A Second Report on Emergent Phonology

Bart de Boer
AI-lab
Vrije Universiteit Brussel

Abstract
This report presents a simulation of the emergence of a sound system in a group of simulated agents that try to imitate each other's sounds. The agents can articulate and perceive vowel sounds. Agents engage in *imitation games* in which they try to imitate each other by using their own phonemes to approximate a sound that was made by another agent.
By iterating the imitation games in a population of agents, sound systems emerge that look remarkably like human vowel systems. It is argued that language-like interactions and local updates of the individual agents' vowel systems are sufficient to explain the emergence of coherent sound systems. Because of the acoustical constraints on the system phonological universals emerge. The sound systems that are generated will only contain vowels that can easily be distinguished from each other.
The model is compared and related to other work that has used computer models to investigate universals of vowel systems. It is argued that the present model is more psychologically plausible than the previous work and that in contrast with the previous work, it can also be used to explain the learning of sound systems.

Keywords:
language origins, phonological universals, communicating agents, cultural evolution

1. Introduction

The languages of the world contain a surprising number of different speech sounds. In total some 600 phonetically different speech sounds can be identified[9,13]. However, no language uses more than a small subset of these possible sounds, usually between 20 and 37, according to Maddieson[13]. The minimum number of sounds is 11 (Rotokas and Mura) and the maximum is 141 (!Xũ)[*].

These observations indicate that phonology cannot be innate. Most languages use a small number of sounds, but these are not at all the same for all languages. Children apparently learn the speech sounds of their parents. However, a number of regularities in the sound systems of different languages, called phonological universals, can be observed. Some sounds occur much more frequently than others. The sounds [a] and [u], for example, are much more common than the sound [ʏ]. Also, sound systems usually show a great deal of internal regularity. If for example, a language has voiced stops at certain places of articulation, it almost always also has the corresponding unvoiced stops, i.e. if it has a [d] it also has a [t]. As we have already observed that phonology cannot be innate, there must be a functional explanation for these phonological universals.
A number of researchers, notably Liljencrants and Lindblom[10], Vallée[17], Boë et al.[2] and Carré et al.[3] have been able to explain universals in (mainly) vowel systems with computer models using criteria of auditory distinctiveness and articulatory simplicity. However, they study phonology in an essentially static way. They only look for stability criteria that explain the form of sound systems that are found in the world's languages. They are not able to explain how these stable systems are reached in the inherently dynamic and noisy interactions of language users.
Recently Berrah et al.[1] have attempted to build a dynamic model that explains phonological universals, in modelling interactions between agents that use language, and by using selection and evolution. The work presented here is related to their work in that a model is provided that can explain phonological universals from dynamic interactions between language users. However, the underlying mechanisms are quite different, but in the basic functions that implement articulation and perception of the agents, the work of the Grenoble group is followed as closely as possible, in order to make the results maximally comparable.
The experiments that will be described here are based on a population of simple agents that can produce and perceive speech sounds, and that try to imitate each other's speech sounds. Speech sounds are gen-

[*]Although an analysis of the clicks with secondary articulations in this language as clusters instead of single phonemes could possibly reduce the number.

erated, shifted or discarded according to the outcome of these imitation games[†]. Coherence and communicative efficiency is reached through self-organisation. Evolution in the traditional, biological sense of the word does not play a role. One could, however, say that there is some kind of cultural evolution.
The work described here is related to Luc Steels'[15] research into the origins of language. According to Steels, the origins of language should not be explained by biological evolution of an innate language ability, such as advocated by Pinker[14]. His hypothesis is that language is an emergent phenomenon of a population of intelligent agents that can benefit from communication. Coherence and complexity are achieved through processes such as co-evolution, self-organisation and level-formation. Steels has implemented his ideas in computer simulations in which populations of artificial agents engage in language-like interactions. His research has mainly been in the area of the development of lexicons and the development of meaning. The work presented here tries to apply these ideas to the area of phonology, in trying to explain the emergence of coherent sound systems with constraints in a population of communicating agents.
The first report[7] contained a description of a system that used an articulatory synthesiser and dynamic articulations, as well as a description of a system that used binary feature based phonemes. The first system did show some promising results, but it was so complicated that evaluating the significance of these results was extremely difficult. The second system showed clearly that the idea of self-organisation through interactions between agents in the context of speech sounds was valid, but the phonemes that were used were not comparable to the phonemes used in human speech.
The system that will be described in this report is a step backward from the first system, and a step forward compared to the second system. It does use a natural articulatory synthesiser, but focuses only on vowels, which are uttered in isolation. No dynamic articulators are used, so no co-articulation effects are present. The current system is heavily influenced by research results of Björn Lindblom[12] of Stockholm university and of the Institut de la Communication Parlée in Grenoble[2].
The next section will describe the architecture of the agents used in the simulations. Their articulatory and perceptual apparati will be described in considerable detail. Section 3 will describe the rules of the interactions between the agents. The design process that has led to the particular imitation game used in this paper will also be described. In section 4 the experimental setting will be explained and a number of experimental results will be presented. In section 5 the results will be compared with those of similar research that uses computational means to explain universals of vowel systems. In section 6 future extensions of the research are presented and some conclusions are drawn from the present results.

2. The Agents

The agents that are used in the computer simulation use vowels to "communicate" with each other. For this purpose, each agent has its own list of vowels. The lists of vowels for each agent are initially empty, and will be filled as the agents engages in interactions with other agents. This process is described in more detail in section 3. The vowels themselves are represented by the three main parameters that are used for describing vowels: tongue position, tongue height and lip rounding. The three parameters can have any value between zero and one. For tongue position, zero means front, and one means back. For tongue height, zero means low and one means high, and for lip rounding, zero means unrounded and one means rounded.
The agents are able to produce any "simple" vowel. Variations on vowels, such as nasalisation, pharyngealisation or change of voicing type cannot be simulated by the present system. However, the system is completely language-independent. No bias towards the vowel system of any language is present in the agents.
The vowels that are present in the agents are produced by a synthesiser and are recognised by a perception unit. The communication process, and the evaluation of vowels is regulated by a special control unit. The internal architecture of an agent is illustrated in figure 1. The role of the evaluation and control unit is to implement the rules of the language game, which will be discussed in the next section. Here we will describe the synthesiser and the process of perception.

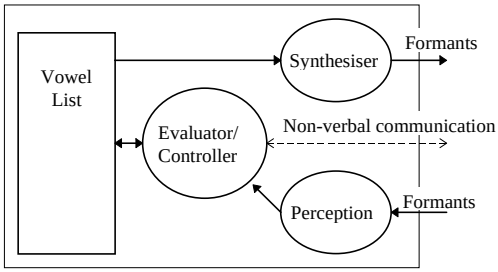

figure 1: Agent architecture

[†] My imitation games should not be confused with Suzuki's[16] imitation games. Both the way they are played as well as their purpose are quite different.

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

figure 2: Synthesiser equations

The synthesiser is a simple articulatory synthesiser that is based on a second order interpolation of a number of artificially synthesised vowels. The input of the synthesiser consists of the three articulatory parameters and the output consists of the frequencies of the first four formants of the vowel associated with this particular articulation. The basic data for the formants have been taken from Vallée's thesis[17, pp. 162–164]. The vowels:[i,ɨ,ɯ,e,ə,ɤ,ạ,a,ɑ,y,ʉ,u,ø,ɵ,o,ɒ] were used as basic data points, while the formant frequencies of [œ̣,œ] where estimated. After this a second order interpolation of this data was calculated. The resulting function is shown in figure 2. This interpolation function is used to calculate the formant frequencies from the articulatory parameters. The function works reasonably well and requires much less calculation time than any other articulatory synthesiser. The disadvantage is, that it is limited to the vowel space that was used to define the data points on which the interpolation is based.

In some experiments a certain amount of noise is added to the formant frequencies that are produced by the agents. The adding of noise consist of multiplying the formant frequencies by:

1) $1 \pm U(a)$,

in which $U(a)$ is a random variable uniformly distributed over [-*a*,*a*], where *a* varies for different experiments. The addition of noise makes the imitation games more natural, as in reality it can not be expected that sounds will always be produced and perceived accurately. It also makes it impossible for the agents to copy each other's phonemes perfectly, thereby forcing them to create sound systems in which the phonemes are not too close together, as well as opening the possibility of change and language evolution.

The agents also have to be able to interpret sounds they hear in terms of their phonemes. The synthesis function is rather hard to invert. Also, it would not be biologically plausible to equip the agent with a function that inverts the articulation function; children have to learn how to reproduce a speech sound they hear. Therefore the agents should learn how to recognise their own phonemes. This could possibly be done by training a neural network, but this would be very costly computationally, and would introduce a number of extra uncertainties that have to do with the training parameters of the neural network. Therefore, it was decided to use prototype vectors for recognising phonemes.

For each phoneme an agent creates, it generates the formants of an ideal articulation of this sound. This ideal articulation is called the *prototype vector* and it is stored together with the articulatory description of the phoneme. Every time an agent hears a sound, it calculates the distance between the prototype vectors of all the phonemes it knows and the formants of the sound it just heard. The phoneme with the prototype vector that is closest to the sound that was heard, is considered as the recognised phoneme. This whole process could in principle be implemented using neural networks, thereby increasing the biological plausibility.

The distance measure that is used to compare phonemes is of crucial importance to the form of the vowel systems that will be generated by the agents. In order to get natural vowel systems, and in order to be able to compare the results of the experiments with those of at least one other group, a distance measure that has been defined by Boë et al.[2]was used in a slightly modified form. The distance measure takes into account that the human auditory system distinguishes vowels by their formant frequen-

cies, lower formants having a greater influence, that it does not distinguish well between formants that are very close together and that it works in an essentially logarithmic manner.
The original distance measure of Boë et al. is a weighted distance in the F_1-F_2' space, where F_1 is the frequency of the first formant (expressed in Bark[‡]) and F_2' is the weighted average of the second, third and fourth formants (also expressed in Barks). This weighted average is calculated as follows: if the distance between the second and third formant is higher than a critical value (chosen to be 3.5 Bark) than F_2' is taken to be the second formant. If the distance between the second and third formant is less than the critical distance, but the distance between the second and the fourth formant is more than the critical distance, then F_2' is taken to be the weighted average of the second and the third formant, where the formants are weighted according to their relative strengths. If the distance between the second and the fourth formant is also less than the critical distance, then F_2' is taken to be the weighted average of the second and the third formant if the third formant is closer to the second formant than to the fourth. If the third formant is closer to the fourth, the weighted average of the third and the fourth formant is taken to be F_2'.
As our synthesiser does not calculate the strengths of the formants, another way of calculating the weighted average had to be used. The weights in our function are not based on the strengths of the formants, but on their distance. This is physically plausible, as the strengths of the formants usually drops as their frequency gets higher, and as the distance to the previous formant gets bigger.
Now two weights can be calculated:

$$2)\ w_1 = \frac{c - (F_3 - F_2)}{c}$$

$$3)\ w_2 = \frac{(F_4 - F_3) - (F_3 - F_2)}{F_4 - F_2}$$

Where w_1 and w_2 are the weights, F_1-F_4 are the formants in Bark and c is the critical distance.
The value of F_2' can now be calculated as follows:

$$4)\ F_2' = \begin{cases} F_2, & \text{if } F_3 - F_2 > c \\ \dfrac{(2 - w_1)F_2 + w_1 F_3}{2}, & \text{if } F_3 - F_2 \le c \text{ and } F_4\text{-}F_2 > c \\ \dfrac{w_2 F_2 + (2 - w_2)F_3}{2} - 1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 < F_4 - F_3 \\ \dfrac{(2 + w_2)F_3 - w_2 F_4}{2} - 1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 \ge F_4 - F_3 \end{cases}$$

The values of F_1 and F_2' for a number of vowels is shown in figure 3. We can see from this figure that the distribution of the vowels through the acoustic space is quite natural. However, as it is a 2-dimensional projection of an essentially 3 dimensional space, not all distances between all phonemes can be represented accurately. This is especially the case with the distinction rounded-unrounded. Unfortunately this is difficult to avoid in any system.
The distance between two vowels, *a* and *b* can now be calculated using a weighted Euclidean distance:

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

This again, in accordance with the work of Boë et al.[2]. The value of the parameter λ is chosen to be 0.5 for all experiments that will be described.

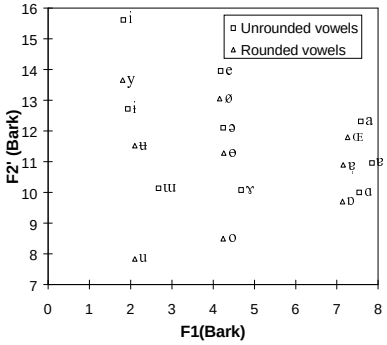

figure 3: Vowels in F1-F2' space

With the articulator function and the perception function that have been described in this section, the agents can produce and perceive speech sounds in

[‡] The Bark scale is an (approximately) logarithmic frequency scale, based on the peculiarities of the human auditory system. An identical distance in Bark between frequencies indicates that the interval between the frequencies is perceived as equal.

a way that is sufficiently human, so that the results that are generated with this systems can at least to some extent be compared to the results of research into human sound systems.

3. The Imitation Game

The experiments presented in this work are concerned with the emergence of a coherent and useful phonology in a population of initially empty agents. In order to investigate how this can happen, the agents engage in exchanges of sounds, so-called imitation games, the goal of which is to learn each other's speech sounds. If necessary, speech sounds are invented, in order to get the communication started, and also in order to introduce more possible sounds in the population.

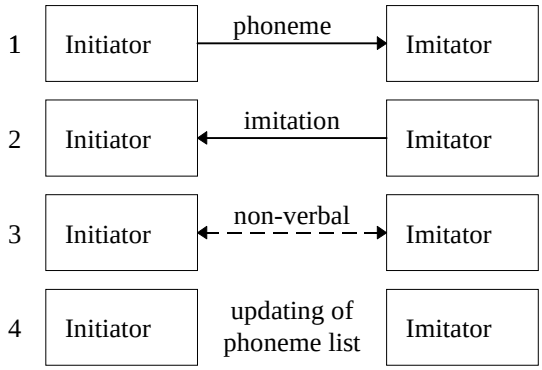

figure 4 The imitation game

The basic rules of the imitation game that is played by two agents, are very simple. Two agents are selected from the population of agents. One of the agents, which we will call the *initiator*, selects one of its phonemes and says this to the other agent. The other agent, which we will call the *imitator*, interprets this sound in terms of its phonemes, and then produces the phoneme it thinks it has recognised. The other agent listens to this imitation, and also interprets it in terms of its phonemes. If the phoneme it recognises is the same as the one it just said, the imitation game is considered to be successful. If it is not equal, the game is unsuccessful. There follows a non-verbal communication, in which the imitator gets to know if its imitation was correct or not. The whole process is illustrated in figure 4.

For each phoneme in the phoneme list of both the initiator and the imitator, the number of times it is used and the number of times it was successful are kept. Every time a phoneme is uttered in a language game, its use count is increased. Every time it was successfully imitated, its success score is increased. If it was not successfully imitated, nothing happens to the success score. The quality of a phoneme is this success score divided by the number of times it was used.

Depending on the course of the language game, the initiator and imitator can change their repertoire of phonemes. The phoneme lists of the agents are initially empty, so at first the initiator has to choose a random articulator position, and use this as its first phoneme. If the phoneme list of the initiator is also empty, it tries to make an imitation of the sound it just heard, by saying sounds to itself, and using a hill-climbing heuristic in order to approach the sound it just heard. It then adds this imitation to its phoneme list.

If the initiator already has a list of phonemes, it picks one of these at random and utters it, or creates a new phoneme with a very small probability. If the imitator already has a list of phonemes, it picks the closest match (as described above) and uses this as imitation. If the imitation was successful, the imitator tries to approach the sound it just heard even more by shifting the phoneme it said a bit closer to it, again using a hill-climbing heuristic. If the imitation was not successful, and if the quality of the phoneme was low, the phoneme is also shifted, in order to try to make its imitation better. However, if the quality of the phoneme was high, the phoneme was not shifted, as from its high score we can infer that it must be a good imitation of another phoneme. Therefore, we create a new phoneme (using again a hill-climbing heuristic) that has a sound that is similar to the sound that had to be imitated.

Two other processes are going on. Firstly, phonemes that have low quality for a long time are removed from the phoneme list. With a certain probability, the initiator's phonemes that have a quality score that is below a certain threshold, are removed. Secondly, phonemes that are too close together, are merged. Phonemes are considered too close together if they are so close together that they can be confused through the noise that is added to the formant frequencies. The fusing is done by taking the articulator position of the phoneme with the highest score as the new articulator position. The success and use counts of the new phoneme are then calculated by adding the success, respectively the use counts of the old phonemes.

All the steps of the language game as have been described above, are both necessary and could in principle be performed by humans. The creation of new phonemes is necessary in order to cope with new phonemes the agent hears from the other agents. Agents need to shift phonemes closer to the sounds they heard for two reasons. The first is to increase coherence in the population. This is also observed in

human language. People do not just imitate sounds sufficiently, they actually make quite accurate copies of the sounds they hear from the people they learn their language from. The second reason phonemes are shifted towards each other, is to accommodate as many phonemes as possible. If the phonemes stay at the positions where they were created, they start interfering if more phonemes are added to the phoneme list. If there are few phonemes, it does not matter if corresponding phonemes in different agents are not quite the same. The more phonemes there are, however, the higher the probability becomes that new phonemes get confused with the original ones. The shifting of phonemes is implemented using hill-climbing, which means that phonemes are shifted a small step in a direction that improves the accuracy of the imitation. This is something humans can do in principle, although they probably use a more complicated, neural mechanism.

Phonemes that are too close together are merged. This is necessary in order to prevent confusion of phonemes that are good in principle. It is possible that a new phoneme is generated close to an existing phoneme, or that a phoneme is shifted quite close to another phoneme. Although both phonemes might be good, it is likely that they will interfere with each other and cause the agent to confuse sounds that it hears. By merging the phonemes, this problem is solved. However, if the two phonemes that are merged correspond to two phonemes in the other agents, the agent will now not be able to imitate one of these phonemes properly. But a new phoneme will quickly be added through the mechanism that adds a phoneme if a phoneme with a high quality score was used in an unsuccessful language game. The quality score of the phoneme that results from the merge will necessarily be high, if the two original phonemes corresponded reliably with phonemes in the other agents.

Phonemes with a low quality score are discarded. This is necessary, because apparently they do not play any useful role, and because they take up a place in the articulatory space, possibly degrading the quality of other phonemes.

4. The Experiments

In this section we will present a number of experiments that have been conducted with the language games and the agents described above. The goal of the experiments was to investigate whether it was possible to develop a successful sound system in a population of initially empty (*tabula rasa*) agents, and what form this sound system would take under different conditions of noise, and in different population sizes. The experiments that were conducted consisted of a large number of iterated imitation games in a homogeneous population of agents. This means that any two agents have an equal chance of becoming engaged in an imitation game. Two measurements were made during the experiments: the success of the imitation games and the form of the sound systems of the agents at the end of the run. The runs were ended after an arbitrary number of imitation games were done.

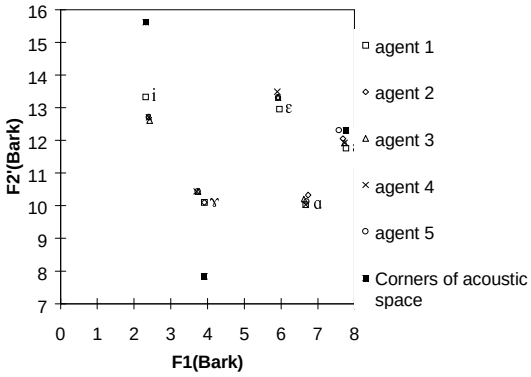

figure 5: Sound systems of five agents

Of course a lot of development and preliminary experimenting was necessary to come up with the present system. The form of the present language games, and the architecture of the agents is therefore at least as interesting as the experimental results.

The first experimental results are presented in figure 5. It shows the sound systems that were developed in a population of five agents after 1000 imitation games were played. The acoustic realisation of the phonemes was subject to 10% noise (a=0.1 in equation 1). It is clear from the clusters in the figure that the five agents share the same phonemes. The corresponding phonemes for the different agents are close together, while the phonemes within one agent are far apart. This is optimal for a sound system that is meant for communicating different sounds between agents. It can also be observed that the phonemes are spread through the available acoustic space in a way that is reminiscent of the way vowels of human languages are spread through acoustic space, even though the vowel system that was arrived at: [i,ɛ,a,ɑ,ɤ] probably does not appear in any human language.

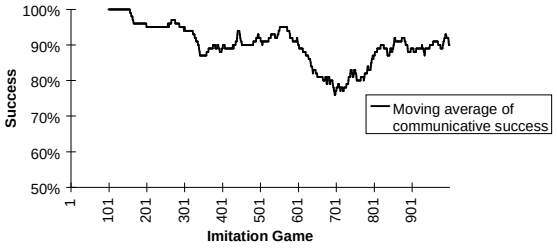

figure 6: Success of agents with 10% noise

The communicative success of the agents, as illustrated in figure 6 is constantly between 70% and 100%. The

success starts at the 100% level in the beginning of the experiment, because at that time the agents only have one phoneme each and confusion is not possible. As soon as the agents start creating new phonemes, however, the success score drops, because phonemes are being confused. After a while, the agents succeed in making copies of the phonemes, and the success score returns to near 90%. The result shown are of the same run that resulted in the sound systems of figure 5, and are representative for the runs that are normally generated by the simulation.

If the amount of noise in the formant frequencies is increased, the area over which phonemes are "smeared" in acoustical space will also increase, and the number of phonemes that can coexist without confusion in the agents' vowel systems will decrease. We therefore expect smaller vowel systems and more variation within the realisation of individual vowels. This is illustrated in figure 7, which represents a typical sound system[§] of the agents after 5000 imitation games. This number is higher than in the previous experiment, as the agents apparently take longer to develop multiple phonemes if there is more noise. This is logical as newly generated phonemes have a higher chance of interfering with existing phonemes. Note that the realisation of a formant can be shifted as much as 2 Bark down or 1.5 Bark up by 30% noise, so any phoneme can be realised in a significant part of the acoustic space.

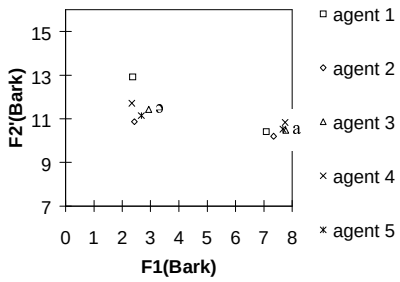

figure 7: Sound systems with 30% noise

When one agent starts using a new phoneme, this can be adopted by the other agents in the population. First one agent invents a new phoneme at random. When it uses this phoneme in an imitation game, the imitation game is bound to fail. However, if a language game fails in an agent whose phonemes otherwise have a good quality score, a new phoneme will be generated that is like the phoneme that was just heard, as has been described in section 3. If this new phoneme does not interfere with the phonemes that are already present, it will be accepted by the population of agents, and will become successful as well. This process can be observed in figure 8. Here one of the agents, agent 1, seems not to have the phoneme marked with 2, that is otherwise shared by all other agents, but it does seem to have an extra phoneme, marked with 1, which it shares with one other agent, agent 5. Actually these two facts are unrelated. The phoneme marked with 2 is a phoneme that has been created by another agent than agent 1, some time before the moment at which figure 8 was made. Agent 1 has not yet had the opportunity to make a successful copy of this phoneme. The phoneme marked with 1, however, has been recently created by agent 1. The only agent that has had the opportunity to make a successful imitation of this phoneme is agent 5. It can be observed that new phonemes are created in gaps between existing phonemes in the acoustical space. Phonemes that are created outside such gaps, will quickly be merged with the existing phonemes, or will interfere with existing phonemes, and be removed from the sound systems, because their quality scores will remain too low.

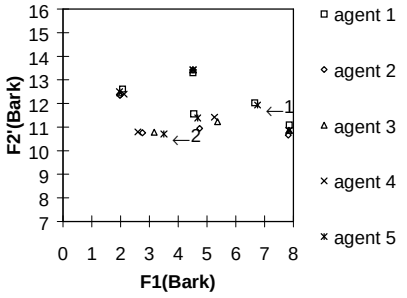

figure 8: Sound systems in transition

A last observation that will be made, is what happens when the agent population is made larger. For this experiments with 12[**] agent have been conducted. The experiments were run for 3000 cycles and had 10% noise on the acoustic space. The success score of a typical experiment is shown in figure 9. We can see that the score stays above 80%, although it does seem to be decreasing a bit over time. This is undoubtedly due to the increasing number of phonemes in the population of agents. But there do not seem to be any significant differences between figure 9 and figure 6, which showed the success score of a population of five agents.

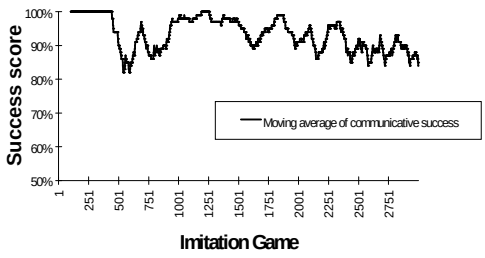

figure 9: Score of twelve agent experiment

[§] The vowel system, consisting of [a] and [ə], coincidentally is similar to the vowel system of Oubykh, a West-Caucasian language[6].

[**] The number of 12 agents was chosen because this is the number Berrah et al.[1] use in their experiments.

The phonemes of the twelve agents of this experiment after 3000 imitation games are shown in figure 10. We can observe that there are four to six clusters of phonemes. Three of these are compact and unambiguous. Another cluster, which can be found between 4 and 6 Bark on the F1 axis and around 11 Bark on the F2' scale is also unambiguous, but much more dispersed. This cluster is quite close to another diffuse cluster, which can be found between 2 and 3 Bark on the F1 scale and 9 and 12 Bark on the F2' scale. This cluster could also be considered as two separate clusters, as some agents (for example agents 6, 10 and 11) have two phonemes near the densest points of this cluster, whereas other agents (3 and 7) have only one phoneme in the centre of this cluster. This could indicate that the cluster represents a phoneme in the process of splitting. More research is needed, however, in order to make this clear.

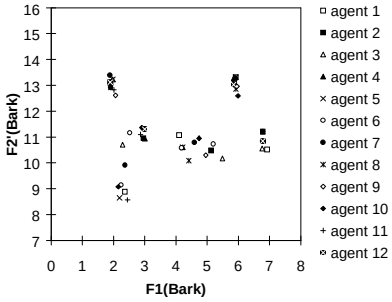

figure 10: Phonemes of 12 agent experiment

In any case, it does not seem that the increase in the number of agents influences the success of the communication very much. Of course, there is bound to be some influence, as an agent will communicate with more other agents, so that its phonemes get shifted in more different directions and therefore converge less quickly to a common point.
The fact that the number of agents does not greatly influence the communicative success is promising, as for realistic experiments the number of agents has to be much larger than the five or twelve used in the present experiments. Fortunately, the simulations are not computationally intensive, so it should be possible to increase the number of agents to about a hundred times the number of agents that were used in the experiments presented here.

5. Comparison to other Systems

The system that was used for the research presented in this paper departs quite radically from the traditional ideas about the acquisition and development of phonology. It does not a priori assume the existence of *distinctive features*. Distinctive features are (usually) binary and minimal properties of phonemes that can cause a contrast in meaning. Examples of features are, for example [voiced], as in the example of the English words *pit* and *bit*. The only difference is the voicing of the initial consonant, which is [-voiced] in the first word and [+voiced] in the second. An example of a feature of vowels is [back], as in the English words *feel* and *fool*, where the only difference is that in the first word the vowel is [-back] and in the second word it is [+back]. Thus phonological features can distinguish meaning.
Distinctive feature theory was introduced by the Prague School (see for example Jakobson and Halle[8]), and advocated by Chomsky and Halle[5]. The theory said that all languages used a universal set of features for building up a repertoire of phonemes, although not all languages used the same subset of this universal set. It also claimed that children use these features in learning the sound system of a language. They are supposed to determine which features are used in their mother tongue and set the parameters of their universal grammar accordingly. This theory was quite successful in predicting the sequence in which phonemes are learnt by children and in explaining regularities in the sound systems of languages; if a feature was used in one series of phonemes, it would probably also be used in another series. Hence, for example, the regular occurrence of voiced/unvoiced pairs of phonemes on the same place of articulation, as already has been mentioned in section 1.
However, distinctive feature theory runs into problems if it has to explain the fact that there are so many different phonemes in the languages of the world, while individual languages only use a small subset of these possible sounds. To cope with this problem a much larger number of features would have to be introduced[9, ch. 11], and the mechanism to select the appropriate features would become more complex. The theory also runs into difficulties explaining why children imitate their parents in a much more accurate fashion than would be necessary to make the minimal contrasts in meaning. Surely, if children only set a number of features in their universal grammars, this would not be necessary. Still, one can easily hear the difference between English *coo*, French *cou* (neck), Dutch *koe* (cow) and German *kuh* (cow) even though all these words would be described as a high back voiceless consonant followed by a high back rounded vowel. Furthermore it runs into difficulties explaining the irregularities that are found at least as often in the sound systems of human languages as the regularities. Regularities can be explained by the use of the same feature in different contexts, but how to explain that a feature is not used in certain other contexts?
Apparently, although features are undoubtedly a very useful abstraction for giving compact descriptions of individual phonemes and of sound systems of languages, they are also not more than abstractions.

Their psychological reality seems to be quite limited and it appears to be necessary to rather invoke mechanisms of perception and productions and criteria of efficiency and effectiveness to explain the universals that are found in phonology.

The role of acoustics and perception in the formation of sound systems (especially vowels) has been pursued by Liljencrants and Lindblom[10,12] and has been elaborated by the people of the Institut de la Communication Parlée in Grenoble[2,17]. René Carré et al.[3,4] have done research into the role of articulatory economy in the formation of words and syllables.

The research into the role of perception in the formation of vowel systems has focused on defining various energy functions. These give a high value for vowel systems that contain combinations of vowels that are difficult to distinguish and low values for systems that contain vowels that are easy to distinguish. One can explore the minima of this energy function for fixed numbers of vowels, or one can calculate for existing vowel systems whether they are stable and how they would evolve if they are not stable. The research has focused in part on finding the appropriate energy functions. It has been quite successful in predicting the forms of vowel systems for small to intermediate numbers of vowels (3-9).

Apart from the perceptual contrast, also the idea of articulatory ease has been explored. According to Lindblom[12], vowel systems tend to exploit a certain number of contrasts in the ordinary vowel space, before using extra parameters, such as length, nasalisation, pharyngealisation etc. His theory is that these extra parameters require extra movements, and that for reasons of articulatory economy these tend to be avoided. He claims that the same arguments apply to systems of consonants. No computer experiments have been done to investigate this, but a survey of consonants of human languages has shown that elaboration of articulations tends to take place only after the simple articulations have been exhausted[11].

The work of Carré et al.[3,4] has also focused on the role of articulatory economy, next to perceptual distinctiveness, both in the area of the structure of vowel systems, as well as in the area of vowel sequences. He claims that both can be explained by a combination of acoustical constraints and simplicity of formation, using a minimal number of necessary gestures to go from one vowel to the next.

Although these efforts have been able to pinpoint a number of constraints that play a role in the determination of sound systems in human language, they have not been able to provide a mechanism that explains how these constraints are implemented by a community of individual speakers. The constraints are actually constraints on a very abstract entity: the sound system of a language. Every speaker knows the phonemes of his or her language, but the optimality functions that can so easily be calculated of a sound system as a whole, are never as such calculated by the speakers of the language. Only through the interactions between the individual speakers can these constraints be implemented. This, however, is completely ignored by most of the aforementioned research.

Berrah et al.[1] have sought to implement the constraints through language-like interactions in a population of simulated robots and through the use of a genetic algorithm. The robots engage in exchanges of vowels. During these exchanges the vowel systems of the robots are updated by shifting the individual phonemes to match more closely with the perceived sounds. After a number of exchanges, the fitness of each robot is calculated, and the fittest robots are allowed to reproduce. The offspring of these robots will have vowel systems that are like the vowel systems of their parents. After a while a population emerges that has more or less natural and coherent vowel systems.

Their work incorporates the ideas of acoustical (and articulatory) constraints that shape vowel systems, instead of innate distinctive features, and it provides a mechanism of how these constraints can be implemented through interactions between individual agents. However, there are still a number of assumptions that make it not quite realistic. The most important of these assumptions is that sound systems are transferred from parents to children in a genetic way. Of course, the authors know that this is not the way things happen in humans and they probably just made this assumption in order to be able to use the powerful techniques that have been developed in genetic algorithms research. Unfortunately, this assumption, simple as it may seem, renders their work much less useful for linguistic research.

First of all, their robots are limited to a fixed number of phonemes. In principle, one could avoid this by using a genetic algorithm with a variable number of chromosomes, although this makes the use of genetic algorithms rather more complicated. However, a variable number of chromosomes would not likely lead to a variable number of phonemes. Whenever an agent “invented” a new phoneme through either mutation or crossover, this would lead to no increase in fitness, as no other agent would be able to cope with the new phoneme, unless, through rather remote chance, the mutation would have happened in two agents simultaneously. A second weak point is that their algorithm requires a global match between the sound systems of two agents every time a new generation of agents is calculated. However, in order to be psychologically plausible, such global operations would have to be avoided.

The system that has been presented in this paper tries to incorporate both acoustical and articulatory constraints as factors that determine the shape of vowel systems, as well as an active and psychologically plausible mechanism for explaining how these constraints can be implemented through local, language-like interactions between agents, without having to resort to genetic encoding of sound systems. The articulatory and perceptual systems of the agents, as well as the rules of the language game have been described in the previous sections.

The system presented in this paper *can* work with a variable number of phonemes. If one agent invents a new phoneme, the other agents can take it over through direct imitation. If the phoneme can not be confused with already existing phonemes, it will be successful and will be kept in the population. Through the pressure on the system to keep good phonemes and to throw away bad ones, as well as the pressure to take together phonemes that are similar and through the noise that causes phonemes to take up a non-zero amount of acoustic space, the phonemes will arrange themselves automatically into systems that would be considered optimal or at least sufficient by the energy functions of Liljencrants and Lindblom[10] or Boë et al.[2].

In essence, the process that is taking place in the system presented here is a cultural evolution, instead of a biological evolution. The theory of Luc Steels[15] is based on this, and claims that not only phonology, but also other aspects of language, such as lexicon and syntax can be viewed as an adaptive process of cultural evolution. As this cultural evolution makes use of the same mechanisms that are used for learning language, there would be no qualitative difference between the origin of language in a population of agents and the learning of language in an individual agent. This observation can be supported by evidence from, for example, the emergence of Creole languages in a community of pidgin speakers or the spontaneous emergence of a sign language in a group of deaf children.

Although the complexity of the system presented here is incomparable with the complexity of human language, it can also be used to investigate both the origin of vowel systems in a group of agents as well as the learning of a vowel system by an individual, empty agent that is inserted in a population that already has a fully developed sound system. It can also be used to investigate how sound systems change over time; how non-optimal, but sufficient systems acquire or lose phonemes or how they evolve into more stable systems.

6. Conclusions and Future Work

The main conclusion that can be drawn from the work that has been presented here is that it is possible to create near-natural vowel systems by using interactions between individual agents that use simple, local rules to update their vowel inventories. It has also been shown that this can be done with a relatively small number of interactions. The results do depend on the amount of noise that is present in the acoustic space, but this does not prevent the population of agents from reaching a high level of success in imitating each other under widely differing noise levels. We can conclude that neither innate distinctive features, nor global calculations are necessary to reach useful vowel systems that exhibit certain universal characteristics.

We could speculate that, although this would possibly be stretching the available evidence a bit, that genetic evolution does not seem to be a mechanism that is necessary to explain human phonological universals. Rather, cultural evolution and adaptation to other speakers led by a need for successful communication appear to be sufficient. This lends support to Steels' theory that the whole of human language has originated through similar mechanisms.

Without stretching the available evidence, however, the present model does seem to be able to explain the universals that are found in human vowel systems, as far as it is able to properly produce and perceive the different sounds that can be produced by humans, of course. This is an improvement over theories that have to resort to innate processes and a useful addition to the theories that provide functional models to explain phonological universals, but that do not provide a model of how these universals will appear in the individual speakers.

Of course the present system is not very natural, and a lot can be done to improve the naturalness of the production and perception of the phonemes. Also, no sequences of vowels and no consonants were investigated. In a future project vowel consonant sequences, or vowel consonant vowel sequences could be implemented and investigated. The co-articulation effects that would appear in such sequences would complicate the recognition and correct imitation of the sequences considerably. Also material on the possible sequences in human languages seems to be much rarer than material on possible vowel systems.

Another variation on the system would be to introduce a dynamic population and to investigate what would happen if new agents are inserted in a population that already has a working sound system, if a

spatial distribution of agents is added, in which agents have a smaller chance of communicating if they are far apart, or if two populations with different sound systems are merged.
The population of agents that learns a set of speech sounds through language-like interactions is a powerful method to investigate in a simplified way the complexities of language acquisition, -change and -origin.

7. Acknowledgements

The work presented in this report has been done in part at the AI-lab of the Vrije Universiteit Brussel in Brussels, Belgium and in part at the Sony Europe computer science laboratory in Paris, France. It forms part of ongoing research project into the origins of language. It was financed in by the Belgian federal government FKFO project on emergent functionality (FKFO contract no. G.0014.95), the IUAP 'Construct' project (no. 20) and Sony Europe. I thank Luc Steels for valuable suggestions on- and discussion of the ideas that are fundamental to the work.

8. Literature

1. Berrah, Ahmed-Reda, Hervé Glotin, Rafael Laboissière, Pierre Bessière and Louis-Jean Boë,(1996) From Form to Formation of Phonetic Structures: An evolutionary computing perspective, in: Terry Fogarty and Gilles Venturini, eds. *ICML '96 workshop on Evolutionary Computing and Machine Learning*, Bari 1996, pp. 23–29
2. Boë, Louis-Jean, Jean-Luc Schwartz and Nathalie Vallée(1995), The Prediction of Vowel Systems: perceptual Contrast and Stability, in: Eric Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*, John Wiley, pp. 185–213
3. Carré, René, Marc Bourdeau and Jean-Pierre Tubach(1995), Vowel-Vowel Production: Yhe Distinctive Region Model (DRM) and Vowel Harmony, *Phonetica* **52**, pp. 205–214
4. Carré, René and Mohamad Mrayati, Vowel transitions, vowel systems, and the Distinctive Region Model, in: C. Sorin et al. (eds.) *Levels in Speech Communication: Relations and Interactions*, Elevier pp. 73–89
5. Chomsky, Noam and Morris Halle (1968) *The sound pattern of English*, MIT Press, Cambridge, Mass.
6. Colarusso, John (1988) *The Northwest Caucasian Languages, A Phonological Survey*, Garland Publishers, NY
7. de Boer, Bart(1996), *A first report on Emergent Phonology*, Vrije Universiteit Brussel AI-memo 96-10
8. Jakobson, Roman and Morris Halle (1956) *Fundamentals of Language*, the Hague: Mouton & Co.
9. Ladefoged, Peter and Ian Maddieson (1996) *The Sounds of the World's Languages*, Blackwell.
10. Liljencrants, L. and Björn Lindblom (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast, *Language* **48** pp. 839–862.
11. Lindblom, Björn and Ian Maddieson (1988), Phonetic Universals in Consonant Systems, in: Hyman, Larry M. and Charles N. Li (eds.) *Language, Speech and Mind*, pp. 62–78.
12. Lindblom, Björn(1996), Systemic constraints and adaptive change in the formation of sound structure, in: James R. Hurford, Michael Studdert-Kennedy and Chris Knight, *Evolution of Language: Social and Cognitive Bases for the Emergence of Phonology and Syntax*
13. Maddieson, Ian,(1984) *Patterns of sounds*, Cambridge University Press.
14. Pinker, Steven, *The language instinct*, Penguin
15. Steels, Luc(1996), Synthesizing the Origins of Language and Meaning using Co-evolution, Self-organization and Level Formation, in: James R. Hurford, Michael Studdert-Kennedy and Chris Knight, *Evolution of Language: Social and Cognitive Bases for the Emergence of Phonology and Syntax*
16. Suzuki, Junji, Kunihiko Kaneko,(1994) Imitation Games, in: *Physica D* **75** pp. 328–342
17. 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)