Embodiment and Self-organization of Human Categories: A Case Study for Speech

Luc Steels and Bart de Boer

Abstract

The paper considers explanations for the kinds of categories that have been found to be involved in human behavior. It insists that embodiment not only plays an important role in shaping these categories, but also that the collective dynamics generated by social interaction is of equal importance. A case study is developed for speech sounds. We show through computer simulations how a group of autonomous agents equipped with a (sufficiently) realistic perceptual and auditory apparatus can arrive at a shared repertoire of vowels and that these vowel systems exhibit the same universal trends that are found in human vowel systems. It is significant that this happens without innate a priori categories.

Keywords: Embodiment, speech, self-organization, categorization.

1. Introduction

The term "embodied" is being adopted by a growing number of researchers interested in cognition (Ziemke 2003), but often with different meanings. At least three senses of embodiment are being used:

- The first one emphasizes that cognitive processes, such as those required for vision or language, are to be implemented in terms of networks that have neural plausibility, both in terms of their computational microstructure and in terms of the general architecture of the brain. While there is of course a wide consensus that cognitive processes must somehow be mapped onto brain hardware, the "strong embodiment" hypothesis claims more, namely, that without this neural realism many cognitive phenomena cannot be understood. This line of reasoning finds its most enthusiastic defenders among philosophers like Searle and those who argue that the brain should not be viewed as an information processor at all (e.g. Edelman 1990).

- The second type of embodiment research first manifested itself in “behavior-based AI” (Steels & Brooks 1995). It argues that part of our interaction with the world is not explicitly controlled by any kind of information processing. Instead, the physical properties of the body, the physical characteristics of sensors and actuators, the structure of the task, and the physical properties of the environment all play an important role in shaping actual behavior. This line of research has proven to be very fruitful for designing and building autonomous robots that interact with real-world environments in real time (Pfeifer & Iida 2004). A neat example is the robot Stumpy that exhibits locomotion without any kind of explicit control or sensing (Iida, Dravid & Paul 2002).
- The third sense of embodiment refers to the grounding and expression of human concepts. It argues that many concepts, even abstract mathematical concepts, find their ultimate foundation in grounded interaction with the world (Lakoff 2000) and that this is still visible in the metaphors and expressions used to communicate using these concepts. For example, the temporal preposition “back” (as in “back in time”) is historically derived from a body part (the back) which first became used to indicate a spatial relation or area (as in the back of the car) and then a temporal relation (see also Núñez 1999). It follows from this position that conceptualization and concept acquisition must be embedded in a developmental history of embodied interaction with the world.

This paper explores the second view on embodiment, as manifested in “embodied AI”, and tries to demonstrate what kind of explanatory power can be drawn from it. We take the domain of speech sounds as a case study because we believe that without studying concrete cognitive tasks and showing what embodiment can bring, the discussion may become too divorced from reality. Section 2 contrasts classical theories of speech that do not take embodiment into account with theories grounded in the function of speech that do take embodiment into account. Section 3 presents theoretical considerations about what constitutes an explanation in cognitive science as well as an overview of the phonological phenomena under consideration. Section 4 presents a case study in which the role in explaining universals of speech sounds of embodiment and situatedness are investigated with a computer model. Section 5 presents a summary and the conclusion.

2. Embodied versus disembodied speech research

The domain of speech is known to be extraordinarily difficult to understand and model because it not only involves complex fine-grained motor control of the vocal articulators for producing speech sounds, but also complex acoustic signal processing and categorization for recognizing speech sounds. Therefore, there is at least the potential for a strong role of embodiment. At the same time speech also seems to imply abstract categories (in the sense that some distinctions are relevant in a language and others are not) and abstract rules (such as “a voiced consonant becomes unvoiced at the end of a word”, a rule which holds for Dutch but not for English).

There is a long tradition in linguistics, shaped by the Prague structuralists, with Roman Jakobson (Jakobson & Halle 1956) as its most prominent member, and Chomsky and Halle's influential work on abstract phonology (Chomsky & Halle 1968), that completely ignores the embodied aspects of speech. Instead, these linguists move immediately to an abstract level in terms of distinctive features (such as +/– voiced, +/– labial, +/– aspirated, etc.). These distinctive features are assumed to be both grounded in articulation, as well as in acoustic recognition and categorization. For example, voiced means that the vocal chords are vibrating while a sound is being produced, such as in the final consonant of “bad” (voiced) versus “bat” (voiceless). Yet it is never specified in detail how exactly this grounding is realized. Abstract phonologists simply use the abstract features to formulate phonological constraints (e.g. which combinations of consonants and vowels are allowed in a language). But it turns out that it is very hard to recognize these features in the acoustic data and that they are not sufficiently precise to realistically drive motor control. Moreover this branch of phonology often assumes that the features are genetically predetermined as part of the innate language acquisition device (LAD) and argues that systematicity is genetically encoded in terms of a set of principles with parameters to be set by cues from the environment (Dresher 1992).

Abstract phonology is therefore a prime example of disembodied cognitive science and it contrasts with work by another group of phoneticians and phonologists who employ what can be considered an embodied approach to speech (see e.g. Lindblom & Maddieson (1988); Browman &

Goldstein (1992)). They start from the actual acoustic signals and from sophisticated articulatory models, taking constraints coming from the physics of the world and embodiment into account. For example, many co-articulation phenomena are no longer assumed to be governed by rules but simply follow from the dynamics of motor control. Grounding now plays a role in explaining the acquisition of speech sounds or why certain sounds or sound combinations are occurring (MacNeilage 1998). Perception is always coupled tightly to production (Plaut & Kello 1999). The research discussed in this paper fits within this embodied view of speech. Section 3 presents theoretical considerations about what constitutes an explanation in cognitive science

3. Explanatory theories

Before delving into the concrete case study presented in this paper, it is worthwhile to make some general methodological remarks about the nature of theories in cognitive science. It is well known that categorization is a central topic in cognition; indeed we can say that a behavior becomes cognitive only when some form of categorization is implied. Traditionally, linguistics, psychology and anthropology have collected a large number of facts about human categorization. For example, Berlin and Kay (1969) collected data about the kinds of color categories found in human languages by doing naming and memory experiments, as first pioneered by Lenneberg and Roberts (1956). Subjects are shown a series of color samples (the Munsell chips) and are asked to show the best representative for a specific color like “red”. Berlin and Kay found that there are some surprising universal trends in focal colors and their work has generated a wide body of literature that seeks to identify additional evidence which confirms (Kay & Regier 2003) or refutes (Davidoff 2001) these basic human color categories.

Another domain that has been studied extensively is space. Here again research has shown that there are universal tendencies in spatial categories. For example, IN, ON, LEFT OF, etc. are very common categories in human spatial reasoning and language (Herskovits 1986) and so it has been argued that they are largely innate. But on the other hand, comparative studies of the expression of spatial categories has shown significant intercultural differences. Also, children only gradually acquire spatial categories under the influence of their native language (Bowerman & Choi 2001).

Similarly the categories (and hence the articulatory gestures) used in sound systems of human languages show remarkable regularities. Languages do not use random subsets of the many possible speech sounds, but rather certain sounds occur more often than others. Although humans are able to produce and distinguish an amazing number of different speech sounds, the average number of distinctive speech sounds that is used in the languages of the world lies between 20 and 37 (Maddieson 1984). Some sounds, such as [a], [m] or [k] occur almost universally, while others occur only rarely. For example, English does not use the rather rare vowel [y] (pronounced as in French "rue") whereas French (and Dutch) do. Phoneme inventories tend to exhibit symmetries as well. If, for example, a repertoire contains [o] it is very likely to also contain [e]. Such symmetries are also found in consonant systems (Lindblom & Maddieson 1988). Languages furthermore constrain the set of possible sound combinations. For example, in English [mb] can occur at the end of a word as in "lamb" but not in the beginning, whereas in some African languages this is possible (as in Swahili "mbali" (far)). So sounds fall into classes and the classes form a combinatorial system. Again there appear to be universal tendencies in these sound combinations although they have been less adequately surveyed.

Regularities like these (and many others can be demonstrated, including universal tendencies in grammar) demand an explanation. Linguists and psychologists in the disembodied tradition often propose that these universal tendencies are based on innate mechanisms (e.g. the distinctive phonological features and their markedness is assumed to be part of the innate language faculty), and so they are said to be universally shared by all human beings, the same way each of us has (normally) five fingers on each hand. But that means that these theories simply pass the buck to evolutionary biologists. However, biologists are often not so keen about ascribing complex human behavior to innate features, partly because it seems rather implausible that the micro circuitry of the brain is under detailed genetic control (Edelman 1990), and partly because we would still need an evolutionary scenario to explain for instance why [y] (as in French "rue") is not common and [a] (as in English "bar") is, or why a language with four color terms usually has terms for either green or yellow.

So simply saying that human categories are (to a large extent) innate is not a sufficient explanation; we also need to show which neural structures can perform the categorization, how these neural structures could be genetically determined in development and how the relevant genes could have become universally present in human populations. All of these are nontrivial tasks that remain to be carried out. Another difficulty for the genetic position is that without exception, human categorization is not absolutely universal across all human populations and cultures. Rather it shows significant variation. Although this variation is supposedly captured in parameters set during development, it is unclear how such a parameterized system could evolve under realistic selection pressures based on language communication.

If genetics is not very satisfactory as a source of explanations, we should perhaps investigate other ways in which certain categories like colors, spatial categories or speech sounds can become shared (universally or locally). Three possible determining factors have been discussed:

(1) Either human embodiment is so constraining that it is not necessary to make the categories innate because they could never be perceived or the motor movements could never be made. For example, the human vision system restricts perception to a certain range of the spectrum; the human vocal tract is limited to certain number of sounds, the human body suggests a division into front and back, etc. This enables and limits the set of possible solutions for a specific task.

(2) Human beings interact socially and this creates a collective dynamics that may also constrain individual choices, pushing them in the direction of a shared system. For example, in Britain cars drive on the left side of the road and in continental Europe they drive on the right side. This is not because British drivers have “left driving genes” so that they are born to drive on the left, or that their embodiment makes driving on the right impossible, but rather because a cultural choice was made and this choice is enforced by the subsequent behavior of all involved. Even if somebody would like to drive on the right in England, they would be quickly sanctioned or maybe even lose their life.

(3) It could be that the regularities in the environment are such that a statistical learning process picks them out easily and so there would be no need to encode the category in the genome. For example, the chromatic distribution of colors in the world might show enough statistical regularity that statistical clustering algorithms could detect them.

In the work being carried out by our research group, we have been exploring explanations for human categorization in which these three factors play a role. A good example is our research on color categories, in which we have attempted to explain the universal tendencies in color categories using embodiment, the collective dynamics generated by social interaction, and the statistical structures inherent in the world (Steels & Belpaeme 2005). In the domain of language in general and speech sounds in particular, the statistical regularity is not present in the world independently of the agents. So only factors (1) and (2) are relevant, particularly when a totally new system is being bootstrapped.

In the case study that follows, originally described in (de Boer 1999, 2000), embodiment and the dynamics of language use in a population are investigated as factors explaining regularities in systems of human speech sounds. As vowels are the simplest speech sounds to model, this paper focuses on vowels, but there also exists work on syllables and syllable structures (Steels & Oudeyer 2000; Redford, Chen & Miikkulainen 2001).

Our main point is that embodiment in itself is not sufficient to explain many universal tendencies or culturally accepted norms. Earlier embodiment approaches have shown that embodiment constrains the possibilities, e.g. that there are limitations of actuators and sensors which prevent or discourage humans from using certain sounds. In addition, these approaches demonstrate that typical speech categories are in some sense optimal (see e.g. Lindblom, MacNeilage & Studdert-Kennedy. 1984). Specifically the sounds in a language are claimed to have a particular distribution and structure that is optimized for the task of reliably producing and perceiving sounds, and hence the sounds are maximally contrastive. If that is the case, then embodiment and the constraints of the task are sufficient to explain why human languages show particular sounds and not others.

Nonetheless, observations of real phonetic data show that suboptimal phonetic systems also exist in many languages. For example, the UPSID database of human sound systems lists two examples with suboptimal 3-vowel systems, namely with the sounds [e] [a] [o] instead of the more common optimal [i] [a] [u]. They come from a North-American Indian language, Alabaman (Rand 1968) and a Peruvian language, Amuesha (Fast 1953). These 3-vowel systems are suboptimal because they do not provide the maximal contrast in the acoustic space. Therefore embodiment (and the constraints of the task) are in themselves not sufficient to explain the vowel systems one finds in human languages. One also has to take the collective dynamics of language use in a population into account. All users have to conform to the system of speech sounds that has historically been chosen in order to achieve communicative success. This system is partly arbitrary and may be suboptimal.

4. Experiments in self-organization of speech sounds

To examine whether the structure of vowel systems is sufficiently determined by self-organization in a population under constraints of embodiment, we need to find a mechanism that incorporates these two factors, and then show that this indeed leads to emergent vowel systems that reflect the universal tendencies observed in human vowel systems. In our case, we use a population of artificial agents that can produce, perceive and remember speech sounds in a human-like way. 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. These interactions are called imitation games, and are in the same line as the "language games" we have used in our work for the past decade (see Steels 2003). The agents can update their vowel repertoires depending on the outcome of the interactions in such a way that the expectation of future imitation success is maximized. The agents' goal in life is to imitate the other agents as well as possible with a repertoire of vowels that is as large as possible. However in doing this, the agents only use local information, and do not carry out any explicit optimization. The behavior of the agents is summarized in table 1 at the end of section 4.1.

4.1. Embodiment constraints on perception and production

The repertoire of speech signals that humans can produce is limited by the physics and the physiology of the vocal tract (figure 1). The repertoire of sounds that they can distinguish is limited by the properties of the auditory apparatus. In this example the focus is on vowels, and therefore only the articulatory and perceptual properties of vowels are integrated. These properties are also better understood than those of consonants. This has to do with the fact that vowels are static signals, while many consonants are dynamic, i.e. their articulation and their signal do not remain constant over time.

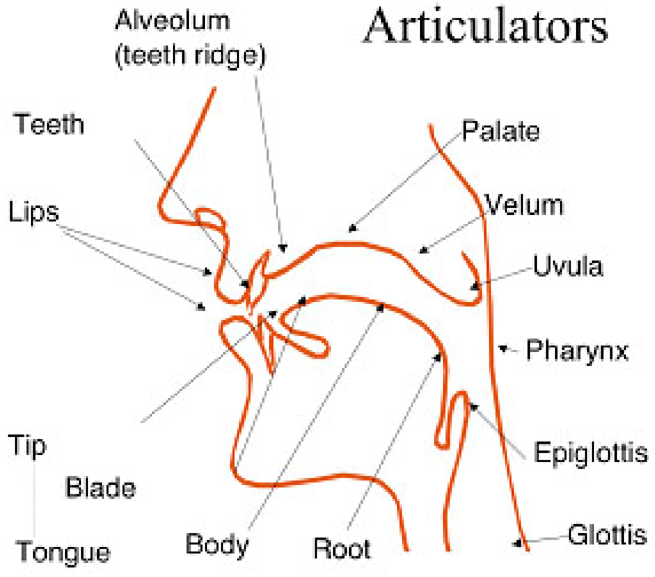

Figure 1. Structure of the human vocal tract. Speech sounds are produced by very fine-grained movements of the articulators

The acoustic signal of a vowel can be described relatively straightforwardly by the resonances of the vocal tract. These resonances cause prominent peaks in the frequency spectrum of a vowel signal. The frequencies at which these peaks occur are called the formant frequencies. This is illustrated for the vowel [a] (as in "father") by the black line in figure 2. For different vowels, these peaks occur in different places. Vowels can therefore be uniquely characterized by their formant frequencies. In practice, only the three or four resonances at the lowest frequencies are relevant.

In the simulations discussed here, no acoustic signals are generated, although it would be straightforward to add the extra complexity. Vowels are represented by their first four formant frequencies. If an agent wants to generate a given vowel, it synthesizes these formants using an articulatory synthesizer. This synthesizer takes as inputs the three major vowel parameters: tongue position, tongue height and lip rounding (Ladefoged & Maddieson 1996, Ch. 9). The outputs of the synthesizer are the first four formant frequencies of the corresponding vowel. The inputs are modeled as real values in the range 0 to 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 the formant frequencies (708, 1517, 2427, 3678) Hertz. The parameters (1, 1, 1) correspond to the vowel [u] and to the formant frequencies (276, 740, 2177, 3506) Hertz. Formant values were taken from (Vallée 1994). This synthesizer is able to generate all possible basic vowels.

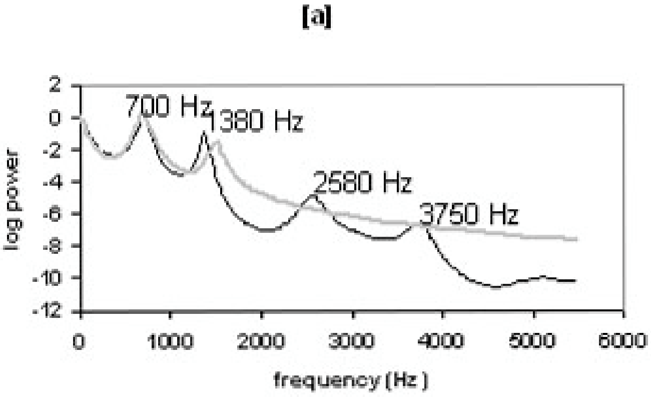

Figure 2. The first four formants of the vowel [a] black line, and its perception in terms of the effective second formant (gray line)

Humans also perceive vowels based on their formant frequencies. These can be used to calculate a perceptual distance between vowels. Unfortunately, the distance calculation is not as simple as calculating a Euclidean distance between two formant vectors. As the bandwidth of the sound receptors at higher frequencies is greater than those at lower frequencies, blurring of spectral detail takes place at higher frequencies. This means that the formant peaks at higher frequencies can generally be compressed into one broader peak. The center frequency of this peak is called the effective second formant. The first peak in the spectrum is usually perceived as remaining in the same place, and is still referred to as the first formant. The perception in terms of the first and effective second formant of the signal

[a] is shown as the gray line in figure 2. Note that this is an idealized view to illustrate the process.

While there are several ways to calculate the effective second formant, the one adopted in the research described here has been developed by Schwartz et al. (1997b). It is a non-linear weighted average of the 2nd, 3rd and 4th formant frequencies. In order to calculate distances between vowel signals, the first and effective second formant are expressed in the Bark frequency scale. The Bark scale is a perceptually inspired frequency scale that can be considered to be logarithmic for the frequencies that are relevant to formants. Equal frequency differences in Bark are perceived as equal intervals, in contrast to the way frequencies in Hertz are perceived.

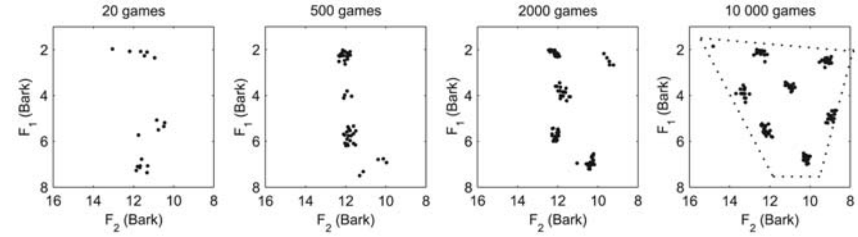

Figure 3. Emergence of a vowel system in a population of twenty agents.

Using the first and effective second formant frequencies in Bark, the distance between two vowels can be calculated as an ordinary weighted Euclidean distance. As differences in the effective second formant are perceptually less important than differences in the first formant, the effective second formant is multiplied by 0.3 when calculating distances.

These distances are then used to determine which vowel in an agent's repertoire is recognized. This is done by calculating the distance between a perceived signal and all the vowels in an agent's repertoire. The vowel that is closest to the perceived signal is considered to be the one that the agent heard.

Table 1. Basic organisation of the imitation game. Explanation of symbols: V is the repertoire of vowels, v, v_{new} and v_{rec} are vowels, u_v a vowel's use, s_v its success and ac_v its acoustic signal.

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

4.2. Results

The agents all start out with an empty vowel repertoire. By playing imitation games with each other, 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 and embodiment are really factors in explaining the structure of human vowel systems.

4.2.1. Emergence of a vowel system

The emergence of a vowel system in a population of twenty agents under 10% acoustic noise is shown in figure 3. In each of the frames of the figure, the acoustic aspects of the prototypes of the agents' vowels in the population are plotted in acoustic space. In this particular acoustic space (based on the first and effective second formant frequencies of a vowel signal), equal distances between points correspond to equal perceptual distances. Each vowel of each agent in the population 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 figure 3.

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).

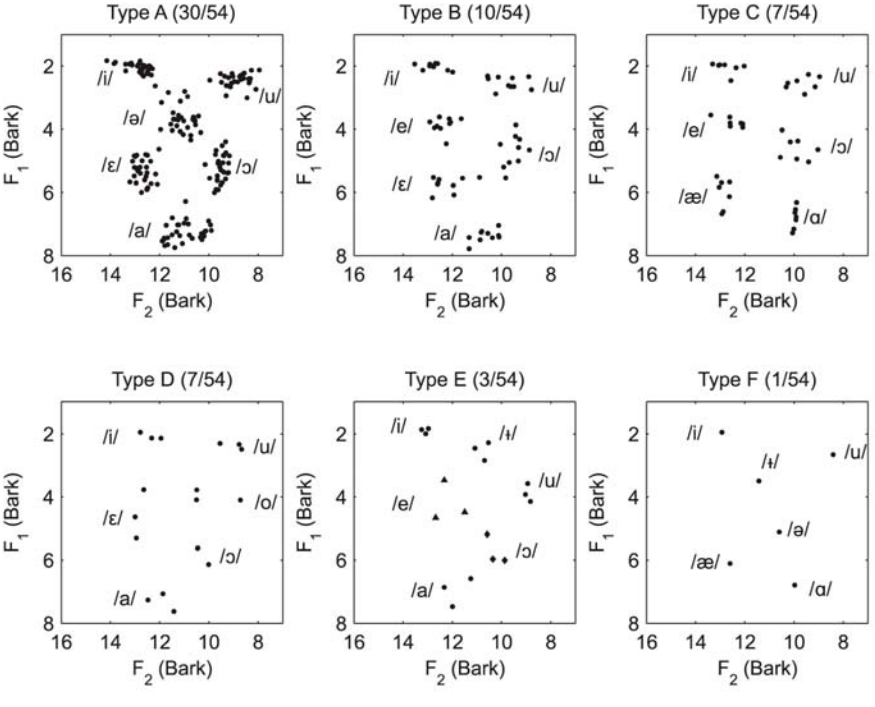

Figure 4. Emerged vowel systems with six vowels.

4.2.2. Evaluation of the emerged vowel system

Are the vowel systems that emerge realistic? 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 by Crothers (1978) and by Schwartz et al. (1997a).

An example of a classification of emerged vowel systems containing six vowels is shown in figure 4. The data in this figure were obtained by running the simulation a hundred times for a given parameter setting (acoustic noise set to 12%). In each run 25 000 imitation games were played.

Note that although the frames in figure 4. look similar to the figures of vowel systems shown previously in figure 3, 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 how self-organization can make 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. When the percentages with which the different emerged systems occur are compared to the percentages with which human vowel systems occur, a good match is found as well. Schwartz et al. (1997a) have measured the occurrence of different vowel system types in the different languages in the UCLA Phonological Segment Inventory Database (UPSID, a database based on speech sound data of 451 languages, Maddieson 1984; Maddieson & Precoda 1990). They find 60 vowel systems with 6 vowels. Although their classification is not exactly the same as the classification shown in figure 4, 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. No systems of type F were found, and two of the systems from UPSID (3%) cannot easily be fitted into the classification used here, but are probably of type A. Equally 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, ch. 6). It seems that the simulation is capable of not only 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.

5. Conclusions

The results of the simulation show that vowel systems can emerge through self-organization in a population of agents whose production and perception is constrained by their embodiment. 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, and optimized for acoustic distinctiveness. The model also predicts the types of vowel systems that occur less frequently and their abundance with a reasonable degree of accuracy. This illustrates that dynamics and embodiment play an important role in determining the structure of phonological and phonetic systems. The importance of embodiment both for constraining the space of possibilities and for pushing towards optimality in the task was known already (see e.g. Lindblom, MacNeilage & Studdert-Kennedy 1984) and the computer model presented here provides additional empirical support for this thesis. At the same time, the model clearly shows that embodiment can (and has to be) complemented with the collective dynamics flowing from social interaction. As the imitation success of agents is partly determined by how well the agents' vowel repertoires conform to the repertoires that are used by the other agents in 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 that are found in human vowel systems are found in the systems that emerge. This indicates that innate rules and constraints are not necessary in order to explain why these tendencies are present; neither is it necessary to assume that they have to be genetically encoded.

More generally (and based on other works such as reported in Steels & Belpaeme 2005) we argue that the universal tendencies found in human categorization, as well as the cognitive systems that use these categorizations for action and decision-making, are a function of multiple forces, and in particular embodiment and collective dynamics. In domains where the statistical structure inherent in the real world is relevant (such as color categories), sensitivity to this structure will also play a role. In domains where there is genuine choice (like driving to the right or left) the regularity is generated by the agents and cannot be explained on the basis of preexisting statistical structure. All this does not exclude the possibility that certain categorizations could not have become genetically assimilated. However, in the case of cultural systems, like language, which undergo relatively rapid change, this seems less likely than is often assumed.

Acknowledgement

The Sony Computer Science Laboratory in Paris funds the work by Luc Steels, with additional funding from the OHLL CNRS project, the OMLL ESF project, and the EU FET ECAgents project. The work of Bart de Boer reported here took place at the University of Brussels AI lab and was funded by a GOA project.

References

Berlin, B. and Kay, P.
1969 Basic color terms: Their universality and evolution. University of California Press.

Bowerman, Melissa and Soonja Choi
2001 Shaping meanings for language: Universal and language specific in the acquisition of spatial semantic categories. In: Melissa Bowerman and Stephen C. Levinson (eds.), *Language Acquisition and Conceptual Development*, 475–511. Cambridge: Cambridge University Press.

Browman, Catherine P. and Louis Goldstein
1992 Articulatory phonology: An overview. *Phonetica* 49: 155-180.

Chomsky, Noam and Morris Halle
1968 *The Sound Patterns of English*. Cambridge, Mass: MIT Press.

Christophe, Anne, Jacques Mehler and Nuria Sebastián-Gallés
2001 Perception of prosodic boundary correlates by newborn infants. *Infancy* 2: 285-394

Crothers, John
1978 Typology and universals of vowel systems. In: Joseph H. Greenberg, Charles A. Ferguson and E. A. Moravcsik (eds.), *Universals of Human Language, Volume 2 Phonology*, 93–152. Stanford: Stanford University Press.

Davidoff, J.
2001 Language and perceptual categorisation. *Trends in Cognitive Sciences* 5(9): 382–87.

de Boer, Bart
1999 Self Organisation in Vowel Systems, PhD Thesis, AI-Lab, Vrije Universiteit Brussel.
2000 Self organization in vowel systems. *Journal of Phonetics* 28(4): 441–465.
2001 *The Origins of Vowel Systems.* Oxford: Oxford University Press.

Dresher, Elan
1992 A learning model for a parametric theory in phonology. In: Robert Levine (ed.), *Formal Grammar: Theory and Implementation,* 290–317. Oxford: Oxford University Press.

Edelman, Gerald
1990 *Neural Darwinism*. New York: Basic Books.

Fast, Peter W.
1953 Amuesha (Arawak) phonemes. *International Journal of American Linguistics* 19: 191–194.

Herskovits, Annette

1986 *Language and Spatial Cognition*. Cambridge: Cambridge University Press.

Iida, Fumiya, Raja Dravid and Chandana Paul
2002 Design and control of a pendulum driven hopping robot. *Proceedings of International Conference on Intelligent Robots and Systems 2002 (IROS 02) Lausanne, Switzerland*, 2141–2146.

Jakobson, Roman and Morris Halle
1956 *Fundamentals of Language*. The Hague: Mouton & Co.

Kay, P. and Regier, T.
2003 Resolving the question of color naming universals. Proceedings of the National Academy of Sciences 100(15): 9085–89.

Ladefoged, Peter, and Ian Maddieson
1996 *The Sounds of the World's Languages*. Oxford: Blackwell.

Lakoff, George and Mark Johnson
1980 *Metaphors We Live by.* Chicago: University of Chicago Press.

Lindblom, Björn, Peter MacNeilage, and Michael Studdert-Kennedy
1984 Self-organizing processes and the explanation of phonological universals. In: Brian Butterworth, Bernhard Comrie and Östen Dahl (eds.), Explanations for Language Universals, 181–203. Berlin: Walter de Gruyter.

Lindblom, Björn and Ian Maddieson
1988 Phonetic universals in consonant systems. In: Larry Hyman. and Charles N. Li (eds.), Language, Speech and Mind, 62–79. London: Routledge.

Lenneberg, Eric H. & Roberts, John M.
1956 The language of experience: A study in methodology. *International Journal of American Linguistics* memoir 13.

MacNeilage, Peter
1998 The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* 21: 499–548.

Maddieson, Ian
1984 *Patterns of Sounds*. Cambridge: Cambridge University Press.

Maddieson, Ian and Kristin Precoda
1990 Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104–111.

Núñez, Rafael
1999 Could the future taste purple? Reclaiming mind, body and cognition. *Journal of Consciousness Studies* 6(11–12): 41–60

Pfeifer, Rolf and Fumiya Iida
2004 Embodied artificial intelligence: Trends and challenges, In: Fumiya Iida, Rolf Pfeifer, Luc Steels and Yasuo Kuniyoshi (eds.),

Embodied Artificial Intelligence, Lecture Notes in Computer Science 3139: 1–26. Berlin: Springer Verlag.

Plaut, D. C. and Kello, C. T.

1999 The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In: Brian MacWhinney (ed.), *The emergence of language*: 381-415. Mahwah, NJ: Erlbaum.

Rand, Earl

1968 The structural phonology of Alabaman, a Muskogean language. *International Journal of American Linguistics* 34: 94–103.

Redford, Melissa A., Chen, Chun C., and Miikkulainen, Risto

2001 Constrained emergence of universals and variation in syllable systems. *Language and Speech* 44: 27–56.

Schwartz, Jean-Luc., Louis-Jean Boë, Nathalie Vallée and Christian Abry

1997 a Major trends in vowel system inventories. *Journal of Phonetics* 25: 233–253.

1997 b The dispersion-focalization theory of vowel systems. *Journal of Phonetics* 25: 255–286.

Steels, Luc

2003 Evolving grounded communication for robots. *Trends in Cognitive Science*. 7(7): 308–312.

Steels, Luc and Tony Belpaeme

2005 Coordinating perceptually grounded categories through language. *Behavioral and Brain Science*. 28(4): 469–529.

Steels, Luc and Rodney Brooks

1995 *The Artificial Life Route to Artificial Intelligence. Building Embodied Situated Agents*. New Haven: Lawrence Erlbaum.

Steels, Luc and Oudeyer, Pierre-yves

2000 The cultural evolution of syntactic constraints in phonology. In M. A. Bedau, J. S. McCaskill, N. H. Packard & S. Rasmussen (Eds.), *Proceedings of the VIIth artificial life conference (alife 7)*: 382–394. Cambridge (MA): MIT Press.

Vallée, Nathalie

1994 Systemes vocaliques: de la typologie aux predictions. Thèse. ICP, Grenoble.

Ziemke, Tom

2003 What's that thing called embodiment? In R. Alterman & D. Hirsch, (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*: 1305–1310. Hillsdale, NJ: Lawrence Erlbaum.