A realistic model of emergent phonology

Bart de Boer
Vrije Universiteit Brussel
bartb@arti.vub.ac.be

Abstract
In this report, ongoing research into the explanation of universal characteristics of human sound systems is presented. The characteristics are explained through a theory which is called emergent phonology. This theory is based on the ideas of Luc Steels (Steels 1995, 1996, 1997) that language is an emergent adaptive system in a population of language users in which coherence is maintained through self-organisation. Complexity is an emergent property of the interactions (called language games) between the agents.
Two simulations are presented. The first is of vowel systems. This simulation has been completed and the system, as well as its results, in which universal characteristics of vowel systems do emerge, are presented. The second simulation is designed for more complex utterances that consist of multiple syllables with consonants. This is ongoing research. The technical details of the production and perception systems are provided in some detail. The first results of the combination of the production and perception parts are presented.

1 Introduction

When one observes the sound systems used in human language, one finds a striking number of regularities. These regularities can be of the relative frequencies of sounds, or of the structure of sound systems. An example of the first kind of regularity is the fact that certain vowels, like [i], [a] and [u] occur almost universally in the world's languages, whereas certain other vowels, such as [y], [œ] and [ɯ] appear much less frequently. An example of the second kind of regularity is that if one finds front rounded vowels, such as [y], [ø] or [œ] in a given language, one almost always finds there unrounded counterparts [i], [e] or [ɛ]. Both kinds of regularities can also be illustrated with consonants. Although these facts are already quite apparent if one studies the sound systems of only a couple of languages, since the work on the Stanford Phonology Archive (Vihman 1976) and the UCLA Phonetical Segment Inventory Database (Maddieson, 1984) reliable statistical analysis of these phenomena could– and has been done by among others (Crothers 1978, Schwarz et al. 1997).
How can these phenomena be explained? Structuralism (Trubetzkoy, 1939; Jakobson & Halle, 1956) and generative theory (Chomsky & Halle, 1968) have explained these phenomena on the basis of *distinctive features* and their *markedness* as well as with rules that operate on these features. Distinctive features are (usually binary) properties that can distinguish one speech sound from another. Although distinctive features are usually assigned some kind of acoustical and articulatory properties, they are in fact rather abstract concepts and most phonologists do not seem to worry much about the exact physical basis of the distinctive features they use. Distinctive features, especially binary ones, can conveniently be used to explain the structure of sound systems, because if a sound system contains sounds that have one setting of a certain distinctive feature, one also expects sounds with the opposite value. For example, if a sound system contains a distinctive feature nasal, one expects to find both nasal and non-nasal phonemes at the same places of articulation. Relative frequencies of sounds can be explained by the markedness of the distinctive features. Distinctive features and settings of distinctive features that are unmarked will be used before the features and the settings that are marked. In our example of the feature nasal, the positive setting is more marked than the negative setting, so one expects less nasal consonants and vowels than non-nasal ones.
But actually, as has already been pointed out by Lindblom (Lindblom, 1984) distinctive feature theory does not explain the observed phenomena. It is just a description of these phenomena. Invoking distinctive features for their explanation would be circular. The distinctive features are deduced from observing the occurrence patterns of sounds in languages. The markedness of a feature is deduced from the frequency with which certain sounds occur. If a sound occurs less frequently, its features are said to be more marked. However, the description one derives in this way should obviously not be used to explain the very facts on which it is based.
Also the production and perception of sounds is such an obviously physical process that one can hardly ignore the functional constraints that apply to this process. Humans can not produce every sound equally easily, nor can they perceive the difference between sounds equally easily. It is therefore to be expected that certain sounds will appear more often in human languages than others. Several researchers, among others Stevens, (Stevens 1972, Stevens & Blumstein 1975, Stevens 1989) Carré (Carré et al. 1995, Carré & Mody 1997) and Lindblom (Liljencrants & Lindblom 1972, Lindblom et al. 1984, Lindblom & Maddieson 1988, Lindblom 1992) have proposed different functional criteria that can explain the structure of sound systems (mainly vowel systems). Stevens' "quantal theory of speech" (Stevens 1972, Stevens & Blumstein 1975, Stevens 1989) focuses on stability of vowels. According to Stevens, certain positions of the human articulatory apparatus are more stable than others,

in the sense that small perturbations of these positions cause small changes in the sound produced. These stable positions will be preferred over less stable positions and will therefore occur most frequently in the world's languages. Carré's "distinctive region model" tries to predict vowel systems (Carré et al. 1995) and (more recently) consonant systems (Carré & Mody, 1997) from physical principles. The vocal tract is simplified to a tube, and Carré postulates that the simplest perturbations that have the maximum acoustic effect will be used as the most basic speech sounds. He shows that the major vowels and consonants can be more or less predicted in this way. Lindblom tries to predict the shape of vowel systems using a maximisation of acoustic distinctiveness. The most frequently found vowel systems can be predicted by starting his computational model with a given number of vowels at random positions, and then gradually having them repel each other, within the acoustic space that can be reached by the human articulatory system. More recently, Lindblom (Lindblom 1992) has extended his ideas to consonant systems (Lindblom 1992). Here he also takes factors of articulatory complexity into account.
Although these theories all provide aspects of good functional explanations of *why* certain sound systems are better than others, they do not provide an explanation of *how* these systems have become the way they are. They all depend on an optimisation of certain acoustical or articulatory properties. However, humans do not do any optimisation when learning a sound system of a language. They just imitate their parents (and their peers). In fact, they imitate their parents much better than would be required for successful communication. Babies are able to hear all the subtle nuances that are used to distinguish meaning in any of the world's languages. Stronger still, they even learn nuances that are not used to distinguish meaning. Speakers of different, closely related, dialects that can understand each other perfectly well, can also hear clearly that the other does not speak the same dialect. Apparently sound systems are learnt by very careful imitation.
How then, can an optimal system of sounds appear in a language if none of the speakers does actual optimisation? Surely, the imitations are never perfect, and all speakers have their own idiolects, but how can a consistent optimisation take place from this random variation? In this paper and elsewhere, (de Boer 1997a, 1997b, 1997c) I show that this can be explained by Steels' theory of language as an emergent adaptive system.
In this theory, Steels (Steels 1995, 1996, 1997) considers language to be a system that is *distributed*, *adaptive* and *emergent* in a population of agents. It is distributed, because no language user (which he calls agents) has perfect or complete knowledge of the language, and none of the agents has central control over the language. It is adaptive, because it serves the purpose of a communication tool and adapts itself constantly to this purpose, constantly changing over time. Language is also an emergent system, because if no agent has central control over the system, the coherence in the language can only be maintained through local interactions (which Steels calls *language games*) between the agents from which the agents can improve and update their knowledge of the language. Emergent global behaviour from local interaction is called *self-organisation*, and this plays a major role in Steels' theory.
While Steels has mainly built computational models of his theory that are concerned with semantics and lexicon formation, (Steels 1995, 1996) his theory is equally applicable to the area of speech sounds. The agents will need to be able to produce and perceive speech sounds in a human-like way. They will try to imitate each other as well as possible. It can then be investigated if the universal tendencies of sound systems of human languages will appear in the sound systems of the agents in the population.
It should be mentioned here that Glotin and others from the Institut de Communication Parlée in Grénoble (Glotin 1995, Glotin & Laboissière 1996, Berrah et al. 1996) have already implemented, independently from the theory of Steels a system in which a population of agents exchanges vowel sounds with each other and develops a system of vowels that shows natural universal tendencies. Unfortunately, this work was not pursued further than a number of preliminary experiments.
This paper will briefly discuss results on vowel systems that have already been achieved, but these are more extensively described elsewhere (de Boer 1997a, 1997b, 1997c). The main part of this paper describes work in progress on the challenging task of applying the theory of emergent phonology and its computational model to complex utterances that contain consonants. The paper is organised as follows: in the next section, we will expand a bit on the theory of emergent phonology. The theory will be explained into more detail and some experimental results on vowels are briefly hinted at. In section 3 it will be explained why it is necessary to expand the model to more complex utterances and what has to be done in order to implement it. In section 4 the articulatory model that will be used in our experiments is described. In section 5 a model for perception is described. Then, in section 6 the work that remains to be done before the complex utterance model can be used is outlined, and finally, in section 7 the work described in this paper is discussed and some preliminary conclusions are drawn.

2 Emergent phonology

Emergent phonology considers speech sounds as the outcome of an adaptive process in which a population of agents develops and maintains a set of sounds with which they communicate under constraints of perception and production. This means that the phonology of a language is not static. In fact, due to the limitations of production and perception, and due to the individual differences between the language users, no language user can learn the

exact same sound system as any of the others. But this is not necessary. Variation that does not disrupt the adaptive (communicative) purpose does not matter, but these variations will tend to be small. Therefore, the sound systems of all the agents will remain similar. However, as the agents learn the sounds from each other, and as agents leave the population (they die or move away) and enter the population (by birth, or by immigration) it might happen that certain fluctuations are amplified, and the sound system of the language as a whole changes. Thus certain changes are amplified and other changes are suppressed through the positive feedback of the agents' learning from each other. This is a characteristic property of a *self-organising system*.

Probably the first time that speech sounds were considered as emergent properties of a self-organising system was in an article by Lindblom, MacNeilage and Studdert-Kennedy (Lindblom et al. 1984). Later the ideas of language as an emergent system were applied to other parts of language by Steels (Steels 1997) who also identified a number of basic mechanisms that cause this emergence: co-evolution, self-organisation and level-formation. Steels also introduced the idea of testing the theories on self-organisation with computer simulations. The ideas of Steels are based for a large part on the research of *artificial life*. In artificial life (see e.g. Langton 1989), one tries to build computer models of diverse life-like behaviours using (mostly) simple, local interactions in large populations of artificial agents. The theory of emergent phonology is a synthesis between the original ideas of Lindblom et al. and Steels' ideas. Just like Steels ideas the theory of emergent phonology is also testable with computer models, as will be illustrated in this paper.

Note that the theory of emergent phonology does not specify how the agents should learn the speech sounds they use, nor how these speech sounds should be represented. However, the theory does demand that the agents are able to imitate other agents sufficiently accurately, and that they also should not have a very strong innate bias towards certain sound systems. If they would, the gradual changes on which the theory of emergent phonology depends would not be possible.

Emergent phonology is rather a theory about how coherent systems of speech sounds can emerge in a population of agents that each are able to produce, perceive and learn these speech sounds. Also the theory seeks to explain how universal tendencies of the structure of sound systems can emerge from the interactions in the population, rather than from optimisations by- or innate constraints on individual speakers. In addition the theory provides a framework with which to investigate sound change.

As the theory and the computational models are mostly concerned with production and perception of actual speech signals, one can ask the question whether it is not rather a theory about emergent phonetics. However, the processes that the theory describes are rather related to the way the available acoustic and articulatory spaces (which are the subject of the study of phonetics) are divided into subspaces or prototypes that the language users can then use to communicate with each other. These subspaces or prototypes can be considered as being phonemes. The theory thus provides an account of how phonology can emerge from phonetic signals, and how phonetic changes can cause phonological changes.

The theory of emergent phonology was first successfully applied to the explanation of the structural regularities one finds in the vowel systems of the world's languages. As has been pointed out in the introduction, these regularities can be seen as the result of an optimisation of articulatory and acoustic distinctions. However, it is also clear that no human learning language does an optimisation of the sound system that he or she learns. Therefore, the theory of emergent phonology proposes that the optimisation is the result of repeated interactions in a population of language users. In these interactions, the language users learn the sound system with (very) slight perturbations. As the sound system has a communicative purpose, the perturbations that increase the communicative success will be favoured over the perturbations that decrease the communicative success. Therefore, in the long run optimal systems will be favoured over sub-optimal systems.

In order to test the validity of this proposal, a computer model was built that models a population of agents that are able to produce, perceive and learn vowels and that try to imitate each other as well as possible. The agents are equipped with an articulatory synthesiser that produces the first four formant frequencies of the vowels, based on the three major vowel features: (Ladefoged & Maddieson 1996, ch. 9) position, height and rounding. This is done by interpolating between a number of data points that were generated by a speech synthesiser based on Maeda's model (Maeda 1989, Vallée 1994 pp. 162–164). In order to make the articulation more realistic, noise can be added to the formant frequencies. They are also equipped with a realistic model of perception of formant frequency patterns based on a model by (Mantakas et al. 1986, Boë et al. 1995). This model is used to calculate a distance between formant patterns. Each agent keeps a list of vowel prototypes. When an agent hears a sound, it finds the closest vowel prototype in its prototype list and assumes that that was the sound it heard.

The interactions between the agents are called *imitation games*. For each imitation game, two agents are selected at random from the population. One agent is the *initiator* of the imitation game, the other is the *imitator*. The initiator randomly selects a vowel from its prototype list and produces it with its speech synthesiser. The imitator listens to this sound, finds the closest match in its prototype list and produces this prototype with its speech synthesiser. The initiator then finds the closest match to this sound, and checks whether the prototype it finds is the same as the one it originally picked. If this is the case, the imitation game is said to be successful, if it is not the case the imitation game is a failure. Depending on whether the imitation game was successful or a failure, vowel

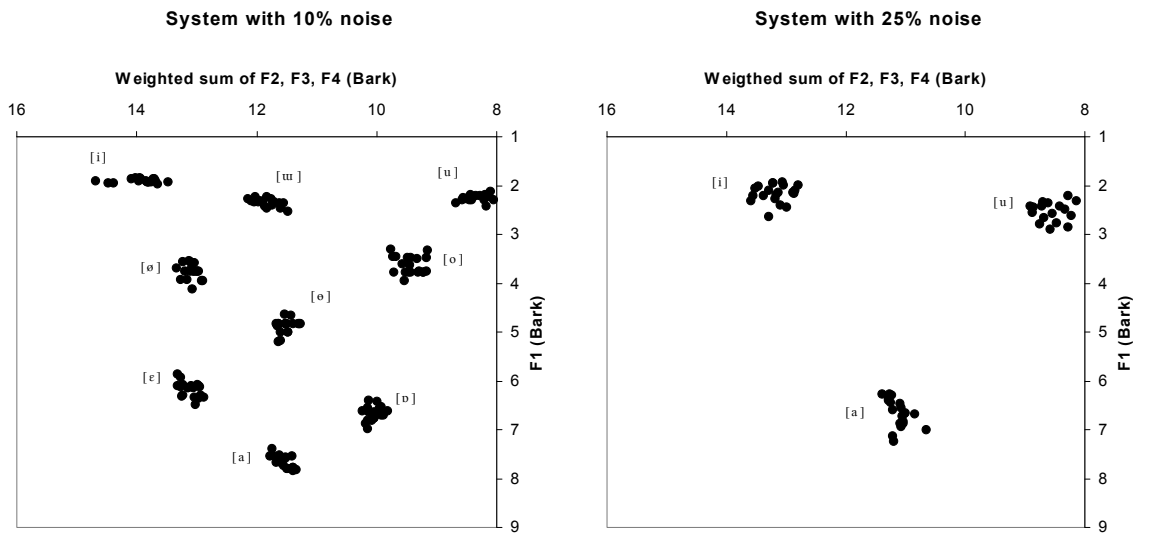

figure 1: Two vowel systems that emerged from the imitation game in a population of twenty agents.

prototypes can be added, removed shifted or merged. This is described in detail in (de Boer 1997a, 1997b, 1997c).

The agents start out with empty vowel lists. For the first few imitation games, the initiating agents will generate vowels at random. The initial vowels will therefore be spread in a random way through the available acoustic space. However, as soon as all agents have added one or more vowels to their prototype lists, the vowels will start to cluster. In order to keep a pressure on the agents' vowel systems, new vowels can still be added at random, but this only happens with very low probability, in the order of once every 100 imitation games. After approximately 2000 imitation games, the agents will have developed realistic vowel systems that contain a number of vowels that depends mostly on the amount of noise that is added to the agents' articulations. Two systems, generated under different conditions of noise, are shown in figure 1. As can be observed, the system that has more noise has less clusters that are more dispersed. Although these systems are nearly optimal, they are not totally stable. The vowel prototypes can shift slightly, existing prototypes can disappear and new ones can appear. This can happen because there is no explicit optimisation; the optimisation is the result of a self-organising process and can therefore be disturbed temporarily by local fluctuations.

3 Towards complex utterances

The experiment described above shows that the theory of emergent phonology has explanatory power and that computer models can be used to test the theory. However, a number of simplifications have been made for implementing the system. The most important of these simplifications is that agents can only produce an utterance that consists of a single vowel. This makes it impossible for the experiment to tell us anything about more complex utterances that obviously play a crucial role in human languages. It can not tell us anything about interesting questions, such as why all human languages have consonants as well as vowels, even though only vowels would be sufficient for making a communication system. Nor can it tell us anything about the regularities one finds in consonant systems (Lindblom & Maddieson 1988) or the regularities with which speech sounds are put together into syllables (Venneman 1988.) For modelling consonants and syllables complex, dynamic utterances are needed.

Another important subject about which the system based on single vowels can not tell us anything is historical change. Although the vowel systems that emerge from the experiments do undergo some change through time, this is only a random drift that is stabilised through self-organisation. Although this undoubtedly plays a role in human languages as well, most historical change is rather caused by the phonetic context in which certain sounds appear. This phonetic context influences the way sounds are pronounced in a systematic way, and can thus trigger sound changes. However, for these sound changes to operate a certain redundancy is necessary. Otherwise the language user will not be able to determine which string of sounds was meant. Both phonetic context and redundancy require a model that works with complex utterances.

Therefore if one wants to proceed with testing the theory of emergent phonology, a more realistic model of articulation and perception that can handle complex and dynamic utterances is needed. This model should a priori be able to generate the same sounds as human speech. We do not want to limit the model to a given subset of speech sounds (say, for example, only to voiceless consonants) because then we can not draw the kinds of conclusions about the universal characteristics of speech sounds that we want to draw. This means that not only the

articulatory model should be accurate, but also the control of the moveable structures that produce the actual speech gestures.
We also want the perception of speech sounds to be as realistic as possible. We do not want to limit ourselves to only specific cues in the speech signals (such as only formant patterns). All the signals that could be important for human perception should be available to the agents. Again this is necessary as not to bias the system.
No system that fulfils these conditions has been built, yet. Some research into functional explanations of the universal tendencies of consonant systems or syllable structure has not used computational models at all. Instead general physical arguments were used (Butt 1992, Carré 1997.) Lindblom (Lindblom 1992) did build a simple computational model to explain the structure of consonant-vowel syllables, but this model was limited to only a discrete (but large) number of consonant-vowel syllables and the perception of consonants was only based on formant patterns and not on the noise bursts that also plays an important role in recognising consonants. Also Lindblom's model consisted of a simple optimisation of articulatory effort and acoustic distinctiveness, which, as has been pointed out above, is not the whole story.

4 The articulatory model

If one wants to explain characteristics of phonological systems through functional criteria and self-organisation, the articulatory model should be as free as possible from any theoretical biases. If the model is biased, it would always be possible that the phenomena that one observes are actually caused by the biases in the model. However, it is obviously impossible to build a model without any theoretical bias, as one always has to make simplifications. In order to reduce the amount of simplification to a minimum, one has to stay as close to the actual physics as possible. Therefore an articulatory model that is based on a physical simulation of the speech signal is required.
The speech signal is caused by two physical processes: the movements of the actuators and the propagation of the sound waves through the vocal tract. Both these processes need to be simulated. We will first describe the modelling of the propagation of the sound waves through the vocal tract into considerable detail, and then we will describe the modelling of the movements of the articulators. Then we will illustrate the performance of the model with some example utterances it can generate.

4.1 The acoustic model

If one wants to know what sound is produced for a given position of the articulators, such as the tongue, lips and pharynx, one needs to know the three-dimensional shape of the vocal tract that is the result of that position of the articulators. However, the complete shape is extremely complex. It is usually approximated by dividing the vocal tract along its length into a number of sections of equal length. The exact shape of the sections does not matter very much for the frequencies that are important for speech signals (0–5000 Hertz.) Therefore, only the average area of each section matters. The sections are usually approximated by a cylinder with an area that is equal to the average area of the corresponding section of the vocal tract.
Several models exist that can calculate the cross-sectional areas of the vocal tract. The two most often used models are by Maeda (Maeda 1989) and by Mermelstein (Mermelstein 1973, Rubin & Baer 1981). Maeda's model is based on a principle components analysis of a large number of frames from Röntgen films of a speaker saying test sentences. This resulted in a model with seven degrees of freedom, roughly corresponding to the articulatory degrees of freedom, from which the areas of the vocal tract can directly be calculated. Unfortunately, the exact numerical relation between the degrees of freedom and the actual cross-sectional areas is nowhere described exactly. Mermelstein's model is based on a geometrical model of the vocal tract, in which the anterior/inferior- and posterior/superior walls of the vocal tract are modelled as a sequence of lines and circular arcs. According to this model, the cross sectional areas of the vocal tract can be calculated from the distances between the two walls, using different functional relations for different positions in the vocal tract. The model has nine degrees of freedom, from which an arbitrary number of cross sectional areas can be calculated. It has been described into considerable detail in (Mermelstein 1973). Although this description is not totally complete, an implementation of the model could be made, based on the text and the accompanying figures. The two-dimensional vocal tract outline, as well as the parameters of the model are illustrated

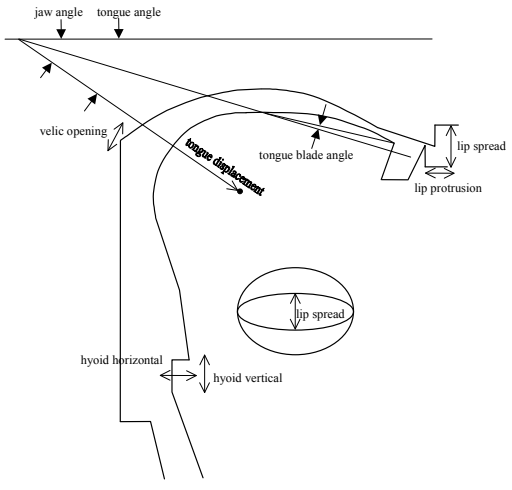

figure 2: The parameters of the articulatory model (adapted from Mermelstein 1973.)

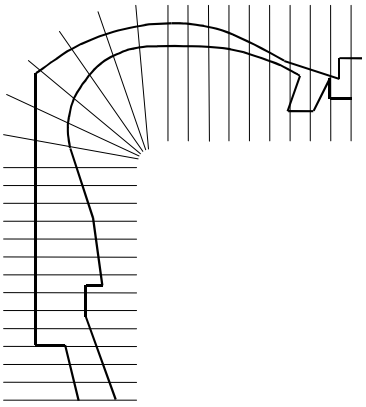

figure 3: The sections of the articulatory model (adapted from Mermelstein 1973).

in figure 2. The nine parameters of Mermelstein's model are: the horizontal and vertical displacement of the hyoid cartilage, the angle of opening of the jaw, the angle of lowering/raising of the tongue body, the displacement in the front/back dimension of the body of the tongue, the angle with which the tongue blade is raised/lowered, the amount of opening of the velum, the spread of the lips and the protrusion of the lips. The model was implemented here with one of the degrees of freedom (the velum) disabled, as no satisfactory data on the nasal passage, which would have to be taken into account if the velum was allowed to be opened, has been found so far. The way cross sections are made in order to calculate the areas is illustrated in figure 3. The output of the model consists of twenty-nine cross sectional areas.

The propagation of sound waves through the sections described with these areas is approximated by a *lossless tube* model. In the lossless tube model, the sections of the vocal tract are considered to be cylindrical with totally rigid walls. This means that sound waves propagate through the tubes without losses. The lossless (Rabiner & Schafer 1977, ch. 3.3) tube model can very easily be approximated with a discrete time model of sound waves propagating through the tubes. An sound wave moving from one section of tube to the next will partly propagate into the next section and partly be reflected back. In a single section two sound waves will therefore be travelling: one from back to front, and a reflected sound wave from front to back. The forward-going sound waves $w_k^+(t)$ in a section of tube k at time t are given by the following formula:

1) $$w_k^+(t) = (1 + r_k) w_{k-1}^+(t-1) + r_k w_k^-(t)$$

Where the backward-going sound waves similarly by:

2) $$w_k^-(t) = (1 - r_{k+1}) w_{k+1}^+(t-1) - r_{k+1} w_k^-(t-1)$$

The reflection coefficient r_k between two tubes of cross sectional area A_{k-1} and A_k is given by:

3) $$r_k = \frac{A_k - A_{k-1}}{A_k + A_{k-1}}$$

The first tube of the vocal tract gets its incoming waves from a simulation of the vibrations of the vocal chords (which will be described below). Backward-going waves are reflected completely at the glottis. To the last section that was calculated by Mermelstein's model, an extra section of a fixed area of 30 cm^2 is added. From this extra section, forward going waves can escape completely, and no reflections come back.

At what frequency should this model be sampled? The sound waves take $\frac{l}{c}$ seconds to propagate through one section of the tube, where l is the length of the section and c is the speed of sound. The model should be sampled every time a sound wave propagates through a section. In our model, where we have 29 tubes for a vocal tract length of 17.5 cm, we have section of 0.60 cm and therefore a sample every 18.3 microseconds, corresponding to a sample frequency of 54.7 kilohertz. As we want to sample the sound at a rate f_s = 11025 Hertz (¼ CD frequency) we need to undersample with a factor of five.

Two more things are taken into account in modelling the propagation of the sound waves through the vocal tract. One is a damping factor that prevents high pitched metallic noises from occurring when tube diameters become too small. This is done by multiplying each wave in a section k of area A_k with:

4) $$1 - \frac{0.0007}{\sqrt{A_k}}$$

which is an adaptation for the damping factor α, described in appendix B of (Rubin and Baer 1981.) This effectively dampens reflections in sections with small cross sectional areas.

It also needs to be taken into account that if the area of sections is sufficiently small, and the pressure of the sound waves is sufficiently large, then turbulence will occur. When the following condition, with w_k^+ the sound wave pressure, A_k the area in section k and φ the threshold, usually set to 300, is met:

5) $$\frac{w_k^+}{A_k} > \varphi$$

the sound wave pressure w_k^+ is multiplied by a random number from a uniform distribution between 0 and 1.

It can be seen from the preceding that cross sectional areas should not become zero. This is prevented in the calculation of areas from the Mermelstein model. Whenever an area becomes smaller than a given threshold, it is set to that threshold. This threshold is usually set to 0.001.
The tube model of the vocal tract has to be excited with vibrating vocal chords for voiced sounds. The vocal chords are modelled as a filtered pulse train. The pulse train is generated as follows: at time zero, an output of (the arbitrarily chosen value of) 1000 is generated. Then the output is set to zero for a number of time steps N, that is given by the following formula:

6) $$N = \frac{f_s}{f_v}(1-u)$$

where f_s is the sampling frequency, f_v is the voice frequency and u is a random value from the uniform distribution between –0.01 and 0.01. Then again an output of 1000 is generated, and the process is repeated. The pulse train is then filtered by a single digital filter with centre frequency of zero Hertz and a bandwidth of 1000 Hertz. Although primitive, this gives a very reasonable approximation of the vibrations of the vocal chords.

4.2 The gestural model

The model described above allows us to calculate the area function of the vocal tract as well as the sound that will be produced by exciting the vocal tract from a given setting of the articulators. However, in order to implement dynamic utterances, we still need a model that describes the movements of the different articulators. These movements should be smooth and realistic. This means that we have to take into account that articulators cannot jump from one position to the next and that they have a certain inertia so that it takes some time before they start or stop moving.
In real speech, different articulatory gestures tend to overlap. For example the /s/ in the English word /soot/ is in fact rounded [s^w] whereas the /s/ in the English word /sit/ is an ordinary [s]. This happens purely because the vowel in the first word is rounded, and the vowel in the second word is not and because in English the difference between [s] and [s^w] cannot distinguish two words in meaning. Therefore the articulatory model should allow for overlapping articulatory gestures if multiple gestures have to be produced.
The requirements of continuous movement and inertia are actually independent. The first one restricts the movements of the articulators to physically plausible ones. The second one is rather a requirement on the possible ways in which the human nervous system can schedule the different articulatory gestures. Both of these requirements will cause certain effects which are called co-articulation effects by linguists. But the first requirement causes language independent co-articulation effects whereas the second requirement causes language dependent co-articulation effects.
Although a lot of research has been done into the way articulators move and how dynamic articulations are made (see e.g. Browman & goldstein 1995, Kaburagi & Honda 1996) these models were either not sufficiently well described or too complex, so it was decided to design a simple but realistic model from scratch. The articulators in this model move according to the following formulae:

7) $$v_{t+1} = v_t + 0.3 \cdot \mu(\mu(g - p_t) - v_t)$$

8) $$p_{t+1} = p_t + 0.3 \cdot v_t$$

where v_t is the speed and p_t is the position at time t. The (inverse) inertia is determined by μ ($0<\mu<1$) and the goal towards which the articulator moves is g. The factor 0.3 was chosen to generate satisfactory movement for the length of time between two updates of the articulators and for articulator positions usually between –1 and 1. The rate with which the articulators are updated in the model is $1/50^{th}$ of the sample rate of 11025 Hertz which is 220.5 Hertz. These formulae cause the articulators to be attracted to the goal position. However, the speed with which an articulator moves cannot change directly, due to the inertia. The maximum speed with which articulators will move is determined by the inertia and by the distance of the articulator from the goal position. The inertia for all articulators in our system is set to 0.5, except for the lips, which are thought to be lighter and faster, where it is set to 0.9.
If a number of articulations has to be performed in sequence, it has to be determined what parts of these articulations can overlap and what parts should not. An articulation always consists of a number of goals for the different articulators. Some of these goals are crucial for the correct realisation of the articulation and some are not. For instance, in the example of the /s/ above lip rounding is not crucial for the realisation of the /s/, whereas other articulator positions, such as tongue tip position, are. Apparently then, articulators that are not crucial for the realisation of an articulation that is being performed, can already assume positions of following articulations.
This is taken care of in the part of the articulatory model that sequences the articulations. If an agent wants to say a sequence of articulations (a "word") it sends the different articulatory goals for these articulations to a sequencer. However, the positions of the articulators that are not crucial for the realisation of the given articulators can be marked as unspecified. Whenever the sequencer sends new goals to the articulators, it takes the specified goals of the present articulation, and for all articulators for which a goal is not specified, it takes the articulator goal of the first articulation for which this articulator is specified and co-articulates it with the present articula-

tion. If there is no goal specified for a given articulator in any of the articulations, it is allowed to move towards a default (rest) position.
For every articulation that is sent to the sequencer, a duration needs to be specified. The duration specifies the time the sequencer waits before sending other goal values to the articulators, after it has started an articulation. The sequencer always waits for this duration, even though possibly most- or all articulators were already moving towards their present goal values, as they could have been co-articulated with previous articulations.
The sequencer has a maximum capacity of articulations in order to limit the maximum amount of co-articulation. The capacity of the sequencer is at present set to five articulations.
A last problem of articulator movement that needs to be addressed is what to do if articulators get blocked. There is only a limited space in the vocal tract, and goals for articulators in this model are not specified relative to other surfaces in the vocal tract, but in an absolute way. It therefore happens quite often that articulators collide with other articulators or with the walls of the vocal tract. Of course this problem is limited to some extent because all articulators have minimum and maximum positions. However it remains necessary to check whether articulators get blocked. This is done in a rather ad hoc way by checking whether the area function gets zero anywhere. This means there is a closure somewhere. As it is hard to determine which articulator is responsible for the closure, all articulators are stopped. Articulators whose goal is forward or downward are allowed to continue, although if their velocity was such that they moved inward or upward, it is set to zero (after which the attraction of the goal will start them moving again). Other articulators are stopped at their present position. Although this method of resolving collisions is rather gung-ho it works well and gives no unrealistic effects.
Besides the articulatory parameters that describe the shape of the vocal tract, the model also has a parameter that describes the voicing. At the moment this parameter does nothing but determine whether the vocal chords vibrate or not. The range of the parameter is [0,1] and the vocal chords vibrate when its value is in the range <0.15,0.95>. The parameter could be used to determine the mode of voicing, (breathy, modal, creaky) the pitch or the volume. At the moment a fixed pitch- and volume contour is used.

4.3 Performance

The articulatory system that has been described above is able to generate a wide variety of speech sounds. The quality of these sounds is reasonable and it is expected that it is good enough for the emergent phonology experiments. The system is able to produce vowels, semivowels, fricative and plosives. The quality of the vowels and semivowels is good, and the quality of the fricatives and plosives is acceptable. Human listeners accept the sounds as human. The sequencing and co-ordination of the articulations is also acceptable, but randomly generated sequences sound rather slurred. Apparently co-ordinating articulations requires rather precise timing. But humans have to learn to co-ordinate their articulations as well, so it is no problem that this has to be learnt for our articulatory system as well.
The system is also able to synthesise a number of different voice qualities. If the system is undersampled with a factor of five, the sound that is produced sounds like a male voice. If it is undersampled with a factor of six or seven, and if the frequency of the vocal chord model is increased, it sounds like a female or child's voice. Adding more or less noise to the vocal chord model makes the sound more or less harsh.
The performance of the model is illustrated in figure 4. Here one can see an example utterance of the model consisting of three syllables, generated from three articulatory goals. One can see that three different vowels are produced, as well as two different voiced plosives and a semivowel. One can also see in the detailed picture of the middle syllable that the onset of the plosive has a small high frequency noise burst, (the jaggedness of the first peaks) just as it should have.
The speed of the model is also good. When the system is running on a standard 133MHz Pentium desktop PC, it is able to calculate a signal as shown in figure 4 in approximately 7 seconds, meaning that it takes about ten times as much time to calculate an utterance than the time it takes to play it. Using faster hardware this could be improved even further. Although this is extremely slow compared to the speeds that can be achieved with the system that works with vowels only, it is still fast enough for playing imitation games.
The model has some limitations as well. As has been said before, the nasal cavities are not modelled. Also, the model is not capable of producing trills and laterals. The quality of fricatives could probably also be increased if a better model for noise generation were to be used. Furthermore, consonants which use airstream mechanisms different from simple pulmonary egressive, (the ordinary way of producing sounds) such as clicks or glottalic ingresive or egressive consonants cannot be modelled. However, these sounds are quite rare in the world's languages and always appear in systems that already have ordinary, pulmonary egressive consonants. We therefore feel confident that, although the model is not complete, it will be useful for first experiments on the emergence of consonant systems and syllable structure.

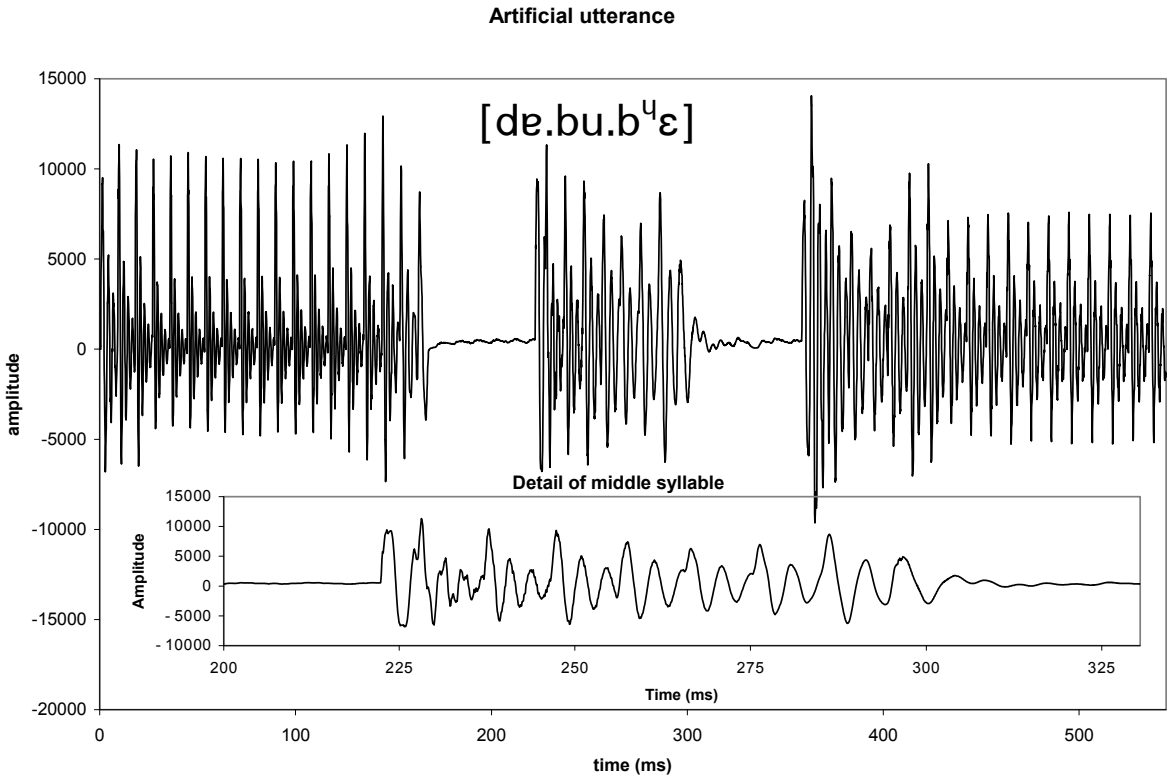

figure 4: An example of an artificially generated random utterance.

5 Models for perception

The agents should not only be able to produce sounds, they should also be able to recognise them. However, being able to recognise sounds is not enough. In order to imitate sounds they hear, they should be able to make an estimate of how a new sound they hear can be produced. In other words, they have to learn a mapping from acoustic signals to articulatory gestures. Although it appears that developmental linguists seem to assume that children are born with the ability to map sounds to articulations it does not seem likely that it is totally innate. It depends on the physical characteristics of the child's mouth and ears, it would require unlikely complex pre-wired neural circuitry and children do learn similarly complex sensory-motor co-ordination when, for example, learning to walk. It is therefore more plausible that it is a skill that is learnt by innately guided learning behaviour. In any case, the mapping from sounds to articulations is too complex to be programmed, so the agents will have to learn it.

This is not a simple problem. The sound that is produced is very indirectly related to the articulatory commands, it changes over time, is influenced by the preceding and following articulations and it is noisy. When articulatory goals are sent to the vocal tract model, it does not immediately assume the shape that is specified by the goals. It takes some time before the desired position is reached (if it is reached at all) and the trajectory depends on the shape of the vocal tract before the movement towards the goals was started. If some of the articulatory parameters are unspecified, goal values for later articulations can be substituted (as was described in section 4.2 above) so that the articulator movement can be influenced by following articulations as well. Noise is caused by frication when articulators are close together and by the fact that the glottal pulse is not totally regular. Another source of random variation is the fact that the tube model never totally returns to the same initial state.

5.1 Sensor pre-processing

An important first step in solving this problem is choosing a good way of pre-processing the sensor data. The sensor information an agent gets is just a time series of amplitudes of the speech signal (an example of which was shown in figure 4.) It is very hard to use this time series directly for recognising and mapping sounds. A good way of extracting useful information, as well as compressing the amount of data slightly, is to make a spectral analysis of the time series. This is, in fact, what happens in human (and animal ears). Due to the shape and the organisation of the inner ear, the auditory neurons are fed with signals that represent the strength of the signal at different frequencies. Although a computer does not have an inner ear, it can split a signal into its constituent frequencies through a discrete Fourier analysis.

The discrete Fourier transform H_n of a sample $h_0 \ldots h_{N-1}$ can be calculated by the following formula (adapted from Press et al. 1995, ch. 12.1):

9) $$H_n = \sum_{k=0}^{N-1} h_k e^{\frac{2\pi i k n}{N}}$$

where $-N/2 \le n \le N/2$. If we assume that h was sampled with frequency f_s, then the so-called periodogram estimate $P(f)$ of the power spectrum of this sample can be calculated as follows (adapted from Press et al. 1995, ch. 13.4):

10)
$$P(f_0) = \frac{1}{N^2}\left|H_0\right|^2$$
$$P(f_k) = \frac{1}{N^2}\left(\left|H_k\right|^2 + \left|H_{-k}\right|^2\right) \quad k = 1,2,\ldots,\left(N/2 - 1\right)$$
$$P(f_{N/2}) = \frac{1}{N^2}\left|H_{N/2}\right|^2$$

where the frequencies f_k, $0 \le k \le N/2$ are given by:

11) $$f_k = f_s \frac{k}{N}.$$

This gives us a reasonable estimate of the power spectrum of the sample, but according to Press et al. (1995) it is not a very good estimate. It can be improved by windowing the sample and by averaging over a number of power spectrum estimates.

Taking a sample from a time series is the equivalent of multiplying the time series with a sampling function that is one at the points where one wishes to take a sample, and zero elsewhere. However, the sharp edges of the sampling function cause the energies of the frequencies of the original signal to leak into neighboring frequencies. It is therefore a much better idea to use a sampling function that is zero at the edges and rises smoothly towards the center of the signal. This reduces the leakage into neighboring frequencies considerably. In the system presented here the sampling function was implemented as follows:

12) $$h_k = d_k\left(1 - \cos\left(2\pi \frac{k}{N}\right)\right)$$

where $h_0 \ldots h_{N-1}$ are the samples that the power spectrum estimator will use and $d_0 \ldots d_{N-1}$ are the actual data points. Using data windowing, a very reasonable power spectrum estimation can be made. However, the estimate is not very good from a statistical point of view. If one considers the power estimates as statistical estimators of the actual power, the standard deviation of these estimators is 100% (Press et al. 1995, ch. 13.4). By averaging each frequency estimate over a number K of consecutive estimates one can reduce this standard deviation by $\sqrt{K}$.

Instead of using the power spectrum estimates for recognition of the speech signal directly, its logarithm is used. The differences in power between the different frequencies is too big to be significant in the original signal. Taking the logarithm is also more realistic, as human hearing works in a logarithmic way as well. Both the original power spectrum as well as the logarithm of an [a] are presented in figure 5. As can be seen from this figure, it was decided to use a power spectrum estimate consisting of 32 elements. Note that for a power spectrum of 32 elements, 64 points from the time series are needed. The estimate was improved slightly by averaging over two

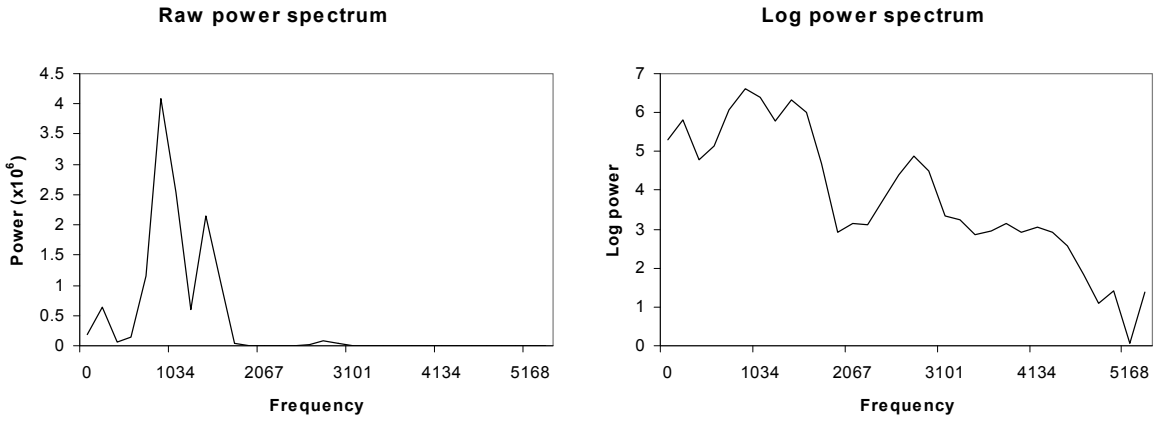

figure 5: The power spectra of [a].

consecutive estimates. Due to the way the average is calculated (copied from Press et al. 1995, ch. 13.4) 160 points from the time series are needed for every power spectrum estimate. At the given sampling frequency of 11025 Hertz, power spectra can therefore be calculated with a frequency of 69 Hertz, which is sufficiently fast to capture most interesting phenomena of speech.

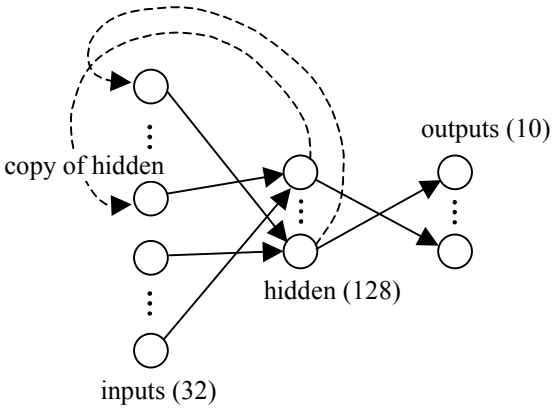

figure 6: Architecture of the network.

5.2 The inverse mapping problem

The main problem that needs to be solved is the mapping of sounds to articulatory gestures. This is called the inverse mapping problem and it has been studied by a large number of speech perception researchers. As has been said above, this mapping is too hard to program directly and should therefore be learnt. Fortunately our agents can train themselves by producing utterances, listening to themselves and learning the mapping between the gestures they make and the sounds they observe.

A learning mechanism that has to learn this mapping has to cope with continuously valued inputs, to take temporal context into account, to handle noisy input and to be resistant to small variations of timing in the input signal. These requirements rule out a large number of possible learning systems. Neural networks, however, are good at coping with continuously valued and noisy inputs. Recurrent networks are also able to cope with data that has temporal structure. One possible candidate is the Elman network (Elman 1990). The Elman network is a variation on the standard back-propagation network (which in this implementation is based on a description in Haykin1994, ch. 6) that takes temporal context into account by mapping the values of its hidden units to its input units. Elman networks have been applied successfully to a number of language-related temporal learning tasks.

The architecture of the network that was used in the system presented here is illustrated in figure 6. Because it was decided to make a power spectrum estimate of 32 points, the network has 32 input nodes. Elman seems to prefer to use four times as many hidden nodes as there are input nodes, so the number of hidden nodes was chosen to be 128. There are ten output nodes. Nine for the articulatory degrees of freedom and one for the setting of the vocal chords. The total number of connections to be updated is: $32 \cdot (128 + 32) + 128 \cdot 10 = 6400$. Also, all nodes have a bias which needs to be learnt. This makes for another 138 degrees of freedom.

The data set for training the network consisted of a sequence of random articulations generated by the articulatory model (with the voicing always on). The sound that was produced was analysed and the power spectra were used as inputs for the network. The output of the network was then calculated. The desired outputs were set to the articulatory goals of that moment. Then the errors were calculated and back-propagated through the network.

A large number of experiments was done in order to determine the best settings of the learning parameters. The Elman network has two learning parameters: a learning rate η and a momentum α. It was found that the best settings of these learning parameters was very low. The best results were achieved with $\eta = 0.001$ and $\alpha = 0$. This might seem very low for a learning method based on back-propagation, but it was necessary, because the examples are not presented in a random order. In fact, the desired output always stays the same for a number of consecutive examples. Therefore, if one sets η and α too high, the network learns how to *copy* the desired output pattern, regardless of its inputs. If no desired output pattern is provided, however, the network's performance drops dramatically. Therefore, one needs to set the learning parameters such, that the network learns too slow to be able to copy the desired output pattern completely.

The performance of the network is illustrated in figure 7. Here the sum squared error of the prediction of thirty random articulations in three periods of the learning process are shown. The first period is right at the beginning, when the network is not trained at all. Performance is very bad, but it is clear that things are learnt. The second period is when the network is trained completely, but when training examples are still provided. Here one sees that performance is much better. At the beginning of every articulation, the error is high, as the network is still in the state of the previous articulation. However, it decreases very rapidly as the articulation proceeds. The third period that is shown, is when the network is fully trained, and when training examples are no longer provided. Here one can see that the performance is worse than if training examples are provided, but it is much better than at the beginning of training.

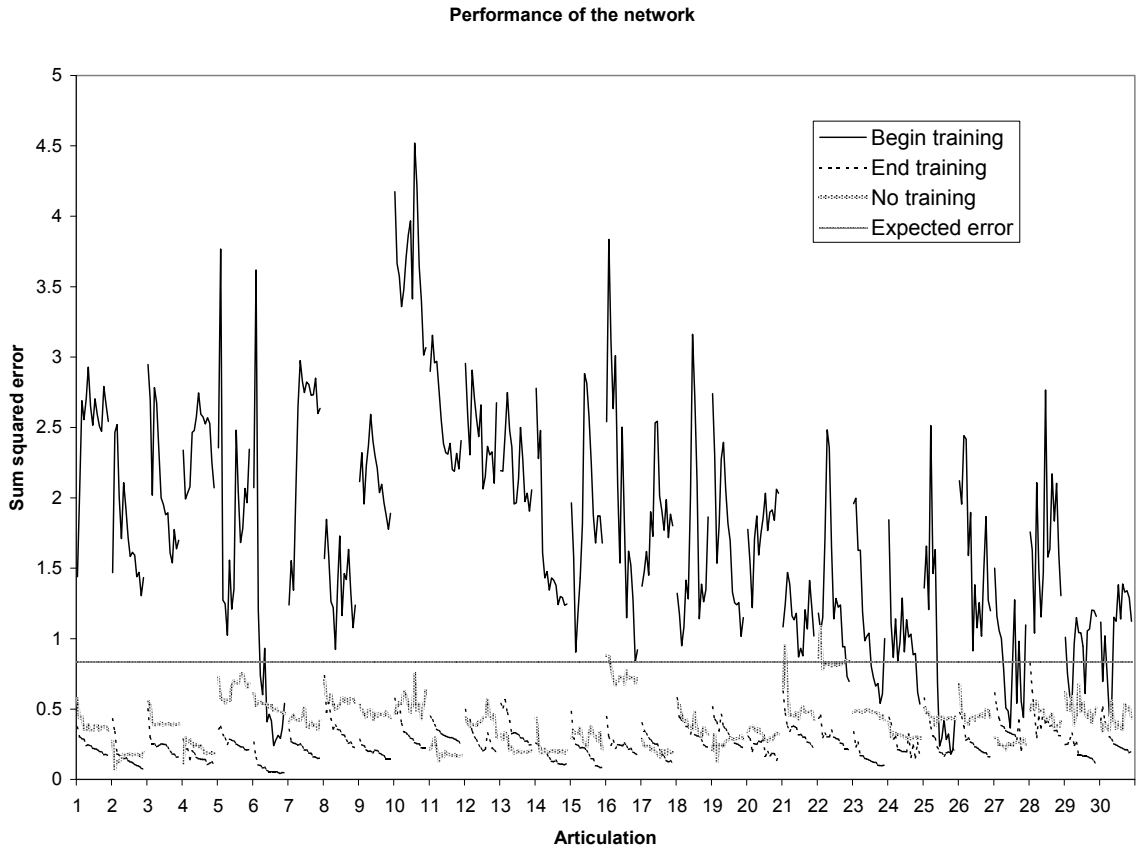

figure 7: Elman network performance. Three times thirty articulations are shown for different periods of training. Note that the articulations are not the same for the three periods.

A question that has to be asked for these result is how the performance compares to the maximum performance that can be achieved if the system did not use information from its inputs at all. The sum squared error E is traditionally calculated with the following formula:

13) $$E = \frac{1}{2}\sum_{i=1}^{N}(o_i - p_i)^2$$

where N is the number of outputs, o_i is the actual output and p_i is the desired output. The best a system that does not use its inputs can do, is always give an optimum value μ. As the value of each parameter that has to be estimated by the system is chosen at random, the expected value of square of the error per parameter with the range $[a,b]$ is:

14) $$\frac{\int_a^b (x-\mu)^2 dx}{(b-a)} = \frac{(b-\mu)^3 - (a-\mu)^3}{3(b-a)}$$

Which has a maximum for $\mu = \frac{(b+a)}{2}$ with the value: $\frac{1}{12}(b-a)^2$.

Substituting this in equation 13) and assuming that all parameters have the range [0,1] we find that the expected error is 0.417 which is the same as the error that is actually found, which fluctuates around this value. No statistically significant difference between the performance of the network and the behaviour that does not use inputs could be found. This means that the network does learn the optimal behaviour that does not use its inputs. Apparently the learning task is too complicated for the Elman network in combination with the 32 frequency bin spectrum estimate.

However, work is in progress on several speech processing techniques (cepstrum analysis and linear predictive coding analysis) for extracting the most important features from the speech signal. These features include: the power of the signal, the presence of voicing and its frequency, the position and bandwidth of the formant frequencies, the rate of change of the formant frequencies and possibly a measure for the presence or absence of noise. The signals that are produced by these feature detectors are much cleaner and map much more directly to the different actions of the articulators, so learning a mapping between these features and the articulator movements should be easier.

6 Future work

A model for producing complex sounds as well as a first model for learning perception have been implemented. These models will have to be integrated into an agent that can then engage in imitation games with other agents. However, a number of obstacles still has to be overcome before the imitation game will be operational. First of all, it must be determined what kind of utterances the agents will learn. Will they be single syllables, or should more complex words be allowed as well? What should the agent store? Obviously, it needs to store a lexicon of the words it knows, but should it also maintain a list of phonemes. Observations of children learning speech seem to indicate that children first learn complete syllables or words, and only later learn the concept of phonemes (Vihman 1996). It would be nice if this could be reproduced in the imitation game, but is as yet not clear how. Also, it should be determined how words are to be compared with each other. The identification of articulations can be done by the Elman network model, but once a sequence of articulations has been determined, how should this then be compared with the items in the lexicon. It is possible that certain articulations have been misinterpreted, or even missed completely, but it might still be possible to identify the most likely item of the lexicon that was uttered by the other agent.

For the rest should the rules of the imitation game remain the same as for the imitation games with simple vowel sounds. The same game where one agents picks an item from its lexicon, with the other agent analysing and trying to imitate it, while the first agent decides whether the imitation was good enough, should be played with these complex utterances. From the imitation games a stable repertoire of speech utterances should emerge.

The linguistic validation of these imitation games would be the appearance of a limited variety of speech sounds, with vowels and consonants. The distribution of the vowels and consonants should obey to rules that are similar to the ones that appear to apply to human languages. Also the allowed combinations of sounds should obey to phonotactic rules that are similar to the human ones. The rules should be explainable through functional criteria, such as ease of production and ease of perception. Obviously, as both the model of perception and production are much simpler than human production an perception, we should not expect to find the same variety an complexity that is found in human languages.

When the behaviour of the system is understood better, we could do experiments with dynamic populations, in which empty agents are added to a population of agent that already have a fully developed sound system. It should then be possible to observe "historical" changes in the sound systems of the agents. These sound changes can be of a more complex kind than the ones that were observed in the vowel systems, as in this model both influencing of sounds by neighbouring sounds as well as disambiguation of changed sounds through context should be possible.

Some more practical problems will have to be tackled as well. A nasal passage should be added to the articulatory model. Nasal sounds play an important role in the sound systems of the world's languages. There is hardly a language that does not have nasal consonants, and the influences of nasal sounds on neighbouring sounds and vice versa are among the best known examples of co-articulation and sound change. The vocal chord model should be improved as well. At the moment the intonation on the utterances is pre-programmed. However, intonation plays an extremely important role in human speech perception. It is therefore important that pitch be regulated by the agents themselves.

7 Conclusion and discussion

Two main conclusions can be drawn from the work done so far. The first is that Steels' (Steels 1997) idea of language as a self-organising system is able to explain the emergence of certain universal characteristics of human sound systems. This has been shown in the experiments with vowel systems (de Boer 1997a, 1997b, 1997c). In these experiments realistic vowel systems emerged in a simulated population of agents that tried to imitate each other's vowel sounds. The second conclusion is that there seem to be no more great practical problems to apply Steels' theory to consonant systems as well. An articulatory synthesiser has been built and tested, and a model of perception seems to work reasonably as well. The theory has already been applied to other parts of language, such as lexicon formation (Steels 1995) and semantics (Steels 1996) and research on syntax is underway. Applying the theory to phonology and phonotactics will extend the experimental foundation of the theory.

Functional explanations for phonetic, phonological and phonotactic universals have been around for a long time. The idea of language as a self-organising system is also not new. However, combining these two approaches and investigating them with computer simulations *is* new. It has been shown that in this way sound systems can emerge that show universal characteristics, and that could be described by distinctive features, although the individual language users do not use (universal) constraints or distinctive features. Although it is a bit early to draw conclusions, this indicates that innate distinctive features and innate constraints are unnecessary in phonological theory.

An interesting question is what the implications of these observations are for other aspects of language and mainly for syntax. The physical character of speech sounds makes it obvious and intuitive to look for functional explanations of universals. The research described here, as well as the research of many others (Berrah et al.

1996, Boë et al. 1995, Butt 1994, Carré et al. 1995, Carré & Mody, 1997, Glotin 1995, Glotin 1996, Liljencrants & Lindblom 1972, Lindblom et al. 1984, Lindblom & Maddieson 1988, Lindblom 1992, Stevens 1972, Stevens & Blumstein 1975, Stevens 1989, Valleé 1994) has shown that functional explanations are indeed valid. However, if self-organisation is taken into account as well, the functional explanations become even more compelling, as self-organisation can make that even very small preferences in any direction will be amplified over the generations of language users. This makes one wonder whether functional explanations and self-organisation are not applicable to many other aspects of language as well.

8 Acknowledgements

This research has been done at the artificial intelligence laboratory of the Vrije Universiteit Brussel. It is part of the 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 during visits to the Sony computer science laboratory in Paris, France. I am grateful to Luc Steels for valuable discussion of the ideas and the work presented here, as well as for providing the research environment of the VUB AI-lab and the hospitality of the Sony computer science laboratory.

9 References

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. Browman, Catherine P. and Louis Goldstein (1995) Dynamics and Articulatory Phonology, in: robert F. Port and Timothy van Gelder (eds.) *Mind as Motion*, MIT Press, Cambridge Mass. pp. 175–194.
4. Butt, Matthias (1992), Sonority and the Explanation of Syllable Structure, in: *Linguistische Berichte* **137**, pp. 45–67.
5. Carré, René, Marc Bordeau and Jean-Pierre Tubach (1995) Vowel-Vowel Production: The Distinctive Region Model (DRM) and Vowel Harmony, *Phonetica* **52**, pp. 205–214
6. Carré, René & Mody, Maria (1997) Prediction of Vowel and Consonant Place of Articulation, in: John Coleman (ed.) *Computational Phonology*, Third Meeting of the ACL Special Interest Group in Computational Phonology, pp. 26–32.
7. Chomsky, Noam and Morris Halle (1968) *The sound pattern of English*, MIT Press, Cambridge, Mass.
8. Crothers, John (1978) Typology and Universals of Vowel Systems, in: Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.) Universals of Human Language, vol2. pp. 93–152.
9. de Boer, Bart (1997a) Generating Vowels in a Population of Agents, in: Phil Husbands and Inman Harvey (eds.) *Fourth European Conference on Artificial Life*, MIT Press, pp. 503–510
10. de Boer, Bart (1997b) Self Organisation in Vowel Systems through Imitation, in: John Coleman (ed.) *Computational Phonology, Third Meeting of the ACL Special Interest Group in Computational Phonology*, 12 July 1997, pp. 19–25
11. de Boer, Bart (1997c) *Learning vowels using self-organization*, Vrije Universiteit Brussel-AI-lab AI-memo 97-20.
12. Elman, Jeffrey L. (1990) Finding Structure in Time, in: *Cognitive Science* 14, pp. 179–211.
13. Glotin, Hervé, (1995) *La Vie Artificielle d'une société de robots parlants: émergence et changement du code phonétique.* DEA sciences cognitives-Institut National Polytechnique de Grenoble
14. Glotin, Hervé, Rafael Laboissière(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
15. Haykin, Simon (1994) *Neural Networks, A comprehensive foundation*, New York: MacMillan.
16. Jakobson, Roman and Morris Halle (1956) *Fundamentals of Language*, the Hague: Mouton & Co.
17. Kaburagi, Tokihiko & Honda, Masaaki (1996) A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes, in: *Journal of the Acoustical Society of America* **99**(5)
18. Ladefoged, Peter and Ian Maddieson (1996) *The Sounds of the World's Languages*, Oxford: Blackwell.
19. Langton, Christopher G.(ed.) (1989) *Artificial Life*, Addison Wesley
20. Liljencrants, L. and Björn Lindblom (1972) Numerical simulations of vowel quality systems: The role of perceptual contrast, *Language* **48** pp. 839–862.

21. Lindblom, Björn, Peter MacNeilage and Michael Studdert-Kennedy (1984), Self-organizing processes and the explanation of language universals, in: Brian Butterworth, Bernard Comrie and Östen Dahl (eds.) *Explanations for language universals*, Walter de Gruyter & Co. pp. 181–203
22. 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.
23. Lindblom, Björn (1992) Phonological Units as Adaptive Emergents of Lexical Development, in: Charles A. Ferguson, Lise Menn and Carol Stoel-Gammon, *Phonological Development*, pp. 131–163
24. Maddieson, Ian,(1984) *Patterns of sounds*, Cambridge University Press.
25. Maeda, Shinji (1989) Compensatrory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal Tract Shapes using an Articulatory Model, in: W.J. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modelling*, Kluwer, pp. 131–149
26. Mantakas, M, Schwartz, J.L. & Escudier, P. (1986) *Modèle de prédiction du 'deuxiéme formant effectif' F_2'—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, pp. 157–161.
27. Mermelstein, P. (1973) Articulatory model for the study of speech production, *The Journal of the Acoustical Society of America*, **53**(4) pp. 1070–1082
28. Press, William H, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery*, Numerical Recipes in C, second edition*, Cambridge University Press, 1995
29. Rabiner, L.R. and R.W. Schafer, (1978) *Digital Processing of Speech Signals*, Prentice-Hall.
30. Rubin, Philip & Baer, Thomas (1981) An articulatory synthesizer for perceptual research, in: *Journal of the Acoustical Society of America*, **70**(2) pp. 321–328
31. Schwartz, Jean-Luc, Boë, Louis-Jean, Vallée, Nathalie and Abry, Christian (1997), Major trends in vowel system inventories, *Journal of Phonetics* **25**, pp. 233–253
32. Steels, Luc (1995) A Self-Organizing Spatial Vocabulary, *Artificial Life* **2**(3), pp. 319–332.
33. Steels, Luc (1996) The Spontaneous Self-organization of an Adaptive Language, in: S. Muggleton (ed.) *Machine Intelligence* **15**.
34. Steels, Luc (1997) Synthesising the origins of language and meaning using co-evolution, self-organisation and level formation, in: J. Hurford (ed.) *Evolution of Human Language*, Edinburgh: Edinburgh University Press.
35. Stevens, K. N. (1972). The Quantal Nature of Speech: Evidence from articulatory-acoustic data. In: E. E. David, Jr. & P. B. Denes (Eds.) Human communication: a unified view. New York: McGraw-Hill.
36. Stevens, Kenneth N. and Sheila E. Blumstein (1975), Quantal aspects of consonant production and perception, *Journal of Phonetics* **3**, pp. 215–233.
37. Stevens, Kenneth N. (1989) On the quantal nature of speech, *Journal of Phonetics* **17**, 1, pp. 3–45
38. Trubetzkoy, N.S. (1939) *Grundzüge der Phonologie*, Travaux du cercle linguistique de Prague 7.
39. 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)
40. Vennemann, Theo (1988) *Preference Laws for Syllable Structure*, Berlin: Mouton de Gruyter.
41. Vihman, M. (ed.) (1976) *A reference manual and user's guide for the Stanford Phonology Archive*, Part 1, Stanford University.
42. Vihman, Marilyn May (1996) *Phonological development: the origins of language in the child.* Cambridge: (MS) Blackwell.