Emergent CV-Syllables

Bart de Boer

AI-lab

Vrije Universiteit Brussel

Abstract

This report describes a system in which a population of agents develops a set of syllables for imitating each other. Initially, the agents do not have a repertoire of sounds. However, they *are* able to produce a range of human like syllables that consist of an onset (consonant) and a nucleus (vowel). Through a pressure to imitate each other successfully and a pressure to increase the number of sounds they are using, the agents develop a consistent set of syllables.

The agents imitate each other in encounters that are called imitation games. In these imitation games one agent utters a sound, which another agent then tries to imitate. Participants of these imitation games are chosen randomly from the population. The imitation games are iterated a large number of times.

The resulting syllable systems are found to be coded phonetically, instead of holistically. This means that a much smaller number of onsets and nuclei are used than there are syllables. The syllables can thus be analysed as consisting of a much smaller set of phonemes. However, these phonemes are not the ones one would expect in human language. This is most probably due to the settings of the many articulatory and perceptual parameters in the system that cause it to be insufficiently like the human articulatory and perceptual systems.

1. Introduction

All human speech is organised in syllables. It is very hard to define in terms of articulatory and acoustic constraints what exactly a syllable is [9, p. 282, 10, p. 113]. A prototypical syllable, however, will start with a closure of the vocal tract, which is released at the onset of the syllable. The release will be followed by a part with a higher acoustic energy and a longer duration. After this the vocal tract can again be closed. We will call the initial opening gesture the *onset* of a syllable, the part with high acoustic energy the *nucleus* and the final closing gesture the *coda*.

The total number of syllables found in any language in particular varies wildly. Maddieson [14, p.22] gives numbers of possible syllables varying from 162 for Hawaiian to 23,638 for Thai. However, in all languages of the world, it is found that the possible sounds that make up the onset, nucleus and coda of the syllables are quite limited. Possible onsets and finals are usually called *consonants* and possible (prototypical) nuclei are usually called *vowels*. Consonants and vowels are the basic building blocks of syllables. For completeness it must be mentioned that in a lot of languages consonants and vowels can also appear in clusters. Also certain consonants (sonorants and fricatives) can appear as syllable nuclei. In this report however, we will concentrate on the simplest possible syllables: sequences of a single consonant followed by a single vowel, CV-syllables in shorthand form. All languages of the world have syllables of this type.

We can now ask the question: "Why do languages have only such a limited number of building blocks for their syllables?" Usually this question is avoided by postulating consonants and vowels, or at least the distinctive features that describe them as innately present in the brain of the language learner [2,8]. However, as Lindblom [12] has pointed out, this only begs the question, as it remains to explain how they became innate and how they originated in the history of language. Lindblom [12,13] gives a satisfactory explanation of why syllables are organised in terms of a *limited number* of consonants and vowels. He says that construction of syllables from a limited number of consonants and vowels is an emergent property of communication under articulatory and acoustical constraints. His computational experiments consisted of picking the "best" syllable (according to a number of articulatory and acoustic constraints) from a large inventory of possible CV-syllables. Although this could in principle result in a *holistic* coding, in which every syllable contains a unique onset and nucleus, in fact it resulted in syllables that were constructed from a small set of onsets and nuclei. This type of coding is called *phonemic* coding. Of course Linblom still assumes that syllables consist of a closure (consonant) followed by a release (vowel), and not the other way around, but this is a much weaker assumption than the assumption that phonemic coding is innate. Lindblom's experiment is discussed in more detail in appendix C.

Although Lindblom's approach gives a good explanation of why syllables can be analysed in terms of phonemes, it does not explain how the process of picking the best syllables from the inventory of possible syllables is implemented in a (human) language community. In human language communities people communicate and children learn language under the same acoustic and articulatory constraints.

However, nobody picks the syllables of his or her language in order to optimise articulatory constraints such as in Lindblom's system. When children learn a language, they just imitate other speakers as closely as possible under acoustical, articulatory and cognitive constraints.
In the research described in this report, it has been tested whether phonemic coding of syllables can emerge in a population of agents that try to imitate each other under acoustic and articulatory constraints. Each agent is capable of producing and perceiving CV-syllables. Together they have to develop a successful set of syllables with which they can imitate each other.
The simulation is based on Steels' [15] ideas on the origins of language. According to Steels, language originates as a self-organising system in populations of communicating agents that have sufficient cognitive capacities. Agents engage in local interactions that he calls *language games*, in which two agents encounter each other, exchange certain linguistic and non-linguistic information, and update their local linguistic knowledge depending on the outcome of the language game.
The rest of this report is organised as follows: in the next section the way consonants are modelled in the simulation is described. In section 3, the implementation of syllables is discussed. In section 4, the particular language game, called *imitation game*, used in the simulations is described. Some preliminary experiments are described in section 5 and these results are discussed in section 6.

2. The Implementation of Consonants

In order to keep the simulation simple, it has been decided to study only simple plosive consonants followed by a simple vowel. This means that nasals, fricatives, liquids etc. can not be produced or accurately perceived by the simulated agents. Also the agents cannot make the difference between voiced and unvoiced consonants. Although this already severely limits the number of possible consonants available, a further restriction had to be imposed by partitioning the articulatory space into discrete consonants.

The production of consonants

Originally, the intention was to find an articulatory model that could produce acoustic descriptions of consonants from articulatory parameters, similar to the articulatory model that has been used in the research into vowel systems. However, as we will see shortly, no simple acoustic correlates with place of articulation could be found. Therefore it was decided to adapt Lindblom et al's [12,13] solution of partitioning the articulatory space into a more or less arbitrary number of discrete consonants. If one takes this number large enough, nearly all possibilities of the human vocal tract will be covered. In this work, an arbitrary number of ten consonants were chosen to represent the articulatory possibilities of the human vocal tract. These consonants were: uvular [q], velar [k], labio-velar [k͡p], palatal [c], retroflex [ʈ], alveolar [t], dental [t̪], linguo-labial [t̼], labio-dental [p̪] and bilabial [p]. Although symbols for unvoiced consonants are used, in fact the voicing of these consonants is unspecified. Note that three more consonants are used here than in the work of Lindblom et al. [12,13]. These are the linguo-labial, the labio-dental and the labio-velar plosives. These were included for filling the articulatory range between dental and bilabial, as Lindblom et al.'s research has no consonants in that range.

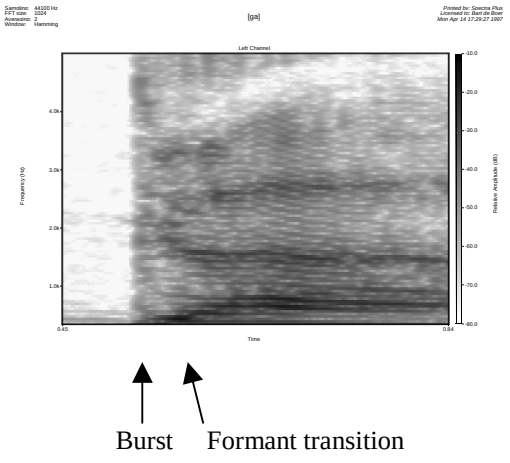

Figure 1: Burst and formant transition in the syllable [kɐ]

The acoustic representation of consonants was inspired by the representation that Lindblom et al. [12,13] have used. In addition, research by different researchers [3,5,6,7,11,16], into the perception of consonants was used, as well as a substantial amount of own observations. It is assumed that two cues are important for the recognition of plosive consonants: a burst of noise that is generated when the closure of the consonant is released, and the transitions of the formant frequencies of the following vowel as the articulators move towards the position of that vowel. The formant pattern at which the transition starts is usually called the *locus* of the consonant. This is illustrated in figure 1.
The values of these parameters for the transition between the above mentioned consonants

and the vowel [ɐ] were measured and compared with the literature. They are presented in table 1. The measurements were performed as follows. First a spectrogram was made from a recording of a sequence of one of the consonants followed by the vowel. Then the parameters were estimated from this spectrogram. These values were used to synthesize the syllable using a simple formant synthesizer (described in more detail in appendix A). If the result was not sufficiently recognizable, the estimates were adjusted. Where available, data from the literature [3,5,6,7,11,12,16] were used for determining the proper values of the parameters. Unfortunately, the data in the table are not always reliable. They have to be verified before further experiments can be undertaken.

	[q]	[k]	[k͡p]	[c]	[t̠]	[t]	[t̪]	[t̠]	[p̪]	[p]	[ɐ]
Burst:	1500	1500	1000	2500	3000	3500	4500	1000	2700	500	
F1:	150	150	150	150	150	150	150	150	150	150	655
F2:	1150	1650	1000	2200	1750	1700	2050	1500	1000	1060	1185
F3:	3000	1950	2300	2500	2250	2700	3080	2500	2000	2270	2585
F4:	3600	2920	3100	3600	2920	3300	3600	3200	3200	3080	3600
position	1	1	1	0	0.5	0.5	0.5	0	0	0	
Extreme?	yes	no	yes	yes	yes	no	no	yes	yes	no	

Table 1: Values of acoustic parameters of different consonants before [ɐ].

Lindblom et al. [12,13] only use the formant transitions (loci) and not the bursts for determining the identity of consonants in their experiments. Although they have succeeded in generating very natural systems of syllables, it seems nevertheless more natural to include information about the burst as well, as bursts happen to be a very salient feature for determining the identity of a consonant. In the research that will be presented in this report, it was decided to use both the formant transitions and the burst.
However, there are still a number of flaws in the data used here. First of all, the properties of consonants in different contexts can differ quite radically. The context (i.e. the following vowel) of a consonant should therefore be taken into account. This has not been done in the present research, but will be done in future research. In addition, the possibility should be considered that there are more variables determining the articulation of consonants. Possibly secondary articulations, such as velarization, labialization and palatalization should be taken into account. All these parameters determine the shape of the mouth and hence determine the sound that will be produced. It might be possible that by taking into account these extra degrees of freedom it will be possible to build a continuous model of consonant articulation. More recent measurements of consonants will be presented in appendix B.
Apart from acoustical properties, consonants also have articulatory properties. Following Lindblom et al. [12,13] the articulations of consonants can be extreme or not extreme. Extreme articulations are articulations where extra effort is needed to perform the articulation. These can be articulations where the articulators have to be moved into a position far away from their resting position, such as [q], [t̠] or [t̪], articulations that are not very stable and easily turn into fricatives, such as [c] and [p̪] or articulations that require complex co-ordination of multiple articulators, such as [k͡p].
The other articulatory parameter of consonants is their inherent position. In this research it is assumed that, because of the shape of the tongue during the articulation of the closure, it is easier to go from a consonant to a vowel with a certain position than to a vowel with a different position. This easiest to reach position is called the inherent position of a consonant. Table 1 contains the values of the inherent positions of the consonants that have been used in the simulations.

The Perception of Consonants

Although syllables are supposed to be perceived as a unit, the implementation still uses a separate comparison for consonants and vowels. Two consonants are compared by calculating a distance between them. The distance *D* is basically the weighted Euclidean distance between the acoustic representations of the consonants and is calculated as follows:

1) $$D_C = \sqrt{(f_b - f_b')^2 + \alpha(f_2 - f_2')^2 + \beta(f_3 - f_3')^2 + \gamma(f_4 - f_4')^2}$$

Where f_b and f_b' are the centre frequencies of the bursts of the two consonants and f_2-f_4 and f_2'-f_4' are the formant frequencies of their loci. All frequencies are in Bark. The constants α,β and γ are weighting factors that determine which formant frequencies are most important for the perception of the consonants. For the present experiments only α was one, the others were zero.

The Bark scale was chosen, because it is a logarithmic frequency scale that is designed to model human perception of sounds accurately. In this paper the conversion between Hertz and Bark is calculated as follows:

$$2) \quad B = \begin{cases} \dfrac{F - 51}{110} & \text{if } F < 271.32 \\ \dfrac{\ln(F/271.32)}{0.1719} + 2 & \text{if } F \geq 271.32 \end{cases}$$

Where B is the frequency in Bark and F is the frequency in Hertz. Note that for lower frequencies the conversion is linear instead of logarithmic. As most formant frequencies are higher than the threshold anyway, this does not influence the perception very much.

The perception of consonants is thus based on the centre frequency of their noise burst and on the starting value of their formant frequencies (their locus). The position of the locus is assumed to be independent of the following vowel, which is probably unrealistic. No attention is paid to the temporal structure of the noise burst or the formant transitions. This might be unrealistic as well, as it appears that the duration of the noise burst and possibly of the formant transitions also affects the perception of the consonants by human subjects.

3. The Implementation of Syllables

In real speech syllables are extremely hard to define in acoustical or articulatory terms. Usually, a typical syllable will start with an onset with low acoustic energy, will have a nucleus with low acoustic energy and will again have a coda with low energy. Ideally, a sequence of syllables would therefore consist of alternating peaks of low and high acoustic energy. Unfortunately, things are not that simple. [9: p. 282, 10: p. 113] Several complicating factors exist. One of these factors is stress. Some syllables are stressed and others are not. The nuclei of unstressed syllables can in principle have as little acoustic energy as onsets or codas of stressed syllables. Another complicating factor is the definition of acoustic energy. Calculation of acoustic energy might seem straightforward at first, but in fact the real acoustic energy of a sound is not quite the same as the acoustic energy of a signal as perceived by human observers. A last difficulty is how to distinguish between the coda of a syllable and the onset of a following syllable.

For these reasons, the segmentation of a continuous speech signal into syllables is very hard to perform automatically. It is therefore avoided in the research that is described in this report. In this report we will assume that syllables are spoken in isolation and that segmentation into nucleus and onset has already taken place. Syllables are described in terms of the acoustic properties of the onset (burst and formant transitions, as described in the previous section) and those of the nucleus (formant frequencies, as described in more detail in [4]). Temporal factors are completely ignored. To what extent these assumptions are unrealistic remains to be investigated, although it does not seem unlikely that the required parameters can reliably be derived from CV-syllables spoken in isolation.

In the system that is described in this report, syllables are implemented as a pair (*C*,*V*) consisting of an onset *C* (one of the ten possible consonants) and a nucleus *V* (a vowel described with the three continuous parameters position, height and rounding). When it is necessary to say a syllable, these parameters are converted to an acoustic description, consisting of the frequency of the burst and locus of the consonant and the first four formant frequencies of the vowel. The agents use this acoustic description to communicate with each other. It must be noted that each agent only has a list of syllables. There is no separate list of consonants (except, of course, for the global list of possible consonants) nor is there a separate list of vowels.

The syllables are perceived in a way that is similar to the way that either consonants or vowels are perceived separately. In order to analyse a syllable that is heard in terms of the syllables that are already available to the agent, it looks for the syllable that is closest to the syllable it just heard. The distance *D* between two syllables is calculated as a weighted average of the distance D_C between the two consonants and the distance D_V between the two vowels, and is calculated as follows:

$$3) \quad D = \frac{\zeta D_V + \xi D_C}{\zeta + \xi}$$

Where ζ and ξ are constants which in the current simulation are both set to 1.

4. The Imitation Game

The agents engage in imitation games in order to learn each other's speech sounds. The imitation game is quite similar in structure to the imitation games that were played by the agents in the vowel system

simulations [4]. However, the details of the actions that are undertaken by the agents to improve the quality of their syllable inventories are quite different.

Two agents are randomly selected from the population of agents. One of the agents then selects a syllable from its repertoire of syllables. If the repertoire happens to be empty, a new syllable is randomly generated. It produces an acoustic representation of this syllable and transmits it to the other agent. This agent then finds in its repertoire of syllables the closest match to the sound it just heard. If its list of syllables is empty, the agent creates a syllable that is a close imitation of the syllable it just heard using a process that will be described below. The imitating agent in its turn produces an acoustic representation of the syllable it selected and transmits it to the agent that initiated the imitation game.

This agent also analyses the sound in terms of its own syllables by finding the closest match to the sound it just heard. If this closest match is the same syllable as it originally produced, the imitation game is considered to be successful. If the match is not the same syllable, the imitation game is considered to be unsuccessful.

The reaction of the agents to the language game is also similar to the reaction of the agents in the vowel system experiments. Both agents increase the use count of the syllables they produced. Also they both increase the success count of the syllables if the language game has been successful. Furthermore, the imitating agent moves the syllable it just said closer to the sound it just heard if the imitation game was successful. If the game was successful and the score of the syllable it used was low, the syllable is also moved closer. However, if the score of the syllable is high and the game was unsuccessful, a new syllable that is close to the sound that was heard is generated.

Also syllables that are too similar are merged and syllables with a success/use ratio that is too low are discarded. The limit of this ratio below which syllables are discarded was 0.7 in the simulations that will be presented here. Two syllables are judged to be too similar if the noise that is added to the formant frequencies of the consonants and vowels could confuse them too much. This is implemented as a closest allowable range of:

4) $$r = \frac{\ln(1+\sigma) - \ln(1-\sigma)}{0.1719 \cdot 2},$$

where σ is the standard deviation of the percentage of noise that is added to the formants. The constant 0.1719 is related to the way frequencies in Hertz are converted into frequencies in Bark (the noise is added to the formant frequencies in Hertz). If the distance *D*, as calculated in equation 1) between two syllables in one agent becomes smaller than this limit, the two syllables are merged.

Another merging process is also taking place in the agents. In preliminary experiments it was observed that syllables in one agent tended not to share the same nuclei (vowels). This was caused by the fact that although there was pressure between the agents to move matching syllables closer together, there was no such pressure in the syllable system of one agent. Agents thus tended to develop systems of successful syllables with highly diverse nuclei. It was concluded that an extra pressure is needed to ensure that nuclei *within* an agent also cluster, as this is the situation in human syllable systems. This would point to the fact that humans perceive onsets and nuclei of syllables as separate categories, over which can be generalised, independent of the syllable in which they occur.

The merging of nuclei is done before the merging of the complete syllables. Nuclei are also merged if the distance between them becomes smaller than the limit *r*. This process of merging of nuclei ensures that the same process of generalisation of nuclei that takes place in humans takes place in the agents.

New syllables can be generated by three different processes. The first generates a random syllable. This is only done if the syllable repertoire of an agent who is starting an imitation game is empty. The second process generates a new syllable in an agent that already has a repertoire of syllables. In order to decrease the probability that the syllable will be merged directly with already existing syllables, a number of random syllables is generated and the one that has the highest sum of all the distances to the syllables that are already present, is selected. This process is invoked with a very low probability (one-percent).

The third, and most important process that adds new syllables to the repertoire of an agent is active when a syllable needs to be added as a reaction to a failed imitation game. The added syllable has to be similar to the sound that was produced by the other agent. In addition, it has to be as simple as possible in articulatory terms. The syllable that will be added will thus be a compromise between acoustical similarity and articulatory ease.

This compromise is reached by a hill-climbing procedure that tries to find the syllable with the highest possible quality. The hill-climbing procedure starts with a random syllable of which it calculates the quality. It then tests different neighbouring syllables and selects the one with the highest quality. If no neighbouring syllable with a higher quality can be found, the procedure terminates. As long as the search space is bounded and the number of neighbouring syllables that is tested is finite, the procedure is guaranteed to terminate.

The quality of a syllable is calculated as the quotient of the articulatory effort needed to produce it and of its acoustic distance to the target sound. The acoustic distance is as defined in equation 3). The articulatory effort is calculated as follows:

$$5) \quad E = \begin{cases} 5 + \vartheta|p_v - p_c| + \mu|h_v - 0.5|, & \text{if consonant not extreme} \\ 11 + \vartheta|p_v - p_c| + \mu|h_v - 0.5|, & \text{if consonant extreme} \end{cases}$$

Where p_v and p_c are the positions associated with the vowel and the consonant, respectively, and h_v is the height of the vowel. The constants 5 and 11 have been chosen to tune the influence of the extremeness of the consonant. The constant 0.5 is supposed to be the neutral vowel height, i.e. for reaching this height no muscular activity is needed. The constants ϑ and μ are weights for the influences of the position and height of the vowel on the articulatory complexity of the syllable. They are both set to one in the simulations that will be presented here.

The quality of a syllable Q is now the quotient of the articulatory effort E and the acoustical distance D to the sound that has to be imitated:

$$6) \quad Q = \frac{E}{D}$$

The imitation games with their modifications of the syllable inventories of the agents are iterated. This is supposed to lead to coherent inventories of natural-looking syllables in the population of agents.

5. Preliminary Results

A number of preliminary experiments have already been performed with the system. These simulations have resulted in systems consisting of syllables that are made up from a smaller number of onsets and nuclei. An example of an acoustic representation of such a system is given in figure 2. In this figure the onsets (consonants) are represented by the frequency (in Bark) of the second formant of their locus and by the frequency of their burst. Nuclei (vowels) are represented by the frequency of their first formant and the weighted average of their second to fourth formants. A line joins the onset and nucleus of every vowel. Note that the scales for the onsets and the nuclei are different.

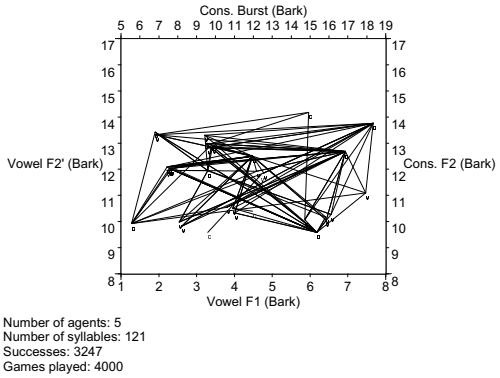

Figure 2: Syllable systems of five agents after 4000 imittion games.

From the figure it is clear that syllables are not picked randomly from the possible syllable space. It is clear that some onsets (marked with a c in the figure) are used more often than others. Also it is clear that the possible nuclei tend to cluster. Furthermore it can be observed that most clusters of nuclei and most onset have links with more than one other cluster. This means that the possible sounds in the agents are used more than once. This indicates that syllables are coded phonemically instead of holistically.

This is also the case in human languages. However, the similarities are mainly qualitative. The actual positions and frequencies of the onsets and nuclei of syllables in human languages can be quite different. This could be due to the fact that the articulatory and acoustic constraints of the agents are different from the ones of human production and perception of speech. The system is quite sensitive to changes in parameters. If one increases the relative importance of the articulatory efficiency, hardly any onsets with extreme articulations will be generated. If one eases this constraint extreme articulations will be preferred, as these are more acoustically distinctive. This is comparable to the results Lindblom obtained from his simulations. He writes: “When simulations are rerun removing articulatory factors…onsets such as [d̻] and [ɢ] are favored” [13, p.144].

In order to illustrate the generated syllable inventories more fully, the complete inventories of the agents are presented in table 2. In the table it can be seen that some onsets and some nuclei are used much more often and with much more success than others are. The table also illustrates nicely that the agent's syllable systems can only be compared in a qualitative way with those of human languages. The agents in the simulation very productively use some onsets and nuclei that are quite rare in human languages. This probably has to do with the settings of the parameters.

agent 1			[kɨ]				[ko]	[kɯ]							[ci]	[cø]	
agent 2		[ku]	[kɨ]	[kɑ]	[ki]		[ko]		[kø]	[kɔ]				[k͡pœ]			
agent 3		[ku]	[kɨ]	[kɑ]	[ki]	[kɐ]	[ko]										
agent 4	[qa]	[ku]	[kɨ]	[kɑ]			[ko]				[ke]				[ci]		
agent 5		[ku]		[kɑ]			[ko]					[kʉ]	[kœ]				

agent 1	[t̪ɨ]	[t̪ø]	[t̪ɵ]	[t̪ɑ]	[t̪o]	[t̪ɯ]						[t̪ɵ]	[t̪ɑ]		[t̪ø]		
agent 2	[t̪ɨ]		[t̪ɵ]		[t̪o]		[t̪u]	[t̪ɔ]				[t̪ɵ]	[t̪ɑ]	[t̪ɯ]	[t̪ø]		
agent 3	[t̪ɨ]	[t̪ø]			[t̪o]		[t̪u]		[t̪ɐ]			[t̪ɵ]	[t̪ɑ]		[t̪ø]		
agent 4	[t̪ɨ]			[t̪ɑ]	[t̪o]		[t̪u]	[t̪ɔ]				[t̪ɵ]	[t̪ɒ]			[t̪ɨ]	
agent 5		[t̪ø]			[t̪o]		[t̪u]	[t̪ɔ]		[t̪ʉ]	[tɶ]	[t̪ɵ]	[t̪ɑ]				[t̪ʉ]

agent 1		[t̪ø]			[t̪ɑ]				[p̪ɨ]	[p̪o]	[p̪ø]	[p̪i]	[p̪ɒ]				
agent 2	[tɔ]	[t̪ø]	[t̪i]	[t̪o]						[p̪o]	[p̪ø]	[p̪i]		[p̪ʉ]	[p̪ɐ]		
agent 3	[tɔ]		[t̪i]						[p̪ɨ]	[p̪o]	[p̪ø]				[p̪ɐ]	[p̪ɔ]	
agent 4						[tʌ]	[t̪a]		[p̪ɨ]	[p̪o]						[p̪ɔ]	[p̪e]	[pu]
agent 5			[t̪i]					[t̪ɒ]		[p̪o]	[p̪ø]	[p̪i]		[p̪ʉ]	[p̪ɐ]		

agent 1	[p̪ɨ]			[p̪ɐ]		
agent 2					[p̪ʉ]	
agent 3	[p̪ɨ]					
agent 4	[p̪ɨ]	[pa]	[po]			
agent 5				[p̪ɐ]	[p̪ʉ]	[pi]

Table 2: Syllables of a population of five agents after 4000 imitation games.

Furthermore, we can see from the table that agents do not always use exactly the same sounds for syllables that they perceive as similar. For example, the agents two to five have a syllable with phonetic representation [ku], whereas agent one has a syllable with phonetic representation [kɯ]. Still these syllables are perceived as similar. More syllables in the table that appear to be isolated can be analysed as idiolectal variations of syllables that are pronounced slightly differently by other agents. Of course there are also real isolated syllables in the agents' repertoires. These are unsuccessful syllables that were created by agents, but that never caught on in the populations, and that are waiting to be removed from the agents' repertoires. An example of such a syllable is [t̪a] in agent four, which has only been successful in one of five imitation games.

The onsets that seem to be used in a regular and successful way are: [k], [t], [t̪] (which might be analyzed as an allophone of [t]) and [p̪]. Other onsets that are used are [q], [c], [k͡p], [t̪] and [p]. However, the first three of these are so rare that they can be considered as accidental outcomes of the random creation of syllables process. The last two are less rare, but they do not have reached consistency throughout the population of agents, yet. It is therefore unclear whether they are idiolectal variants, allophones or just consequences of the random nature of the syllable creation process. The nuclei that appear most consistently are [u] with idiolectal allophone [ɯ], [i] with idiolectal allophone [ʉ], and [o], [ɑ], [ø] and [ɵ].

6. Discussion of Results

The results of the simulations presented here are at once realistic and unrealistic. They are realistic in the sense that the resulting syllable systems seem to consist of phonemically coded syllables, instead of holistically coded ones. However, the actual phonemes that can be analysed from the syllables are not the ones one would expect in human sound systems. Also the coding seems to be more random and less parsimonious than in human sound systems.

The first problem is probably caused by the settings of the parameters in the system and by the way consonants are coded in the system. If we take a look at the way distances between consonants are calculated in the system, we see that some consonants are close to each other in our system whereas they would not be perceived as close to each other by humans. This is illustrated in figure 3. Here the frequency of the burst and of the first formant of the locus of every consonant in the system are plotted. The distance between two consonants is now approximately proportional to the distance between the

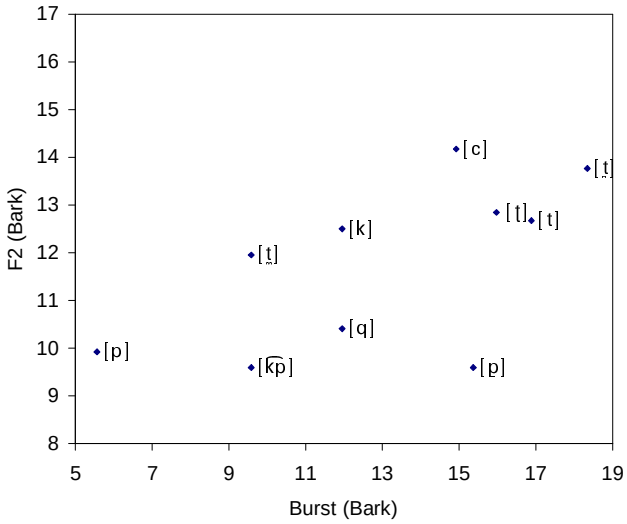

Figure 2: Values of the burst frequencies and second formant frequencies of consonants.

positions of these consonants in the figure. It is clear that the descriptions of consonants that were used in the simulation were not quite adequate. One way of solving this problem, better data, is presented in appendix B.

The second problem is more intimately related to the algorithm that was used. Apparently new syllables are created too quickly for the system to reach complete coherence. Also old useless syllables appear not to be thrown away often enough. This causes the agents' syllable repertoires to contain too many syllables that do not appear in any other agent. One way to solve this is by making the probability of adding new syllables much lower. Another way to solve it is to choose the syllables that are to be used in an imitation game in a way which is not random, but that makes use of the knowledge that the agent has about the performance of a syllable.

The system shares its sensitivity to parameter settings with Lindblom et al.'s [12,13] experiments. These experiments were duplicated as a preparation to the research described here. They are described in more detail in appendix C. In their system, too, the shape of the syllable systems that originated depended quite heavily on the representation of the consonants, on the way the distance between consonants was calculated and on the relative importance of articulatory ease versus acoustic distinctiveness. Furthermore, with certain settings of parameters, their system also generated lots of syllables that contained onsets or nuclei that occurred only in those syllables and not in others, effectively making the syllable systems less phonemically and more holistically coded.

One difference between Lindblom et al.'s system and the system presented in this report is that in the latter a special mechanism is needed to make the nuclei of syllables stay together. As was reported in section 4, nuclei of syllables have to be merged separately from the merging of entire syllables. Although it might be cognitively plausible to treat onsets and nuclei of syllables as separate entities, this does beg the question of the origin of the phonemic coding somewhat. On the other hand, a strong point can be made for the cognitive plausibility of merging sounds that are very close together. If the agents cannot reliably make the difference between two entities, they cannot continue to consider them as separate entities. However, the same system should also affect the onsets. This is unfortunately impossible with the discrete way in which they are presently encoded.

How plausible is the system from a cognitive point of view? Do children learn the syllables of their mother tongue in a way similar to the way agents in the system learn their syllables? In fact nothing happens in the agents that could not be done by human brains in principle, but the way in which it is implemented differs quite radically. Where agents are capable of creating a syllable that sounds like a syllable they never heard before, human children are probably not able to do that. They have to struggle much longer to make a proper imitation of certain sounds. Also, when past a certain age, humans can no longer learn new phonemes easily, whereas the agents in the simulation are always able to do that. This could be an important difference between the system presented here and human learning of syllables.

In the future a number of improvements has to be made to the system. First of all, the representation of consonants needs to be improved. At the moment a consonant always looks the same, independent of which vowel follows it. This is unlike the way consonants work in human languages. Already a number of measurements have been performed to improve the production of consonants and to make them dependent on which vowel follows them.

Another change that has to be made to the system is that it should also be able to handle syllables that consist of nothing but a vowel. In this way it can be investigated whether adding a consonant to a syllable improves the understandability of the syllable. Probably some study into the perception of syllables should be done before this can be implemented in a reasonably realistic way. Ultimately the system should be able to handle sequences of syllables. For this it would need to be able to segment the sequences into the constituent syllables. Because in human language syllables are not spoken in isolation,

any system that wants to predict possible syllables, should take into account that the syllables occur in a sequence of other syllables that influence each other.
The number of agents that is simulated has to be increased. At the moment the imitation games take place in a population of five agents. This number is too low. With more agents, syllables that are bad will probably be removed from the population faster and the syllables will probably also converge towards a common value better. It remains to be seen what size of populations can be handled without increasing the computational load too much.
The preliminary results presented in this report show that it is possible to obtain phonetically coded syllables in a population of agents that try to imitate each other under articulatory and acoustical constraints. Although the resulting phonemes are not quite like the ones found in human language, they do form a pseudo-optimal set for the agents themselves. A better tuning of parameters will probably make it possible to generate more humanlike syllable systems. This seems to support the theory (previously advocated by Lindblom and by Steels) that innate mechanisms play a smaller role in determining the shape of the syllable- and phoneme inventories of human languages than is usually assumed.
When the system has been made more realistic and the dynamics of the system with more agents are more fully understood, research can be done into the changes of the agents' syllable inventories over time, or even over space if the agents are given a spatial position. This can then be used as a powerful model to learn more about the dynamics of language change and variation, especially if sequences of syllables can also be handled successfully.

7. Acknowledgements

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

8. References

1. Allen, Jonathan, M. Sharon Hunnicutt and Dennis Klatt, (1987) From text to speech: The MITalk system, Cambridge University Press.
2. Chomsky, Noam and Morris Halle (1968) *The sound pattern of English*, MIT Press, Cambridge, Mass.
3. Cooper, Franklin S, Pierre C. Delattre, Alvin M. Liberman, John M. Borst and Louis J. Gerstman (1976), Some Experiments on the Perception of Synthetic Speech Sounds, in: D.B Fry (ed.) *Acoustic Phonetics*, Cambridge University Press, pp. 258–283
4. de Boer, Bart (1997) *A Second Report on Emergent Phonology*, Vrije Universiteit Brussel Artficial Intelligence Laboratory AI-memo 97-04.
5. Delattre, Pierre C., Alvin M. Liberman and Franklin S. Cooper (1976) Acoustic Loci and Transitional Cues for Consonants, in: D.B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press.
6. Fant, Gunnar (1973), *Speech Sounds and Features*, The MIT Press.
7. Halle, M, G.W. Hughes and J.-P. A. Radley (1976) Acoustic Properties of Stop Consonants, in: D.B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press.
8. Jakobson, Roman and Morris Halle (1956) *Fundamentals of Language*, the Hague: Mouton & Co.
9. Ladefoged, Peter and Ian Maddieson (1996) *The Sounds of the World's Languages*, Blackwell.
10. Laver, John (1994) *Principles of phonetics*, Cambridge University Press
11. Liberman, Alvin M. Pierre C. Delattre, Franklin S. Cooper and Louis J. Gerstman (1976) The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants, in: D.B. Fry (ed.) *Acoustic Phonetics*, Cambridge University Press.
12. 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
13. Lindblom, Björn (1992) Phonological Units as Adaptive Emergents of Lexical Development, in: Charles A. fergusion, Lise Menn and Carol Stoel-Gammon, *Phonological Development*, pp. 131–163
14. Maddieson, Ian,(1984) *Patterns of sounds*, Cambridge University Press.

15. 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.
16. Stevens, Kenneth N. and Sheila E. Blumstein (1975), Quantal aspects of consonant production and perception, *Journal of Phonetics* **3**, pp. 215–233.

Appendix A: The formant synthesiser

The syllables that are used by the agents consist of an onset and a nucleus. The nucleus is in fact a vowel and is characterised by four formant frequencies. These formant frequencies can be considered the poles of the transfer function of the vocal tract if the nucleus is articulated.
The onset, which is in fact a plosive consonant, is characterised by a noise burst and a locus pattern of formant frequencies. These formant frequencies are the poles of the transfer function of the vocal tract if it is in the position just after the release of the closure that is characteristic for the consonant in question. In reality this locus pattern is dependent on the vowel that follows the consonant as the vowel is co-articulated with the closure for the consonant. However, in the present system we always take the locus pattern for co-articulation with the vowel [ɐ].
A syllable is generated from the description of the onset and the nucleus. First the burst of noise is generated. The centre frequency of this noise is as specified for the consonant, the bandwidth of the noise is taken to be equal to one tenth of the centre frequency. The burst of noise has the same length (of 0.01s) and the same strength for every consonant. After the noise burst, there is a silence of 0.01s. Although the fixed length and strength are unrealistic, they do not influence the perception of consonants by human subjects very much.

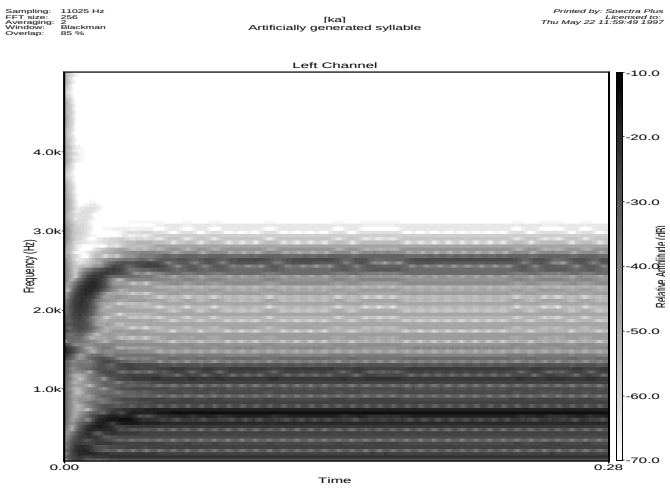

Figure 4: The spectrogram of an artificial syllable [kɐ]

The formant transition is generated by letting the formant values in the locus pattern of the onset move towards the formant values of the nucleus. In order to do this on the computer, the time needs to be divided in slices. In the present system time slices of 1 millisecond are used. In every time slice, a constant formant pattern is assumed. Between time slices, the formant pattern changes discontinuously. However, these changes are so small as not to influence the quality of the sound too much. From the formant values in every time slice, a digital filter is calculated that is excited by a simulated glottal pulse.
The formant frequencies in every time slice are calculated as follows:

7) $$F_{i,t+1} = (1-\alpha)F_{i,t} + \alpha F_{i,\infty}$$

Where $F_{i,t}$ is the value of formant i at time t, $F_{i,\infty}$ is the goal value of formant i (the value of this formant in the nucleus of the syllable) and α is a constant that determines the speed of the transition and that is set to 0.1 in the present system. The value of $F_{i,0}$ is set to the value of formant i in the locus of onset of the syllable.
The glottal pulse is calculated by low-pass filtering an impulse train. This impulse train has a fixed frequency with slight random fluctuations. The effects of radiation at the mouth are accounted for by taking the first difference of the signal that comes out of the vocal tract digital filter. These methods have been inspired by the Klatt synthesiser from the MITalk system [1].
The resulting sound is illustrated with a spectrogram in figure 4. This figure should be compared with figure 1. Figure 4 is the syllable that was artificially synthesised from the data of figure 1 with the method described above. The purpose of synthesising the syllables in the present system is to enable the researcher who runs the simulation to hear what is going on. However, in future applications the dynamic aspects that are introduced by actually synthesising the syllables could be useful for implementation of sequences of syllables that influence each other in a natural way. Furthermore, rendering and analysis of syllables is necessary if one wants to use the system for practical applications in the domains of speech synthesis and speech processing.

Appendix B: New data on consonants

Since the system was built, new measurements of the values of the bursts and onsets of consonants in different contexts have been made. These measurements have been made in the same way as the original measurements that were used in the simulations described above. First, different consonants in the context of the three vowels [i], [a] and [u] were recorded. Then these recordings were analysed for the centre frequency of the burst of the consonant, as well as for the first four formant frequencies of their locus pattern. These consonants were then synthesised in the appropriate context using the synthesiser described in appendix A. The parameters that describe the consonants were tuned until they were easily recognisable by a human subject (the author).

The consonants that were used were the same as the ones used in the simulations, except for [k͡p]. It was judged that double articulations should not be part of the system, as they require more complex articulatory processes than ordinary consonants. They could therefore not fairly be compared with other consonants that only have a single articulation.

	[ɢ]	[g]	[ɟ]	[d̠]	[d]	[d̪]	[d̻]	[b̪]	[b]	[i]
Burst:	1600	2000	3000	2500	2500	3000	2400	2100	1600	
F1:	300	300	300	250	300	250	250	200	200	310
F2:	1700	2100	2100	2000	2100	1500	1600	1600	1600	2300
F3:	2400	2700	3000	2400	2700	2200	2300	2300	2300	2800
F4:	3800	3400	3700	3000	3500	3500	3000	3400	3200	3600

Table 3: Consonants before [i].

	[ɢ]	[g]	[ɟ]	[d̠]	[d]	[d̪]	[d̻]	[b̪]	[b]	[a]
Burst:	1330	1800	2800	2500	3200	3800	2500	2800	1750*	
F1:	500	340	280	310	370	310	400	400	400	700
F2:	1200	1800	2000	1800	1800	1800	1600	1500	1500	1450
F3:	3000	2100	2100	2100	2600	2700	2600	2500	2400	2700
F4:	3500	3200	3300	3100	3600	3500	3300	3900	3500	3800

Table 4: Consonants before [a] (new values)

	[ɢ]	[g]	[ɟ]	[d̠]	[d]	[d̪]	[d̻]	[b̪]	[b]	[u]
Burst:	900	900	2300-1700	1500	2500	2500	1500	2500	2000	
F1:	350	250	250	350	400	400	400	400	400	250
F2:	900	900	1700	1200	1500	1900	1600	900	900	900
F3:	2800	1800	1900	2100	1700	2200	2200	2400	2400	2400
F4:	3700	3600	3500	2800	3700	3700	3200	3000	3300	3300

Table 5: Consonants before [u].

The results of the measurements are given in tables 3–5. It should be noted that the margin of error is quite large in these measurements, especially in the measurement of the first formant and in the measurement of the centre frequencies of some of the bursts. However, the values that are presented are the ones that were perceived most clearly by a human subject. In addition, some processes that seem to play an important role in the perception of consonants, such as volume and length of the noise burst as well as the actual shape of the formant transitions, have been ignored. No attempts have been made so far to construct a model that can predict the formant values from articulatory parameters.

In order to learn more about the validity of the measurements, they can be compared with similar measurements that can be found in the literature. Unfortunately, in the literature [6,3] only measurements of the plosives of Swedish [b], [d] and [g] have been found. These values are shown in table 6.

	[gi]	[di]	[bi]	[gɑ]	[dɑ]	[bɑ]	[gu]	[du]	[bu]
Burst:	3000	3500	1500	1500	3500	500	500	3500	1500
F1:	200	200	200	200	200	200	200	200	200
F2:	2000	1850	1700	1600	1400	1000	1100	1600	900
F3:	3000	2600	2300	1900	2600	2400	2300	2400	2200
F4:	3500	3400	3200	3100	3300	3050	?	3200	3200

Table 6: Fant's data on onsets of CV-syllables and Cooper et al.'s data on bursts.

* Actually, it appears as if [pa] is best perceived without a burst.

From the table we can see that our measurements and the measurements of Fant and Cooper et al. diverge quite much in certain respects (especially on the burst frequencies). This undoubtedly has to do with the difficulties of performing these measurements, the differences in subjects (both their native language and the acoustical properties of their vocal tracts) and the different methodologies. As it stands, we can only conclude that further research on the acoustic properties of CV-syllables is needed.

Appendix C: Linbloms' simulations

Lindblom et al.'s experiments [12,13] were partly replicated as a preparation for the research presented in this report. Their experiment consisted of composing a syllable system that was near optimal under certain articulatory and acoustical constraints. This was done as follows: first the repertoire of syllables was initialised with a random syllable. Then the best syllable, out of 133 (seven consonants and nineteen vowels) that could be added to the existing syllable system was calculated and added. The calculating and adding of an optimal syllable was iterated until a system with 25 syllables was reached. This resulted in realistic looking, phonemically coded vowel systems. Also, it did not matter very much which syllable was taken as initial syllable, as the final syllable systems tended to be very similar, independent of the choice of the initial syllable.

The calculation of which syllable was the best to add was based on a minimisation of articulatory effort and a maximisation of acoustic distinctiveness. The acoustic distinctiveness was calculated as the sum of the acoustic differences between the new syllable and all the other syllables. The distance between two syllables was calculated by, roughly speaking, summing over time for every formant frequency the surface of the difference between the values of these formants in the two syllables. The distance was thus based on the formant transitions alone.

The articulatory effort was higher for syllables that required long movements of the articulators and for syllables that required extreme articulations. The first criterion made a syllable like [gi] harder to articulate than [ɟi]. The second criterion makes retroflex, uvular etc. articulations harder than more straightforward ones like alveolar, velar or bilabial ones.

In order to test the methods by which articulatory effort and acoustic distinctiveness were calculated in the system described in this report, we used them to calculate optimality of syllables within the framework of Lindblom et al. It was found that it was possible to make phonemically coded syllable systems that looked more or less human. The systems were a little less nicely coded than the systems that Lindblom et al. managed to generate. It was found, however, that by tuning the parameters the shape of the systems could be improved. In this way the settings of the parameters for the actual simulations were determined.

It is suspected that Lindblom et al.'s system is also quite sensitive to parameter settings. Unfortunately, they do not really report on the actual settings of their parameters, nor do they report on the precise data they have used for coding the onsets of their consonants. This makes it hard to assess whether their system is fundamentally better than the system presented here, or that it is just tuned more optimally.

An example of a system that was obtained by picking 25 near optimal syllables in this way is given in table 7. Here it can be seen that a definite phonemic coding has developed. Although the vowel space was continuous, the vowels are only represented by discrete symbols. Thus there were 24 possible vowel symbols and 10 possible consonants. Of these only four consonants (the ones without extreme articulations) and twelve vowels have been used. The vowels that occur in more than one syllable form a symmetric system that could appear in a natural language. Also, we could say that there is some kind of allophonic variation. For [k] and [p] there are two unrounded low vowels: [a] and [ɑ]. For the two apical articulations [t] and [t̪], there is only one: [ɐ]. This one vowel is actually a central vowel, which is predictable from the fact that for moving from an apical articulation towards a central vowel, a minimum of articulatory effort is needed.

	[i]	[ɨ]	[ɛ]	[a]	[ɐ]	[ɑ]	[ɔ]	[ʉ]	[u]	[ʌ]	[ə]	[o]
[k]	•	•		•		•			•	•		
[t]	•				•		•		•		•	
[t̪]	•	•	•		•		•	•	•			•
[p]	•		•	•		•	•	•				

Table 7: Results of Lindblom et al.'s method and the distance and effort functions from this report.

The system obtained here is similar to the system obtained in section 5, although there are less spurious syllables, but it is clear that imitation games can generate similar systems to optimisation algorithms.