel prototypes of all agents in the population are plotted in the acoustic space formed by the first and effective second formant. The first formant is plotted on the vertical axis and the effective second formant is plotted on the horizontal axis. The scales of the axes are in Barks. Note that the direction of the axes is reversed with respect to the usual direction of axes in graphs. This has been done in order to put the vowels in positions that correspond to the positions that they are usually given by linguists, with front vowels in the left- and high vowels in the upper part of the graphs. Note also that articulatory limitations cause vowels to be produced only in a roughly triangular region, with the apex at the bottom of the graph.

The leftmost frame of the figure shows the system after 25 imitation games. It can be seen that the distribution of the agents' vowel prototypes is still quite random, although vowel prototypes tend to occur in pairs. This is because the main factors at work are the random addition of vowel prototypes and the direct imitation of these. After 500 imitation games, in the second frame of figure 1, the main factor at work is a clustering of the agents-vowel prototypes. All agents in the population already have a vowel prototype near one of these clusters. Most imitation games will therefore be successful. In response to this the agents will shift their vowel prototypes closer to the corresponding vowel prototypes of the other agents. Because of the noise with which vowels are produced, however, the clusters remain a certain size and do not reduce to points. Between 1000 and 5000 imitation games, the number of clusters increases until the available acoustic space is filled evenly with vowel clusters. The resulting vowel system consists of [i], [e], [a], [o], [u], [ɨ] and [ə] a system that is natural and that occurs for example in the Sa'ban language of Borneo. The artificial vowel system can be compared with measurements of a real vowel system in figure 2 (note that the scales in this figure are linear!) The system keeps on changing from this stage on, even though the changes are much less rapid. Vowel clusters might change position new and vowel clusters sometimes appear, get merged or split. But the appearance of the system remains the same.

Not all simulations with the same parameter settings result in the same vowel system. Sometimes the number of clusters is smaller, and their position might be different. This is illustrated in figure 3. This figure was generated by running 1000 times a run of 5000 imitation games with the same parameter settings as were used for figure 1. It shows the frequencies of the average sizes of the vowel systems of the agents in each of the populations that resulted from every run of 5000 games. Peaks occur at different integer values. This indicates both that systems of different sizes emerge and that the average size of the population's vowel systems tends towards integer numbers. This is because agents in the same population usually have the same number of vowels, indicating that the emerged vowel systems are coherent.

Vowel systems that emerge for the same parameter settings do not only have different sizes, but within the same system size, different distributions of the vowel prototypes in the acoustic space are found. Figure 4 shows this for systems with five vowel prototypes. The

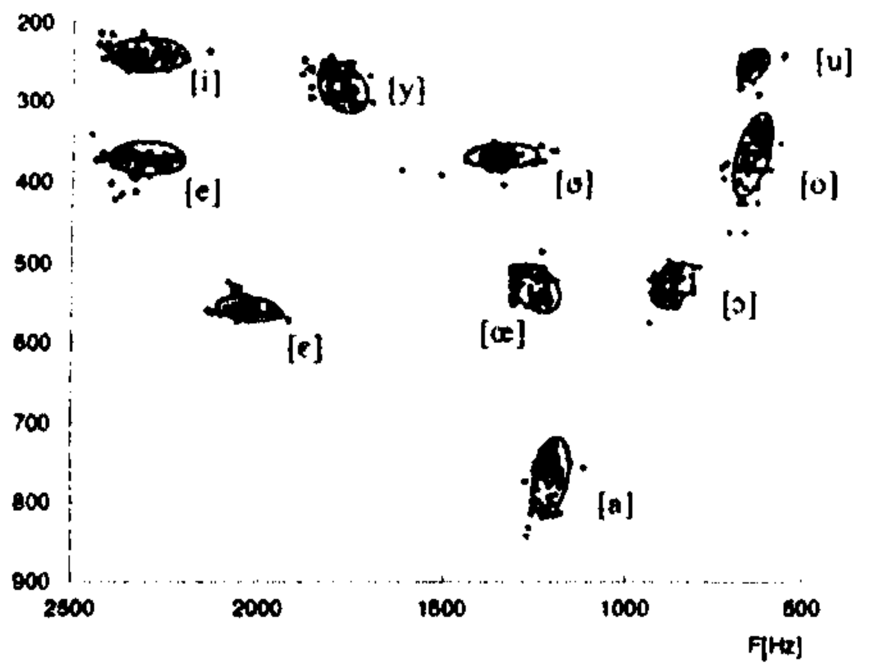

Figure 2: Vowel system of French, [Rober-Ribes 1995]

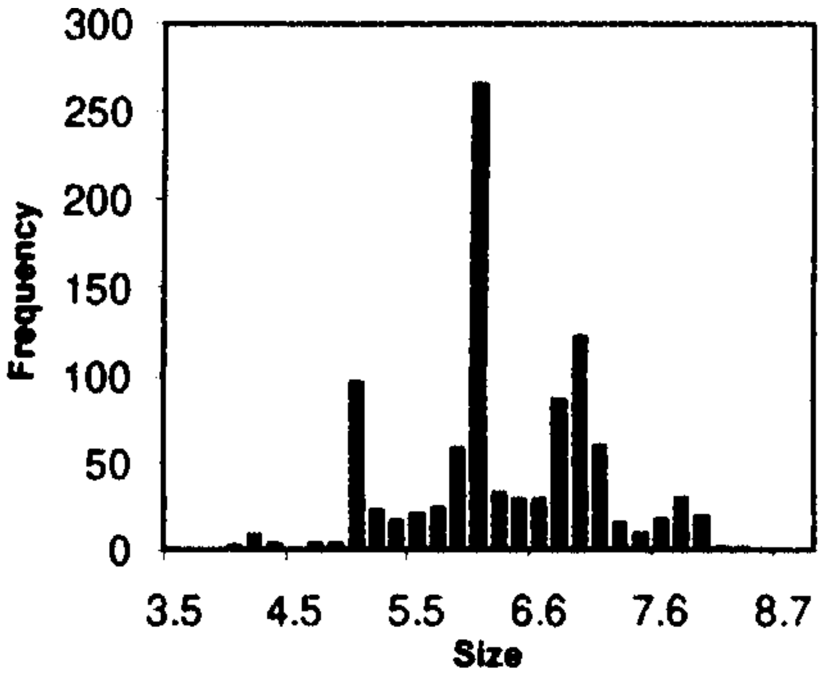

Figure 3: Size distribution of 10% noise systems.

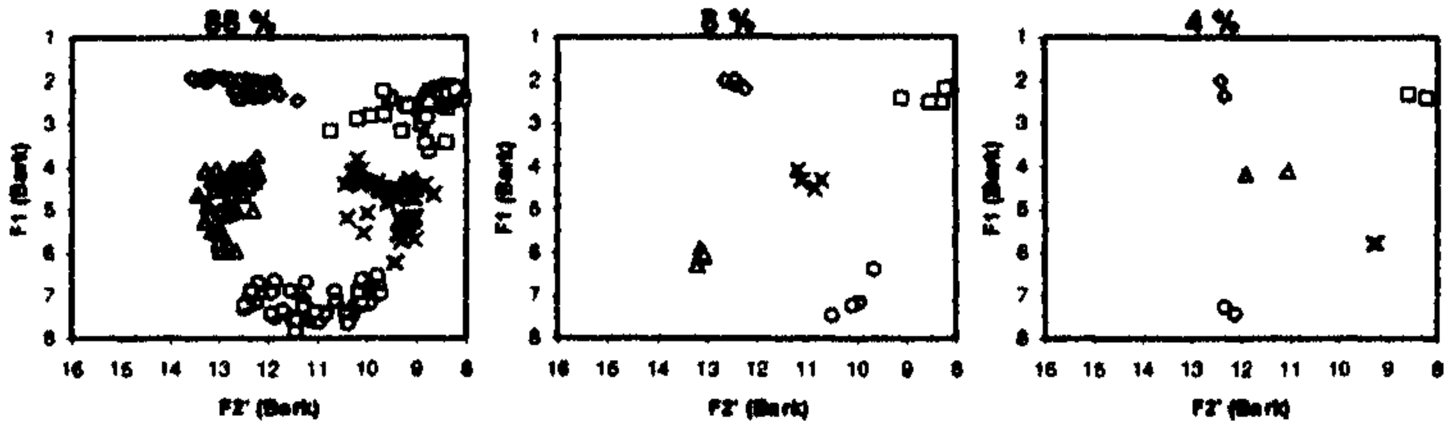

Figure 4: Vowel configurations for five vowel systems.

systems were obtained from running the simulation with 15% acoustic noise, for 25 000 imitation games. Of the 100 runs, 49 resulted in populations with on average five vowels per agent. From each of these populations, one agent with the average number of vowels was taken at random. The vowel systems of these agents are shown in the figure, classified by type. It is found that the symmetric type occurs in 88% of the cases, the type with a central vowel and more front vowels occurs in 8% of the cases and the type with a central vowel and more back vowels occurs in 4% of the cases. This agrees very well with what has been found in natural languages. Schwartz *et al.* [1997a] found that in a previous version of UPSID (with 317 languages) 89% of the languages had the symmetric system, while the two types with the central vowel each occur in 5% of the cases. For different system sizes similarly good matches between emerged systems and human vowel systems are found, except for the smallest inventories (of three and four vowels) where discrepancies occur for the less frequent systems.

The outcome of the simulations does not depend very sensitively on the settings of the different parameters. Although the number of vowel clusters and their distribution are different for different parameter settings, their distribution is realistic in the sense that they could occur in human languages. Unfortunately space is too limited to show this in detail (see de Boer, *in preparation).*

A further observation of human languages is that they have a preference for vowel systems consisting of five vowels, and especially the symmetric system shown in figure 4. This is remarkable, because five is neither the minimum, nor the maximum number of vowels found in human languages. Apparently the frequency with which vowel system sizes occur is non-monotonic with respect to the number of vowels. This same phenomenon appears in the simulations. Simulations were run for values of the noise parameter ranging from 8% to 24% with increments proportional to the noise value (so that each parameter change has equal influence). The frequencies of the different vowel system sizes are plotted. This is shown in figure 5. The solid line shows the frequency of sizes of actual human vowel systems and the dashed line shows the frequency (which is of course relative to the total sample size) of sizes of emerged vowel systems. Both lines show a peak, but unfortunately, the peak for human systems occurs at 5 vowels, while the peak for artificial systems occurs at 4 vowels. This can probably be explained by the fact that the perception model is not perfect, so that high front vowels tend to be centered too much. This is probably also the explanation for the fact that predictions for configurations with 3 and 4 vowels are not accurate.

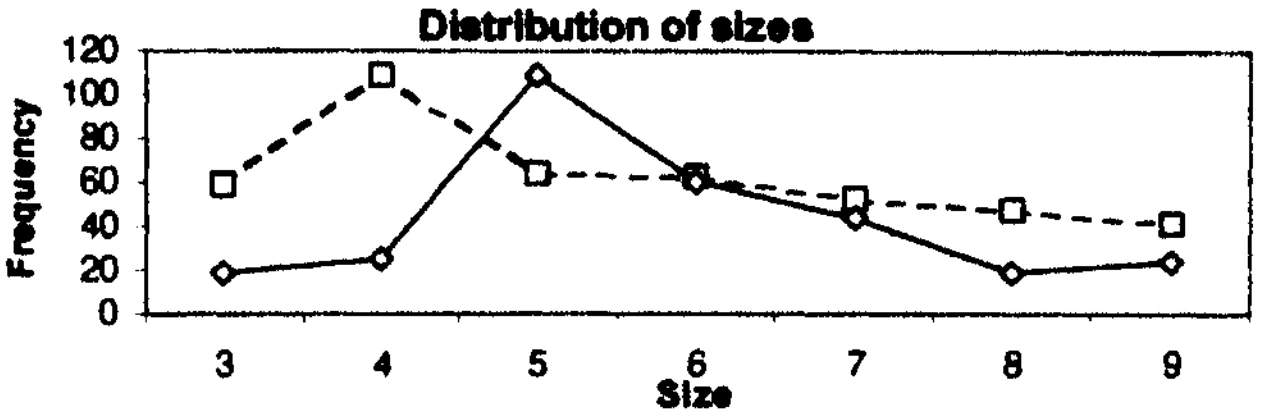

Figure 6: Size distribution in real and artificial systems.

It has now been shown that self-organization can predict the vowel systems that occur in human languages to a large degree of accuracy. But would it really be as robust as Steels' [1997, 1998] theory claims? It has already been shown that it is robust against changes in the language itself It is also robust against changes in the population. This is shown in figure 6. The gray squares in this figure show the starting vowel system of a population of 50 agents. The population was then run for 15 000 imitation games. There was a probability of 1% per language game of taking an old agent from the population or inserting a new (empty) one in the population. The black circles show the system afterwards. By that time the whole population has been replaced. The vowel system has simplified a bit, but has remained mostly the same. It can thus be concluded that vowel systems are robust against changes in the population.

4 Conclusion

The simulations of populations that develop vowel systems clearly show that self-organization under constraints of perception and production is able to explain the structure of the vowel systems in human languages. The agents and their interactions form a dynamic system, in the sense described by Steels' [1995, 1997, 1998] theories. The most frequently occurring systems can be considered attractors of this dynamical system. Due to the random influences—noise on the articulations, random choice of agents—the populations never quite settle in exactly one of these attractors. They can settle in several different near-optimal configurations, just as human languages do not always have the optimal systems as predicted by optimization models [Liljencrants & Lindblom 1972; Schwartz *et al* 1997b]. Although the agents do not have an innate predisposition towards certain vowel configurations, some seem to be preferred over others. This

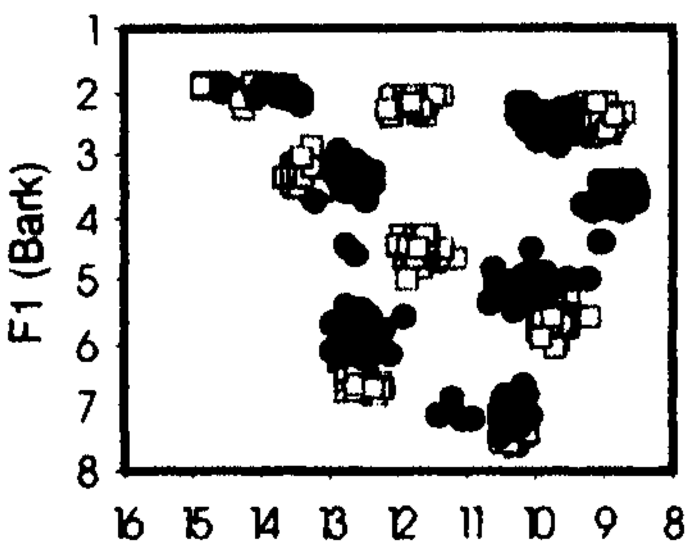

Figure 6: Vowel system conservation under population replacement.

is not, as a naive observer might think, the result of innate rules and representations, but of self-organization in the population.

Just like human languages, emerged vowel systems also never quite stop changing. The systems that emerge are also robust to changes in the language and to changes in the population, just as required by any realistic model of language. The fact that sound systems can be transferred reliably from one generation to the next opens the possibility of cultural evolution.

Many things still need to be investigated: more complex utterances (so that not only acoustic constraints have to be taken into account, but also articulatory ones) and more realistic signals (so that the predictions match even better with real languages) are the ones that come to mind first. Nevertheless, these simulations already lend strong support to Steels' theory that language is a complex dynamic system and that self-organization and cultural evolution have played important roles in its emergence.

References

[Baldwin 1896] J. Mark Baldwin (1896) A new Factor in Evolution, *The American Naturalist* 30 (June 1896) pp. 441-451,536-553.

[Berrah 1998] Ahmed Reda Berrah *Evolution Artificielle d'une Societe d'Agents de Parole: Un Modele pour l'Emergence du Code Phonetique,* These de I'Institut National Polytechnique de Grenoble, Speciality Sciences Cognitives, 1998

[Chomsky 1980] Noam Chomsky, Rules and representations, in *The behavioral and brain sciences* 3, pp. 1-21, 1980

[de Boer, *in preparation]* Bart de Boer *Self Organisation in Vowel Systems,* Vrije Universiteit Brussel AI-Lab Ph. D. thesis, to be defended June 1999.

[Glotin 1995] Herve Glotin *La Vie Artificielle d'une socidtd de robots parlants: emergence et changement du code phonttique.* DEA sciences cognitives-Institut National Polytechnique de Grenoble 1995

[Johnson 1997] Mark H. Johnson *Developmental Cognitive Neuroscience,* Oxford: Blackwell, 1997

[Ladefoged & Maddieson 1996] Peter Ladefoged and Ian Maddieson *The Sounds of the World's Languages,* Oxford: Blackwell, 1996

[Liljencrants & Lindblom 1972] L. Liljencrants and Bjorn Lindblom (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast, *Language* 48: 839-862, 1972

[Lindblom & Maddieson 1988] Bj6rn Lindblom and Ian Maddieson, Phonetic Universals in Consonant Systems, in: Larry M. Hyman & Charles N. Li (eds.) *Language, Speech and Mind,* pages 62-78, 1995

[Maddieson 1984] Ian Maddieson *Patterns of sounds,* Cambridge University Press, 1984 [Mantakas *et al* 1986] Mantakas, M, J.L. Schwartz & P. Escudier *Modele de prediction du 'deuxiime formant effectif* F_2*'—application a l 'etude de la labialite des voyelles avant du frangais.* In: Proceesings of the 15th journees d'etude sur la parole. Societe Francaise d'Acoustique, pages 157-161.

[Pinker & Bloom 1990] Steven Pinker and P. Bloom Natural Language and Natural Selection. The *Behavioral and Brain Sciences* 13: 707-784, 1990

[Rober-Ribes 1995] Jordi Rober-Ribes *Modeles d'integration audiovisuelle de signaux linguistiques.* These de docteur de I'Institut National Polytechnique de Grenoble, 1995

[Schwartz *et al..* 1997a] Jean-Luc Schwartz, Louis-Jean Boe, Nathalie Vallee and Christian Abry (1997a), Major trends in vowel system inventories. *Journal of Phonetics* 25:233-253, 1997

[Schwartz *et al..* 1997b] Jean-Luc Schwartz, Louis-Jean Boe, Nathalie Vallee and Christian Abry (1997b), The Dispersion-Focalization Theory of vowel systems, *Journal of Phonetics* 25: 255-286, 1997

[Senghas 1994] Ann Senghas Nicaragua's Lessons for Language Acquisition, in: *Signpost* 7(1), pp. 32—39,1994

[Steels 1995] Luc Steels A Self-Organizing Spatial Vocabulary, *Artificial Life* 2(3): 319 332, 1995

[Steels 1997] Luc Steels, The Synthetic Modelling of Language Origins, *Evolution of Communication* 1(1): 1—34, 1997

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

[Steels & Kaplan 1998] Luc Steels and Frederic Kaplan (1998) Spontaneous Lexicon Change. In: *Proceedings of COLJNG-ACL 1998, Montreal,* pp. 1243 1249, 1998

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

[Vallee 1994] Nathalie Vallee, *Systemes vocaliques: de la typologie aux predictions.* These preparee au sein de TInstitut de la Communication Parlee (Grenoble-URA C.N.R.S. no 368) 1994