Emergence of sound systems through self-organisation

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

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 effectively decreases the number of linguistic phenomena that have to be explained by biological evolution.

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 Maddieson 1984) the vowels [i], [a] and [u] appear in 87%, 87% and 82% of the languages, respectively while the vowels [y], [œ] and [ɯ] occur in only 5%, 2% and 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]).

Not only the inventories of speech sounds of languages show great regularities. Regularities are also found in the way speech sounds are strung together into syllables. It is said that all languages have syllables consisting of either a vowel (V) or a consonant followed by a vowel (CV). Syllables that end in a consonant are rarer, as are clusters of consonants at the onset or the end of a syllable. When consonants occur in clusters, certain sequences occur much more frequently than others (Vennemann 1988). For example, a plosive followed by a nasal, e.g. [gŋ] occurs much more frequently than the inverse sequence at the *beginning* of a syllable. However, at the *end* of a syllable, the reverse is true.

Sometimes these universal characteristics are explained by innate properties of the brain (Jakobson & Halle 1956; Chomsky & Halle 1968). 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 (e.g. Carré et al. 1995, Liljencrants & Lindblom 1972; Lindblom 1992; Stevens 1972). Some of these have been tested with computer simulations. These criteria can be summarised as articulatory ease, acoustic distinctiveness and minimum effort of learning.

However, 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 explains 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 (Trudgill 1995).

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 (Glotin 1995; Glotin & Laboissière 1996) later followed by Berrah (Berrah *et al.* 1996; Berrah 1998). 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 mostly based on Steels' (Steels 1996, 1997, 1998) ideas on the origins of language, but fits in the larger recent tradition of studying the origins of language using computer simulations of populations (see also Hurford, this volume and Kirby, this volume). 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. In section 4 work in progress on extending the system to more complex utterances are presented. Finally, in section 5 conclusions 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. The architecture of an agent is illustrated in figure 1. 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.

(Figure 1 approximately here.)

An agent (illustrated in figure 1) 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 $S: Ar \rightarrow Ac$, 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 (Ladefoged and Maddieson 1996: ch. 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, $F_4 \in$ 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 among 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:

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

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 (Maeda 1989; Vallée 1994 pp. 162–164). The equations for calculating the synthesiser function are shown in figure 2. 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*.

(Figure 2 approximately here.)

A vowel 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 way to represent vowels. Research in human perception of speech sounds (e.g. Cooper *et al.* 1952; Liberman *et al.* 1954) 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, (e.g. 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. In other parts of language, such as syntax and semantics prototypes appear to be used as well (Comrie 1981; Lakof 1987).

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} (we will describe below how and when this is decided), 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: $Ac^2 \rightarrow R$ (which will be described below). It will then assume that it has recognised the vowel v_{rec} with the minimum distance to *A*:

$$2) \left\{ v_{rec} \middle| v_{rec} \in V \cap \neg \exists v_2 : (v_2 \in V \cap \mathrm{D}(A, ac_{v2}) < \mathrm{D}(A, ac_{v\,rec})) \right\}$$

It should be stressed that the acoustic representations of the vowels are only stored in order to decrease the number of calculations needed for vowel recognition. Whenever an agent wants to say a vowel to another agent, it takes the *articulatory* prototype from the list and transforms it into an acoustic representation using the synthesis function S; it does not use the acoustic prototype.

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 logarithmic frequency scale) 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.* (1986) (also described in Boë *et al.* 1995). 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:

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

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

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:

$$5)\ F_2' = \begin{cases} F_2, & \text{if } F_3 - F_2 > c \\ \frac{(2-w_1)F_2 + w_1F_3}{2}, & \text{if } F_3 - F_2 \le c \text{ and } F_4\text{-}F_2 > c \\ \frac{w_2F_2 + (2-w_2)F_3}{2}\text{-}1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 < F_4 - F_3 \\ \frac{(2-w_2)F_3 + w_2F_4}{2}\text{-}1, & \text{if } F_4 - F_2 \le c \text{ and } F_3 - F_2 \ge F_4 - F_3 \end{cases}$$

(Figure 3 approximately here.)

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 4-dimensional space, not all distances between all phonemes can be represented accurately. This is especially the case with the distinction rounded-unrounded.

The distance between two signals, a, $b \in Ac$ can now be calculated using a weighted Euclidean distance:

$$6)\ D(a,b) = \sqrt{\left(F_1^a - F_1^b\right)^2 + \lambda\left(F_2'^a - F_2'^b\right)^2}$$

The value of the parameter λ is 0.5 for all experiments that will be described.

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 to allow 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 pro-

ducing 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 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.

(Table 1 approximately here.)

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. First of all, the use and success counts u and s of the vowels that were used are updated. The use count u is increased every time a vowel is used. The success count s is only increased if the imitation game, in which the vowel was used, was successful.

(Table 2 approximately here.)

If the imitation game was successful, the vowel that was used for imitation is shifted a little closer to sound more like the signal that was heard. This is done by finding the neighbour of this vowel whose sound is closer to the signal that was heard. The neighbours of a vowel are the six vowels that differ by a certain small value, which was fixed to 0.05 in all experiments described in this paper, in only one of the three articulatory parameters. The reason for this shift is as follows: if the imitation game was successful, the vowel that was used is the same as the

vowel that was used by the other agent. Shifting it to sound more like the signal that was just heard increases cohesion in the population.

If the imitation game was a failure, however, and if the vowel that was used was successful in previous imitation games (its use to success ratio being higher than a certain threshold, 0.8 in all games presented) then the reason the imitation game failed is probably that the vowel was confused. It is likely that the other agent distinguished two vowels where this agent distinguished only one. The confusion between the two vowels caused the imitation game to fail. It is therefore a good idea to add a new vowel, which sounds like the signal that was heard. This is done using the *find phoneme* procedure, shown in table 2.

However, if the imitation game was a failure and the vowel that was used has a low use-to-success ratio, the vowel was probably not a good imitation of any other sound. It is therefore shifted towards the signal that was heard in the hope that it will become a better imitation.

(Table 3 approximately here.)

The phoneme is not thrown away. This is done in the *other updates* routines, described in table 3. This routine does three things: it throws 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). Furthermore, 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 from emerging at a position where only one good vowel would be required. This has been 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 (de Boer 1997a, 1997b). 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 universal characteristics of human vowel systems and their functional explanations are more thoroughly described than those of other speech signals.

(Figure 4 approximately here)

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 4. 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 F1 and the weighted sum of the second, third and fourth formants (F2'). 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 an artifact). 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 to have a maximal number of vowels (caused by the occasional random insertion of new vowels in the agents' repertoires). 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 (from UPSID, through Vallée 1994).

(Figure 5 approximately here.)

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 5, 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: [i], [e], [ɛ], [a], [ɔ], [o], [u] and [ɯ] while the one with 25% noise is the canonical three vowel system, consisting of [i], [a] and [u]. Note that the vowel system that obtained under 10% noise in this simulation run is not the same as the one that obtained in figure 1. This is because the population does not converge to one optimal solution, rather it converges 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.

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.

(Figure 6 approximately here.)

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 done in figure 6. 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%.

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. 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 Towards complex utterances

The experiments with vowel systems show that it is possible for coherent and realistic sound systems to emerge in a population and that the possible structures of these sound systems are determined by the functional constraints under which they are produced, perceived and learnt. However, interesting linguistic change is not really possible with this system. The vowel repertoires rapidly converge towards near-optimal systems and change relatively little after that. Some drift may occur in the positions of the vowel clusters, and clusters might even merge or split, but this is not the way in which human sound systems generally change.

Human sound change is often caused by the phonetic environment in which sounds occur. For example, nasalised vowels almost always derive from non-nasal vowels that are followed by a nasal consonant. Context is also necessary for the spread of sound changes. If an agent learns to pronounce a certain sound differently than other agents, it can only use this sound to successfully imitate other agents when the sound appears in a context that allows the other agents to disambiguate it. If there were no context, the sound could not be imitated, would become unsuccessful and would be discarded. Free variation of sounds in the population, and therefore sound change, is only possible when there is sufficient context. Therefore it is necessary to extend the system to handle longer and more complex utterances.

This is also necessary if one wants to investigate universal characteristics of consonants and syllable structure. As was said in the introduction, the same general tendencies that exist for vowels also exist for consonants and syllable structure. If we want to investigate whether these can be explained with the same mechanisms of self-organisation in a population, we need to build a simulation of an agent that is able to generate, perceive and learn complex utterances

(for another approach to investigating syllable structure with computer simulations, see Redford *et al.*, this volume).

Work is in progress to build agents that are able to handle complex utterances. The basic imitation game will remain the same, but the architecture of the agents will be different. Their sound production system will consist of an articulatory synthesiser, based on Mermelstein's model (Mermelstein 1973). The degrees of freedom of this model correspond roughly to the different articulators (tongue, lips, teeth, etc.) of the human vocal tract. The movements of the articulators are simulated dynamically, taking into account their inertia. The agent's utterances are modelled as gestures (Browman & Goldstein 1995.) Different articulatory gestures can be scheduled to occur in sequence, influencing each other where necessary. This system is already operational and an example output is shown in figure 7.

(Figure 7 approximatelty here.)

Perception will be based on extracting features from the speech signal. These features might be the formant frequencies and their rates of change, presence of voicing, presence of noise, presence of silence, strength of the signal, etc. Associations between the different articulatory gestures and these features will have to be learnt by the agents, so that they can find articulatory gestures that correspond to the acoustic signals they hear. A perception model is nearly operational. The extraction of features such as formant frequency, voicing frequency, voicing prominence and power of the signal are demonstrated in figures 8 and 9.

(Figures 8 and 9 approximately here.)

The learning of the agents is the most difficult to model. The simple use of prototypes as with the vowel system is no longer sufficient. At least two levels of storage are needed. One level for the possible words (sequences of phonemes) the agents know and one level for the articulatory gestures and their acoustic correlates (phonemes) from which these words are built up. The model will have to conform to what is known about how children learn sound systems (Vihman 1996), although much of this is controversial. Agents will first learn words as holistic gestures, and split these up into phoneme-like constituents under pressure of minimal storage requirements.

Once the agents have been built, it will first be tested whether a population of these agents is able to generate a coherent system of speech sounds. Then experiments can be run that investigate the sound changes that can take place and the extent to which the results can be compared to the way human sound systems behave.

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 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 evolutionary account of the shape of the human vocal tract and 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 have 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' (Steels 1997, 1998) thesis that the same mechanism that is responsible for the ability to learn language is responsible for the emergence of language in the first place. The use of computer simulations makes 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 explain 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 for building agents that can produce and perceive complex utterances. 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 computer science laboratory 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 computer science laboratory 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

Berrah, A.-R., Glotin, H., Laboissière, R., Bessière, P. & Boë, L.-J.(1996) From Form to Formation of Phonetic Structures: An evolutionary computing perspective. In Fogarty, T. & Ventur-

ini, G. (eds.) *ICML '96 workshop on Evolutionary Computing and Machine Learning*, Bari 1996: 23–29

Berrah, A.-R. (1998) *Évolution Artificielle d'une Société d'Agents de Parole: Un Modèle pour l'Émergence du Code Phonétique*, Thèse de l'Institut National Polytechnique de Grenoble, Spécialité Sciences Cognitives.

Boë, L.J., Schwartz, J.-L. & Vallée, N. (1995), The Prediction of Vowel Systems: perceptual Contrast and Stability. In Eric Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition*, John Wiley:185–213

Browman, C. P. & Goldstein, L. (1995) Dynamics and Articulatory Phonology. In Port, R.F. & van Gelder, T. (eds.) *Mind as Motion*, Cambridge (MS): MIT Press:175–194.

Carré, R., Bordeau, M. & Tubach, J.-P. (1995) Vowel-Vowel Production: The Distinctive Region Model (DRM) and Vowel Harmony, *Phonetica* 52, 205–214

Chomsky, N. & Morris, H. (1968) *The sound pattern of English*, Cambridge (MS): MIT Press.

Comrie, B. (1981) *Language universals and linguistic typology*, Oxford: Blackwell.

Cooper, F. S., Delattre, P. C., Liberman, A.M., Borst, J. M. & Gerstman, L. J. (1952), Some Experiments on the Perception of Synthetic Speech Sounds. Journal of the Acoustical Society of America, 24: 597—606. Reprinted in Fry, D. B. (ed.)(1976) *Acoustic Phonetics*, Cambridge: Cambridge University Press: 258–283

de Boer, B. (1997a) Generating Vowels in a Population of Agents. In Phil Husbands, P. & Harvey, I. (eds.) *Fourth European Conference on Artificial Life*, MIT Press: 503–510

de Boer, B. (1997b) Self Organisation in Vowel Systems through Imitation. In Coleman, J. (ed.) *Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology*, July 12, 1997: 19–25

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

Glotin, H., Laboissière, R. (1996) Emergence du code phonétique dans une societe de robots parlants. *Actes de la Conférence de Rochebrune 1996 : du Collectif au social*, Ecole Nationale Supérieure des Télécommunications – Paris

Hurford, J. (This volume) *Social Transmission Favours Linguistic Generalization*.

Jakobson, R. & Morris, H. (1956) *Fundamentals of Language*, The Hague: Mouton & Co.

Kirby, S. (This volume) Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages*, Oxford: Blackwell.

Lakof, G. (1987) *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: Chicago University Press.

Liberman, A. M., Delattre, P. C., Cooper, F. S. & Gerstman, L. J. (1954) The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants, Psychological Monographs, 68(8). Reprinted in D.B. Fry (ed.)(1976) *Acoustic Phonetics*, Cambridge University Press: 315—331.

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

Lindblom, B. (1992) Phonological Units as Adaptive Emergents of Lexical Development. In: Ferguson, C.A., Menn, L. & Stoel-Gammon, C. (eds.) *Phonological Development*, York Press: 131–163

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

Maeda, S. (1989) Compensatrory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal Tract Shapes using an Articulatory Model. In Hardcastle W. J. & Marchal, A. (eds.) *Speech Production and Speech Modelling*, North-Holland: Kluwer: 131–149

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

Mermelstein, P. (1973) Articulatory model for the study of speech production, *The Journal of the Acoustical Society of America*, 53(4): 1070–1082

Redford, M. A., Chen, C. C., Miikkulainen, R. (This volume) Constrained evolution of syllable systems.

Steels, L. (1996) The Spontaneous Self-organization of an Adaptive Language. In Muggleton, S. (ed.) *Machine Intelligence* 15.

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

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

Stevens, K. N. (1972). The Quantal Nature of Speech: Evidence from articulatory-acoustic data. In David, E. E. Jr. & Denes P. B. (Eds.) *Human communication: a unified view*. New York: McGraw-Hill: 51—66

Trudgill, P. (1995) *Sociolinguistics: An Introduction to Language and Society*, Penguin books.

Vallée, N. (1994) *Systèmes vocaliques: de la typologie aux prédictions*, Thèse préparée au sein de l'Institut de la Communication Parlée (Grenoble-URA C.N.R.S. no 368)

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

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

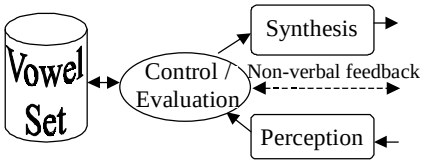

Figure 1: Agent architecture.

$$
\begin{aligned}
F_1 \;=\; & \left(\left(-392+392z\right)y^2+\left(596-668z\right)y+\left(-146+166z\right)\right)x^2+\\
& \left(\left(348-348z\right)y^2+\left(-494+606z\right)y+\left(141-175z\right)\right)x+\\
& \left(\left(340-72z\right)y^2+\left(-796+108z\right)y+\left(708-38z\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
F_2 \;=\; & \left(\left(-1200+1208z\right)y^2+\left(1320-1328z\right)y+\left(118-158z\right)\right)x^2+\\
& \left(\left(1864-1488z\right)y^2+\left(-2644+1510z\right)y+\left(-561+221z\right)\right)x+\\
& \left(\left(-670+490z\right)y^2+\left(1355-697z\right)y+\left(1517-117z\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
F_3 \;=\; & \left(\left(604-604z\right)y^2+\left(1038-1178z\right)y+\left(246+566z\right)\right)x^2+\\
& \left(\left(-1150+1262z\right)y^2+\left(-1443+1313z\right)y+\left(-317-483z\right)\right)x+\\
& \left(\left(1130-836z\right)y^2\left(-315+44z\right)y+\left(2427-127z\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
F_4 \;=\; & \left(\left(-1120+16z\right)y^2+\left(1696-180z\right)y+\left(500+522z\right)\right)x^2+\\
& \left(\left(-140+240z\right)y^2+\left(-578+214z\right)y+\left(-692-419z\right)\right)x+\\
& \left(\left(1480-602z\right)y^2+\left(-1220+289z\right)y+\left(3678-178z\right)\right)
\end{aligned}
$$

Figure 2: Synthesiser equations

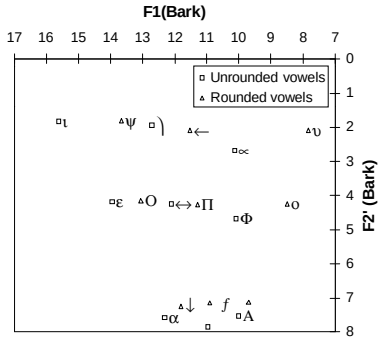

Figure 3: Vowels in F1-F2' space

Table 1: Basic organisation of the imitation game.

initiator	imitator
if ($V = \varnothing$) 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 ($V = \varnothing$) 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.

Table 2: Actions performed by the agents

Shift closer (v, A); return v_{best}	Find phoneme (A); return v_{best}	Update according to feedback signal

{
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}
}
	{
	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})
	}
		{
		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)	Do other updates of 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
}
	{
	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.
	}

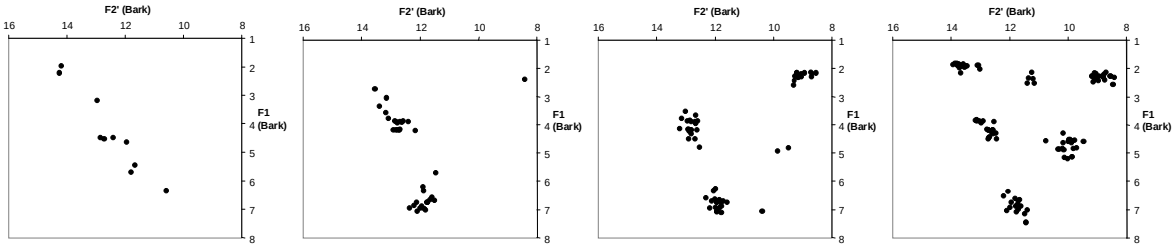

Figure 4: Vowel system after 20, 200, 1000 and 2000 games, 10% noise

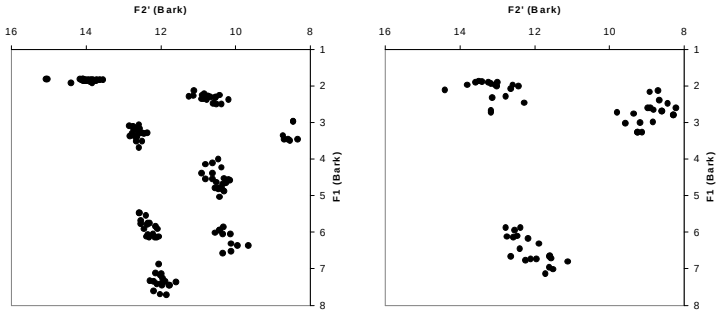

Figure 5: Systems with 10% and 25% noise

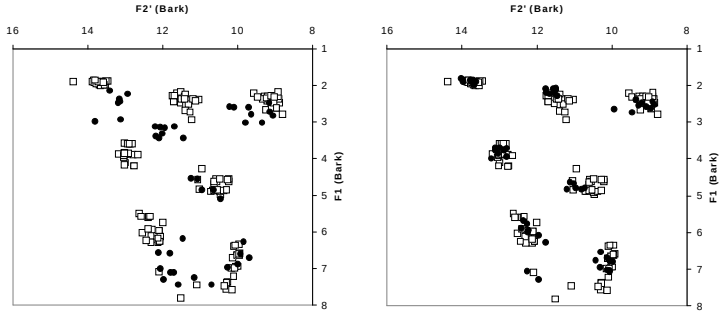

Figure 6: Systems after population replacement.

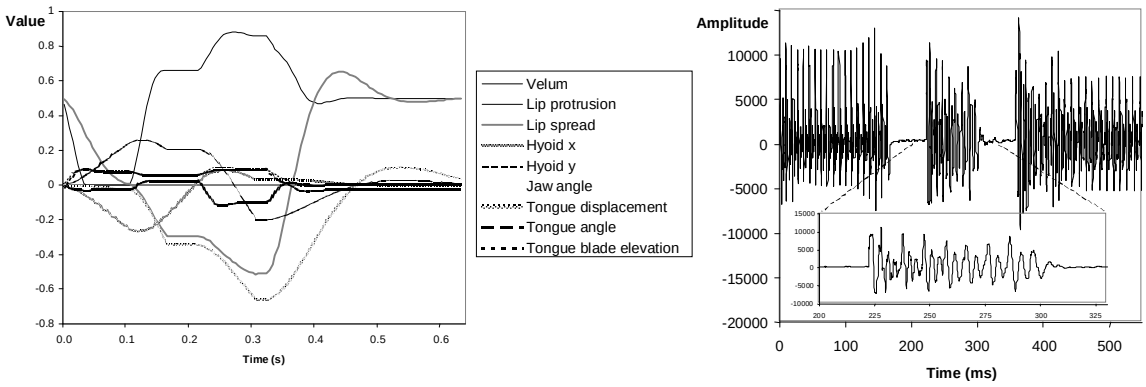

Figure 7: Gestural score for articulation (left) and acoustic signal: [dɐ.bu.bᵚɛ] (right)

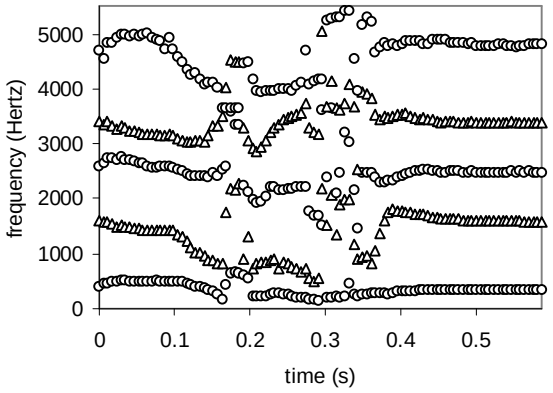

Figure 8: Formants extracted from artificial utterance.

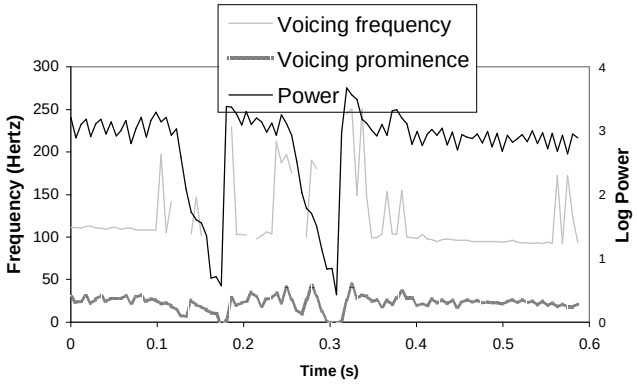

Figure 9: Other acoustic features