Emergence of vowel systems through self-organisation

Bart de Boer
Artificial Intelligence Laboratory
Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussel
bartb@arti.vub.ac.be

Abstract

This paper describes a model of the emergence and the universal structural tendencies of vowel systems. Both are considered as the result of self-organisation in a population of language users. The language users try to imitate each other and to learn each other's vowel systems as well as possible under constraints of production and perception, while at the same time maximising the number of available speech sounds.

It is shown through computer simulations that coherent and natural sound systems can indeed emerge in populations of artificial agents. It is also shown that the mechanism that is responsible for the emergence of sound systems can be used for learning existing sound systems as well.

Finally, it is argued that the simulation of agents that can only produce isolated vowels is not enough. More complex utterances are needed for other interesting universals of sound systems and for explaining realistic sound change.

Introduction

The research described in this paper tries to explain the emergence and structure of systems of speech sounds. It investigates how a coherent system of speech sounds can emerge in a population of agents and how the constraints under which the system emerges impose structure through self-organisation. If self-organisation can explain structure, then innate and biologically evolved mechanisms are not necessary. This would decrease the number of linguistic phenomena that have to be explained by biological evolution.

This research is a small part of research into the origins of intelligence and language using computer simulations (see e.g. [12,24]). In this respect it belongs to the branch of artificial intelligence that uses computers to increase the understanding of human intelligence, rather than to the branch of artificial intelligence that tries to build more intelligent computer programs.

What are the phenomena that have to be explained by a theory of the emergence of speech sounds? The systems of speech sounds in the world's languages show remarkable regularities. First of all, certain sounds occur much more frequently than others. In the UPSID (UCLA Phonological Segment Inventory Database), a database that contains the phoneme inventories of 451 languages (the first version with 317 languages is described in [17]) the vowels [i], [a] and [u] appear in 87%, 87% resp. 82% of the languages, while the vowels [y], [œ] and [ɯ] occur in only 5%, 2% resp. 9% of the languages. This holds even more for consonants. Some consonants, e.g. [m] (94%), [k] (89%) or [j] (84%) appear very frequently, while others, e.g. [ʀ] (1%), [ʃ'] (1%) and [ʡ] (1%) appear very rarely.

The sound systems of languages also display a fair amount of symmetry. If a language has a front unrounded vowel of a given height, for example an [e] (occurring in 27% of the languages), it is quite likely that it also has the corresponding back rounded vowel [o] (which occurs in 29% of all languages, but in 85% of the languages with [e]). In the case of consonants, if a language has a voiced stop at a given place of articulation, e.g. [d] (27%) it usually also has a [t] (40% in whole sample vs. 83% in languages with [d]).

Sometimes these universal characteristics are explained by innate properties of the brain [3,11]. However the question then becomes how these innate properties have evolved. Also, if there are innate constraints it is not clear why there is still such huge variation between different languages. It is clearly preferable to have an explanation that does not need innate mechanisms.

Functional explanations of the above mentioned phenomena are more satisfying. A number of articulatory, perceptual and cognitive criteria have been proposed [2,14,15,25]. Some of these have been tested with computer simulations. These criteria can be summarised as articulatory ease, acoustic distinctiveness and minimal effort of learning.

These functional explanations are not the full explanation, either. They assume that the systems of speech sounds one finds are the result of an optimisation of one or more of the proposed criteria. However, it is not clear who is doing the optimisation. Certainly children that learn a language do not do an optimisation of the system of speech sounds they learn. Rather, they try to imitate their parents (and peers) as accurately as possible. This accounts for the fact that people can speak the same language with different accents, from which one can identify their place of birth or their social group.

If none of the individual speakers does an explicit optimisation of their sound system, but still (near-) optimal sound systems are found more frequently than non-optimal ones, it is clear that the optimisation must be an emergent property of the interactions in the population. Therefore, if one wants to explain the sound systems that are found in the world's languages, one has to model populations of agents that imitate and learn each other's sounds under acoustic, articulatory and cognitive constraints.

A first attempt at building a computer model of a population of interacting agents for explaining the shape of vowel systems was undertaken by Glotin [10] later followed by Berrah [1]. Both methods have the drawback that the population is subject to some genetic evolution and that the agents still do local optimising by pushing the vowels in their vowel systems away from each other. Also the number of vowels in every agent has to be fixed beforehand in these simulations.

In this paper a system is presented in which a population of agents that are each able to produce, perceive and learn vowels, develops a coherent system of vowel sounds that conforms to the tendencies of vowel systems in human languages. The number of vowels need not be fixed beforehand and there is no genetic evolution of the agents. Although the agents are able to change their repertoire of vowels in order to optimise the successfulness of imitation they only do this in reaction to interactions with other agents. They also cannot change the positions of their vowels in any global way. The emerging vowel systems are therefore truly the result of the interactions between the agents. The research is based on Steels' [22,23,24] ideas on the origins of language. Steels considers language as the result of a process of mainly cultural evolution, while the universal tendencies of language can be explained as the results of self-organisation under constraints of perception and production. Steels has applied his ideas mainly to lexicon and meaning formation, and is now working on syntax.

In the next two sections, the agents and their interactions are described in considerable detail. In section 3 some results of the simulations that were performed with this system are presented. Finally, in section 5 conclusions, future work and a discussion of the work are presented.

1 The agents

The agents are equipped with an articulatory synthesiser for production, a model of human hearing for perception and a prototype list for storage of vowels. All the elements of the agent were constructed to be as humanlike as possible, in order to make the results of the research applicable to research in linguistics and in order to make it possible to use the agents to learn *real* human vowels.

An agent consists of three parts (S, D, *V*) where S is the synthesis function, D is the distance measure and *V* is the agent's set of vowels. The synthesiser function is a function $\mathrm{S} : Ar \rightarrow Ac$, where *Ar* is the set of possible articulations and *Ac* is the set of possible acoustic signals. For the agents presented in this section the set of possible articulations is the set of articulatory vectors (*p*, *h*, *r*) where *p*, *h*, *r* are real numbers in the range [0,1]. Parameters *p*, *h* and *r* are the major vowel features [13, Chapter 9] *position*, *height* and *rounding*. Position corresponds (roughly) to the position of the highest point of the tongue in the front to back dimension, height corresponds to the vertical distance between the highest part of the tongue and the roof of the mouth and rounding corresponds to the rounding of the lips. Position zero means most fronted, height zero means lowest and rounding zero means that the lips are maximally spread. The parameter values for the high, front, unrounded vowel [i], such as in "leap" are (0,1,0). For the high, back rounded vowel [u], such as in "loop" they are (1,1,1). For the low, back, unrounded vowel [ɑ] such as in "father" they are (1,0,0).

The set *Ac* of possible outputs of the synthesiser function consists of vectors (F_1, F_2, F_3, F_4) where $F_1, F_2, F_3, F_4 \in \mathbf{R}$ are the first four formant frequencies of the generated vowel. These formant frequencies correspond to the peaks in the power spectrum of the vowel. When agents communicate with each other, they exchange only the formant values, not a real signal. This is done to reduce the amount of computations. A certain amount of noise is added, however. This noise consists of a random shifting of the formant frequencies, according to the following formula:

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

Figure 1: Vowel synthesiser equations.

1) $F_i \leftarrow \left(1+\frac{Noise\%}{100}U(-0.5{,}0.5)\right)F_i$.

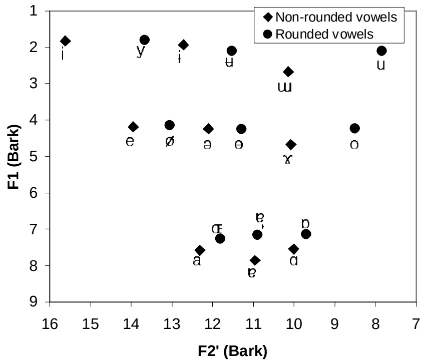

Figure 2: Vowels in F1-F2' space

In which *U*(-0.5,0.5) is a random number drawn from the uniform distribution between –0.5 an 0.5, *Noise%* is the noise percentage (a parameter of the system) and F_i represents the formants.

The formant frequencies are generated by a three dimensional quadratic interpolation between sixteen data points that have been generated by Maeda's articulatory synthesiser [18,26 pp. 162–164]. The equations for calculating the synthesiser function are given in figure 1. As an example, the formant values for [i] are (252, 2202, 3242, 3938), for [u]: (276, 740, 2177, 3506) and for [ɑ]: (703, 1074, 2356, 3486). An important property of the synthesis function is that it is easy to calculate the formant frequencies from the articulatory description, but that it is very hard to calculate the articulatory description from the acoustic description. With this synthesiser all basic vowels can be generated. It is therefore *language-independent.*

A vowel prototype *v* consists of elements (*ar*, *ac*, *s*, *u*), where $ar \in Ar$ is the articulatory prototype, $ac \in Ac$ is the corresponding acoustic prototype and *s*, *u* are the success and use scores, (which will be explained with the imitation game) respectively. The vowels are represented as prototypes as this seemed to be both a realistic and computationally effective representation. Research in human perception of speech sounds (e.g. [4]) seems to indicate that humans perceive speech sounds in terms of prototypes. If human subjects are presented with acoustic signals that vary continuously from one speech sound to another, (i.e. from [ga] to [ba]) they tend to perceive these signals as either the one category [ba] or the other [ga], never as something "in between". Perception suddenly switches somewhere in the middle.

An agent's vowels are stored in the set *V*, which we will call the vowel set. When an agent decides it has encountered a new vowel v_{new} it adds both the acoustic and the articulatory descriptions of v_{new} to *V*: $V \leftarrow V \cup v_{new}$. A sound *A* that the agent hears will be compared to the acoustic prototypes ac_v of the vowels *v* in its vowel set, and the distance between *A* and all ac_v ($v \in V$) is calculated using the distance function $D:(Ac)^2 \rightarrow \mathbf{R}$. It then considers that it has recognised the vowel v_{rec} that has minimal distance to *A*.

The distance between two vowels is determined by using a weighted distance in the F_1-F_2' space, where F_1 is the frequency of the first formant (expressed in Bark, a frequency scale based on human perception of pitch, logarithmic in the relevant frequency range) and F_2' is the weighted average of the second, third and fourth formants (also expressed in Barks). This distance measure is based on the distance measure described by Mantakas *et al.* [19] and has been designed to model the way in which humans perceive vowel signals. The distance measure is based on weighting formant peaks differently depending on their distance relative to a critical distance *c*, which is taken to be 3.5 Bark. In order to calculate F_2' two weights have to be calculated:

2) $w_1 = \frac{c-(F_3-F_2)}{c}$, $w_2 = \frac{(F_4-F_3)-(F_3-F_2)}{F_4-F_2}$

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

3) $$F_2^{\prime} = \begin{cases} F_2, & \text{if } F_3 - F_2 > c \\ \dfrac{(2-w_1)F_2 + w_1 F_3}{2}, & \text{if } F_3 - F_2 \leq c \text{ and } F_4\text{-}F_2 > c \\ \dfrac{w_2 F_2 + (2-w_2)F_3}{2}\text{-}1, & \text{if } F_4 - F_2 \leq c \text{ and } F_3 - F_2 < F_4 - F_3 \\ \dfrac{(2-w_2)F_3 + w_2 F_4}{2}\text{-}1, & \text{if } F_4 - F_2 \leq c \text{ and } F_3 - F_2 \geq F_4 - F_3 \end{cases}$$

The values of F_1 and F_2’ for a number of vowels is shown in figure 2. We can see from this figure that the distribution of the vowels through the acoustic space is quite natural, as vowels that are perceptually far apart appear far apart in the space. However, as it is a 2-dimensional projection of an essentially 4-dimensional space, not all distances between all phonemes can be represented accurately. The distance between two signals, a, $b \in Ac$ can now be calculated using a weighted Euclidean distance:

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

The value of the parameter λ is 0.3 for all experiments that will be described. There is independent evidence [16] that this value is realistic for describing human perception of vowels.

With the synthesis function and the distance measure that have been described in this section, the agents can produce and perceive speech sounds in a human-like way. The results that are generated with this system can therefore be compared with the results of research into human sound systems.

2 The imitation game

The imitation game was designed for allowing the agents to determine the vowels of the other agents and to develop a realistic vowel system. The imitation game is played in a population of agents (size 20 in all the experiments presented here). From this population two agents are picked at random: an *initiator* and an *imitator*. The initiator starts the imitation game by producing a sound that the imitator has to imitate. The imitator listens to the sound, and tries to analyse it in terms of the sound prototypes it already knows. It then produces the acoustic signal of the prototype it found. The initiator then listens to this signal and analyses it in terms of its prototypes. If the prototype it finds is the same as the one it used to produce the original sound, the game is considered *successful*. Otherwise it is a *failure*. This is communicated to the imitator. The exact steps of the imitation game are illustrated in table 1. Note that non-verbal feedback is needed to indicate whether the game was a success or a failure. If one draws the parallel with human communication, the non-verbal feedback can be compared to gesture or facial expression or the failure to achieve a communicative goal. Making the imitation game dependent on non-verbal communication might seem like introducing a very unrealistic element in the agents’ learning. To human children it is hardly ever directly indicated whether the sounds they produce are right or wrong. However, there are more indirect ways of discovering that the right sound was not used, such as a failure to achieve the desired goal of the communication. But our imitation game abstracts from this and assumes that a feedback signal is somehow available.

Depending on the outcome of the imitation game, the imitator can alter its vowel inventory. The way this is done is described in table 2, together with a number of other routines that are used. The imitation game could also have been implemented such that both imitator and initiator update their inventories. This has not been investigated, however.

Table 1: Basic organisation of the imitation game.

initiator	**imitator**
if ($V = \emptyset$) Add random vowel to *V*	
Pick random vowel *v* from *V* $u_v := u_v + 1$ Produce signal $A_1 := ac_v$	
	Receive signal A_1. **if** ($V = \emptyset$) v_{new} := Find phoneme(A_1) $V := V \cup v_{new}$ Calculate v_{rec}: $v_{rec} \in V \wedge \neg\exists v_2 : (v_2 \in V \wedge D(A_1, ac_{v2}) < D(A_1, ac_{v\,rec}))$ Produce signal $A_2 := ac_{vrec}$
Receive signal A_2. Calculate v_{rec}: $v_{rec} \in V \wedge \neg\exists v_2 : (v_2 \in V \wedge D(A_2, ac_{v2}) < D(A_2, ac_{v\,rec}))$ **if** ($v_{rec} = v$) Send non-verbal feedback: *success.* $s_v := s_v + 1$ **else** Send non-verbal feedback: *failure.*	
Do other updates of *V*.	Receive non-verbal feedback. Update *V* according to feedback signal. Do other updates of *V*.

Basically, if the imitation game was successful, the vowel prototype that was used is shifted closer to the signal that was perceived. If it was a failure, either a new prototype can be added, or the original one can be shifted, depending on whether the success/use ratio is high or low, respectively. The reason of this is that if an imitation game fails, but the prototype that was used was good, the failure was probably not caused by the bad quality of the prototype, but because it caused confusion between two prototypes of the other agent.

Some periodical updating of the agents' vowel inventories is also done independently of the imitation games. This is done in the *other updates* routines, described in table 3. These routines do three things: they throw away bad vowels that have been tried at least a minimum number of times (five times in all experiments presented). Vowels are considered bad if their use-to-success ratio is less than a threshold (0.7 in all experiments presented). Also, vowels that are too close in articulatory and acoustic space can be merged. This is done in order to prevent a cluster of bad phonemes to emerge at a position where only one good vowel would be required. This has been

Table 2: Actions performed by the agents

```
Shift closer ( $v$, $A$ ); return $v_{best}$
{
$v_{best} := v$
for (all six neighbors $v_{neigh}$ of $v$) do:
        if (D($ac_{vneigh}$, $A$) < D($ac_{vrec}$, $A$) )
            $v_{best} := v_{neigh}$
}
```

```
Find phoneme ( $A$ ); return $v_{best}$
{
vowel $v$:
    $ar_v$ = ( 0.5, 0.5, 0.5 )
    $ac_v$ = S( $ar_v$ )
    $s_v = 0$
    $u_v = 0$
do
    $v_{best} := v$
    $v$ := Shift closer($v_{best}$, $A$ )
until( $v = v_{best}$ )
}
```

```
Update according to feedback signal
{
$u_{vrec} := u_{vrec} + 1$
if (feedback signal = success)
    $v_{rec}$ := Shift closer( $v_{rec}$, $A_1$ )
    $s_{vrec} := s_{vrec} + 1$
else
    if( $u_{vrec}/s_{vrec}$ > threshold )
        $v_{new}$ := Find phoneme( $A_1$ )
        $V := V \cup v_{new}$
    else
        $v_{rec}$ := Shift closer( $v_{rec}$, $A_1$ )
}
```

Table 3: Other updates of the agents' vowel systems

```
**Merge**( $v_1, v_2, V$ )
{
**if** ( $s_{v1}/u_{v1} < s_{v2}/u_{v2}$ )
    $s_{v2} := s_{v2} + s_{v1}$
    $u_{v2} := u_{v2} + u_{v1}$
    $V := V - v_1$
**else**
    $s_{v1} := s_{v1} + s_{v2}$
    $u_{v1} := u_{v1} + u_{v2}$
    $V := V - v_2$
}
```

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

observed in experiments without merging. The articulatory threshold for merging is the minimal distance to a neighbouring prototype set to be 0.03 in all experiments. The acoustic threshold for merging is determined by the noise level. If two vowels are so close that they can be confused by the noise that is added to the formant frequencies, they are merged. The last change agents can make to their vowel inventories is adding a random new vowel. This is done with a low probability (0.01 in all experiments presented). The values for the articulatory parameters of the new vowel are chosen randomly from a uniform distribution between 0 and 1.

The imitation game contains all the elements that are necessary for the emergence of vowel systems. There are different mechanisms causing variation and innovation: the noise, the imperfect imitations and the random insertions of vowels. Other mechanisms take care of (implicit) selection of good quality vowels: vowels are only retained if they exist in other agents as well, otherwise no successful imitations are possible, and their success score will drop. Unsuccessful vowels will eventually be removed. The merging ensures that phonemes will stay apart, so that sufficiently spaced vowel systems emerge. Note that all the actions of the agents can be performed using local information only. The agents do not need to look at each other's vowel systems directly.

3 Vowel experiments

So far, only experiments with vowels have been done. These experiments have already been partly described in [6,7]. The first aim of the experiments was to show that a coherent sound system can indeed emerge in a population of agents that are in principle able to learn such a sound system, but that do not have a sound system at the beginning. The second aim was to show that the system that is learnt has the same characteristics as human sound systems. Vowels were the signals of choice, as they are easy to represent, generate and perceive and because the univer-

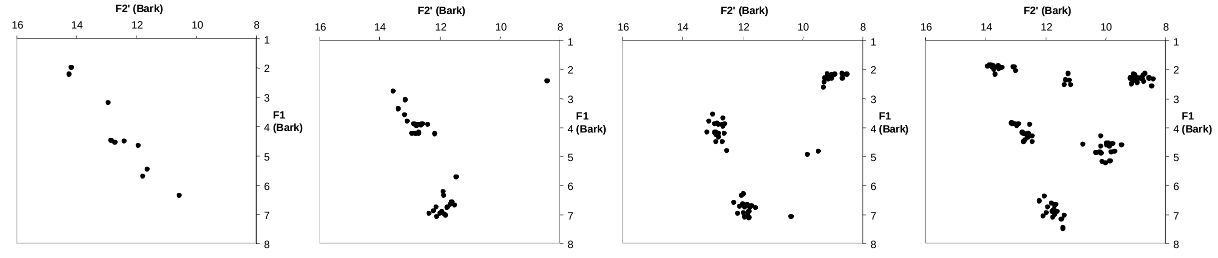

Figure 3: Vowel system after 20, 200, 1000 and 2000 games, 10% noise

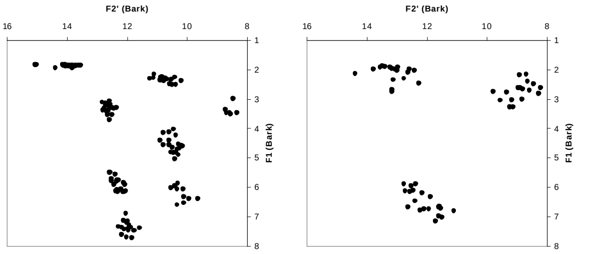

Figure 4: Systems with 10% and 25% noise

sal characteristics of human vowel systems and their functional explanations are more thoroughly described than those of other speech signals.

A typical example of the emergence of a vowel system in a population of twenty agents with maximally ten percent noise is illustrated in figure 3. In this figure the vowel systems of the agents in the population are shown after different numbers of imitation games. All vowels of all agents in the population are plotted on top of each other. They are plotted in the acoustic space consisting of the first formant F_1 and the weighted sum of the second , third and fourth formants (F_2'). The frequency of the formants is shown in the Bark frequency scale. Note that due to articulatory limitations the acoustic space that can be reached by the agents is roughly triangular with the apex at the bottom of the graph.

In the leftmost graph the agents' vowels after 20 imitation games are shown. One can see hardly any structure at all; the vowels are dispersed through the acoustic space (the apparent linear correlation is just coincidence). This is caused by the fact that initially vowels are mostly added at random. After 200 imitation games, clusters emerge. This happens because the agents try to imitate each other as closely as possible while at the same time there is a pressure of having a maximal number of vowels. Almost every agent in the population now has two vowels: one in each cluster.

After 1000 imitation games the available acoustic space starts to get full, and the clusters become tighter. Every agent in the population now has at least three vowels. Some agents have more (the isolated dots in the graph), other agents have not had the opportunity to copy these, yet. Finally, after 2000 imitation games, the available acoustic space is completely covered. The system that emerges consists of tight clusters that are approximately equally spaced. The vowels that emerge are [i], [e]-[ø], [a], [o], [u] and [ɨ] which, except for the rounding of the front mid segment, is a possible six-vowel system (such as found, for example in the Saami language of Lapland).

The noise level determines the number and size of the clusters. If the noise level is higher, the number of clusters will be lower and they will be more widely dispersed. This is shown in figure 4, where a system with 10% noise is compared with a system with 25% noise. Note however, that the clusters are still spread near-optimally through the available acoustic space. Both systems are also natural. The one with 10% noise has eight vowels, while the one with 25% noise is the canonical three-vowel system, consisting of [i], [a] and [u]. Note that the vowel system that is obtained under 10% noise in this simulation run is not the same as the one that is obtained in figure 3. This is because simulation runs do not converge to one optimal solution, but they converge to a good system, which might, apparently, consist of 6 or 8 vowels. Both systems, however, show similar characteristics of symmetry and spread of vowel clusters. It has been shown that coherent and successful vowel systems emerge for a large range of parameter settings [9].

These experiments show that a coherent sound system can emerge in a population of agents and that these sound systems show the same universal characteristics as sound systems from natural languages. However, there is no transfer from one generation of speakers to the next, yet. In real language communities speakers enter (they are born) and leave (they die or move away) the community constantly. Still, the language remains relatively stable. The simulation presented here can be used to test whether it is possible to transfer the sound system in a stable way from one generation to the next.

Succession of generations can be modelled by adding and removing agents from the population at random. These processes model birth and death of language users. After a sufficiently long period of time, all the original agents in the population will have been replaced and the new agents will have learnt their sound system from the original population. The sound system in the population of new agents can then be compared with the original sound system. This is illustrated in figure 5. The white squares represent the positions of the original agents' vowels and the black circles represent the positions of the vowels after 2000 imitation games. On average every 50 imitation games an agent was removed from- or added to the population. The original population consisted of 20 agents, the final population consisted of 11 agents for the left graph and 14 agents for the right graph (the number of agents was not fixed, due to the independence of adding and removing agents.) The noise level was a constant 10%.

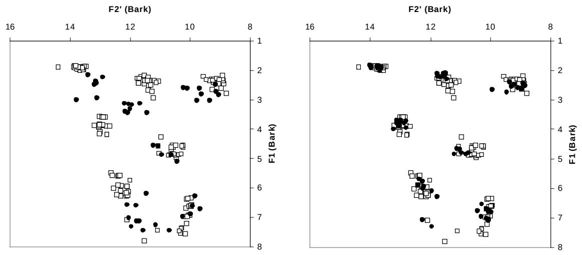

Figure 5: Systems after population replacement.

In the simulation that resulted in the left graph, agents could learn equally well, independent of how long they were already present in the population. For the right graph, agents were used that could change their vowel repertoire more easily when they were young than when they were old. This was implemented by allowing younger agents to make larger changes to their vowel prototypes while approaching a given signal (as in the Shift closer and Find phoneme routines described in table 2). The total number of improvement steps that an agent could make was limited to 10. Comparing the two graphs, it can be observed that both systems preserve the approximate positions of the clusters. However, in the left graph the clusters have become more dispersed, have moved slightly, and even two clusters in the upper left corner have merged. In the right graph, the positions and number of clusters has hardly changed at all.

Apparently cultural transfer of sound systems is possible in both simulations. Extra stability is ensured when older agents can change their vowel systems less easily than younger agents. Apparently the older agents provide a stable target to which the younger agents can adapt their vowel systems.

4 Comparison with human vowel systems

The acid test of a theory that claims to explain the structure of human vowel systems is whether its predictions actually agree with what is found in human languages. The vowel systems that result from the simulations do look quite similar to the vowel systems that are found in human languages, such as the one of French, shown in figure 6 (but note that in this figure the axes are not logarithmic.) However, this similarity is impressionistic and hard to judge objectively. Another

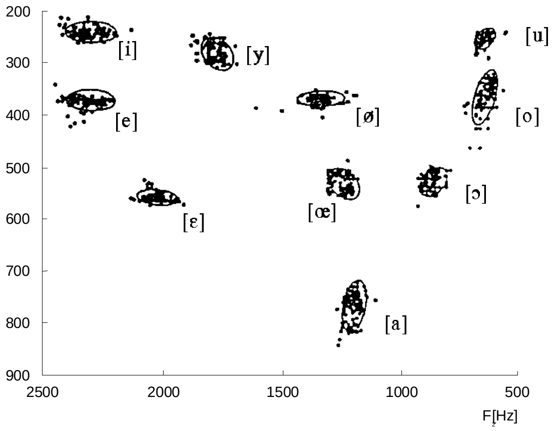

Figure 6: Vowel system of French, from [20] through [10].

means for judging the similarity between the emerged systems and human vowel systems is necessary. The *classification* of vowel systems is such a means. Several linguists have made classifications of human vowel systems, but the two that will be used here are by Crothers [5] and by Schwartz *et al.* [21]. Classification of human vowel systems is done by abstracting away from too much phonetic detail, and by paying attention to the positions of the different vowels in the system relative to each other. By comparing the vowel systems of many different languages from different geographical regions and from different language families, one can derive *universals* of human vowel systems. Crothers [5] has made a list of 15 universals, 12 of which are applicable to the artificial vowel systems studied here (the rest have to do with extra features, such as nasality and vowel length, that were not implemented). Schwartz *et al.* [21] who used a larger database of languages derive a similar list of universals and also present detailed frequencies of the different possible vowel systems. These universals specify which vowels are likely to occur together, and which ones are to be expected in systems of a given size. The universals are not really universal, but rather strong tendencies. There are always exceptions, but these are rare.

Almost all the vowel systems that emerge from the simulations conform to both Crothers' and Schwartz *et al.*'s universals. Furthermore the percentages with which the different vowel systems occur conform to the percentages with which corresponding vowel systems occur in human languages (except for systems with a very small or a very high number of vowels). This is a unique property of the simulations presented here. Previous simulations (e.g. [1, 10, 14, 26]) succeeded in predicting the most frequently occurring vowel systems for a given number of vowels, but not in predicting less frequently occurring systems. However, in human languages it is quite possible that for a given number of vowels, different systems occur almost equally often.

An illustrative example is given in figure 7. In this figure, vowel systems with six vowels that emerged from the simulations are classified. These vowel systems were obtained by running the simulation 100 times for 25 000 imitation games with an acoustic noise of 15% and the parameter λ set to 0.3. Of the 100 runs, 54 ended in systems with 6 vowels. Out of each population a random agent that had a number of vowels that was equal to the average number of vowels per agent

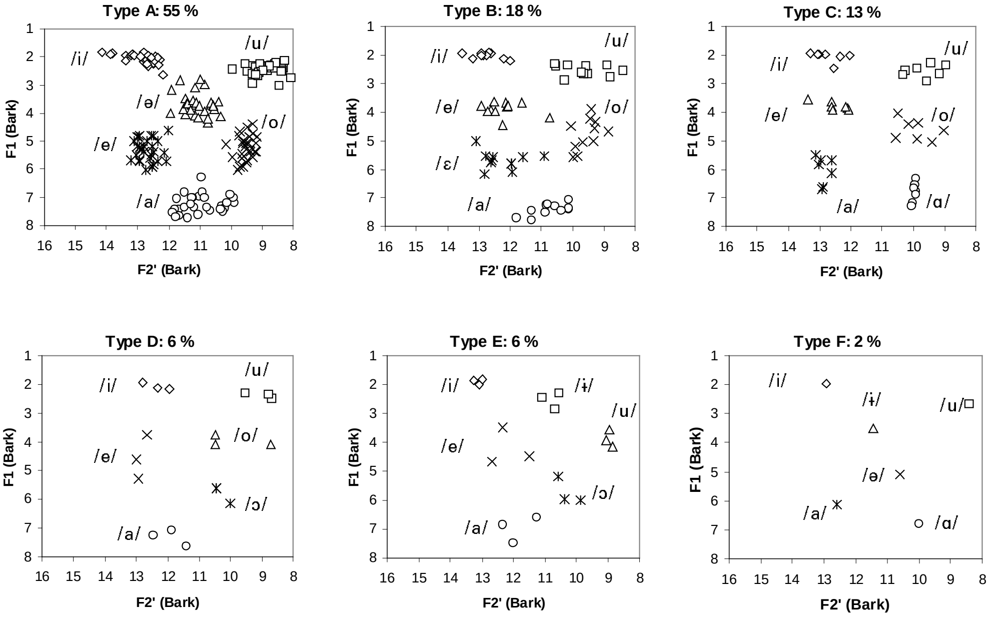

Figure 7: Classification of six vowel systems.

in that population, was chosen and its vowel system was plotted in the F_1-F_2' space. Note that although these graphs look like the ones shown previously, they present something quite different. The previous ones showed the vowel systems of the members of a *single* population. These graphs show systems of members of *different* populations.

Systems of type A, B and C (for a total of 86%) conform to all applicable universals in Crothers' list. The frequencies with which the systems are predicted conform well to what Schwartz *et al.* have observed. Types A and E (which they consider as one type) occur in 68% of the 60 languages with six vowels in their data. Type B occurs in 20%, type C occurs in 5% and type D occurs in 7% of the cases. Type F does not occur at all in their data. Similar results were obtained for other numbers of vowels. For more details see [9].

5 Conclusions and discussion

The results of the simulations show clearly that coherent sound systems can emerge as the result of local interactions between the members of a population. They also show that the systems that emerge show characteristic tendencies similar to the ones that are found in human sound systems, such as more frequent use of certain vowels and symmetry of the system. This means that we do not need to look for (evolutionary) biological ways of explaining the universal tendencies of vowel systems. Apparently the characteristics emerge as the result of self-organisation under constraints of perception, production and learning. The systems that are found can be considered attractors of the dynamical system that consists of the agents and their interactions. Of course we still need an account of the biological evolution of the shape of the human vocal tract and of the performance of human perception, but we do not need any specific innate mechanisms for explaining the structure of the vowel systems that appear in human languages.

It has also been shown that the vowel systems can be transferred from one generation of agents to the next. For this, no change in the interactions and the behaviours of the agents has to be made, although the transfer from generation to generation is improved if older agents are made to learn less quickly than young agents. Apparently the same mechanism can be used to learn an existing vowel system as well as to produce a sound system in a population where no sound system existed previously. This lends support to Steels' [23,24] hypothesis that the same mechanism that is responsible for the ability to learn language is responsible for the emergence of language in the first place. Computer simulations make it easy for the researcher to perform experiments like these, and thus provides an extra means to test and fine-tune linguistic theories.

The ability to model the emergence, the learning and the universal structural tendencies of sound systems as the result of local interactions between agents that exist in a population is a remarkable result. It indicates that not all aspects of language need to be explained through biological evolution. This makes it easier to explain that language evolved in a relatively short time.

It needs to be tested, however, whether these results also hold for more complex utterances than isolated vowels. Work is in progress [8] on building agents that can produce and perceive complex utterances, using articulatory synthesisers, dynamically moving articulators, models of human perception and models of infant learning of speech.

In any case, modelling aspects of language as the result of interactions in a population seems to be a promising way to learn more about the origins of language, especially so because it provides an extra mechanism next to biological evolution for explaining the complexity and structure of language.

6 Acknowledgements

This research was done at the artificial intelligence laboratory of the Vrije Universiteit Brussel. It is part of an ongoing research project into the origins of language and intelligence. Funding was provided by the GOA 2 project of the Vrije Universiteit Brussel. Part of the work was done at the Sony CSL in Paris, France. I thank Luc Steels for valuable discussion of the ideas and the work presented here and for providing the research environment of the VUB AI-lab. I thank Edwin de Jong, Tony Belpaeme and Paul Vogt of the VUB AI-lab for their comments and suggestions and I thank the people of the Sony CSL for their hospitality. I also thank Björn Lindblom, Christine Ericsdotter and the other people of the phonetics laboratory of Stockholm University for the opportunity to present my work there and for their feedback and suggestions.

References

1. A. R. Berrah, Évolution artificielle d’une société d’agents de parole: Un modèle pour l’émergence du code phonétique, Thèse de l’Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives, 1998
2. R. M. Carré, R, M. Bordeau and J.-P. Tubach, Vowel-vowel production: The distinctive region model (DRM) and vowel harmony, *Phonetica* **52**, (1995), 205–214
3. N. Chomsky and M. Halle, *The sound pattern of English*, MIT Press, Cambridge, Mass, 1968
4. F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst and L. J. Gerstman, Some experiments on the perception of synthetic speech sounds, in: *Acoustic phonetics*, D.B Fry (ed.) Cambridge University Press, 1976, pp. 258–283
5. J. Crothers, Typology and Universals of Vowel systems. *In Universals of Human Language*, Volume 2 Phonology, J. H. Greenberg, C. A. Ferguson and E. A. Moravcsik (eds.) Stanford: Stanford University Press, 1978, pp. 93–152.
6. B. G. de Boer, Generating vowels in a population of agents, in: *Fourth European Conference on Artificial Life*, P. Husbands & I. Harvey (eds.) MIT Press, Cambridge, Mass. 1997, pp. 503–510
7. B. G. de Boer, Self organisation in vowel systems through imitation, in: *Computational Phonology, Third Meeting of the ACL SIGPHON*, J. Coleman (ed.), July 12, 1997, pp. 19–25
8. B. G. de Boer, *A realistic model of emergent phonology*, Vrije Universiteit Brussel AI-lab AI-memo 98-04, 1998
9. B. G. de Boer, *Self Organisation in Vowel Systems*, Ph. D. thesis, Vrije Universiteit Brussel, 1999
10. H. Glotin, La Vie artificielle d'une société de robots parlants: émergence et changement du code phonétique. DEA sciences cognitives-Institut National Polytechnique de Grenoble, 1995
11. R. Jakobson and M. Halle, *Fundamentals of language*, The Hague: Mouton & Co, 1956
12. S. Kirby and J. R. Hurford, Learning, Culture and Evolution in the Origin of Linguistic Constraints. In: *Fourth European Conference on Artificial Life*, P. Husbands and I. Harvey (eds.) Cambridge (MS): MIT Press, 1997, pp. 493–502
13. P. Ladefoged, and I. Maddieson, *The sounds of the world’s languages*, Oxford: Blackwell, 1996
14. L. Liljencrants, and B. Lindblom, Numerical simulations of vowel quality systems: The role of perceptual contrast, *Language* **48**, (1972), 839–862.

15. B. Lindblom, Phonological units as adaptive emergents of lexical development, in: *Phonological Development*, C. A. Ferguson, L. Menn and C. Stoel-Gammon, (eds.) 1992, York Press, Timonium, Md. pp. 131–163

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

17. I. Maddieson, *Patterns of sounds*, Cambridge University Press, 1984

18. S. Maeda, Compensatrory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model, in: *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal (eds.) Kluwer, 1989, pp. 131–149

19. M. J. Mantakas, L. Schwartz and P. Escudier, Modèle de prédiction du ‘deuxiéme formant effectif’ F2’—application à l’étude de la labialité des voyelles avant du français. In: *Proceesings of the 15th journées d’étude sur la parole*. Société Française d’Acoustique, 1986 pp. 157–161.

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

21. J. Schwartz, L. Boë, N. Vallée and C. Abry, Major trends in vowel system inventories. *Journal of Phonetics* **25**, (1997), 233–253

22. L. Steels, The spontaneous self-organization of an adaptive language, in: *Machine Intelligence* **15**. S. Muggleton (ed.), *to appear*

23. L. Steels, The synthetic modelling of language origins, *Evolution of Communication* **1**(1): (1997) 1–34

24. L. Steels, Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation, in: *Approaches to the evolution of language*, J. R. Hurford, M. Studdert-Kennedy & C. Knight (eds.) Cambridge: Cambridge University Press, 1998, pp. 384–404

25. K. N. Stevens, The quantal nature of speech: Evidence from articulatory-acoustic data. In: *Human communication: a unified view*. E. E. David, Jr. and P. B. Denes (eds.) New York: McGraw-Hill. 1972 pp. 51–66

26. N. Vallée, 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), 1994