Imitation games for complex utterances

Bart de Boer

AI-Lab

Vrije Universiteit Brussel

bartb@arti.vub.ac.be

Abstract

This paper presents a preliminary experiment in modelling the emergence of repertoires of complex words in a population of agents. Agents that can produce words, which are strings of (abstract) phonemes, try to imitate each other as well as possible. For this they do not only have to learn each other's words, but also each other's phonemes. But they cannot know a priori which sounds are actually used to distinguish words, and which sounds are just the result of the sound changes and co-articulation that are caused by the pronunciation rules. A simple learning heuristic is presented that can extract phonemes from words.

Experiments have shown that this imitation game can result in successful and expanding vocabularies of words that do show regularities in the ways sounds are combined. However, due to the simplicity of the simulations, the results do not yet have a large linguistic relevance.

1. Introduction

Can self-organisation explain the regularities that are found in the sound systems of human languages? All human languages contain vowels as well as consonants. Of both types of sounds there is an enormous variety, and an enormous amount of ways in which they can be combined. Still, there is reason to the madness. Some sounds appear more often than others, repertoires of sounds in sound systems of languages tend to be symmetrical and combinations of sounds in words are restricted. For vowels in isolation it has been shown in previous papers (de Boer 1999; de Boer & Vogt 1999) that self-organisation in a population under (mostly) pressure of acoustic distinctiveness could predict the vowel systems that occur in human languages. However, vowels in isolation are not quite representative of the kind of linguistic utterances that are generally used. The question that surfaces is then whether and how self-organisation can also explain other universal tendencies of human sound systems.

But let us first explore some of the regularities that are found in systems of more complex speech sounds. These regularities can be divided into two types: those of repertoire and those of combination. The regularities of repertoire are comparable to the ones that were investigated in vowel systems, but more complicated. The diversity of possible consonant sounds in human languages is much larger than the diversity of vowel sounds. The number of different consonants ranges from 6 in the East-Papuan language Rotokas (Firchow & Firchow 1969) to 95 in the Khoisan language !Xũ (Snyman 1970). Still, certain consonants appear more often than others, so that sounds such as [m], [k] and [l] appear in the majority of the world's languages while others are very rare. Certain symmetries are observed, so that, for example, languages are more likely to have both voiced and unvoiced stops of the same articulation (such as [d] and [t]) rather than either one of them alone. Also, gaps in repertoires tend to be predictable to some extent. For example, if a language has both voiced and unvoiced plosives ([b,d,g] and [p,t,k] respectively) then if it lacks a voiced plosive, this will almost always be [g] (such as in Dutch, where it has been replaced by a fricative) while if it lacks an unvoiced plosive, this will almost always be [p] (such as in Arabic). Finally, it has been found (Lindblom & Maddieson 1988) that consonant sounds can be divided into three classes: basic ones, elaborated ones and complex ones. The basic ones are the ones that are used in the smallest repertoires, only above a certain size the elaborate ones are used while only in the largest repertoires the complex ones are used. In most Western-European languages only basic and a limited number of elaborated consonants are used. Complex articulations occur mostly in the really large repertoires of certain Caucasian languages, native American languages and certain South-African languages that contain clicks.

Regularities of combination are regularities in the way sounds can co-occur in words, and what sounds are preferred in which positions in words. A most interesting result in this respect is the sonority hierarchy (Vennemann 1988). This hierarchy is based on the observation that in a syllable, there tends to be a definite preference for certain sounds to be at the nucleus (typically these would be vowels) and other sounds to be at the periphery (typically these would be consonants). Types of sounds can be (partially) ordered in a hierarchy indicating the sequence in which they are most likely to occur in a syllable. However, this is not all. Certain types of syllables occur more often than others. Syllables of the type CV (consonant followed by a vowel) and V (a single vowel) occur in every language. Languages with syllables of the type VC are already rarer, while the types of complex syllables that tend to occur in Germanic languages, such as Dutch, English or German are quite rare cross-linguistically. Also, in certain languages some sounds tend not to occur in certain positions in a word, while they do occur in others. In Dutch and German, for example, voiced fricatives and plosives become unvoiced at the end of a word. In other languages, plosives that occur between two vowels will always become voiced.

A lot of this large range of phenomena correlate with functional criteria. Some sounds and some combinations of sounds are easier to pronounce than others, and these tend to be the ones that occur most often cross-linguistically. However this cannot be the full explanation. Sounds and combinations of sounds that are sub-optimal from a functional viewpoint do occur in human languages. For example, in Ducth and German final voiced plosives and fricatives are devoiced, but in English and French they are pronounced voiced. Also no easy functional explanation can be given for the fact that languages tend to have both voiced and unvoiced plosives at the same place of articulation. This paper describes a first exploration of the hypothesis that, just as in the case of vowel systems, the consonant systems and the ways in which consonants and vowels are combined into words, are the result of self-organisation in a population under constraints of perception, production and learnability.

For predicting the consonant systems of human languages very realistic models of human production and perception of speech are necessary. Although realistic articulatory (Boersma 1998) and perceptual (e.g. the cochlear model of Pont & Damper 1992) models do exist, these are extremely costly computationally. Also very little is known about the higher cognitive processes of control of articulation or processing of speech signals. It was therefore decided to concentrate on a more abstract system that tries to model the emergence of regularities of combination in the language of a population as emerging from functional constraints on production and perception of individuals. The two main goals of the preliminary research were to find out whether interesting regularities would emerge in a relatively abstract system and how a system of complex utterances could be learnt.

Some other work on computer modelling of the emergence of combinations of speech sounds has been done. Lindblom *et al*. (1984) presented a system that created a set of syllables from a given (large) set of possible consonants and vowels based on criteria of articulatory ease, acoustic distinctiveness and maximisation of distance between the syllables in the repertoire. For the right parameter settings, systems of “phonemically coded” (meaning that only a small subset of the available vowels and consonants would be used in a combinatorial way) realistic syllables would emerge. However, this system was limited to one type of syllable (CV) and used only optimisation. Redford *et al*. (1998) used a genetic algorithm in order to develop sets of words that followed the sonority hierarchy. However, the population that was subject to the genetic evolution did not consists of agents that used the words, but of the words themselves. Perrone (as presented on the Evolution of Language conference 2000 in Paris) uses imitation games to develop repertoires of syllables consisting of three vowels. This work comes closest to the work presented here, and it does use realistic signals. However, the shape of the syllables is pre-determined, and it uses vowels only, but they *are* analysed in terms of their constituent phonemes, and these phonemes are recombined into new words.

2. The simulation

First a word on terminology and notation: “phonemes” are representations of speech sounds in the agents and are written between slashes (/p/). “Sounds” are the actual realisations of these phonemes and are written between square brackets ([p]). The simulation is based on a population of agents that can produce and perceive abstract sequences of speech sounds. These sequences of speech sounds consist of sounds that are described as sets of binary features. Although the idea of binary features is inspired by their use in phonology, (starting with Trubetzkoy (1929) and Jakobson & Halle (1956)), this research does not necessarily subscribe to these theories, nor does it assume the cognitive reality of binary features. They are just adopted as a convenient abstraction.
Neighbouring sounds influence each other during production and perception. The aim of the agents is to imitate the other agents in the population as well as possible. For this they have to learn both which words and which phonemes the other agents are using, as they analyse the sounds they perceive first in terms of the phonemes they know, and then in terms of the words they know.

The agents have a list of words and a list of phonemes (figure 1). In a realistic system there would be an acoustic aspect as well as an articulatory aspect to at least the phonemes, but as this system works with abstract signals only, the phonemes are stored as a set of features. At the moment four features have been implemented, making possible 16 sounds of which only nine are used in the preliminary experiments: one vowel and eight consonants. Vowels and consonants are distinguished by the binary feature *consonantal*. Consonants are distinguished by the feature *voiced*, and two more abstract features making possible four places of articulation. The possible consonants are then: p, t, c, k (voiced) and b, d, j and g (unvoiced). This is a decidedly simplistic repertoire, but as the preliminary experiments described here were mainly meant to test the learning mechanism and the behaviour of the pronunciation rules, this is no problem at present. For further experiments the repertoire will have to be extended.

Words are stored as sequences of pointers to phonemes. Whenever an agent wants to say a word, it produces the sequence of sets of features that corresponds to this word. This sequence is then passed through a rule base that implements the kinds of articulatory constraints that are also encountered in human production of speech sounds. Depending on a parameter called the “articulatory effort” a final speech signal is produced that resembles the intended sequence more when the effort is higher. When the effort is low, the produced sequence will tend towards a default pronunciation. An example would be the devoicing of final plosives. This would be the default behaviour, as it requires least effort. However, the agents are in principle quite able to produce voiced final plosives, given enough articulatory effort.

There are different ways in which sounds can be modified. In all cases, modification of sounds is context dependent. This means that a sound can be pronounced differently depending on the sounds (or word gaps) occurring around it. In this way voicing of plosives surrounded by vowels or devoicing of plosives at the end of a word can be implemented. These two examples are examples of the modification of a sound. However, it is also possible that a sound is deleted or that an extra sound is inserted. These are all processes of variation of pronunciation that are observed in human languages. One important possible variation of pronunciation that has not been implemented is that of harmony: the phenomenon that a feature of pronunciation spreads over different sounds throughout the word. This is a non-local effect and only local effects have been implemented so far.

Recognition of words and phonemes is done with a distance measure (just as in the vowel imitation games (de Boer 1999, de Boer & Vogt 1999)). The distance between two phonemes is calculated as the number of features that is different. No weights are attached to the features, although it is plausible that different aspects of a sound are perceived with different importance in human perception. A sound (consisting of a string of phones) an agent hears is first analysed by finding the phonemes that most closely match the phones Then the distance between two words is calculated (in principle) as the sum of the distance between the individual phonemes, but as phonemes can be deleted and inserted during the pronunciation, the perceived word is warped dynamically (by doubling or deleting constituent phonemes) in order to get a minimal distance. The word with the shortest overall distance is assumed to be recognised.

An imitation game between two agents then proceeds as follows: the first agent (the initiator) selects a random word from its repertoire and pronounces it with a given effort. At the moment the effort is fixed to be 0, so that utterances are maximally deformed. The second agent (the imitator) analyses the sound in terms of its phonemes and finds its word that is closest to it. It then pronounces the word it found, and the initiator analyses it in the same way. If it finds the same word as the one it originally said, the game is successful, if not it is a failure. It communicates this to the imitator using non-linguistic feedback. Both can then update their repertoire of words and phonemes.

In contrast with the vowel imitation game, the agents are not initially empty. They all have one phoneme (the vowel, represented as /a/) and one word, consisting of this one phoneme. This has been done out of convenience of programming, rather than out theoretical considerations: the vowel is the only phoneme that can be said in isolation, so it would be the first to be learned in any case.

Also in contrast with the vowel imitation game, it is possible for the agents to remain silent during a game. Whenever an agent picks a word that would be reduced to nothing by the pronunciation rules (for example a

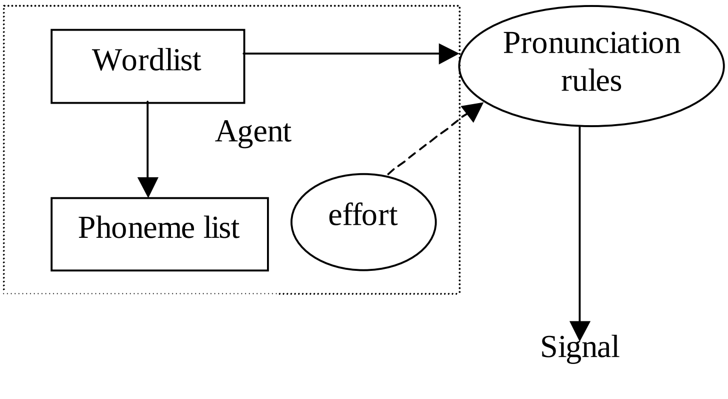

Figure 1: Schematic architecture of an agent with complex utterances.

word consisting of a single consonant) it remains silent. This can happen to both the initiator and the imitator.

In reaction to the imitation game both the imitator and the initiator keep track of the use and success counts of phonemes and words. The use count is updated every time a phoneme or a word is used, while the success count is updated every time it is used successfully. The ratio between the success and the use counts is an estimation of how well a word or a phoneme is shared in the population. Note that whenever a phoneme is used several times in one word, its use and success counts are updated as many times as it appears in the word. Whenever the use count exceeds a certain value (5 in the experiments presented here) and the success/use ratio falls below a certain threshold (0.7 in these experiments) a word is removed from an agent's word list. For the time being phonemes are not removed from an agent's repertoire, although ultimately this should be done.

Another update is the random insertion of words and phonemes. With low probability (1% of the imitation games) a new word is inserted into an agent This word is generated as a random string of the agent's phonemes of a length that is uniformly distributed around the average length of the agent's other words with a standard deviation of 2 phonemes. Whenever the randomly generated word already appears in the agent's repertoire, a random phoneme is generated and inserted if it does not already appear in the agent's phoneme repertoire.

Finally there is the addition of new words and phonemes. Whenever an imitation game was unsuccessful, an agent has the choice of adding a new word and/or a new phoneme. Adding words would be relatively straightforward if an agent would know the other agent's phonemes, but this is not a priori known. The agent therefore needs a criterion when to add a new phoneme. As phonemes are by definition minimal units of sound that can change the meaning of a word, meaning has to be taken into account in order to learn phonemes. In order to simplify the model meaning is here reduced to the non-linguistic feedback an agent receives. The heuristic that is used in these simulations is that whenever an imitation game was a failure, an agent adds the string of phonemes it analysed, unless this string of phonemes is exactly the same as a word that is already in the repertoire. This means that the other agent makes a distinction in sound that this agent is not able to perceive. This agent must therefore add a new phoneme to its repertoire. The new phoneme is derived by finding the set of features that maximally distinguishes the analysed string of phonemes from the sound that was actually perceived. If this set of features does not yet appear in the agent's phoneme repertoire, it is added.

As will be shown below, in this way the agent is capable of learning another agent's phonemes. However, for this process to work it is crucial that the sounds are coupled with non-linguistic feedback. This is what distinguishes this approach from other systems that are used for learning to recognise or imitate speech.

3. First results

The goal of the preliminary experiments was twofold. The first goal was to check whether the agents would be able to develop shared sound systems and vocabularies so that they would be able to play successful imitation games. The second goal was to see in what way the constraints on articulation that were imposed on the agents would influence the emerged vocabularies. For this reason the articulatory effort of the agents was fixed to the smallest possible value, 0.

There were a fair number of articulatory constraints, but it is necessary to list them all in order to get an understanding of the behaviour of the system and the vocabularies that emerge. The simplest constraint was that consonants in isolation are not pronounced. Also, sequences of two or more identical phonemes are reduced to only one. Then there were a number of rules for dealing with consonant clusters. These were as follows:

$$G_1G_2C \to G_1G_2VC$$

$$G_1KG_2 \to G_1VKG_2$$

$$GK_1K_2 \to GK_2$$

$$KGC \to KGVC$$

$$K_1K_2C \to K_1C$$

where a K denotes a voiceless consonant, a G denotes a voiced consonant, a C denotes any consonant and a V denotes a vowel. Whenever a consonant is surrounded by two voiced sounds, at least one of which is a vowel, it is automatically voiced as well.

At the beginning of a word, the voiced velar /g/ is always turned into unvoiced [k] while voiced consonants that appear before an unvoiced consonant at the beginning of a word are also devoiced. At the end of a word, a vowel gets added after a voiced consonant.

Note that these rules are just a subset of the possible and necessary rules for making a description of the way humans tend to simplify their pronunciation when they speak fast and in an informal register. In fact they were mostly tuned in order to get words that did not contain huge unrealistic clusters of consonants.

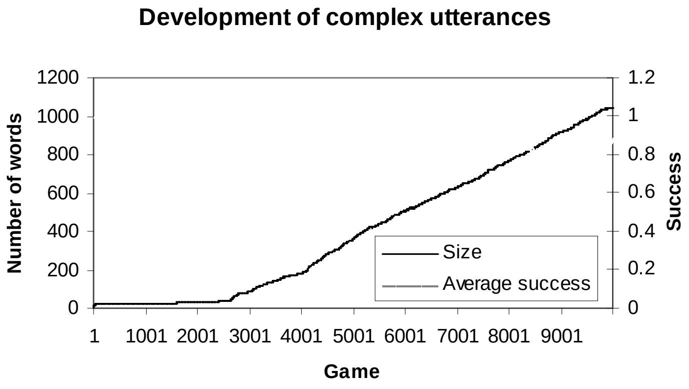

Figure 2: Development of complex utterance vocabularies over time.

The first experiment that will be described was done in order to investigate how the vocabularies and the imitation in the population of agents changes over time. The development over time of a typical experiment is shown in figure 2. This experiment was run with a population of 20 agents. The figure shows the running average of the success of the imitation games and the total number of words in the population. The running average is calculated as 0.01 times the present outcome of the imitation game (0 if failure, 1 if success) added to 0.99 times the previous running average.

Three stages can be observed in the development of the vocabularies. The first stage, consisting of the first 1500 games or so, shows a perfect success, but vocabularies consisting of only one word per agent; the word consisting of the single vowel with which the agents were initialised. Then a first transition occurs: the number of words increases (not always as slow as shown in this example), while the average success drops dramatically (in this example below 50%). The third stage shows a linear increase in the number of words, while the success is restored to around 80%.

In the first stage, the agents do not have any shared phonemes, except for the single vowel phoneme. Therefore, possible new words are doomed to be mostly indistinguishable from existing words and in any case to be inimitable by the other agents. These words will therefore be removed quickly from the agents' repertoires. However, any newly generated phonemes will be retained. This makes it possible for the second stage to take place: new phonemes have been inserted in all agents, so longer words can now be imitated in principle. However, these words will consist of randomly generated strings of phonemes (which are, moreover, not universally shared between agents) that will generally be changed beyond recognition by the two stages of pronunciation in the imitation game. This is the reason the success drops. However, because of all the failures in the imitation game, agents are forced to update their repertoires of words and phonemes. This initiates the third stage. In this stage, agents all have almost the same repertoires of phonemes (they have copied them off each other) and will learn words from each other, rather than generate them randomly. As the words that are learnt have already passed through the pronunciation rules, they will tend not to be changed beyond recognition by another pass through these rules. From this moment on, vocabularies can expand quickly, as all agents have the necessary phonemes, and there is no penalty on longer words.

An example of a vocabulary that emerged from the imitation games is given in table 1.

ga	ga	ga	ga	ga	ga	ga	ga		ga	ga	ga	ga	ga	ga	ga	ga	ga	ga	ga	0.959	0.95
			gab																	0.000	0.05
					gaba															0.400	0.05
gada											gada									0.433	0.1
											gba									0.667	0.05
													gdacda					gdacda		0.300	0.1
	gja					gja		gja		gja			gja					gja		0.690	0.3
ja	ja	ja	ja	ja	ja	ja	ja			ja	ja		ja		ja	ja	ja		ja	0.916	0.7
jaga	jaga		jaga		jaga	jaga				jaga	jaga					jaga	jaga	jaga	jaga	0.893	0.6
					jba		jba		jba			jba						jba	jba	0.667	0.3
jc	jc		jc	jc		jc	jc	jc	jc		jc	jc			jc	jc		jc	jc	0.899	0.7
						jcpb														0.000	0.05
											jda									0.667	0.05
		jga		jga	jga	jga	jga	jga	jga	jga	jga			jga	jga				jga	0.886	0.6
			jkpa																	0.000	0.05

jpa	jpa											jpa	jpa						jpa	0.838	0.25

Table 1: Words from the emerged vocabulary with their success and their relative abundance.

In order to save space, only the words starting with phonemes /j/ and /g/ are shown. In the rightmost two columns the average success/use score of the words and their relative abundance (number of agent in which the word occurs divided by total number of agents) are shown. It can be observed that some words are shared by all agents in the population. Although not all these words appear very likely to humans (/jc/ would be a particularly unlikely word) the ones that are shared by multiple agents do conform to the pronunciation rules. Ones that do not conform (such as /jkpa/ or /jcpb/) will not spread through the population. Furthermore, it can be observed that words that are shared by a large number of agents are more successful (correlation 0.67). However, some words that occur in only one agent do appear to be successful (/gba/ for example). This is probably caused by the fact that other agents in the population have words for which this word is the closest imitation or that are the closest imitation for this word.

The emerged vocabularies are slightly disappointing in that they always tend to converge to the same kinds of words. This is understandable, because effort is fixed and the rules therefore always operate in the same way.

4. Discussion

Apparently successful repertoires of words and shared repertoires of phonemes can be developed in a population of agents by playing imitation games. However, in contrast with the vowel imitation games, the results do not have much linguistic relevance, yet. The most important result of these simulations so far is that agents are able to extract phonemes from words by using the feedback of imitation. This indicates that the idea of playing imitation games with more complex utterances is viable in principle. The simulation can be made more linguistically relevant by making a number of changes. These changes include making the sets of pronunciation rules more realistic. Care must be taken that the rules represent phenomena that are actually observed in human behaviour. Necessity of adding rules that do not correspond to actual behaviour of individuals in order to make realistic vocabularies emerge would constitute a falsification of the theory that self-organisation in a population is responsible for the emergence of regularities of the combinations of sounds in human languages. It would also be interesting to have sets of rules of which the output will not stay constant whenever the rules are applied a second time. This would force the agents to pronounce words with some effort.

This leads to the second weakness of the present simulation: the agents' articulatory effort is fixed during the simulation. For this reason, vocabularies always tend to converge to the same kinds of words. It would be more desirable to have the agents adapt their articulatory effort dynamically. This extra feedback loop would increase the complexity of the system and make it possible for it to converge towards different final states (or to move from one temporarily stable state to the next) instead of to only one. It can be imagined that one final state would consist of short, but complex words, while the other final state would consist of longer, but less complex words. For this to happen though, a penalty must be introduced for longer words.

Finally, it could be interesting to have perceptual constraints as well as articulatory constraints, so that certain sounds and combinations of sounds can be confused more easily than others. However, it must also be kept in mind that the present simulation is inherently limited because it is based on an abstract representation based on distinctive features that is not directly suited for modelling all the subtleties of human production and perception of speech sounds. Whenever it turns out that this way of modelling yields interesting results, we must face the consequence that we have to move on to a more realistic model of speech production and perception.

References

Boersma, Paul (1998) *Functional Phonology* The Hague: Holland Academic Graphics.

de Boer, B. G. (1999) *Self-organisation in vowel systems*, Ph. D. Thesis, Vrije Universiteit Brussel.

de Boer, B. G. & Vogt, P. (1999) Emergence of speech sounds in changing populations, In: D. Floreano, J-D. Nicoud & F. Mondada (eds.) Advances in Artificial Life, *Lecture Notes in Artificial Intelligence* **1674**, pp. 664–673

Firchow, Iwin & Jacqueline Firchow (1969) An abbreviated phoneme inventory. *Anthropological Linguistics* **11**, pp. 271–276.

Jakobson, Roman & Morris Halle (1956) *Fundamentals of Language*, the Hague: Mouton & Co.

Lindblom, Björn, Peter MacNeilage & Michael Studdert-Kennedy (1984) Self-organizing processes and the explanation of language universals. In Brian Butterworth, Bernard Comrie & Östen Dahl (eds.) *Explanations for language universals*, Walter de Gruyter & Co. pp. 181–203.

Lindblom, Björn & Ian Maddieson (1988), Phonetic Universals in Consonant Systems. In Larry M. Hyman & Charles N. Li (eds.) *Language, Speech and Mind*, pp. 62–78.

Pont, M. J. & Damper, R.I. (1992) A Computational model of afferent neural activity from the cochlea to the dorsal acoustic stria, *Journal of the Acoutical Society of America* **89** pp. 1213–1228

Redford, Melissa Annette, Chun Chi Chen & Risto Miikkulainen (1998) Modeling the Emergence of Syllable Systems. In Morton Ann Gernsbacher & Sharon J. Derry (eds.) *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society* (COGSCI-98) pp. 882–886.

Snyman, J. W. (1970) *An introduction to the !Xũ̃ (!Kung) language*, Cape Town: Balkema.

Trubetzkoy, N. S. (1929) Zur allgemeinen Theorie der phonologischen Vokalsysteme. *Travaux du cercle linguistique de Prague* **7**, pp. 39–67.

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