Multi-Agent Simulations of the Evolution of Combinatorial Phonology

Bart de Boer
Kunstmatige Intelligentie
Rijksuniversiteit Groningen
b.de.boer @ ai.rug.nl

Willem Zuidema
ILLC
University of Amsterdam
jzuidema@science.uva.nl

Abstract

A fundamental characteristic of human speech is that it uses a limited set of basic building blocks (phonemes, syllables), that are put to use in many different combinations to mark differences in meaning. This paper investigates the evolution of such "combinatorial phonology" with a simulated population of agents. We first argue that it is a challenge to explain the transition from holistic to combinatorial phonology, as the first agent that has a mutation for using combinatorial speech does not benefit from this in a population of agents that use a holistic signaling system. We then present a solution for this evolutionary deadlock. We present experiments that show that when a repertoire of holistic signals is optimized for distinctiveness in a population of agents, it converges towards a situation in which the signals can be analyzed as combinatorial, even though the agents do not use this structure. We argue that in this situation adaptations for productive combinatorial phonology can spread.

1 Introduction

Human speech is combinatorial. This means that it combines a limited number of basic sounds into a potentially infinite set of complex utterances that all differ in meaning. Languages can be extremely complex in the repertoire of speech sounds they use and in the way they are combined. For example, the Khoisan language !Xóõ is analysed to use 186 different speech sounds [1] while the Caucasian language Georgian allows words such as *prtskvna* "to peel" [2].

Despite this apparent complexity, children acquire the system of speech sounds of their native language remarkably quickly. At six month of age infants already learn which speech sounds are important in their native tongue, even before they can properly produce speech or understand the complete meaning of utterances [3].

Compared to our closest relatives, these are impressive feats. Although some primates use utterances that are built up from smaller units [4, 5] changing the order of the units does not substantially alter the meaning of the utterances. And although there are indications that chimpanzee vocalizations are partly learned [6], chimpanzees only have very limited abilities for vocal imitation [7].

These facts suggest that the last common ancestor of humans and other great apes did not use combinatorial phonology in semantic communication. Hence, the ability for learning and using combinatorial systems of speech sounds must have evolved in the hominid lineage since. We should thus ask the question how the transition from holistic to combinatorial repertoires of speech sounds could have taken place.

Although combinatorial systems are in general more robust against noise and therefore preferable from an information theoretic point of view [8], this does not in itself constitute an evolutionary explanation. Crucially, evolutionary explanations must provide a path of ever increasing fitness, where each new variant can *invade* in a population

where it is initially infrequent. One can imagine that once a holistic repertoire is established in the population, adaptations for combinatorial speech never have a chance to spread. After all, if all the other agents in the population are using holistic systems, a mutant agent will not benefit from being able to produce, perceive or learn more combinatorial utterances. A recurring problem in language evolution research is that the usefulness of an innovation depends on how many agents in the population can process it: fitness in language evolution is typically "frequency-dependent" [9].

Explanations that propose that the combinatorial nature is an exaptation of existing behavior, such as repetitive motion of the jaw in chewing and breathing [10] are only partly satisfactory, as they only explain the syllable structure of speech, and not the internal combinatorial nature of syllables.

The challenge is therefore to show that even when agents do not make use of a combinatorial system of internal representations for speech, systems of utterances evolve culturally towards showing aspects of combinatorial systems. Such systems can be said to be *superficially* combinatorial (as opposed to *productively* combinatorial systems where language users make use of the combinatorial nature of the speech sounds in learning, storing, perceiving and producing them). In a population that uses superficially combinatorial speech sounds, a mutation for using this combinatorial structure would have a chance of spreading, thus providing a pathway of continuously increasing fitness from holistic to combinatorial utterances. Adaptations for learning combinatorial structure could then evolve.

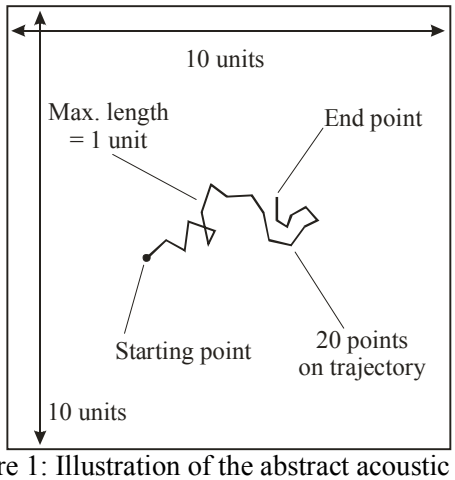

Figure 1: Illustration of the abstract acoustic space and a trajectory in it.

We propose that optimizing a system of speech sounds that are extended in time (trajectories in acoustic space) for acoustic distinctiveness results in superficially combinatorial structure. We investigate this using a simulated population of agents that try to imitate each other as well as possible. The agents produce and perceive signals purely holistically. This makes them different from the agents used in [11] where agents make use of the combinatorial structure of the utterances in their repertoire, and the combinatorial structure of the utterances is built in. Also in the work of [12] the combinatorial structure is built in, the only choice is between the building blocks that are used.

2 The Model

The model that was used in this research was an individual- or agent-based computer simulation. In individual-based models of language and its evolution, a population of language users is modeled with a set of simplified models of the individuals (called agents) in the population. Each agent is able to engage in certain language-like interactions with other individuals of the population. This paradigm has been used by other researchers of language evolution and is sometimes referred to as "language games" [13, 14] or "iterated learning" [15-17]. The aim of the agents is to develop a repertoire of signals with which they can communicate as well as possible with the other agents in the population. For this it is required that the agents in the population agree upon a repertoire and that the signals in the repertoire are as different from each other as possible.

Each agent has a repertoire of trajectories and is able to produce and perceive these as signals in an abstract "acoustic" space. In the experiments presented here, this space was chosen to be two-dimensional, but it could in principle have any number of dimensions. The dimensions can be imagined as features of the acoustic signal. For human perception, such features could be the pitch, the loudness or the formants (peaks in the frequency spectrum) of the signal. In the model presented here, the space is abstract. The dimensions do not correspond to any real feature of the signal. We decided to use an abstract space, as our aim is

to investigate a general property of trajectories used for signaling, independent of the actual properties of perception and production. Using more realistic features would result only in an alteration of the shape of the acoustic space. Our acoustic space is a square with sides of length 10 in all simulations presented here.

Trajectories in this space consist of a fixed number (N) of points, representing acoustic signals with fixed duration. Points on a trajectory can be considered as samples of that trajectory, taken at fixed time intervals. Points on a trajectory can have any distance between 0 and R to their predecessor and their successor. The values of N and R were taken to be 20 and 1, respectively in the simulations presented here. The acoustic space and a trajectory are illustrated in figure 1.

Distances between two trajectories T_1 and T_2 are calculated as the sum over all distances between corresponding points of the trajectories, corresponding to the following equation:

$$d = \sum_{i=1}^{N} \left\| \mathbf{t}_{1,i} - \mathbf{t}_{2,i} \right\| \qquad (1)$$

where $\mathbf{t}_{1,i}$ and $\mathbf{t}_{2,i}$ are points from trajectories T_1 and T_2, respectively. The double bars give the Euclidean vector distance (the points $\mathbf{t}_{1,i}$ and $\mathbf{t}_{2,i}$ are of the same dimensionality as the assumed acoustic space). This distance measure is convenient and easy to calculate for trajectories with a fixed number of points. It can be argued that for a variable number of points, a distance measure such as dynamic time warping would be better and that this would probably correspond better to the way humans perceive acoustic signals [18]. Some preliminary experiments with dynamic time warping were performed but no qualitative difference in performance between the two distance measures was found.

When an agent produces an utterance, noise is added. Noise is added in a way that preserves the general shape of trajectories. We found that this leads to faster convergence than adding noise that is independent for each point on the trajectory. Adding noise that is correlated from point to point is realistic, as it implies the existence of disturbances of longer duration. The method adds independent shifts to the first and last points of the trajectory, and interpolates for the points in between. A smaller independent shift is also added to all points. In equation form:

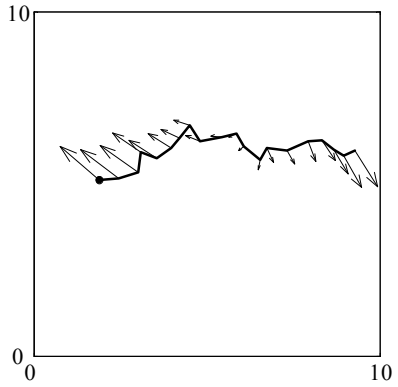

Figure 2: Example of shape-preserving noise. Arrows indicate shift by noise. Note correlation between neighboring shifts.

$$\mathbf{t}_i' \leftarrow \mathbf{t}_i + \alpha_i \mathbf{s}_1 + (1-\alpha_i)\mathbf{s}_N + \mathbf{n}_i\left(0, \frac{\sigma_n}{N}\right) \qquad (2)$$

where $\alpha_i = \dfrac{N-i}{N-1}$ and $\mathbf{s}_1$ and $\mathbf{s}_N$ are vectors that are constant for each trajectory (their components are taken from the normal random distribution with mean 0 and standard deviation σ_n). The vector $\mathbf{n}_i$ is different for each point on the trajectory (each component is taken from the normal random distribution with mean 0 and standard deviation σ_n/N). As this kind of noise preserves the overall shape of a trajectory, it is called *shape-preserving noise*. An example of shape preserving noise is presented in figure 2.

The interaction between agents is an instantiation of the imitation game [19, 20]. There are some important differences, however. The number of trajectories in every agent is set to a fixed number, K, beforehand. Trajectories are initialized randomly at the start of the simulations. The first point is randomly taken from the uniform distribution over a square of size 1×1 ($1/100^{\text{th}}$ of the available acoustic space) at the center of the acoustic space. Subsequent points on the trajectory are generated at distance R from the previous points and with uniformly distributed angle with respect to the previous point. An example of such a randomly initialized system can be found in the left frame of figure 3.

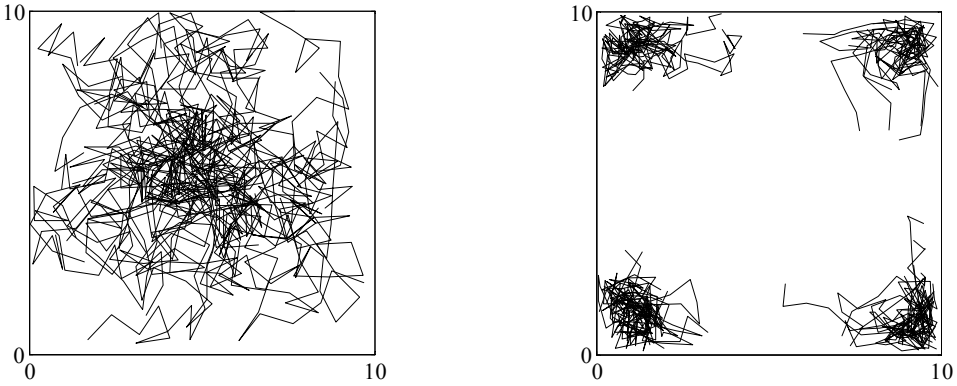

Figure 3: System of trajectories before (left frame) and after (right frame) 60 000 generations. Each cluster in the right frame contains a trajectory for each agent in the population. Note how trajectories become bunched up in the corners

With each trajectory of an agent, a success score is associated. This score measures how successful a trajectory has been in previous interactions. This score is initially set to 0 for each trajectory.

For each iteration of a simulation, one agent (the *initiator*) is selected randomly from the population of agents. This agent selects one trajectory from its repertoire and makes a slight modification to it. The modification is made by selecting a random point and adding a vector to it. The vector has components that are taken from the normal random distribution with mean 0 and standard deviation σ_i. In order to make trajectories stay within the bounds of the available acoustic space, the point to be moved is projected on the nearest edge of the space. Another constraint on the trajectories is that the distance between subsequent points must not be larger than R. If the random shift causes this to happen, previous and following points are shifted towards the shifted point such that their distance becomes equal to R again. This procedure is repeated for all points on the trajectory.

With the modified trajectory, the agent plays repeated imitation games with all other agents in the population (called *imitators*). In an imitation game, the initiator produces the modified trajectory with noise added to it. The imitator then selects the trajectory in its repertoire that is closest to this, and in turn produces it while adding noise. The initiator then checks whether the trajectory in its repertoire that is closest to this is the selected trajectory. If this is the case, the imitation game is successful, else it is a failure. In this way, 50 imitation games are played with each other agent in the population. Finally, the number of successful games is divided by the total number of games played. This number (s_{modified}) is compared to the success score (s_{original}) of the selected trajectory. If it is lower, the modified trajectory is discarded, and the original trajectory is kept. If the score is higher, the original trajectory is averaged with the modified trajectory and this is stored in the agent's repertoire. The calculation is as follows:

$$\mathbf{t}_{i,\text{new}} \leftarrow \beta \mathbf{t}_{i,\text{original}} + (1-\beta)\mathbf{t}_{i,\text{modified}}$$

where β is a weighting constant, whose value was set to 0.5 in the simulations presented here. In the case of both success and failure, the success score of a trajectory is updated similarly:

$$s_{\text{new}} \leftarrow \beta s_{\text{original}} + (1-\beta)s_{\text{modified}}$$

where s_{modified} is the success of the modified trajectory in the imitation games.

This procedure is repeated for a predetermined (but large) number of iterations. Trajectories eventually converge to a local optimum.

3 Results

Running the system results in increasing average success of imitation as well as increasing structure in the repertoire of trajectories that is used in the population. All simulations were run with a population of 10 agents and with a noise standard deviation (σ_n in the equations) of 2. The first thing to note is that success increases over the iterations and converges to an asymptotic value, as illustrated in figure 4. This figure shows the running average

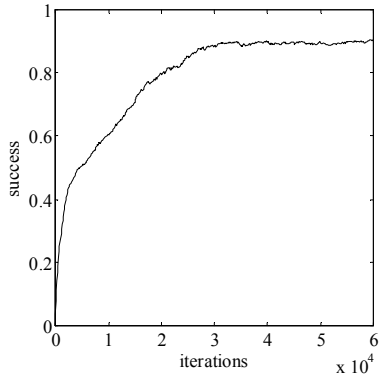

Figure 4: Success over the iterations of a population of 10 agents using 4 trajectories each.

of success (calculated as $\bar{x}_t \leftarrow 0.999 \cdot \bar{x}_{t-1} + 0.001 \cdot x_t$, where x_t is the success at time t, and $\bar{x}_t$ is the running average at time t) for a typical run in which there were 4 trajectories per agent. Between 3000 and 30 000 iterations, the success rises almost linearly, after which a plateau is reached and maintained. For larger numbers of trajectories, success rises slower and a somewhat lower final value is reached. This is to be expected, as only one trajectory gets adapted per iteration, so adapting more trajectories takes more iterations. Also, given a limited acoustic space and a fixed noise level, confusion probability is expected to be greater for larger numbers of trajectories, and success correspondingly lower.

More interesting for the purpose of this paper is what happens to the trajectories themselves. In figure 3, the initial (random) trajectories of all agents are shown in the acoustic space, as well as the trajectories after 60 000 generations. The trajectories form clusters with different shapes and positions. Each agent in the population has a trajectory in each cluster. It is found that the four trajectories of each agent bunch up in the four corners of the acoustic space. This seems understandable, as in this way, trajectories within a cluster are close together, while distance between clusters is maximized. Both properties contribute to higher success in imitation.

The question then arises what would happen if a fifth trajectory were added to the repertoires. It is conceivable that a structure would emerge in which the fifth trajectory is bunched up in the center of the acoustic space, giving a structure similar to the five dots on the face of a die. This would be analogous to the structures found when optimizing vowel systems, where the acoustic distances between vowels is maximized [i. e. 19, 20-23]. This is not what is found in our experiments, however.

When adding a fifth trajectory to the repertoire, it generally become stretched out from one corner of the acoustic space to another. This is shown in figure 5. It can be observed that four trajectories are still bunched up in the corners of the acoustic space, but that the fifth trajectory is now on the diagonal. For clarity the repertoire of a single agent is shown in the right frame of the figure. This shows the identical structure.

In fact, in hindsight it is understandable that trajectories would stretch out over the diagonal, as this results in a larger distance between the points on the stretched out trajectory and the trajectories in the corners than for a trajectory bunched up in the center. The average distance of points on the

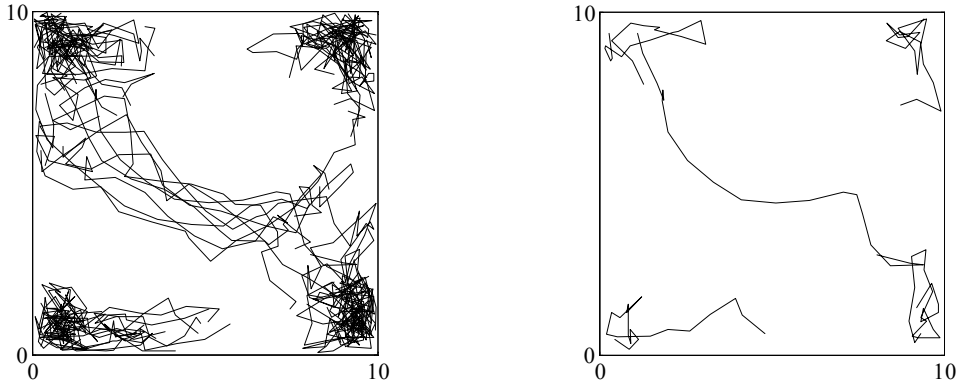

Figure 5: Population of 10 agents with 5 trajectories each after 60 000 generations. The left frame shows the whole generation, the right frame shows the repertoire of one agent from the population.

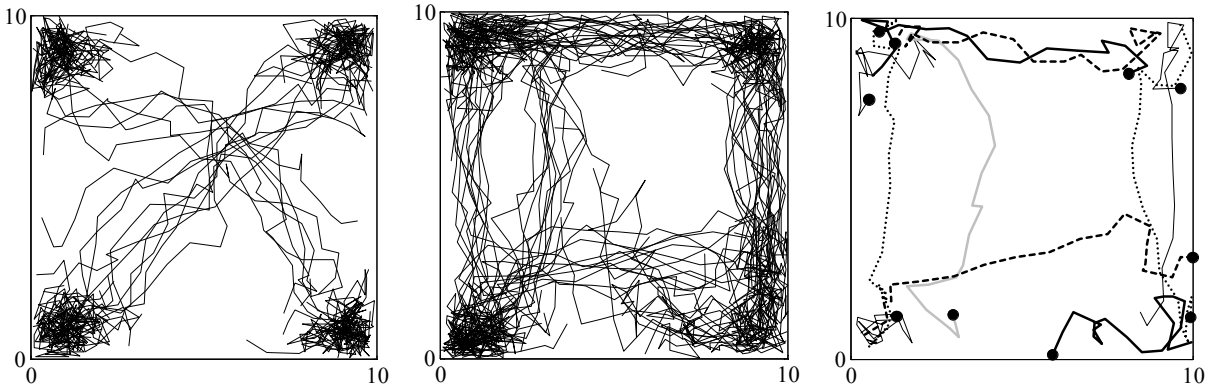

Figure 6: Results for larger number of trajectories. Shown are a system with six trajectories(left frame), a system with 10 trajectories (middle frame) and the repertoire of an agent from the population with 10 trajectories. Trajectories in the rightmost frame have been given different styles to make them easier to follow. Note that trajectories that appear to follow a similar path have starting points (indicated with a black dot) in different corners.

trajectory to the two corners it visits remains equal, while the average distance to the two corners it does not visit increases [24].

The situation is less clear for larger numbers of trajectories. Therefore the simulation was run for other numbers of trajectories as well. As the complexity of the imitation games increases proportionally to the square of the number of trajectories, and the number of games it takes to converge increases at least linearly with the number of trajectories per agent, the total time for the simulations to run increases with at least the cube of the number of trajectories. Running these simulations is therefore very time consuming, and only 6 and 10 trajectories were tried. These are shown in figure 6. The system with six trajectories uses two diagonal trajectories in addition to the four trajectories in the corners. The system with 10 trajectories is more complicated, and therefore an individual agent is shown in the rightmost frame of the figure. It can be observed that all trajectories make use of the corners of the acoustic space, and that trajectories that make use of the same two corners, have different directions. It is also noteworthy that there is no trajectory bunched up in the upper right corner. Apparently stretched out trajectories are equally distinctive.

4 Conclusion

All experiments with more than four trajectories resulted in repertoires with trajectories that were stretched out through the available acoustic space. Only the corners were occupied by trajectories that were bunched up into point-like signals. Moreover, the trajectories in the repertoires reused the same start- and endpoints. They can therefore be considered to be constructed from a limited number of building blocks. In this sense, they are combinatorially coded.

The emerged repertoire with ten trajectories that is shown in figure 6, for example, can be analyzed as having four basic building blocks: the four corners of the acoustic space. The ten trajectories can be coded by their starting corner and their ending corner. The trajectories then become: *tl–tl, tl–tr, tl–bl, tr–tl, tr–br, bl–bl, bl–tl, br–br, br–tr* and *br–tl* (*tl* is top left, *tr* is top right, *bl* is bottom left and *br* is bottom right). Although some information about the actual path of the trajectory is lost in this way, these description are sufficient for playing successful imitation games. The system of ten trajectories can therefore be considered to be built up from four phonemes.

It must be stressed that the agents that use these repertoires of trajectories are not aware of this. They store, perceive and produce the trajectories in a completely holistic way. The repertoire is therefore only *superficially* combinatorial.

As the structure is present in the repertoire of trajectories, however, it becomes advantageous for agents to exploit it. Agents can evolve learning mechanisms that learn utterances in a combinatorial way instead of holistically. In this way, agents

can evolve towards using combinatorial structure productively, like in human speech.

In this way, the cultural evolution of sound systems towards superficial combinatorial structure through the pressure of increasing acoustic distinctiveness allows biological evolution of the learning mechanisms to take place. Even though an agent might be the only one in the population to actively use combinatorial structure, it will still have an advantage over the other agents, as the combinatorial structure is already superficially present. This is completely different from the situation where an agent mutates towards the ability to use combinatorial structure in a population where no speech with combinatorial structure is used.

There are still a number of things that need to be improved about the work presented here. A weak point is that emerged repertoires of trajectories can only be characterized in a qualitative way, and not in a quantitative way. For quantitative analysis, a measure of the extent in which trajectories are combinatorially coded would be needed. Unfortunately, such measures appear not to exist in the phonological literature, and this is perhaps understandable as all human languages are combinatorially coded. Studies of animal calls [such as e. g. 4] do not provide measures either. In animal work, it is sometimes investigated *whether* systems of calls are combinatorially coded, but not to what extent. There are thus no existing measures of the extent in which systems of signals are combinatorially coded.

Another point that could be improved is the acoustic space that was used. In the experiments it is a square without internal structure. It would be interesting to investigate spaces that are more plausible, both from the perspective of production and perception. This would alter the shape of the allowed space. As the emerged trajectories in the experiments presented here appear to exploit the shape of the available space (trajectories tend to use corners) it is likely that trajectories in a more complex space would exploit its features to maximize distinctiveness. These features could conceivably be used to explain universal properties of human sound systems.

In the most basic case investigated in this paper, however, it has been shown that trajectories that are maximized for distinctiveness develop a structure that can be interpreted as combinatorial and that can be exploited by agents that learn speech combinatorially. This creates an interaction between cultural evolution of a repertoire and biological evolution that allows combinatorial learning of speech to evolve.

References

[1] A. Traill, *Phonetic and phonological studies of!Xóõ bushman*. Hamburg: Helmut Buske Verlag, 1985.

[2] J. C. Catford, "Mountain of tongues: the languages of the Caucasus," *Annual Review of Anthropology*, vol. 6, pp. 283–314, 1977.

[3] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom, "Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age," *Science*, vol. 255, pp. 606–608, 1992.

[4] J. C. Mitani and P. Marler, "A phonological analysis of male gibbon singing behavior," *Behaviour*, vol. 109, pp. 20–45, 1989.

[5] A. Arcadi, "Phrase structure of wild chimpanzee pant hoots: patterns of production and interpopulation variability," *American Journal of Primatology*, vol. 39, pp. 159-178, 1996.

[6] C. Crockford, I. Herbinger, L. Vigilant, and C. Boesch, "Wild Chimpanzees Produce Group-Specific Calls: a Case for Vocal Learning?" *Ethology*, vol. 110, pp. 221–243, 2004.

[7] W. T. Fitch, "The evolution of speech: a comparative review," *Trends in cognitve science*, vol. 4, pp. 258–267, 2000.

[8] M. A. Nowak and D. Krakauer, "The evolution of language," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 8028–8033, 1999.

[9] L. L. Cavalli-Sforza and M. W. Feldman, "Paradox of the evolution of communication and of social interactivity," *Proceedings of the National Academy of Sciences*, vol. 80, pp. 2017-2021, 1983.

[10] P. F. MacNeilage and B. L. Davis, "On the Origin of Internal Structure of Word Forms," *Science*, vol. 288, pp. 527–531, 2000.

[11] P.-Y. Oudeyer, "Phonemic coding might be a result of sensory-motor coupling dynamics," in *Proceedings of the International conference on the simulation of adaptive behavior (SAB)*, J.

Hallam, Ed. Edinburgh: MIT Press, 2002, pp. 406–416.

[12] B. Lindblom, P. MacNeilage, and M. Studdert-Kennedy, "Self-organizing processes and the explanation of language universals.," in *Explanations for language universals*, M. Butterworth, B. Comrie, and Ö. Dahl, Eds. Berlin: Walter de Gruyter & Co., 1984, pp. 181–203.

[13] L. Steels, "The Synthetic Modelling of Language Origins," *Evolution of Communication*, vol. 1, pp. 1–34, 1997.

[14] L. Steels, "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, S.-K. Michael, and C. Knight, Eds. Cambridge: Cambridge University Press, 1998, pp. 384–404.

[15] S. Kirby and J. R. Hurford, "The Emergence of Linguistic Structure: An overview of the Iterated Learning Model.," in *Simulating the Evolution of Language*, A. Cangelosi and D. Parisi, Eds. London: Springer Verlag, 2001, pp. 121–148.

[16] K. Smith, S. Kirby, and H. Brighton, "Iterated Learning: a framework for the emergence of language," *Artificial Life*, vol. 9, pp. 371–386, 2003.

[17] W. Zuidema, "How the poverty of the stimulus solves the poverty of the stimulus," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 51–58.

[18] H. Sakoe and S. Chiba, "Dynamic programming optimization for spoken word recognition," *IEEE transactions on acoustics, speech and signal processing*, vol. 26, pp. 43–49, 1978.

[19] B. de Boer, "Self organization in vowel systems," *Journal of Phonetics*, vol. 28, pp. 441–465, 2000.

[20] B. de Boer, *The origins of vowel systems*. Oxford: Oxford University Press, 2001.

[21] J. Liljencrants and B. Lindblom, "Numerical simulatons of vowel quality systems," *Language*, vol. 48, pp. 839–862, 1972.

[22] P.-Y. Oudeyer, "Coupled neural maps for the origins of vowel systems," in *Proceedings of the International Conference on Artificial Neural Networks, Lecture Notes in Computer Science 2130*, G. Dorffner and K. H. Bischof, Eds. Berlin: Springer Verlag, 2001, pp. 1171–1176.

[23] P.-Y. Oudeyer, "The self-organization of speech sounds," *Journal of Theoretical Biology*, vol. 233, pp. 435–449, 2005.

[24] W. Zuidema, "The major transitions in the evolution of language," in *Theoretical and applied linguistics*. Edinburgh: University of Edinburgh, 2005.