JPHO 20000125 PROD.TYPE: TYP ED: Jayashree
pp. 1–25 (col.fig.: NIL) PAGN: Terence – SCAN: Yoga

Journal of Phonetics (2000) **27,** 000–000
doi.10.006/jp2000.0125
Available online at http://www.idealibrary.com on IDEAL®

Self-organization in vowel systems

Bart de Boer*

AI-lab, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium

Received 20th September 1999, and accepted 9th October 2000

This paper presents a computer simulation of the emergence of vowel systems in a population of agents. The agents (small computer programs that operate autonomously) are equipped with a realistic articulatory synthesizer, a model of human perception and the ability to imitate and learn sounds they hear. It is shown that due to the interactions between the agents and due to self-organization, realistic vowel repertoires emerge. This happens under a large number of different parameter settings and therefore seems to be a very robust phenomenon. The emerged vowel systems show remarkable similarities with the vowel systems found in human languages. It is argued that self-organization probably plays an important role in determining the vowel inventories of human languages and that innate predispositions are probably not necessary to explain the universal tendencies of human vowel systems.

1. Introduction

The sound systems of human languages show remarkable regularities. Although humans are able to produce and perceive many different speech sounds, languages do not use random subsets of these sounds. In UPSID, the UCLA Phonological Segment Inventory Database (Maddieson, 1984; Maddieson & Precoda, 1990) which now contains 451 languages, 921 different speech sounds are found. However, the average number of phonemes that is used in the languages in this sample lies between 20 and 37 (Maddieson, 1984). The minimal number of phonemes in any language in the sample is 11 in Rotokas (Firchow & Firchow, 1969) and Pirahã (Everett, 1982) and the maximum number is 141 in !Xũ (Snyman, 1970). This means that human languages generally use only a small subset of the available speech sounds.

Within these subsets, some speech sounds occur more often than others. Some sounds, such as [a] (in 87% of the languages in UPSID_{451}) [m] (94%) or [k] (89%) occur almost universally while others, such as [œ] (2%) [ʀ] (1%) or [ʡ] (1%) occur only rarely. Moreover, phoneme inventories tend to exhibit certain symmetries. If, for example, a repertoire contains an [ɔ] (in 36% of the languages of UPSID_{451}) it is almost 5 times more likely to contain an [ɛ] (in 41% of the total sample) than if it does not contain [ɔ].

* E-mail: bartb@arti.vub.ac.be.

0095–4470/00/000000 + 00 $35.00/0

Of the languages with [ɔ], 83% also contain [ɛ], whereas of the languages without [ɔ], only 18% contain [ɛ]. Symmetries are also found for vowels of different heights, and for consonant systems. If a language contains a voiced plosive at a certain articulatory position (e.g. [d], which occurs in 27% of the languages of UPSID$_{451}$) it is much more likely to contain the corresponding voiceless plosive ([t], which occurs in 40% of the entire sample, but in 83% of the languages with [d]).

Finally, the sequences in which sounds occur in human languages are not random either. All languages contain syllables that consist of a vowel only (V) or of a consonant followed by a vowel (CV). Other types of syllables, for example with clusters of consonants or with consonants at the end of the syllable are rarer (Venneman, 1988). In general, speech consists of alternating vowels and consonants. Whenever consonants appear in clusters, different constraints also apply. At the beginning of a syllable an obstruent followed by a sonorant (e.g. [pl]) is much more common than the other way around, whereas at the end of a syllable, a sonorant followed by an obstruent (e.g. [lp]) is more likely. Apparently, both the repertoire and the use of speech sounds in human languages are severely constrained.

Regularities like these demand an explanation. Possible explanations can be divided into two classes: those that are based on innate human cognitive capacities and those that are based on the functional constraints of a good communication system. Such constraints derive from the fact that linguistic communication has to be robust and learnable: ideally, there should be redundancy, predictability, and speech sounds should be easily distinguishable and produced. Explanations based on innate capacities for language postulate properties of the human brain, such as distinctive features and their markedness (e.g. Jakobson & Halle, 1956; Chomsky & Halle, 1968) based on observations of linguistic data. However, such explanations have a certain risk of circularity (Lindblom, MacNeilage & Studdert-Kennedy, 1984) and they generate two important new questions: how did these capacities become innate and how are they implemented in the neural circuitry of the brain? Such questions are extremely hard to answer, and it is therefore preferable to first explore explanations that do not depend on innate properties of the brain.

Explanations of phonetic phenomena that do not depend on innate properties of the brain have been based on independent evidence from physics and communication theory. They start from the assumption that speech is meant to provide communication over an unreliable channel using the human vocal tract and perceptual system. They then take one or more aspects of the communication process (such as acoustic distinctiveness or articulatory stability) and see whether optimization of these aspects can predict properties of human systems of speech sounds. An example is the optimization of acoustic distance between vowels, as has been done by Liljencrants & Lindblom (1972). They found that by optimizing an abstract energy function (which was based on the acoustic distance between the vowel phonemes, see Equation (6) in Section 4) over vowel systems with a fixed number of elements, they could predict the most frequently found vowel systems in human languages. Other functional explanations for properties of systems of human speech sounds have been provided for example by Stevens (1972, 1989) and by Carré (1994).

It has been found that the sound systems of human languages are often optimized for criteria such as acoustic distinctiveness or articulatory ease. Maximization of acoustic distinctiveness seems to play an important role in vowel systems (Liljencrants & Lindblom, 1972; Schwartz Boë, Vallée & Abry, 1997*b*). Minimization of articulatory effort seems to play an important role in determining the consonant inventories of the world's languages (Lindblom & Maddieson, 1988).

However, one important aspect is still missing in such explanations, and that is *how* sound systems become optimized. When children learn the sound system of their native language, they do not explicitly optimize it. Rather, they learn to produce sounds that will eventually become extremely similar to those of their parents and peers. The sounds they learn are much closer to the sounds the other speakers use than is necessary for successful communication. This can be observed in the way different dialects of a single language can coexist. The phonological system of the different dialects can be the same (although this is not necessarily the case) but the pronunciation of different sound categories may differ consistently between two dialects. Although speakers of the different dialects can understand each other perfectly (i.e., the similarity is large enough to allow successful communication) children growing up in a given community will always learn the pronunciation of *their* community (for a discussion of the importance of vowel pronunciation in recognizing social class (see Trudgill, 1995, pp. 37–38).

The hypothesis explored in this paper is that optimization is caused by self-organization in the population. Self-organization (Nicolis & Prigogine, 1977) is the phenomenon that order on a global scale can emerge spontaneously in a group of interacting entities. Although the entities and their interactions may be quite simple, it is possible that complex organization emerges. A good example of an ordered system that emerges through self-organization is the honeycomb. Individual bees each working on a single cell cause a hexagonal grid to emerge in the absence of global control. The idea that self-organization plays a role in the sound systems of human languages is not new (see e.g. Lindblom *et al.*, 1984). However, because of the iterated interactions between large numbers of entities, the behavior of self-organizing in any system is often hard to predict. The best way to study them is through computer models. Therefore, in the research described in this paper, such a computer model was implemented and investigated.

The implementation described in this paper differs in an important respect from many previous computer simulations (although not all—see the next section) in that it simulates a population of language users instead of the linguistic behavior of a single individual. In this respect, the work presented here does fit with the computer simulations by Steels and others (see e.g. Steels, 1998) on the origins of language. The goal of the members of the population is to imitate each other as well as possible using a repertoire of speech sounds that is as large as possible. It will be shown that under these circumstances, vowel systems emerge that show remarkable similarities with vowel systems that are found in human languages.

Self-organization is a notion that is generally used in a rather loose way for widely different phenomena, and sometimes only in a metaphorical way. It is not possible to apply a strict mathematical definition of self-organization to vowel systems. The best one can do is to define a number of properties that a system must have in order to be called self-organizing. First of all, *organization on a global scale* must emerge. In the case of vowel systems this means that most agents in a population should have the same vowels. This implies that they should be able to imitate each other's vowels successfully. Secondly, the emergence of coherence should be due to *interactions* between the agents, rather than to actions (e.g., optimization) within the individual agents. Finally, there should not be a non-local influence on the agents' behavior. In the case of vowel systems, this means that agents cannot observe each other's vowel systems directly, and are not provided with pre-wired knowledge, (e.g., features or innate dispositions) or have their vowel systems initialized beforehand. In other words, the agents start as *tabula rasa*.

The next section describes two previous attempts at building a simulation of a population of agents for explaining the universals of vowel systems. Section 3 describes the simulation that was used in this paper in sufficient detail such that the experiments can be re-implemented. It also discusses the reasoning behind the different design decisions that were made in implementing the simulation. Section 4 describes various experiments done with the simulation and agrues that the emerging vowel systems are very close to the vowel systems found in human languages. Section 5 discusses to what extent the results are relevant to the study of human languages, in which ways the present results are limited and what can safely be concluded from the experiments.

2. Previous work

The first attempt at building a simulation of a population for explaining the properties of human vowel systems was made by Hervé Glotin and others of the *Institut de la Communication Parlée* in Grenoble (Glotin, 1995; Glotin & Laboissière, 1996; Berrah, Glotin, Laboissière, Bessière & Boë, 1996). This work was based on a population of agents (which Glotin calls *carls*) each of which had a fixed number of vowels. These vowels were represented in both an acoustic and an articulatory space. The interactions take place between two agents. One agent (A) picks a random vowel from its repertoire and transmits its acoustic description to the other agent (B). This agent (B) then finds the closest vowel in its repertoire and calculates the amount of effort needed to move it towards the perceived signal. It then moves its vowel towards the perceived position, by calculating articulatory changes that need to be made in order to change the acoustic signal. It then produces the new signal and the first agent (A) in turn adapts its vowel repertoire. At the same time the other vowels of the agents (both A and B) are moved away from the vowel that is being used. Over a number of interactions, the "articulatory" cost of moving the vowels is calculated. With this cost an evolutionary fitness (which is inversely proportional to the cost) is calculated. On the basis of this fitness function, agents are selected from the population and allowed to create offspring. The initial vowel repertoires of these offspring are determined by crossing the vowel systems of the parents.

There are two major problems with this work. The first is that calculating the mapping from acoustic differences onto articulatory changes is computationally very demanding. Glotin has therefore not been able to do experiments with large populations and large numbers of vowels. The second problem is more fundamental, in that Glotin adds a genetic element to the dynamics of the population. New agents in the population inherit the vowel systems of their parents. He explains this as a simplification of the learning process by infants (Glotin, pers. comm) and as a means to increase coherence in the population. However, it complicates matters enormously, and makes it very hard to determine which results of his simulations are caused by self-organization, which by the genetic algorithm and which by optimization. Nevertheless, the work that is described in this paper, and especially the organization of the interactions between the agents, is in large part based on Glotin's work.

Glotin's work has been elaborated upon by Berrah (1998). Berrah uses only an acoustic representation of the vowels, and his system is therefore much faster than Glotin's. He has therefore been able to do experiments with more agents and more vowels. He also extended the model to include an extra feature dimension (such as length) in order to investigate how and when his agents would start using this extra feature. However, the main force operating in his agents is repulsion between the vowel prototypes in an individual agent. In fact, it turns out that his system is equivalent to the Liljencrants & Lindblom (1972) optimizing system. The interactions between the agents do not play a role in determining the shape of the vowel systems that emerge and only cause coherence in the population. However, since this coherence is also caused in part by Darwinian selection and reproduction, it is hard to assess the role of the interactions between the agents.

In both Glotin's and Berrah's work, the role of the interactions between the agents is not quite clear as their influences are obscured by the global optimization of the agent's vowel systems through the repulsion of the vowels. The role of agent interactions is also obscured by the action of Darwinian selection. These mechanisms will therefore be avoided in the model that is used in this paper.

3. The simulation

The hypothesis that is investigated in this paper is that the structure of vowel systems is determined by self-organization in a population under constraints of perception and production. Self-organization is a phenomenon that is very hard to predict from just the description of a system. Therefore, any theory about the role of self-organization in a system is best tested with a computer simulation. Such a computer simulation should be sufficiently simplified to make implementation feasible and at the same time capture the essential details of the phenomenon that is being investigated. Building computer simulations of life-like phenomena is the domain of science that has been called *artificial life* since the mid-1980s (Langton, 1989). Artificial life is in fact not so much a science as a methodology. Instead of investigating complex natural phenomena by analyzing them, it attempts to gain understanding by trying to synthesize these phenomena. Whenever a phenomenon can be successfully synthesized, the model that was used can be considered a candidate mechanism for explaining the phenomenon. Of course, in practice, analysis and synthesis go hand in hand. The analysis of a phenomenon results in several candidate hypotheses. Based on these hypotheses, computer simulations are built and the one that best reproduces the phenomenon is accepted as the most likely explanation. Further analysis can then be undertaken in order to extend the hypothesis and the model.

The computer simulation investigated in this paper is based on a population of agents that can produce, perceive, and remember speech sounds in a human-like way. For this purpose, each agent is equipped with an articulatory synthesizer, a model of human perception for calculating the distances between different signals and an associative memory for storing vowel prototypes. Also, each agent can interact with other agents (following a fixed pattern) by imitating them. The agents can update their vowel repertoires depending on the outcome of the interactions. The agents' (implicit) goal is to accurately imitate the outer agents with a repertoire of vowels that is as large as possible. The architecture of the agents is illustrated in Fig. 1.

The agents' articulatory synthesizer takes as inputs the three major vowel parameters: tongue position, tongue height, and lip rounding (Ladefoged & Maddieson, 1996, Chapter 9). The outputs of the synthesizer are the first four formant frequencies of the

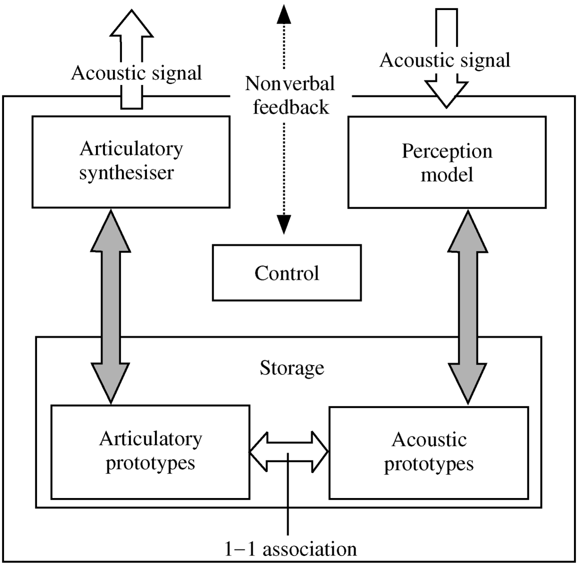


Figure 1. Agent architecture.

corresponding vowel. The inputs are modeled as continuous values in the range 0–1. For tongue position, 0 means most to the front and 1 means most to the back. For tongue height, 0 means lowest and 1 means highest. For lip rounding, 0 means least rounded and 1 means most rounded. Thus (0, 0, 0) corresponds to the vowel [a] and to frequencies of 708, 1517, 2427, 3678 Hz for F1–F4. The parameters (1, 1, 1) correspond to the vowel [u] and to formant frequencies of 276, 740, 2177, 3506 Hz. This synthesizer is in principle able to generate all possible basic vowels. It should be noted, however, that it cannot produce secondary distinctions, such as voicing type, nasalization, length, etc.

The synthesizer is based on interpolation (quadratic in the dimensions of height and position and linear in the dimension of lip rounding) between the formant frequencies of 16 artificially generated vowels and two estimates by the author. The artificial vowels were generated by Maeda's (1989) articulatory model. The actual data were taken from Vallée (1994, pp. 162–164). The value of the articulatory parameter position was defined as 0 for all front vowels in the data set, 0.5 for all central vowels and 1 for all back vowels. Height was defined as 1 for all high vowels, 0.5 for all mid-vowels and 0 for all low vowels. Rounding was defined as 0 for all unrounded vowels and 1 for all rounded vowels. The data points are presented in Table I (the ones estimated by the author are marked with an asterisk) and the resulting articulatory formulas are presented in Table II. In these formulas, p corresponds to tongue position, h to tongue height and r to lip rounding. This articulatory synthesizer allows for quick calculation of formant frequencies from articulatory parameters, thus making it possible to play a large number of iterated imitation games in a reasonable amount of time.

In order to make the interactions between the agents more interesting, noise can be added to the formant frequencies. This is done by shifting the formant frequencies a random percentage. This percentage is bound to a maximum, called the acoustic noise.

TABLE I. Data points for articulatory synthesizer

Vowel	p	h	r	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)	V_4 (Hz)
[a]	0	0	0	708	1517	2427	3678
[œ]*	0	0	1	670	1400	2300	3500
[ɐ]	0.5	0	0	742	1266	2330	3457
[ɐ̹]*	0.5	0	1	658	1220	2103	3200
[ɑ]	1	0	0	703	1074	2356	3486
[ɒ]	1	0	1	656	1020	2312	3411
[e]	0	0.5	0	395	2027	2552	3438
[ø]	0	0.5	1	393	1684	2238	3254
[ə]	0.5	0.5	0	399	1438	2118	3197
[e]	0.5	0.5	1	400	1267	2005	2996
[ɤ]	1	0.5	0	430	1088	2142	3490
[o]	1	0.5	1	399	829	2143	3490
[i]	0	1	0	252	2202	3242	3938
[y]	0	1	1	250	1878	2323	3447
[ɨ]	0.5	1	0	264	1591	2259	3502
[ʉ]	0.5	1	1	276	1319	2082	3118
[ɯ]	1	1	0	305	1099	2220	3604
[u]	1	1	1	276	740	2177	3506

TABLE II. Synthesizer equations

$$
\begin{aligned}
F_1 = {}& ((-\,392 + 392r)h^2 + (596 - 668r)h + (-\,146 + 166r))p^2 \\
&+ ((348 - 348r)h^2 + (-\,494 + 606r)h + (141 - 175r))p \\
&+ ((340 - 72r)h^2 + (-\,796 + 108r)h + (708 - 38r))
\end{aligned}
$$

$$
\begin{aligned}
F_2 = {}& ((-\,1200 + 1208r)h^2 + (1320 - 1328r)h + (118 - 158r))p^2 \\
&+ ((1864 - 1488r)h^2 + (-\,2644 + 1510r)h + (-\,561 + 221r))p \\
&+ ((-\,670\cdot + 490r)h^2 + (1355 - 697r)h + (1517 - 117r))
\end{aligned}
$$

$$
\begin{aligned}
F_3 = {}& ((604 - 604r)h^2 + (1038 - 1178r)h + (246 + 566r))p^2 \\
&+ ((-\,1150 + 1262r)h^2 + (-\,1443 + 1313r)h + (-\,317 - 483r)) \\
&+ ((1130 - 836r)h^2(-\,315 + 44r)h + (2427 - 127r))
\end{aligned}
$$

$$
\begin{aligned}
F_4 = {}& ((-\,1120 + 16r)h^2 + (1696 - 180r)h + (500 + 522r))p^2 \\
&+ ((-\,140 + 240r)h^2 + (-\,578 + 214r)h + (-\,692 - 419r))p \\
&+ ((1480 - 602r)h^2 + (-\,1220 + 289r)h + (3678 - 178r))
\end{aligned}
$$

The shift of the formant frequencies is calculated as follows:

$$\bar{F}_i = F_i(1 + v_i) \tag{1}$$

where F_i is the frequency of the ith formant as calculated by the synthesizer, $\bar{F}_i$ is the frequency of this formant after shifting and v_i is the shifting factor which is randomly chosen from the uniform distribution in the range $-\,\psi_{ac}/2 \leq v_i < \psi_{ac}/2$, where ψ_{ac} is the maximal noise allowed, a very important parameter of the simulation.

A perception function is needed to calculate the distance between a perceived signal and a signal that is stored in an agent's list of acoustic prototypes. The distance is calculated as a weighted Euclidean distance in a two-dimensional acoustic space. The dimensions of this space are the first formant of the signals and their *effective second formant*. The effective second formant of a signal is a nonlinearly weighted sum of the higher formants of the signal. Its calculation is based on the observation that for many natural vowel signals (that have multiple formants) a signal with only two formants can be found that is perceived by human subjects as being identical to the original signal (Carlson, Granström & Fant, 1970). It has been found that the first formant of such signals is equal to the first formant of the original signal, but that the second formant of the artificial signal has a more complex relation with the higher formants of the original signal. The fact that signals with multiple formants sound the same as signals with two formants is probably due to the lower resolution with which higher frequencies are detected in human hearing (Schroeder, Atal & Hall, 1979). Therefore, signals of higher frequency are blurred.

The effective second-formant frequency is calculated based on a method by Mantakas, Schwartz & Escudier (1986). This method was also used in the work of Glotin, so it was adopted in order to make the results of this work more comparable to Glotin's results. It is based on a critical distance c. In Glotin's work this critical distance was 3.5 Bark and this is the value that is used here as well. The Bark scale is a frequency scale that is based on human perception and it is logarithmic for high frequencies and linear for low ones:

$$Bark = \begin{cases} \dfrac{\ln(Hertz/271.32)}{0.1719} + 2, & Hertz > 271.32 \\ \dfrac{Hertz - 51}{110}, & Hertz \leq 271.32 \end{cases}$$

(interpolated from a table of Hertz–Bark pairs found on the Internet). Roughly, if formant peaks are closer together than this critical distance, the value of the effective second formant is taken as their weighted average. More exactly, the following formulas are used:

$$F'_2 = \begin{cases} F_2 & \text{if } F_3 - F_2 > c \\ \dfrac{(2 - w_1)F_2 + w_1 F_3}{2} & \text{if } F_3 - F_2 \leq c \quad \text{and} \quad F_4 - F_2 > c \\ \dfrac{w_2 F_2 + (2 - w_2)F_3}{2} - 1 & \text{if } F_4 - F_2 \leq c \quad \text{and} \quad F_3 - F_2 < F_4 - F_3 \\ \dfrac{(2 + w_2)F_3 - w_2 F_4}{2} - 1 & \text{if } F_4 - F_2 \leq c \quad \text{and} \quad F_3 - F_2 \geq F_4 - F_3 \end{cases} \tag{2}$$

where w_1 and w_2 are the weighting factors, which in the original formulation by Mantakas *et al.* (1986) depend on the strength of the formant peaks. The articulatory synthesizer used in this research did not calculate the strengths of the formant peaks. Therefore, the weights were estimated depending on the distance between the formant peaks, as it was found that in the vowels generated by the model, the closer together formant peaks are, the stronger they are. They were calculated as follows:

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

and

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

This is a rather *ad hoc* solution, and it is probably better to take constants as weighting factors, (see, for example, Schwartz *et al.*, 1997*b*) as calculating the weights in this way tends to introduce discontinuities in the distance function. Although this corresponds to a real perceptual phenomenon, it causes problems for the imitation of high front vowels in the computer simulations. As this discontinuity interferes with the algorithms that were used for learning vowels, a term -1 is added to the last two clauses of the original equations to counterbalance it somewhat.

The total distance D between two signals a and b can now be calculated as follows:

$$D = \sqrt{(F_1^a - F_1^b)^2 + \lambda(F_2^{a'} - F_2^{b'})^2} \tag{5}$$

where λ determines the relative weight of F_2 with respect to F_1. It is taken to be 0.3 for all experiments in this paper, as this was thought to be the most realistic by Vallée (1994) and Schwartz *et al.* (1997*b*). There is also some independent evidence for this value of 0.3 from experiments in human articulation (Lindblom & Lubker, 1985).

The agents store vowels as *prototypes*. As the research described here is based on computer simulations, the term prototype is used in the machine learning sense of the word. In this interpretation a prototype is the central point of a category. If an object has to be classified, the category to which it is assigned is the one whose prototype is closest to that object according to some predefined distance measure. If an object has to be reproduced, the prototype of the object's categories is reproduced, rather than an exact copy of the object itself. Prototypes in the machine learning sense are not static. They can be moved if new information about the classification becomes available. Also, new prototypes can be added and old ones removed.

The prototypes have an articulatory and an acoustic aspect. The articulatory aspect is used for (re-) production and the acoustic aspect is used for perception (classification). The acoustic aspect is calculated by synthesizing the articulatory aspect without adding noise.

For every signal that is perceived, the agents find the prototype that is closest. That is considered as the one that is recognized. Perception is therefore in terms of prototypes, that is, categorical rather than gradual. This kind of perception is probably realistic. Research into human perception (for a critical overview of research on vowels, see, e.g., Frieda, Walley, Flege & Sloane (1999); for older work on consonants see, e.g., Cooper, Delattre, Liberman, Borst & Gerstman (1952) and Liberman, Delattre, Cooper & Gerstman (1954)) has shown that perception in terms of categories or prototypes is biologically plausible. Research into other areas of language and cognition has shown that prototypes play an important role there as well (e.g., Comrie, 1981; Lakoff, 1987).

The list of vowels of an agent is updated depending on the outcome of the imitation games. It must be stressed that the number of vowels of an agent is not constant and that an agent's vowel repertoire is initially empty.

3.1. *The interactions between the agents*

The interactions between agents are called *imitation games*. This name has been inspired by Steels' (1998) interpretation of Wittgenstein's (1967) term *language games*. For each imitation game, two agents are randomly chosen from the population. One of these agents is assigned the role of *initiator* and the other agent is assigned the role of *imitator*. The actions of the agents during an imitation game are illustrated in Tables III–V. In

Table III. Basic organization of the imitation game

	Imitator	Imitator	
1	If ($V = \emptyset$) Add random vowel to V Pick random v vowel v from V $u_v \leftarrow u_v + 1$ *Produce signal* $A_1 : A_1 \leftarrow ac_v + noise$		
		Receive signal A_1 If ($V = \emptyset$) Find phoneme (v_{news}, A_1) $V \leftarrow V \cup v_{new}$ Cacalculate v_{rec}: $v_{rec} \in V \wedge \neg\exists v_2 :(v_2 \in V \wedge D(A_1, ac_{v2}) < D(A_1, ac_{vrec}))$ Produce signal $A_2 : A_2 \leftarrow ac_{vrec} + noise$	2
3	Receive signal A_2. Calculate v_{rec}: $v_{rec} \in V \wedge \neg\exists v_2 :(v_2 \in V \wedge D(A_2, ac_{v2}) < D(A_2, ac_{vrec}))$ If ($v_{rec} = v$) Send nonverbal feedback: *success.* $s_v \leftarrow s_v + 1$ Else Send nonverbal feedback: *failure.*		
		Receive nonverbal feedback Update V according to feedback signal.	4
5	Do other updates of V.	Do other updates of V.	5

these tables, the actions of the agents are described in a rather formal way, which might not be very familiar to phoneticians. This has been done in order to make the description as unambiguous as possible so that the system can be re-implemented, but for a global understanding of the algorithm, the tables are not crucial. The information in the tables will be explained in an informal way in this section. The notation follows standard conventions of pseudo-code, mathematical logic and set theory as much as possible. Boldface indicates a subroutine that is described in one of the other tables or in a different column of the same table. In these subroutines, the first argument in parentheses is returned to the calling routine after the function has finished.

The initiator selects a random vowel v from its vowel repertoire V (if it is not empty; if it is empty, the initiator creates a random new vowel and adds this to the repertoire) and synthesizes it, using the articulatory aspect ar_v. The imitator receives the frequencies of the first four formants of this signal (A_1 in the table) and finds the "recognized" vowel v_{rec} in its vowel repertoire whose acoustic aspect (ac_v in the table) is closest. If its repertoire is empty, the agent finds a vowel that is a close approximation to the perceived signal by talking to itself and then adds this to its repertoire (described as "Find phoneme" in Table IV). The imitator then synthesizes a signal A_2 with the articulatory aspect of the

TABLE IV. Actions performed by the agents

```
Shift closer ($v$, $A$)
$v_{best} \leftarrow v$
For (all six neighbors $v_{neigh}$ of $v$)
do:
    If ($D(ac_{vneigh}, A) <$
        $D(ac_{vrec}, A)$
        $v_{best} \leftarrow v_{neigh}$
$v \leftarrow v_{best}$
```

```
Find phoneme ($v_{new}$, $A$)
$ar_v \leftarrow (0.5, 0.5, 0.5)$
$ac_v \leftarrow S(ar_v)$
$s_v \leftarrow 0$
$u_v \leftarrow 0$
Do
    $v_{new} \leftarrow v$
    Shift closer ($v_{new}$, $A$)
Until ($v = v_{new}$)
```

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

TABLE V. Other updates of the agents' vowel systems

```
Merge ($v_1$, $v_2$, $V$)
If ($s_{v1}/u_{v1} < s_{v2}/u_{v2}$)
    $s_{v2} \leftarrow s_{v2} + s_{v1}$
    $u_{v2} \leftarrow u_{v2} + u_{v1}$
    $V \leftarrow V - v_1$
Else
    $s_{v1} \leftarrow s_{v1} + s_{v2}$
    $\mathrm{u}_{v1} \leftarrow \mathrm{u}_{v1} + \mathrm{u}_{v2}$
$V \leftarrow V - v_2$
```

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

vowel it found. The initiator in turn listens to this signal and finds the closest vowel in its repertoire. If this vowel is the same as the vowel the initiator initially selected, the imitation game is considered to be successful. If it is not the same, it is considered to be a failure. This information is communicated to the other agent through nonverbal feedback. Direct nonverbal feedback might be considered unrealistic, as human children, when learning a language hardly get any direct feedback about the sounds they produce (although parents do seem to exaggerate the distinctiveness of their vowels, Kuhl, Andruski, Chistovich, Chistovich, Kozhevikova, Rysinka, Stolyarova, Sundberg & Lacerda, 1997). However, such feedback could also be derived from context (e.g., from the failure to achieve a communicative goal or through facial expressions). In any case, the nonverbal feedback is a simplification of a process that is much more complex in reality.

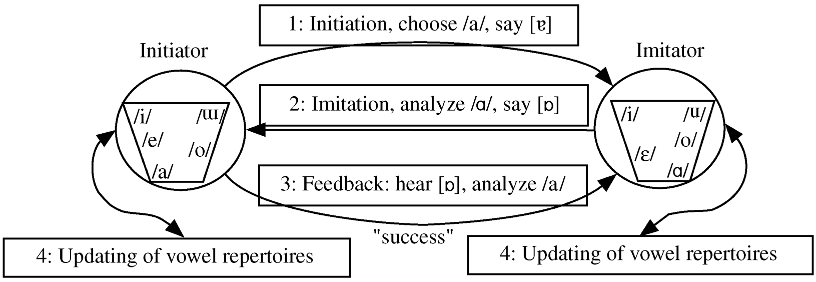

1: Initiation, choose /a/, say [ɐ]
Initiator
Imitator
2: Imitation, analyze /ɑ/, say [ɒ]
3: Feedback: hear [ɒ], analyze /a/
/i/
/ɯ/
/e/
/o/
/a/
/i/
/u/
/o/
/ɛ/
/ɑ/
"success"
4: Updating of vowel repertoires
4: Updating of vowel repertoires

Figure 2. Example of the imitation game. First, the initiator chooses a random vowel (in this case /a/) from its repertoire, produces it with its synthesizer, adding noise (it becomes [ɐ]). Secondly, the imitator analyzes this sound in terms of *its* vowels and synthesizes the recognized vowel (/ɑ/) also adding noise (it becomes [ɒ]). Then the initiator listens to the imitator's sound, analyzes it, and checks if the recognized vowel is the same as the original one (here, [ɒ] is analyzed as /a/, so the game is successful). If the [ɒ] had been perceived closer to /ɔ/, then the game would have been a failure. The vowel systems shown are representative examples. In reality, agents' vowel systems can contain all possible vowels and may contain different numbers of vowels.

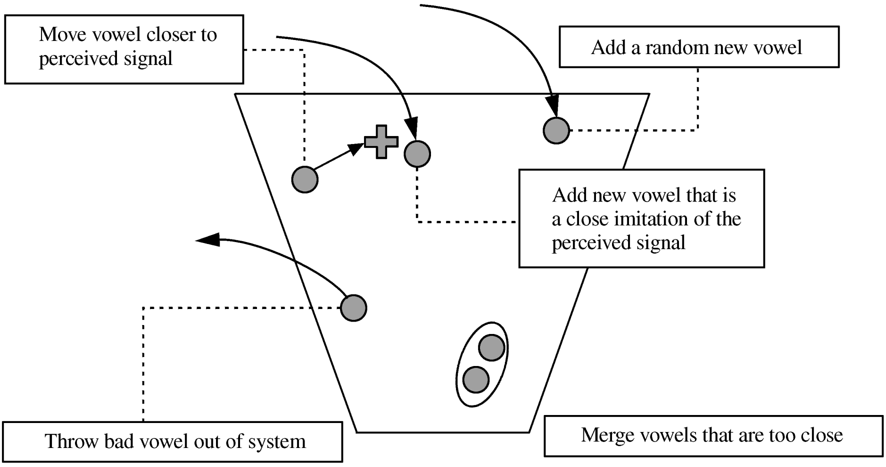

Move vowel closer to perceived signal
Add a random new vowel
Add new vowel that is a close imitation of the perceived signal
Throw bad vowel out of system
Merge vowels that are too close

Figure 3. Changes an agent can make to its vowel system. Circles indicate vowels in the agent's repertoire (both articlatory and acoustic aspects) while the cross indicates the position (in acoustic space) of the signal the agent just perceived.

However, it is thought to capture the essentials for this simulation. The whole process is illustrated with an example in Fig. 2.

In reaction to the imitation game, the agents undertake several actions (described in routine "Update according to feedback signal" in Table IV and illustrated in Fig. 3). Both the imitator and the initiator keep track of the number of times each vowel is used (u_v in the table) and the number of times it was used successfully (s_v). The imitator also changes its vowel repertoire in reaction to the imitation games. If the imitation game was successful, it shifts the vowel it used so that it matches the signal that it heard more closely. This is done by making slight modifications to the articulatory prototype of the vowel, synthesizing this and listening to see whether the resulting acoustic signal is closer to the perceived signal (described as routine “shift closer” in Table IV). In this routine, the six neighbors of a vowel are the vowels that differ from it by adding and subtracting a small value (typically 0.1) to the values of the three articulatory parameters.

If the imitation game was not successful, there are two possible courses of action. If the vowel’s success/use ratio was lower than a certain threshold (chosen to be 0.5 in the simulations) it is probably just a bad vowel. An attempt to improve this vowel will be made by shifting it closer to the signal that was perceived. If, on the other hand, its success/use ratio is higher than the threshold, this means that it has been successful in previous imitation games. It is therefore probably a good imitation of vowels from at least some of the other agents in the population. The most likely cause of the failure of the imitation game is that the other agent had more vowel prototypes at the place where this agent only had one. The imitator therefore adds a new vowel prototype that is a good imitation of the perceived signal, again by talking and listening to itself.

Further modifications of the agents’ vowel repertoires are made independently of the imitation games. First of all, the agents regularly discard vowels whose success/use ratio is below a preestablished threshold (chosen to be 0.7 in the simulations). In order to give vowels a fair chance to be tested in a number of imitation games, they are only discarded if their score is still too low after they have been used at least 5 times. Also, vowels whose prototypes come so close together in either articulatory or acoustic space that they will be confused too easily with each other because of the noise that is added, are merged. This is done by discarding the vowel with the lowest success/use ratio and keeping the one with the highest ratio. The new use and success counts are the sums of the original counts. Finally, in order to keep pressure on the agents to increase the size of their repertoires, random vowels are added with a small probability. This probability is set to be 1% for most experiments in this paper.

All the actions in the imitation game make use of only local information and the signals the agents can observe. The agents cannot “look inside each other’s heads”. Furthermore, they do not perform any global optimization of their vowel systems. The actions that the agents perform are therefore cognitively plausible, meaning that humans could perform them in principle. Although it is not claimed that the model presented here is an accurate model of how humans learn vowel systems, it probably does capture the most important aspects of this process. It is therefore a good model for testing whether self-organization can explain the universal tendencies of human vowel inventories.

4. Results

The agents all start out with an empty vowel repertoire. By playing imitation games with one another, the agents have to develop a vowel system that is as large as possible, that allows for successful communication and that should be realistic if self-organization is really a factor in explaining the structure of human vowel systems.

4.1. *Emergence of a vowel system*

The emergence of a vowel system in a population of 20 agents under 10% acoustic noise is shown in Fig. 4. In each of the frames of the figure, the acoustic aspects of the

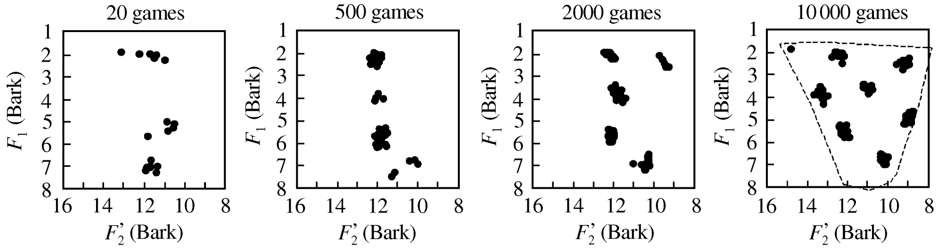


Figure 4. Emergence of a vowel system in a population of 20 agents. The approximate part of the acoustic space that can be reached is indicated in the rightmost frame. All acoustic prototypes of all agents in the population are superimposed.

prototypes of the agents' vowels in the population are plotted in the F_1–F'_2 space. Each prototype is represented by a dot. Note that due to articulatory constraints, only a roughly triangular area of the acoustic space is available to the agents. This is indicated in the fourth frame of Fig. 4.

From the figure it is clear that after the first 20 games the agents still only have very few vowels. The vowels that exist are more or less randomly dispersed through the acoustic space, although some of them already show a tendency to cluster. This is caused by the fact that all agents start out with an empty vowel repertoire. In order to get the imitation games started, random vowels are inserted. However, the imitating agents in the games try to make imitations that are as close as possible and add these to their vowel repertoires. This accounts for the clustering. After some 500 imitation games, shown in the second frame, the clustering has become more pronounced. The most important process at this moment is the compacting of the clusters due to the fact that the agents move their vowel prototypes closer to the signals they perceive. However, there is still sufficient room in the auditory space for extra vowels, so the random addition of new vowels also plays a role. After 2000 games, the available vowel space becomes filled more evenly with vowels and the shape of the vowel system becomes more realistic. After 10 000 imitation games, the available acoustic space has become more or less filled up with vowels and the vowel system has become realistically symmetric and dispersed. After this has happened, the vowel system remains stable. However, it is not static. The vowel prototypes of agents (and therefore the clusters) tend to move, and it is even possible that they merge or that new clusters are formed (if they do not interfere with other clusters).

The way in which vowel systems emerge in the simulation is not realistic. Children probably learn vowel systems in a different way (e.g., see Kuhl & Meltzoff (1996), where it appears that infants' vowel systems gradually expand from a homogenous beginning towards the language's vowel system). Also, human vowel systems evolve in more complex ways than do the vowel systems that emerge in the simulation. The purpose of the simulations, however, is not to model the historical *evolution* of vowel systems. Many more complex mechanisms play a role in historical language change than can be modeled by a simple computer simulation. Its purpose is to show that interactions between *individual* speakers can cause organization on the scale of the *population* and that the organization that emerges is similar to the organization one finds in human vowel systems.

4.2. *Evaluation of the emerged vowel system*

Two questions immediately arise. Are the vowel systems that emerge realistic and how sensitive are they to the settings of the parameters of the simulation? The first question can be answered partly by looking at different properties of the emerged vowel systems. The second question will be addressed in the next section.

For assessing the properties of the emerged vowel systems, a number of objective measures must be defined. The average size of the agents' vowel systems can be measured easily. The average number of times that the imitation games played by all agents were successful (*success*) can be determined in a straightforward way as well. The realism of the emerging vowel systems, on the other hand, is more difficult to measure. Two methods of assessing the realism of emerged vowel systems will be used in this paper. The first is an energy function, the same that was used by Liljencrants & Lindblom (1972) for optimizing vowel configurations. The second is classification and comparison with human vowel systems. This will be done in Section 4.4.

The energy of a vowel system is calculated as the sum over the reciprocal of all squared distances between all vowels in the system:

$$E = \sum_{i=1}^{n-1} \sum_{j=0}^{i-1} \frac{1}{r_{ij}^2} \tag{6}$$

where r_{ij} is the distance between vowels i and j. This distance is calculated with formula (5). The value of λ is 0.3. The average energy of the vowel systems of all the agents in the population can be used as a measure of the realism of the vowel systems in the population. This is because minimizing the energy function has been shown to result in realistic vowel systems. Systems that have a near-minimal energy will therefore be expected to be realistic as well. The values of these measures for the parameter settings that were used in Fig. 4 are shown in Fig. 5. The graphs were obtained from running the simulation 1000 times for 5000 imitation games and collecting the measures at the end of these 5000 imitation games. It can be seen that the success is always very high. It lies somewhere between 0.88 and 1.00 with the highest peak at perfect success. This means that the imitation games are always very successful. The size distribution of the emerging vowel systems is more interesting. Although many different values of average sizes appear, the distribution shows very clear peaks at integer values of the sizes 4–8. This is

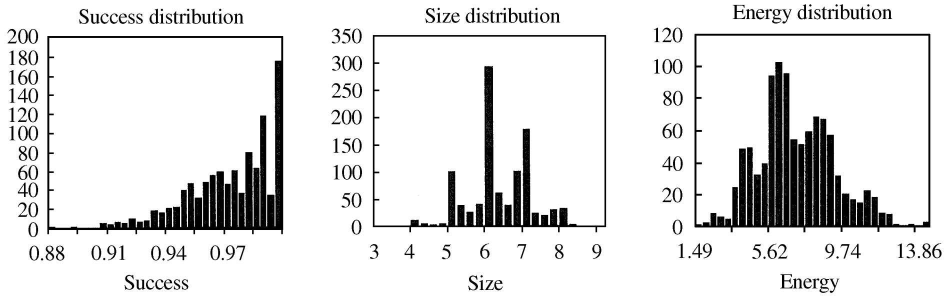


Figure 5. Distribution of measures for simulations with 10% acoustic noise. Note the high success, and the preference for integer average vowel system sizes with the corresponding peaks in the energy distribution. This is caused by the fact that the system converges to systems where all agents have nearly the same vowel system.

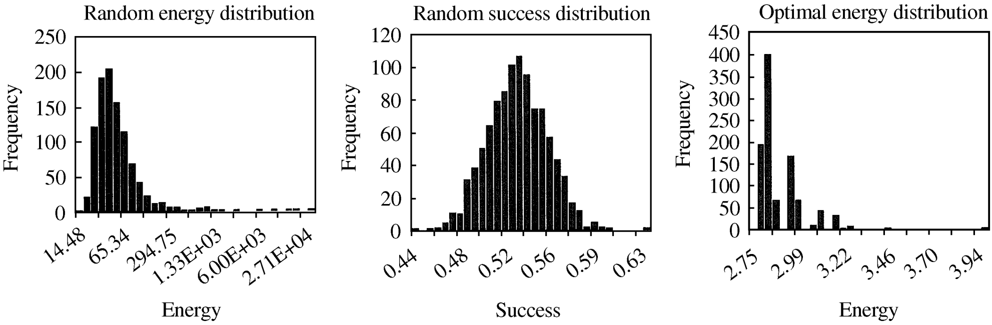


Figure 6. Comparison of random and optimal six-vowel systems. Note that optimal energy is orders of magnitude lower than random energy, but that imitation success in random vowel systems is around 50%.

because populations tended to have agents that all had the same number of vowels. This indicates uniformity of the vowel systems within the populations, but apparently the simulation did not always lead to the same vowel systems. Furthermore, the vowel inventories are reasonably large, with six vowels being the most frequently occurring size (but as will be shown later, this depends on the parameter settings). The energy distribution also shows peaks. These peaks correspond to the different vowel system sizes and to the different configurations for one given size. But the important question is how these energy values compare with the optimal vowel systems and randomly generated systems. The distributions of energy of random and optimal systems with six vowels are presented in Fig. 6. Optimal systems have been determined in the same way as Liljencrants & Lindblom (1972) optimized their systems, except that the vowels were moved in *articulatory space* and not directly in acoustic space. The distribution of success of random systems is also given in this figure. It can be seen that the energy of the emerged systems is much lower than the energy of the random systems and compares favorably with optimal systems.

From detailed comparison of emerged, optimal and random systems (de Boer, 1999, Section 4.2, Appendices B and C), it can be concluded that the energy, and therefore the dispersion of the emerged vowel systems is much better than random and only slightly worse than optimal. The emerged vowel systems are therefore likely to be realistic. Also, the size of the emerged systems is sufficiently large to be realistic and the communicative success of the agents is much better than would be the case if the vowel systems were chosen randomly. These results are highly statistically significant ($p < 0.01$) if one compares the distributions with the Kolmogorov–Smirnov test (tests that assume a normal distribution cannot be used with the kinds of distributions that are found here). These results are interesting, because they show that the emerging vowel systems are optimized with respect to size and success of imitation. This is an emergent result, as no explicit optimization is being done.

4.3. *Sensitivity to parameter changes*

However, the question remains whether these results are due to fine-tuning of parameters or whether they are an inevitable result of the interactions between the agents. In order to test this, the different parameters of the system were changed and the resulting vowel

JPHO 20000125

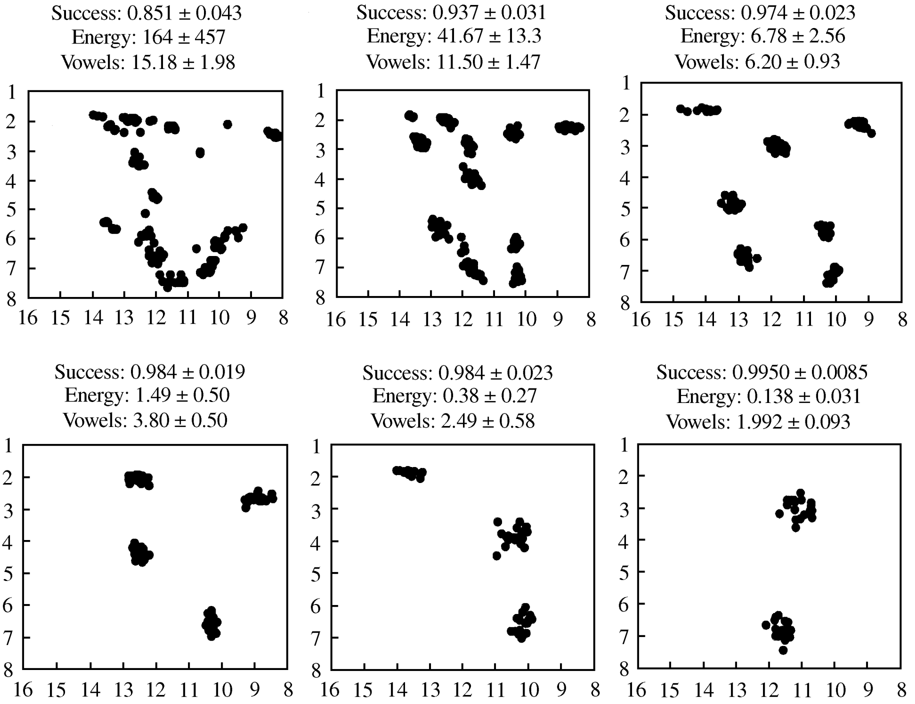


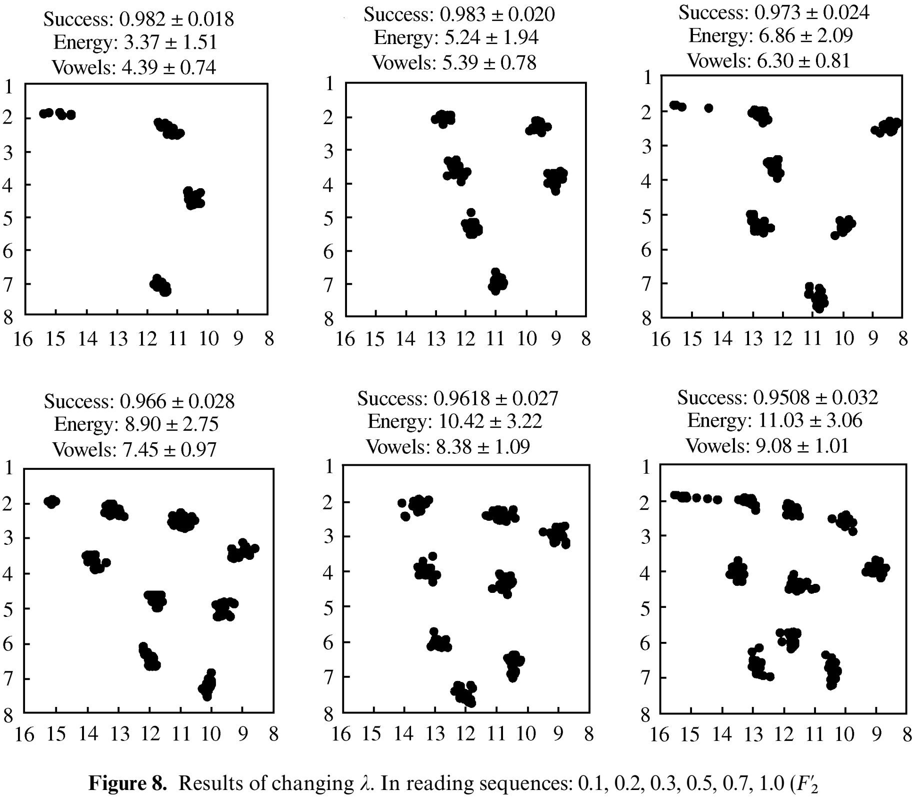
Figure 7. Results of changing acoustic noise. In reading sequence: acoustic noise of 0, 5, 10, 15, 20 and 25% (F'_2 in Bark on x-axis and F_1 in Bark on y-axis). Increasing noise decreases the number of vowels the agents can distinguish reliably.

systems were investigated. As there are quite a few parameters in the system that can be changed, this cannot be presented in detail in this paper. The interested reader is therefore referred to de Boer (1999, Chapter 4) for more details. Three parameters are especially interesting for understanding the simulations. These are the acoustic noise level, the weighting of the first formant relative to the effective second formant and the size of the population.

The result of changing the acoustic noise level is shown in Fig. 7. The projections of the vowel systems in this figure were made in the same way as those in Fig. 4. They show representative vowel systems that emerged for the given noise levels. The value of the formant weighting was set to 0.3 and the population size was 20. The measures were calculated (and the pictures generated) after running 5000 imitation games in the population. The values of the measures that are shown above the graphs were calculated as the average over 100 runs of the simulation for the given parameter settings. The standard deviations are also shown, but it must be kept in mind that the distributions of the measures are *not* normal. It can be seen that the number of vowels per agents decreases with increasing noise level. The success of imitation is lowest for the lowest noise level. Apparently, even though the agents can produce their vowels perfectly, the imitation games result in too many vowels and this causes confusion (hence the high energies). Regarding realism, simulations with acoustic noise between 10 and 20% result in vowel systems that have a realistic number of vowels and that have low energy (note that due to the way energy is calculated, realistic systems with more vowels will have higher energies). The energy is always close to the optimal value and much lower than the values obtained for random systems (for more details, see de Boer, 1999, Appendix B). Apparently, the simulations result in successful vowel systems with low energy for a reasonable range of acoustic noise values.

Another important parameter that determines the shape of the emerging vowel systems is the relative weight of the effective second formant with respect to the first formant (λ in Equation (5)). If this value is low, agents can make fewer distinctions in the effective second formant dimension than in the first formant dimension. If it is high, the situation is reversed. It has been found that humans generally make more distinctions in the height of vowels than in their position in the front–back dimension (Crothers, 1978, universal 9). Therefore, only values for λ that were lower than 1 were tried.

The results are shown in Fig. 8. The pictures and the numerical data have been obtained in the same way as those in Fig. 7. However, this time λ was changed and the acoustic noise was kept constant at 10%. It can be seen that for all values of λ, vowel systems with low energy and high success are obtained. However, the number of vowels tends to increase with λ. It can be observed that the number of vertical distinctions remains approximately the same. It is the number of horizontal distinctions that increases and that causes a corresponding increase in the total number of vowels. With respect to realism, one could say that the vowel systems obtained for $\lambda = 0.1$ and 1.0 make unrealistically few and many horizontal distinctions, respectively. The rest of the


Success: 0.982 ± 0.018
Energy: 3.37 ± 1.51
Vowels: 4.39 ± 0.74

Success: 0.983 ± 0.020
Energy: 5.24 ± 1.94
Vowels: 5.39 ± 0.78

Success: 0.973 ± 0.024
Energy: 6.86 ± 2.09
Vowels: 6.30 ± 0.81

Success: 0.966 ± 0.028
Energy: 8.90 ± 2.75
Vowels: 7.45 ± 0.97

Success: 0.9618 ± 0.027
Energy: 10.42 ± 3.22
Vowels: 8.38 ± 1.09

Success: 0.9508 ± 0.032
Energy: 11.03 ± 3.06
Vowels: 9.08 ± 1.01

Figure 8. Results of changing λ. In reading sequences: 0.1, 0.2, 0.3, 0.5, 0.7, 1.0 (F'_2 in Bark on x-axis and F_1 in Bark on y-axis). Increasing λ increases the number of horizontal distinctions. Values from $\lambda = 0.2$ to 0.7 result in realistic systems.

vowel systems could occur in human languages, but the most realistic results probably occur for $\lambda = 0.3$. This was therefore the value that was used for all the other experiments discussed in this paper.

It might seen problematic to tune with data from human vowel systems a parameter that is used in a model for explaining the properties of these very human vowel systems. This introduces a certain circularity. However, there are three reasons why this is, in fact, not a problem. First of all, the realism of the emerging vowel systems does not depend critically on the value of λ, as can be seen in Fig. 8. Secondly, the model's primary contribution is to elucidate the role of self-organization, and not to predict the vowel systems of human languages exactly. Tuning the parameter therefore serves to bring the model's results more in line with what is found in human languages, but does not alter them qualitatively. Finally, and most importantly, there is language-independent psychophysiological evidence that humans are approximately 3 times as good at perceiving and producing vowel height distinctions compared to vowel position distinctions (see, e.g., Lindblom & Lubker, 1985). Therefore, the value of 0.3 for λ can also be derived independently.

It is also very important to investigate the influence of the population size on the results of the simulation. Is the organization of the vowel systems really caused by the interactions between the agents or is it caused by the individual actions of the agents? In the latter case, one cannot really speak of self-organization. Such was the case in Berrah's (1998) work. Therefore, experiments must be done with different population sizes. In order to compare populations of different sizes in a fair way, some small changes must be made in conducting the experiments. First of all, it is fairer to keep the number of games per agents, rather than the absolute number of games, constant. Therefore, the smaller populations play fewer games in total than the larger populations. But this means that the probability that new random vowels are added to the population (set to 1% per game played in the populations with 20 agents) must also be made dependent on the number of games per agent, rather than the absolute number of games played to keep the average number of random insertions constant. The probability of adding new vowels was therefore set to $0.2/N$, where N is the number of agents in the population. This formula was chosen such that it results in a 1% insertion probability for a population of 20 agents, as was used in the previous experiments.

It was found that changing the population size did not so much change the shape of the emerging vowel systems, as their stability. For this reason the evolution over time of the vowel systems of two populations is presented in Fig. 9. It can be seen that the population consisting of only two agents does not converge towards a stable system, whereas the vowel system of the population of 20 agents remains completely stable. Apparently, stable vowel systems can only emerge in a larger population, thus stressing the importance of the interactions between the agents. The better performance of the larger populations is also illustrated by the statistical results presented in Table VI. The success of all population sizes is comparable, but the vowel system size of small populations is smaller than that of large ones, reflecting the lower stability. This result is significant for the Kolmogorov–Smirnov test, $p < 0.01$.

4.4. *Comparison with human vowel systems*

So far, emerged vowel systems have only been compared with real human vowel systems in an indirect way. Their energy was compared with that of optimal and random systems

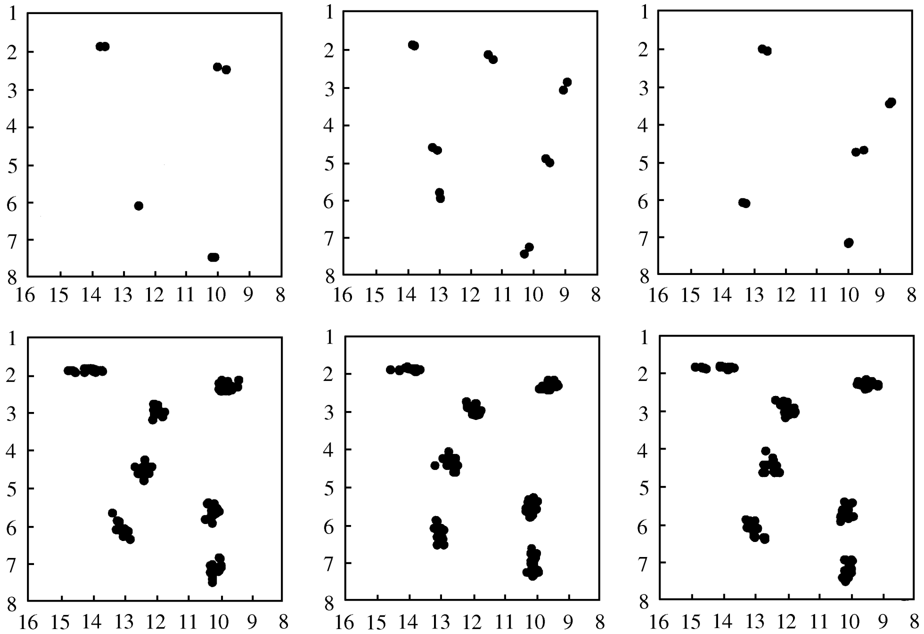


Figure 9. Populations of 2 (top) and 20 (bottom) agents shown at 250, 350 and 450 games per agent from left to right (F'_2 in Bark on x-axis and F_1 in Bark on y-axis). Note the relative instability over time of the vowel system in the small population.

TABLE VI. Quality measures for different population sizes

Pop. size	Success	Energy	Size
2	0.975 ± 0.016	6.06 ± 2.83	5.99 ± 1.16
5	0.971 ± 0.021	6.95 ± 2.40	6.36 ± 0.89
10	0.969 ± 0.026	6.72 ± 2.25	6.23 ± 0.81
20	0.978 ± 0.020	6.61 ± 2.25	6.18 ± 0.85
50	0.974 ± 0.022	7.68 ± 2.44	6.53 ± 0.81
100	0.975 ± 0.023	7.85 ± 2.67	6.51 ± 0.97

with the same number of vowels. Although there is a link between the realism of vowel systems and low energy, low energy does not necessarily mean that a given vowel system is also likely to occur in human languages. Therefore, a classification of the emerged vowel systems was made and compared with similar classifications of human vowel systems. If in the emerged systems the same types of vowel systems are found in the same proportions as in human languages, they are realistic. The classifications of human vowel systems that were used as a reference were the ones of Crothers (1978) and of Schwartz, Boë, Vallée & Abry (1997*a*). These classifications are based on the most frequent phonetic realizations of phonemes in a language, rather than on all possible allophonic realizations. This means that a certain abstraction of the actual signals has to be made before one can classify. Especially in Crothers' work, the relative positions of vowels in a vowel system are used for classification rather than their actual absolute phonetic

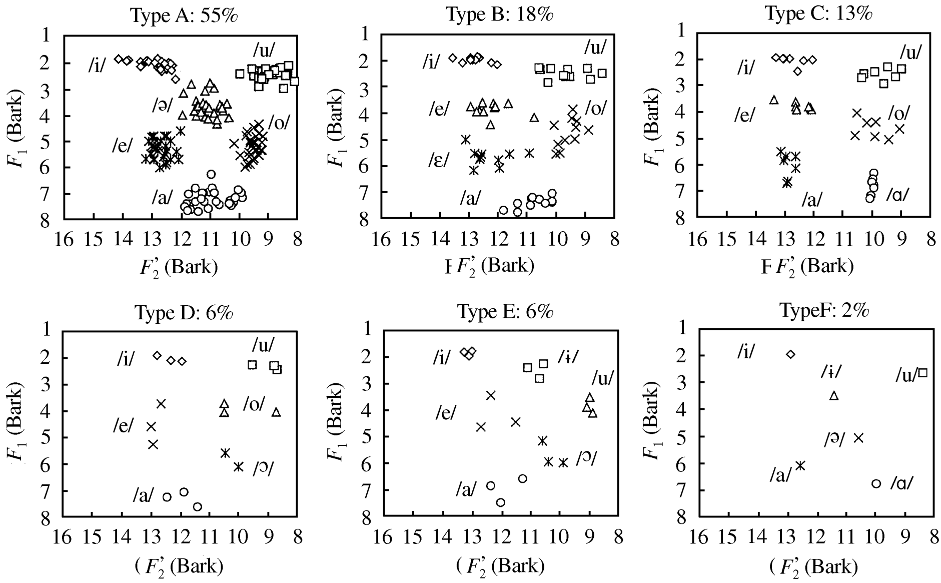


Figure 10. Classification of six-vowel systems. In contrast with the previous figures, the agents plotted in each frame come from different populations, and have never interacted. The fact that their vowel systems are still very similar indicates that the above systems are attractors of the imitation game. These systems and their frequency are remarkably similar to those found in human languages.

realizations. This has also been done in the classification of the emerged systems. Thus, the two three-vowel systems consisting of [i], [a], [u] and [e], [a], [o], respectively, are assigned to the same category, as they are both symmetrical triangular systems consisting of three vowels.

An example of a classification of emerged vowel systems containing six vowels is shown in Fig. 10. The data in this figure were obtained by running the simulation 100 times for a given parameter setting (acoustic noise set to 12%). In each run, 25 000 imitation games were played. From the results, the populations whose agents had on average six vowel prototypes were selected (there were 54 of these). From these populations a random agent with six vowel prototypes was selected. This agent's vowel system was then classified with the systems of other agents from other populations according to its shape. Note that although the frames in Fig. 10 look similar to the figures of vowel systems shown previously, there is a crucial difference. In previous figures, the agents shown in one frame were all members of the same population (and had therefore interacted with each other). In the present figure, all agents shown are members of different populations. The fact that there still is a large amount of similarity between the vowel systems of agents from different populations is a strong demonstration of the self-organization that makes populations converge towards similar vowel systems.

The emerging systems are realistic. Most of them conform to the universals that Crothers (1978) found for human vowel systems. Emerged vowel systems of types A, B, C, and E (a total of 92%) conform to all the universals. Types D and F do not conform to Crothers' universal 4 (they do not contain /e/ or /i/). However, not all human vowel systems conform to all of Crothers' "universals" either. When the percentages with which the different emerged systems occur are compared with the percentages with which human vowel systems occur, a good match is found as well. Schwartz *et al.* (1997*a*) have measured the occurrence of different vowel system types in the different languages in UPSID. They find 60 vowel systems with six vowels. Although their classification is not exactly the same as the classification of Fig. 10, there is good agreement. Of the systems they found in UPSID, 43% is of type A, 20% is of type B, 5% is of type C, 7% is of type D and 20% is of type E (if [ɨ] nd [ɯ] are classified as the same). No systems of type F were found, and two of the systems from UPSID (3%) cannot easily be classified in the classification used here, but are probably of type A. Similarly, good agreement was found for systems of 4 and 5 vowels and reasonable agreement was found for systems of 7 and 8 vowels (de Boer, 1999, Chapter 6). For systems of 3 vowels, too many "vertical" systems were found. This might have to do with a discontinuity in the perception function (Equations (2)–(4)) which made it more difficult for agents to discover the unrounded high front vowel [i]. However, it does seem that the simulation is capable not only of predicting the most frequently found vowel systems in human language (as was already possible with systems that optimize acoustic distinctiveness), but also of predicting the less frequently occurring vowel systems and approximately their relative abundance.

Human vowel systems have a very strong tendency towards systems with five vowels. The question can be asked whether the self-organizing vowel systems in the simulation presented here also have a preference for a certain size. This is indeed the case. However, the preference is not for systems with five vowels, but for systems with four vowels. This is illustrated in Fig. 11. The solid line shows the distribution of sizes for human vowel systems, based on data from Schwartz *et al.* (1997*a*). The dashed line shows data obtained from running the simulation for a large number of different values for the acoustic noise and by calculating the frequency with which different vowel system sizes emerged. The simulation was run for acoustic noise values of 8–24% with steps of 1%. For each value, 100 runs consisting of 25 000 imitation games were done. At the end of each run, the number of vowels in the emerged vowel system was measured and stored. The graph shows the numbers of times each size occurred. It can be observed that the distribution has a peak at systems of size four. What this result means and why the preferred size is four and not five are not quite clear (although the preference for systems of size 4 might be caused by the above-mentioned discontinuity in the perception function). Perhaps by tuning the value for λ the preferred size might be increased from four to five. This has

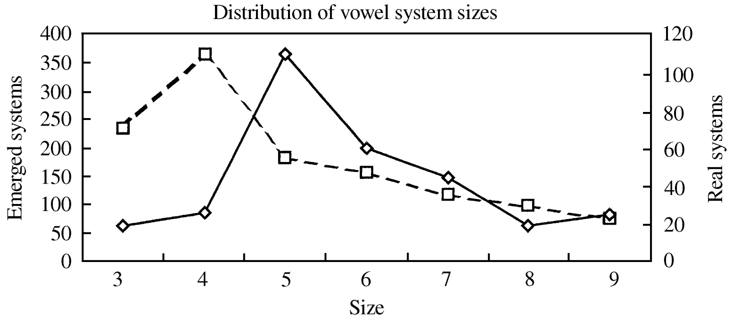


Figure 11. Distribution of vowel system sizes. Continuous line indicates real systems, dashed line indicates emerged systems. Both distributions have a peak that does not fall at the extremes.

not been done, because this would be manipulating the simulation to obtain the results one wants to explain. The simulation results do seem to indicate, however, that the preference for a vowel system size that is intermediate between the absolute maximum and the minimum number of vowels can also be explained as a result of self-organization.

5. Conclusions

The results of the simulation show that vowel systems can be predicted as a result of self-organization in a population of agents. Although the agents start with an empty vowel repertoire and do not have any constraints on the kinds of vowel systems they can learn, the vowel systems that emerge in the population tend to be symmetric and dispersed, just as predicted by explicitly optimizing methods. The performance of the self-organizing model is even better than that of optimizing models, as it also predicts the types of vowel systems that occur less frequently and with a reasonable degree of accuracy their relative abundance. This illustrates that self-organization might play an important role in determining the structure of phonological and phonetic systems. This was already suspected for a long time (see, e.g., Lindblom *et al*., 1984), but the computer model presented here provides powerful empirical support for this thesis, as well as a means by which self-organization of vowel systems in a population can be studied systematically.

Self-organization in a population under constraints of perception and production causes systems of vowels that are acoustically dispersed to be favored over systems that are less dispersed. However, as the imitation success of agents is also determined by how well they conform to the population, different (sub-optimal) configurations can emerge and be maintained. Some configurations can be considered stronger attractors than others in the dynamical system that is defined by the agents and their interactions.

The same universal tendencies as found in human vowel systems are found in the systems that emerge. This indicates that innate rules and constraints are not necessary to explain them. They are an emergent result of the way human perception and production work. Fewer innate cognitive structures make it easier to construct an evolutionary account of the origins of speech (and language).

However, it is not claimed that production and perception are the only factors that determine the shape of human vowel systems, let alone the shape of the complete sound systems of human languages. Cognitive factors (e.g., learnability) and processes (e.g., analogy), as well as historical factors (the way a language changed over time), all play an important role in explaining the structure of human sound systems. However, self-organization "filters" the output of these processes and causes them to work in certain directions more often than in others.

Many issues have not yet been addressed by the work presented here. First of all it is limited to vowel systems, and to vowels uttered in isolation. This is obviously a very strong limitation and arguably unrealistic. Ideally, the model should be tested with more complex utterances (e.g., combinations of vowels and consonants). More complex utterances and speech sounds produced in sequence would also make it possible to investigate more realistic developments of the emerging sound systems. Unfortunately, the task of implementing this in a computer simulation has so far proven to be too complex. However, in the light of the success of the present experiments with vowels, this is definitely a goal worth pursuing.

The work described here has shown that it is possible to explain the universal tendencies of vowel systems as a result of self-organization in a population under constraints of perception and production. This eliminates the need to postulate an innate predisposition towards certain vowel systems, as well as the need for explicit optimization by language users. The work has also shown that computer simulations of language use in a population can make valuable contributions to the understanding of language in general, and more specifically of the structure of sound systems. It makes it possible to do many controllable and repeatable experiments in a short time, thus allowing for a rapid exploration of hypotheses.

This research was done at the Artificial Intelligence Laboratory of the Vrije Universiteit Brussel. It is part of an ongoing research project into the origins of language and intelligence. Funding was provided by the GOA 2 project of the Vrije Universiteit Brussel. Part of the work was done at the Sony Computer Science Laboratory in Paris, France.

I would like to thank Francesco Lacerda, Jean-Luc Schwartz and Amanda Walley for providing very many helpful suggestions and comments for improving the manuscript.

References

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

Berrah, A. R., Glotin, H., Laboissière, R., Bessière, P. & Boë, L. J. (1996) From form to formation of phonetic structures: an evolutionary computing perspective. In T. Fogarty & G. Venturini (editors), ICML '96 workshop on Evolutionary Computing and Machine Learning, Bari, pp. 23–29.

Carlson, R., Granström, B. & Fant, G. (1970) Some studies concerning perception of isolated vowels, *Speech Transmission Laboratory-Quarterly Progress and Status Report* (*STL-QPSR*), **2–3,** 19–35.

Carré, R. (1994) 'Speaker' and 'speech' characteristics: a deductive approach, *Phonetica*, **51,** 7–16.

Chomsky, N. & Halle, M. (1968) *The sound pattern of English*. Cambridge, MA. MIT Press.

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

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J. (1952) Some experiments on the perception of synthetic speech sounds, *Journal of the Acoustical Society of America*, **24,** 597–606. Reprinted In *Acoustic Phonetics* (D. B. Fry, editor), pp. 258–272. Cambridge: Cambridge University Press.

Crothers, J. (1978) Typology and universals of vowel systems. In *Universals of human language*, (J. H. Greenberg, C. A. Ferguson & E. A. Moravcsik, editors) Vol. 2, Phonology, pp. 93–152. Stanford: Stanford University Press.

de Boer, B. G. (1999) *Self organization in vowel systems*. PhD thesis, AI-Lab, Vrije Universiteit Brussel.

Everett, D. L. (1982) Phonetic rarities in Piraha, *Journal of the International Phonetic Association*, **12**(2), 94–96.

Firchow, I. & Firchow, J. (1969) An abbreviated phoneme inventory, *Anthropological Linguistics*, **11,** 271–276.

Frieda, E. M., Walley, A. C., Flege, J. E. & Sloane, M. E. (1999) Adults' perception of native and nonnative vowels: implications for the perceptual magnet effect, *Perception and Psychophysics*, **61**(3), 561–577.

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

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

Jakobson, R. & Halle, M. (1956) *Fundamentals of language*. The Hague: Mouton & Co.

Kuhl, P. K. & Meltzoff, A. N. (1996) Infant vocalization in response to speech: vocal imitation and developmental change, *The Journal of the Acoustical Society of America*, **100**(4), 2425–2438.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevikova, E. V., Rysinka, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F. (1997) Cross-language analysis of phonetic units in language addressed to infants, *Science*, **277,** 684–686.

Ladefoged, P. & Maddieson, I. (1996) *The sounds of the world's languages*. Oxford: Blackwell.

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

Langton, C. G. (editor) (1989) *Artificial life*. Reading, MA: Addison-Wesley.

Liberman, A. M., Delattre, P. C., Cooper, F. S. & Gerstman, L. J. (1954) The role of consonant–vowel transitions in the perception of the stop and nasal consonants. *Psycholofical Monographs* **68**(8) Reprinted In *Acoustic phonetics* (D. B. Fry, editor), pp. 315–331. Cambridge: Cambridge University Press.

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

Lindblom, B. & Lubker, J. (1985) The speech homunculus and a problem of phonetic linguistics. In *Phonetic linguistics*: *essays in honor of Peter Ladefoged* (V. A. Fromkin, editor), pp. 169–192. Orlando: Academic Press.

Lindblom, B., MacNeilage, P. & Studdert-Kennedy, M. (1984) Self-organizing processes and the explanation of language universals. In *Explanations for language universals* (B. Butterworth, B. Comrie & Ö. Dahl, editors), pp. 181–203. Walter de Gruyter & Co.

Lindblom, B. & Maddieson, I. (1988) Phonetic universals in consonant systems. In *Language*, *speech and mind* (L. M. Hyman & C. N. Li, editors), pp. 62–78.

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

Maddieson, I. & Precoda, K. (1990) Updating UPSID, In *UCLA Working Papers in Phonetics*, **74,** 104–111.

Maeda, S. (1989) Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In *Speech production and speech modelling* (W. J. Hardcastle & A. Marchal, editors), pp. 131–149. Dordrecht: Kluwer.

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 *Proceedings of the 15th journées d'étude sur la parole.* Société Française d'Acoustique, pp. 157–161.

Nicolis, G. & Prigogine, I. (1977) *Self-organization in non-equilibrium systems.* New York: John Wiley.

Schroeder, M., Atal, S. & Hall, J. L. (1979) Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In *Frontiers of speech communication research* (B. Lindblom & S. Öhman, editors), pp. 217–229. London: Academic Press.

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

Schwartz, J. L., Boë, L. J., Vallée, N. & Abry, C. (1997*b*) The dispersion–focalization theory of vowel systems, *Journal of Phonetics*, **25,** 255–286.

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

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

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

Stevens, K. N. (1989) On the quantal nature of speech, *Journal of Phonetics*, **17**(1), 3–45.

Trudgill, P. (1995) *Sociolinguistics*: *an introduction to language and society.* London: Penguin Books.

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

Vennemann, T. (1988) *Preference laws for syllable structure.* Berlin: Mouton de Gruyter.

Wittgenstein, L. (1967) *Philosophische untersuchungen.* Frankfurt: Suhrkamp.

AUTHOR QUERY FORM

HARCOURT PUBLISHERS

JOURNAL TITLE: JPHO DATE: 8/11/2K
ARTICLE NO. : 20000125

Queries and/or remarks

Manuscript Page/line	Details required	Author's response
	Pl check table 3. Matter in shaded area is not clear	