Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 37 (2009) 125–144
Journal of Phonetics
www.elsevier.com/locate/phonetics

The evolution of combinatorial phonology

Willem Zuidema[a,*], Bart de Boer[b,1]

[a]Institute for Logic, Language and Computation, University of Amsterdam Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands
[b]Artificial Intelligence, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Received 1 July 2006; received in revised form 3 September 2008; accepted 27 October 2008

Abstract

A fundamental, universal property of human language is that its phonology is combinatorial. That is, one can identify a set of basic, distinct units (phonemes, syllables) that can be productively combined in many different ways. In this paper, we develop a methodological framework based on evolutionary game theory for studying the evolutionary transition from holistic to combinatorial signal systems, and use it to evaluate a number of existing models and theories. We find that in all problematic linguistic assumptions are made or crucial components of evolutionary explanations are omitted. We present a novel model to investigate the hypothesis that combinatorial phonology results from optimizing signal systems for perceptual distinctiveness. Our model differs from previous models in three important respects. First, signals are modeled as trajectories through acoustic space; hence, both holistic and combinatorial signals have a temporal structure. Second, acoustic distinctiveness is defined in terms of the probability of confusion. Third, we show a path of ever increasing fitness from unstructured, holistic signals to structured signals that can be analyzed as combinatorial. On this path, every innovation represents an advantage even if no-one else in a population has yet obtained it.

1. Introduction

1.1. Natural language phonology is combinatorial

One of the universal properties of human language is that its phonology is *combinatorial*. In all human languages, utterances can be split into units that can be recombined into new valid utterances. Although there is some controversy about what exactly the units of (productive) combination are, there is general agreement that in natural languages—including even sign languages (Deuchar, 1996; Stokoe, 1960)—meaningless atomic units (phonemes or syllables) are combined into larger wholes.

In the traditional view, the atomic units are *phonemes*, or the distinctive features of these phonemes (Chomsky & Halle, 1968). Signal repertoires that are built-up from combinations of phonemes are said to be "phonemically coded" (Lindblom, MacNeilage, & Studdert-Kennedy, 1984). One popular alternative view is that the atoms are *syllables*, or the possible onsets, codas and nuclei of syllables (e.g. Levelt & Wheeldon, 1994). A second alternative theory uses *exemplars*, which can comprise several syllables or even words, as its basic units (e.g. Pierrehumbert, 2001). In this paper we will avoid the debate about the exact level of combination—and the conventional term "phonemic coding"—and instead focus on the uncontroversial abstract property of "combinatorial phonology".[2]

Note that, whichever the real level of combination is, there is no logical necessity to assume that all recurring sound patterns observed in speech are in fact units of productive combination in the speaker's brain. For instance, if one accepts that syllables or exemplars are the units of combination used by the speaker, phonemes are

*Corresponding author. Tel.: +31 20 5256340; fax: +31 20 5255206.
E-mail addresses: jzuidema@science.uva.nl (W. Zuidema), B.G.deBoer@uva.nl (B. de Boer).
[1]Now at the Institute of Phonetic Sciences, University of Amsterdam.

[2]In the animal behavior literature the term "phonological syntax" (coined by Peter Marler, see Ujhelyi, 1996) is often used; Jackendoff (2002, p. 238) uses the term "combinatorial, phonological system" on which our terminology is based.

0095-4470/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2008.10.003

still a useful level of description to characterize differences in meaning. We distinguish between:

(1) *Productively combinatorial phonology*, where the cognitive mechanisms for producing, recognizing and remembering signals make use of a limited sets of units that are combined in many different ways. Productive combinatoriality is a property of the internal representations of language in the speaker.
(2) *Superficially combinatorial phonology*, where parts of signals overlap (that is, occupy the same position in acoustic and perceptual space) with parts of other signals. Superficial combinatoriality is a property of the observable language. Importantly, the overlapping parts of different signals need not necessarily also be the units of combination of the underlying linguistic representations.

This paper is concerned with mathematical and computational theories of the evolution of combinatoriality of human languages at both these levels. It has often been observed that natural language phonology is *discrete*, in that it allows only a small number of basic sounds and not all feasible sounds in between. In this paper we argue that it is important to distinguish between discreteness per se, and superficial and productive combinatoriality. In Section 2, we will review existing models of Liljencrants and Lindblom (1972), Lindblom et al. (1984), de Boer (2001) and Oudeyer (2001, 2002, 2005), and argue that they are relevant for the origins of discreteness, but have little to say about the origins of superficial and productive combinatoriality. Nowak and Krakauer (1999) and Nowak, Krakauer, and Dress (1999) do address the origins of productive combinatoriality, but these models have a number of shortcomings that make them unconvincing as an explanation for its evolution.

In our own model, that we will introduce in Section 3, we address the questions of why natural language phonology is both discrete and superficially combinatorial. We assume, but do not show in this paper, that superficial combinatoriality is an important intermediate stage in the evolution of productive combinatoriality.

1.2. The origins of combinatorial phonology

It is increasingly realized that many examples of bird and cetacean songs (e.g. Doupe & Kuhl, 1999; Payne & McVay, 1971) and, importantly, non-human primate calls are combinatorial as well, (Ujhelyi, 1996). For instance, the "long calls" of tamarin monkeys are built up from many repetitions of the same element (e.g. Masataka, 1987), and those of gibbons (e.g. Mitani & Marler, 1989) and chimpanzees (e.g. Arcadi, 1996) of elaborate combinations of a repertoire of notes.

Such comparative data should be taken seriously, but it is unwarranted to view combinatorial long calls in other primates as an immediate precursor of human combinatorial phonology, because there are some important qualitative differences:

- Although a number of building blocks might be used repeatedly to construct a call, it does not appear to be the case that rearranging the building blocks results in a call with a different meaning.
- It is unclear to what extent the building blocks of primate "long calls" are flexible and whether they are learned.
- In human language, combinatorial phonology functions as one half of the "duality of patterning" (Hockett, 1960): together with recursive, compositional semantics it yields the unlimited productivity of natural language, but it is unclear if the single combinatorial system of primates can be seen as its precursor.

Nevertheless, combinatorial phonology must have evolved from holistic systems by natural selection. There are at least two views on what the advantages of combinatorial coding over holistic coding are:

(1) It makes it possible to transmit a larger number of messages over a noisy channel (e.g. Nowak & Krakauer, 1999). Note that this argument requires that the basic elements are distinct from each other, and that signals are strings of these basic elements. The argument does not address, however, how signals are stored and created.
(2) It makes it possible to create an infinitely extensible set of signals with a limited number of building blocks. Such productivity provides a solution for memory limitations and for generalization (the "productivity argument", a point often made in the generative linguistics tradition, e.g. Jackendoff, 2002).

These views are a good starting point for investigating the question of *why* initially holistic systems (which seem to be the default for smaller repertoires of calls) would evolve toward combinatorial systems. However, just showing an advantage does not constitute an evolutionary explanation (Parker & Maynard Smith, 1990). At the very least, evolutionary explanations of an observed phenotype involve a characterization of (i) the set of possible phenotypes, (ii) the fitness function over those phenotypes, and (iii) a sequence of intermediate steps from an hypothesized initial state to the observed phenotype. For each next step, one needs to establish that (iv) it has selective advantage over the previous, and thus can invade in a population without it. In Section 2 we will criticize some existing models because they lack some of these required components.

In language evolution, fitness will not be a function of the focal individual's traits alone, but also of those of this individual's conversation partners. That is, the selective advantage of a linguistic trait will depend on the frequency of that trait in a population (it is "frequency dependent").

Therefore, evolutionary game theory (Maynard Smith, 1982) is the appropriate framework for formalizing evolutionary explanations for language (Benz, Jäger, & Van Rooij, 2005; Komarova & Nowak, 2003; Nowak & Krakauer, 1999; Smith, 2004). In this framework, the crucial concept is that of an evolutionary stable strategy (henceforth, ESS): a strategy that cannot be invaded by any other strategy (Maynard Smith & Price, 1973). Thus formulated, the challenge is to show that (i) repertoires of signals with a combinatorial phonology are ESSs, and that (ii) plausible precursor repertoires, without combinatorial phonology, are not evolutionarily stable.

There are also theories of the origins of combinatorial phonology that do not explicitly involve natural selection. For instance, Lindblom et al. (1984), de Boer (2001) and Oudeyer (2001, 2002) see "self-organization" as the mechanism responsible for the emergence of combinatorial phonology. Models of self-organization are usually considered by the authors as compatible with natural selection. We agree, and view the two groups of models as detailing *proximate* and *ultimate* causes, respectively (Tinbergen, 1963). Natural selection modifies the parameters of a self-organizing process, while self-organization creates the adaptive landscape for natural selection (Barton & Zuidema, 2003; Boerlijst & Hogeweg, 1991; Oudeyer, 2006; Waddington, 1939).

2. Existing approaches

2.1. Maximizing discriminability

Liljencrants and Lindblom (1972) argued that one can understand the structure of the sound systems in natural language as determined by physical factors, such as perceptual discriminability and articulatory ease, and not as the result of arbitrary settings of abstract parameters (e.g. Chomsky & Halle, 1968). In their paper they focused on the discriminability of vowel repertoires, and proposed the following metric to measure their quality:

$$E = \frac{1}{2}\sum_{i,j\neq i\in R}\frac{1}{d_{ij}^2} = \sum_{i=2}^{T}\sum_{j=1}^{i-1}\frac{1}{d_{ij}^2}, \quad (1)$$

where R is a repertoire with T distinct sounds, and d_{ij} is the perceptual distance between sound i and sound j, determined by the Euclidean distance in the space of the first and the second formant. E is a measure for the quality of the system, where lower values correspond to a better distinguishable repertoire. The E stands for "energy", in analogy with the potential energy that is minimized in various models in physics.

Lindblom and Liljencrants performed computer simulations using a simple hill-climbing heuristic, where at each step a random change to the repertoire is considered, and adopted only if it has a lower energy than the current state. Their results compared favorably to observed data on vowel system distributions. These results were important because they showed that sound systems in natural languages are not arbitrary. However, a number of questions remain. First of all, what in the real world exactly is the optimization criterion meant to be modeling? It is important to realize that the optimization criterion in Eq. (1) is neither maximizing the distances between vowels nor minimizing the probability of confusion. Minimizing $E = \frac{1}{2}\sum_{i,j\neq i\in R} 1/d_{ij}^2$ by changing the configuration of a set of vowels in a restricted acoustic space is not necessarily the same a maximizing the average distance $\overline{d} = (1/T)\sum_{i,j\neq i\in R} d_{ij}$ (or squared distance), nor is it the same as minimizing the average confusion probability $\overline{C} = (1/T)\sum_{i,j\neq i\in R} P(j\ perceived|i\ uttered)$. At intermediate distances, these three criteria behave very similarly. The crucial difference is at distance 0, where Lindblom and Liljencrants's E goes to infinity, and at large distances, when both the E and $\overline{C}$ measure, but not $\overline{d}$, approach 0. In Section 3.2 we will argue that Liljencrants and Lindblom's E behaves unrealistically, and that minimizing the average confusion probability (or equivalently, maximizing the *distinctiveness* $D = 1 - \overline{C}$) is a better criterion.

Second, we should ask which mechanism in the real world is responsible for the optimization. Lindblom himself has referred to both natural selection and self-organization. However, the frequency dependence of language evolution means that natural selection at the level of the individual cannot be equated with optimization at the level of the population (see Zuidema & de Boer, 2003). A game-theoretic analysis must show that every new configuration of signals can *invade* in a population. Models of this type will be discussed in the next section. For self-organization, the mechanism for optimization has been worked out more precisely. de Boer (2000, 2001) has studied a simulation model of a population and showed that similar configurations of vowels emerge as in the Lindblom and Liljencrants model.

Finally, these existing models have little to say about the evolution of superficial and productive combination. Lindblom et al. (1984), and similarly de Boer (1999, Chapter 7), did study models where signals are trajectories, going from a point in consonant space, to a point in vowel space. But those models were still about the emergence of categories, because the sequencing of sounds is taken as given. In this paper, in contrast, we will focus on the emergence of superficial combination.

2.2. Natural selection for combinatorial phonology

Nowak and Krakauer (1999) and Nowak et al. (1999) apply notions from information theory and evolutionary game theory to the evolution of language. They derive an expression for the "fitness of a language". The authors observe that when communication is noisy and when a unique signal is used for every meaning, the fitness is limited by an "error limit": only a limited number of sounds can be used because by using more sounds the successful recognition of the current signals would be

impeded. They further show that in such noisy conditions, fitness is higher when sounds are combined into longer words. These results are essentially instantiations of Shannon's (1948) more general results on "noisy coding", as is explored in a later paper by the same group (Plotkin & Nowak, 2000).

More interesting is the question how natural selection could favor a linguistic innovation in a population where that innovation is still very rare. Neither Nowak and Krakauer (1999) nor Nowak et al. (1999) really address that problem. Nowak and Krakauer (1999) do, however, perform a mathematical, game-theoretic analysis of the evolution of "compositionality", and point out that this analysis can be adapted easily to the case of combinatorial phonology, as is worked out in Zuidema (2005). In such an analysis, all strategies use both holistic and combinatorial signals.

Nowak and Krakauer assume that the confusion between holistic signals is larger than the confusion between combinatorial signals, and that there is no confusion between the two types of strategies. From these assumptions it follows that a more combinatorial language can always invade a population with a less combinatorial language. This is the case, because for languages L and L' (with proportions of combinatorial signals x and x', respectively) it turns out that

$$F(L',L') > F(L,L') > F(L,L) \quad \text{if } x' > x, \tag{2}$$

where $F(a,b)$ is the expected communicative success (fitness) of users of language a communicating with users of language b.

If L' is very infrequent, then all speakers of language L (the "residents") will have a fitness of approximately $F(L,L)$ and the rare speakers of language L' (the "mutants") will have fitness of approximately $F(L,L')$, because for both residents and mutants the vast majority of interactions will be with speakers of language L. Once the frequency of mutants starts to rise, the residents will gain in fitness, that is, move toward a fitness $F(L,L')$. However, the mutants will gain even more by interacting more and more with other mutants, that is, move toward $F(L',L')$. Hence, these calculations show that strategies that use more combinatoriality can invade strategies that use less. This means that the evolutionary dynamics of languages under natural selection should lead to compositionality and combinatorial phonology.

Although this model is a useful formalization of the problem and gives some important insights, as an explanation for the evolution of compositionality—and, by implication, the evolution of combinatorial phonology—it is incomplete. The problem is that the model only considers the advantages of combinatorial strategies, and does not address two disadvantages: (1) by having a "mixed strategy" individuals have essentially two languages in parallel, which one should expect to be costly because of memory and learning demands and additional confusion. Nowak and Krakauer simply assume that the second system is in place, and that the hearer interprets all signals correctly, even if the proportion of combinatorial signals is close to zero, and the number of learning experiences is therefore extremely small; (2) combinatorial signals that consist of two or more sounds take longer to utter and are thus more costly. A fairer comparison would be between holistic signals of a certain duration (where continuation of the same sound decreases the effect of noise) and combinatorial signals of the same duration (where the digital coding decreases the effect of noise). This is the approach we take in the model of this paper, but like Nowak and Krakauer, we will look at invasibility in addition to optimization.

2.3. Crystallization in the perception–imitation cycle

A completely different approach to combinatorial phonology is based on "categorical perception". Categorical perception (Harnad, 1987) is the phenomenon that categorization influences the perception of stimuli in such a way that differences between categories are perceived as larger and differences within categories as smaller or non-existent. For instance, infants of six months are already unable to perceive distinctions between sounds that are not phonemes in their native language, something they *were* able to do at birth (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Apparently, the frequency and position of acoustic stimuli gives rise to particular phoneme prototypes, and the prototypes in turn "warp" the perception.

Oudeyer (2001, 2006) observes that signals survive over time because they are perceived and replicated (as many other speech researchers have noted as well, e.g. Ohala, 1981; Blevins, 2004). Because of noise and categorical perception, the replicated signal will not always be exactly the same as the perceived signal. However, the signals that are produced shape the categories of the other agents. Thus there is feedback between emitted signals, formation of categories and perception. This shapes the repertoire of signals in a cycle from articulation to perception to articulation (the perception–imitation cycle; see also Westermann, 2001).

Oudeyer (2001) presents a model to study this phenomenon. In this model, signals are modeled as points in an acoustic space. The model consists of two coupled neural maps, one for perception and one for articulation. The perceptual map is of a type known to be able to model categorical perception: its categorization behavior changes in response to the input data. In addition, the associations between perceptual stimuli and articulatory commands are learned. Through this coupling between perceptual and articulatory maps, a positive feedback loop emerges where slight non-uniformities in the input data lead to clusters in the perceptual map, as well as weak clusters in the articulatory map, and hence to slightly stronger non-uniformities in the distribution of acoustic signals. Oudeyer

calls the collapse of signals into a small number of clusters "crystallization".

In later publications (e.g. Oudeyer, 2006) he generalizes these results to a model with (quasi-) continuous trajectories in which well-defined clusters also form in the perceptual and articulatory maps. The signals can be analyzed as consisting of sequences of phonemes.

Oudeyer's model gives a completely non-adaptive mechanism for the emergence of combinatorial phonology. However, the question whether recombination increases the functionality of the language, and thus the fitness of the individual that uses it, remains unanswered. In particular, in Oudeyer's (2001) first model, where signals are instantaneous, a large repertoire of signals collapses into a small number of clusters. A functional pressure to maintain the number of distinct signals would thus have to either reverse the crystallization, or combine signals from different clusters.

In his later models signals are continuous trajectories and potentially a much larger distinct repertoire can emerge. However, the functionality of the repertoire is not monitored, and plays no role in the dynamics. The number of "phonemes" that form (the discretization of the acoustic space: categorization) is a consequence of the parameters and initial configuration, and in a sense accidental. The use of a limited number of points through which the trajectories pass (the superficial combination aspect) is built-in in the production procedure. The need for a large and distinctive repertoire, however, is a functional pressure. In Oudeyer's model there is no interaction between the number of phonemes that is created, and the degree of reuse (the number of phonemes per signal) that emerges. This issue, which seems the core issue in understanding the origins of combinatorial phonology, is not modeled by Oudeyer. In our model, in contrast, we ensure that the functionality increases rather than decreases.

2.4. Other models and linguistic theories

All other computational models of the evolution of combinatorial phonology that we are aware of, also assume the sequencing of phonetic atoms into longer strings as given. They concentrate rather on the structure of the emerged systems (de Boer, 2001; Ke, Ogura, & Wang, 2003; Lindblom et al., 1984; Redford, Chen, & Miikkulainen, 2001) or on how conventions on specific combinatorial signal systems can become established in a population through cultural transmission (Steels & Oudeyer, 2000). Theories on the evolution of speech developed by linguists and biologists focus on possible pre-adaptations for speech. MacNeilage and Davis (2000) propose oscillatory movements of the jaw such as used in breathing and chewing as precursors for syllable structure. Fitch (2000) sees sexual selection as a mechanism to explain the shape of the human vocal tract. These models and theories bridge the gap with empirical evidence on how combinatorial phonology is implemented in the languages of the world. However, they do not address the origins of the fundamental, qualitative properties of discrete and combinatorial coding. That is, they leave open the question as to under what circumstances a system of holistically coded signals with finite duration would change into a combinatorial system of signals.

Studdert-Kennedy (1998, 2000) has argued that combinatorial coding is a direct consequence of articulatory constraints. In this theory there is a hierarchy of difficulty of producing speech sounds, which is revealed in development. For instance, children master syllables like *ba* much earlier than syllables like *through*. The reason is that *through* requires a large number of carefully coordinated articulatory movements (gestures). Studdert-Kennedy speculates that the ability to produce such complex sounds is a relatively recent evolutionary innovation, and that the inherent difficulty makes the re-use of motor programs unavoidable. Hence, the combinatorial nature of speech follows from the difficulty of production and the large repertoire of words in human languages.

Consistent with this scenario is the neurological evidence discussed by Deacon (2000) that he believes shows intense selection for "the coupling of precisely timed phonation with rapid articulatory movements of tongue, lips and jaw." If Studdert-Kennedy and Deacon are right, the evolutionary transition to combinatorial phonology is characterized by radical changes in articulatory motor control. Nevertheless, this innovation is driven by the need for a large repertoire of perceptually distinct signals. While leaving open the possibility that the articulatory constraints already impose a form of combinatorial phonology, we do not need this assumption in the model of this paper. Rather, we study its evolution as the result of selection for perceptual distinctiveness alone.

3. Model design

Our model shares features with all existing approaches. Like the Liljencrants and Lindblom (1972) model, it makes use of an "acoustic space", a measure for perceptual distinctiveness and a hill-climbing heuristic. Like the Nowak and Krakauer (1999) model, the measure for distinctiveness is based on confusion probabilities, and our study includes a game-theoretic invasibility analysis. Finally, like Oudeyer (2002), we model signals not just as points, but as trajectories through acoustic space.

In the model, we do not assume combinatorial structure, but rather study the gradual emergence of superficially combinatorial phonology from initially holistic signals. We do take into account the temporal structure of both holistic and phonemically coded signals. We view signals as continuous movements ("gestures", "trajectories") through an abstract acoustic space. We assume that signals can be confused, and that the probability of confusion is higher if signals are more similar. We further assume a functional pressure that maximizes distinctiveness.

Real speech sound repertoires are of course subject to many other pressures and represent a complex compromise between acoustic distinctiveness, articulatory ease, conformity with (historically established) speech norms, frequency effects and interactions with other components of the language faculty. We should therefore not expect real vowel inventories to always maximize distinctiveness—as indeed they do not (e.g., Butcher, 1994). The model in this paper remains abstract and general.

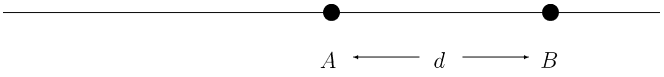
3.1. The acoustic space

The model of this paper will deal with repertoires of signals, their configuration and the similarities between signals. This requires conceptualizing signals as points or movements through a space. An appropriate definition of acoustic space will, as much as possible, reflect the articulatory constraints as well as perceptual similarities, such that signals that cannot be produced fall outside the space, and that points in the space that are close sound similar and are more easily confused.

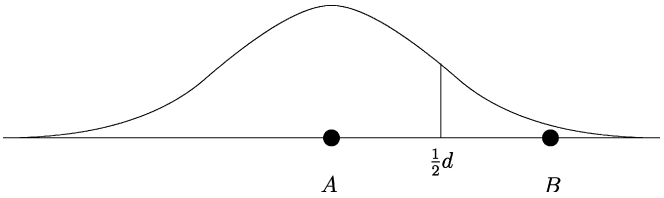
For human perception of vowels, a simple but very useful acoustic space can be constructed by looking at the formants. In contrast, it appears that pitch is a more salient variable in articulations of non-human primates (although at least some primates are able to manipulate and perceive formant frequencies as well, Andrew, 1976). Of course, it is difficult to tell what the appropriate acoustic space is for modeling articulation and perception of early hominids that feature in scenarios of the evolution of language (e.g. Jackendoff, 2002; Lieberman, 1984). However, the considerations that will be presented below remain the same, independent of the exact nature of the underlying perceptual dimensions.

3.2. Confusion probabilities

We now have to define how the distance in perceptual space relates to the probability of confusion. We can get a general idea by first looking at the simple example of a one-dimensional acoustic space with just two prototype signals A and B (modeled as points in that space), and a distance d between them:



Now assume that a received signal X, lying somewhere on the continuum, will be perceived as A or B depending on which is closest. Finally, assume a degree of noise on the emitted signals, such that when a signal, say A, is uttered, the received signal X is any signal drawn from a Gaussian distribution around A:



Now we can calculate the probability that an emitted signal A is perceived as B:

$$P(B\ perceived|A\ uttered) = \int_{(1/2)d}^{\infty} \mathcal{N}(\mu = 0, \sigma = \delta)\, dx = \int_{(1/2)d}^{\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-x^2/2\delta^2}\, dx, \qquad (3)$$

where δ is the standard deviation of the Gaussian. This integral, which describes the surface under the Gaussian curve to the right of the point $\frac{1}{2}d$ (midway between A and B), has a number of important features, as illustrated with the solid curves in Fig. 1 (the points are discussed later).

First of all, at $d = 0$, the confusion probability is not 100%, as the naive first intuition might be, but 50%. That is, even if two signals are identical, the hearer still has 50% chance of decoding them correctly. Second, with increasing d, the confusion probability first rapidly decreases and then slowly approaches 0. These are crucial properties: even though the confusion probability as a function of distance can have many different shapes depending on the exact type of noise and the exact type of categorization, the function will always have these general characteristics at $d = 0$ and in the limit of $d \to \infty$. In contrast, the previously discussed E measure, and summed distance measure, do not have both these properties.

If the acoustic space has more than one dimension, and if there are more than two signals, calculations like in Eq. (3), quickly get extremely complex, and confusion probabilities

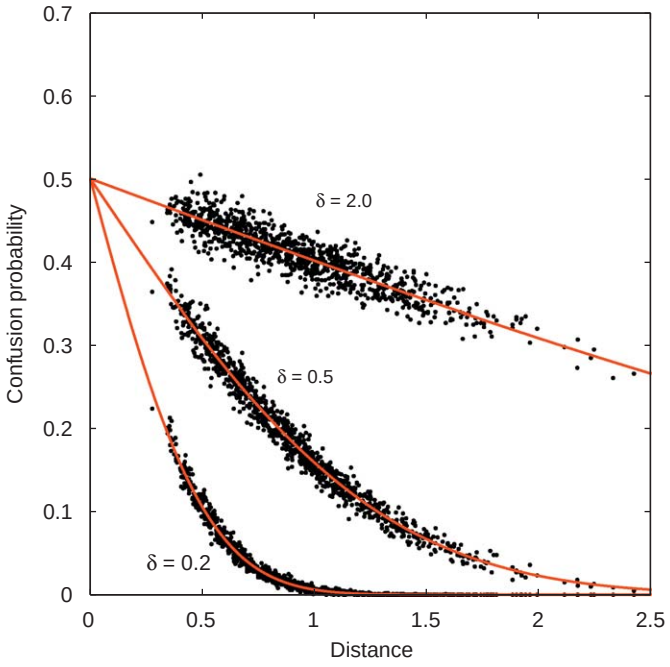


Fig. 1. The probability of confusion as a function of distance for several values of δ. The curves give the theoretical prediction based on the calculations in Section 3.2; the points are data from a computational simulation of the confusion probabilities between two trajectories in a two-dimensional acoustic space (discussed in Section 3.4).

are no longer uniquely dependent on distance. We can, however, assume that the confusion probabilities are generally proportional to a function of distance with a shape as in Fig. 1. Define $f(d)$ to be a function of distance d, parameterized by the noise level δ:

$$f(d) = \int_{(1/2)d}^{\infty} \frac{1}{\sqrt{2\pi}\delta} e^{-x^2/2\delta^2}\, dx, \\ = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{d}{2\delta\sqrt{2}}\right). \quad (4)$$

We assume that confusion probabilities are proportional to their "f-value": $P(B\ perceived|A\ uttered) \propto f(d(A,B))$. But we also know that the probabilities of confusing a signal with any of the other signals in a repertoire (including the signal itself) must add up to 1:

$$\sum_{X\in R} P(X\ perceived|A\ uttered) = 1.$$

Hence, we can estimate the probability of confusing signal A with signal B as

$$P(B\ perceived|A\ uttered) = \frac{f(d(A,B))}{\sum_{X\in R} f(d(A,X))}. \quad (5)$$

From this a measure for the *distinctiveness* D of a repertoire can be defined. Let $D(R)$ be the estimated probability that a random signal t from a repertoire R with T signals is correctly identified (note that $d(R_t, R_t) = 0$):

$$D(R) = \frac{1}{T}\sum_{t=1}^{T} \frac{f(d(R_t,R_t))}{\sum_{t'=1}^{T} f(d(R_t,R_{t'}))}. \quad (6)$$

3.3. Trajectory representation

We can now extend the model to deal with signals that have a temporal dimension. We define temporal signals as *trajectories*: movements through the acoustic space. In a digital computer, continuous quantities need to be discretized, and continuous trajectories will therefore need to be split up in a fixed number of points (this is analogous to the sampling of speech signals). In our approach, a trajectory is a connected sequence of points. Each connection between two points represents the acoustic and perceptual properties of a small interval of the original trajectory.

To illustrate the feasibility of deriving trajectory representations from acoustic data, we show in Fig. 2 a number of trajectories through vowel space that are based on actual recordings. The graph shows the trajectories from a number of recorded vowels, which correspond to more-or-less stationary trajectories in the space, and from recordings of a number of diphthongs, which correspond to movements from one vowel's region to another.

We will take as our starting point a set of trajectories through an abstract acoustic space. The model is based on piece-wise linear trajectories in bounded two-dimensional or three-dimensional continuous spaces of size 1×1 or $1 \times 1 \times 1$. Trajectories are sequences of a fixed number of points (parameter P). Each point has a maximum distance (parameter S) to the immediately preceding and following points in the sequence. This is to prevent signals from changing unrealistically fast. The following and preceding points to a point can lay anywhere within a circle of radius S with that point at the center. This way, there is no bias for straight trajectories. As will be seen in the results, this sometimes causes trajectories to 'jitter' around a point, but this effect is small enough to be considered noise. Trajectories always stay within the bounds of the acoustic space.

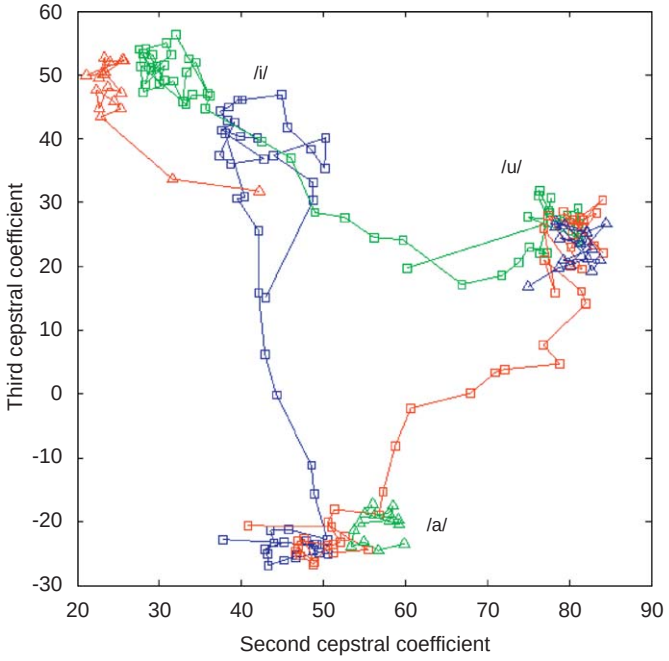


Fig. 2. Trajectory representations derived from recorded acoustic data from an American–English native speaker. Each point on a trajectory is given by the first two cepstral coefficients (Bogert, Healy, & Tukey, 1963) of the frequency spectrum of a short time interval of the signal.

As mentioned above, we need to discretize the trajectories. To ensure that we do not impose the combinatorial structure we are interested in, we discretize at a much finer scale than the combinatorial patterns that will emerge. Hence, the points implement a discretization of a continuous trajectory that can represent both a holistic and a combinatorial signal.

3.4. Measuring distances and optimizing distinctiveness between trajectories

When optimizing trajectories, we measure the distance between complete trajectories and optimize their distinctiveness. In such an approach, there is a role for combinatorial phonology: the confusion probability between two largely overlapping trajectories might be very low, as long as they are sufficiently distinct along one stretch of their length. We define the distance between two trajectories t_i and t_j, as the *average* distance between

the corresponding points on the trajectories:

$$d(t_i, t_j) = \frac{1}{P}\sum_{p=1}^{P} d(t_i^p, t_j^p), \quad (7)$$

where t_i^p is the p-th point on the i-th trajectory in a repertoire, and $d(a,b)$ gives the distance between two points a and b.

This distance measure then provides the input to the distance-to-confusion function that we derived for points (Eq. (6)). For trajectories, it is far from trivial to derive the exact shape of that function analytically, even if the noise and categorization mechanisms were completely known. However, we have performed computational experiments that demonstrate that our approximation is very accurate. In these simulations, noise was simulated using the DISTURBANCE function as will be defined in Section 3.5, and nearest neighbor classification. The dots in Fig. 1 show results relating distance between two trajectories with the probability that they are confused. The results give an excellent fit with the approximation of Eq. (6), and thus indicate that the distance-to-confusion function for points is also applicable to trajectories.

3.5. The hill-climbing heuristic

Now that we have defined a distance metric, it is straightforward to use a hill-climbing heuristic such as in Liljencrants and Lindblom (1972). Hill-climbing is an iterative procedure, where repeatedly a random change to the repertoire is considered, and if it improves the distinctiveness it is applied. In pseudo-code, the procedure looks as follows:

```
% R is a repertoire of signals
% S is the segment length parameter
% ρ is the hill-climbing rate parameter
% δ is the acoustic noise parameter
for i = 1 to I
  R' = CONSTRAIN(R + DISTURBANCE(ρ), S);
  if HILLCLIMBING-CRITERION(R, R', δ) then R = R';
end for
```

Here, DISTURBANCE applies random noise (from a Gaussian with $\mu = 0$ and $\sigma = \rho$), to all of the coordinates of a (uniformly) random point on a random trajectory. CONSTRAIN is a function that enforces that all points on the trajectories fall within the boundaries of the acoustic space, and that all segments have maximum length S. Hence, after a random point t_x is moved to a new random position, the CONSTRAIN function first moves it back, if necessary, within the boundaries of the acoustic space; it then moves the two points on both sides of the moved point, t_{x+1} and t_{x-1}, closer, if necessary, such that the distance to t_x is no more than S. The direction from t_x to t_{x+1} or t_{x-1} remains the same. The same procedure is applied iteratively to the neighbors of t_{x+1} and t_{x-1} until the ends of the trajectory are reached. The HILLCLIMBING-CRITERION(R, R', δ) in the basic model, which we call the "optimization condition" (OP), is defined as follows:

$$\text{OP:} \quad D_\delta(R') > D_\delta(R), \quad (8)$$

where D is the distinctiveness function given in Eq. (6). Note that this criterion is frequency-independent; in Section 4.5 we will consider frequency-dependent criteria.

Trajectories are initialized randomly. In the default initialization, we generate for each trajectory P random points (from a uniform distribution over the acoustic space), and then apply the CONSTRAIN function to it.

Hill-climbing is just an optimization *heuristic*; there is no guarantee that it will find the optimal configuration. The system is likely to move toward a local optimum. This problem is in general unavoidable for systems with so many variables. Hence, also in nature, the optimization of sound systems has not escaped the problem of local optima. The real optimum is therefore not necessarily interesting for describing the patterns in human speech. Instead, we will concentrate on general properties of the local optima we find, and on the gradual route towards them.

4. Results

We have implemented the model in C++ and MatLab.[3] We have run simulations with a large number of parameter combinations and a number of variations of the basic model. In the following we will first briefly give an overview of the general behavior of the model in these simulations by means of a representative example, and then give a detailed analysis of why we observe the kind of results that we do. In Sections 4.5 and 4.7 we will study extensions of the basic model where we test whether innovations can invade in populations where they are rare, and where we evaluate some of the other simplifications made in the basic model.

4.1. An overview of the results

Fig. 3(a) shows an equilibrium configuration of nine point-like signals in an abstract acoustic space, optimized for distinctiveness at an intermediate noise-level ($\delta = 0.1$). This particular configuration is stable: no further improvements of the distinctiveness of the repertoire can be obtained by making small changes to the location of any of the signals. The distinctiveness $D = 0.97$; that is, with the given noise level, our estimate of the probability of successful recognition of a signal is 97%.

Fig. 3(b) shows nine trajectories, consisting of 10 points and hence nine segments each. Each of these trajectories was created by taking 10 copies of one of the points in figure (a) and connecting them. A small amount of noise was added to each point, and the CONSTRAIN function, as described above, was applied to each trajectory, enforcing a maximum distance ($S = 0.1$) between all neighboring

[3]The program is available at http://staff.science.uva.nl/~jzuidema.

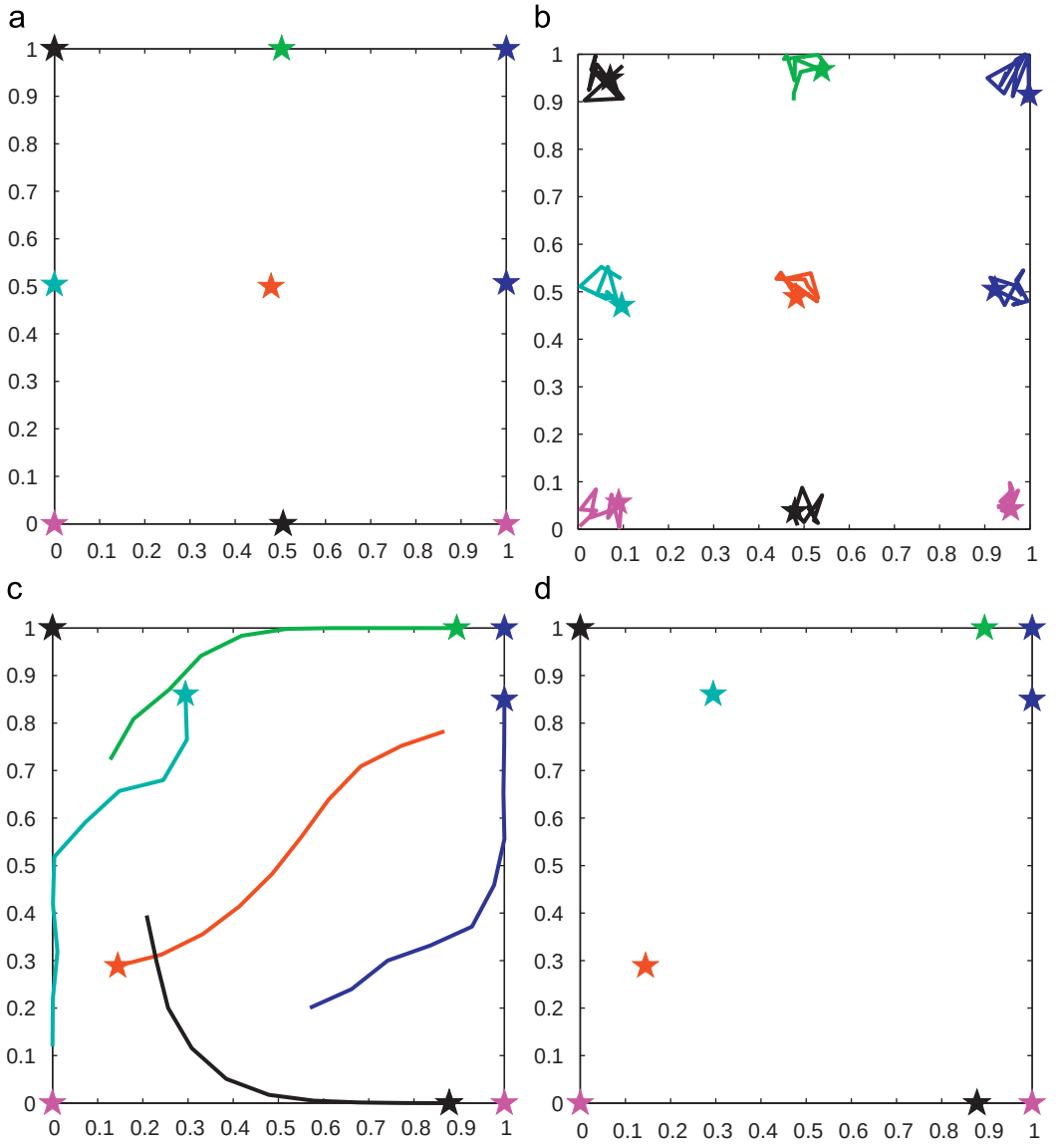


Fig. 3. In a combinatorial phonology, distinctiveness of signals at each particular time-slice is sacrificed for better distinctiveness of the whole trajectory. Instantaneous signal (or equivalently, stationary trajectories) will be organized in patterns like (a) and not like (d) when optimized for distinctiveness. For non-stationary trajectories, the same pattern, as in (b), is not stable, but will—after optimization—instead be organized like (c). Each individual time-slice, as illustrated with the end-points in (d) is suboptimal, but the whole temporal repertoire is at a local optimum.

points on the same trajectory. Due to this perturbation, the distinctiveness of this repertoire of trajectories is somewhat lower, $D = 0.94$, than of the repertoire in (a).

What will happen if we now optimize, through hill-climbing, the repertoire of trajectories for distinctiveness? One possibility is that the applied perturbations are nullified, such that the system moves back to the configuration of (a). That is not what happens, however. Rather, the system moves to a configuration as in Fig. 3(c). This graph shows a number of important features. First, all trajectories start and end near to where other trajectories start and end. The repertoire therefore can be said to exhibit a *superficially combinatorial phonology*: if we label the corners A, B, C and D, we can describe the repertoire as: $\{A, AB, B, CA, BC, C, CD, DB, D\}$. That is, we need only four category labels (phonemes) to describe a repertoire of nine signals. In contrast, the repertoire in (b) is most easily described by postulating nine categories, one for each trajectory.

Second, some trajectories are bunched up in as small a region as possible, but other trajectories are stretched out over the full length of the space. Third, the configuration of the repertoire is in a local optimum.[4] Fourth, at each time-slice the configuration of the corresponding points is in fact suboptimal. For instance, in Fig. 3(d) just the endpoints of the trajectories in (c) are shown. All of these points are closer to their nearest neighbor than any of the points in

[4]No qualitative changes have been observed in many thousands of additional iterations.

(a). Similar results are obtained for random initial conditions. We will now look at a number of simple cases that explain why the optimized repertoires have these features.

4.2. The optimal configuration depends on the noise level

To evaluate the role of the noise parameter δ, it is instructive to first look at a simple, one-dimensional example with signals as points. Consider a situation with three signals, two of which are fixed at the edges of a one-dimensional acoustic space. The third signal is at distance x from the leftmost signal, and at distance $1-x$ from the rightmost signal:

The x that maximizes distinctiveness depends on the noise level δ. Recall that distinctiveness D is defined as the average probability of correct recognition (Eq. (6)). In this case, we have three terms describing the recognition probabilities of each of the three signals. These are

$$P(t_1 \text{ perceived}|t_1 \text{ uttered}) = \frac{f(0)}{f(0)+f(x)+f(1)}, \quad (9)$$

$$P(t_2 \text{ perceived}|t_2 \text{ uttered}) = \frac{f(0)}{f(x)+f(0)+f(1-x)}, \quad (10)$$

$$P(t_3 \text{ perceived}|t_3 \text{ uttered}) = \frac{f(0)}{f(1)+f(1-x)+f(0)}. \quad (11)$$

The values of these three functions, for two different choices of δ are plotted in Fig. 4(a) and (b). If we add up these three curves, we find the curves in Fig 4(c). Clearly, for low levels of noise the optimal value of x is $x=0.5$. For higher noise levels this optimum disappears, and the optimal configuration has $x=0$ or 1. That is, if there is too much noise, it is better to have several signals overlap.

Fig. 5(a) shows a two-dimensional system of nine points optimized for distinctiveness with a high-noise level ($\delta=1$). The optimal configuration under these conditions is to have each signal in one of the four corners: three corners with two signals, and one corner with three signals. With this configuration, the distance between the two or three signals that share a corner is $d=0$, and their mutual confusability is high. But at least the distance to the other signals is high ($d=1$, or $d=\sqrt{2}$).

Maximizing distinctiveness is, because of the high-noise level, equivalent to maximizing summed distance. Consider one of the signals in the top-right corner, and consider moving it to the left, that is, away from the two signals already in that corner. The gain in distance from the top-right corner (Δdtr), will be exactly canceled out by the loss in distance from the top-left corner (Δdtl). The gain in distance from the bottom-right corner (Δdbr), however, will not compensate for the loss in distance from the bottom-left corner (Δdbl). To see why, consider moving the signal a distance ε to the left. The (squared) gain in distance to the top-right is given by

$$\Delta dbr^2 = [\varepsilon^2+1]-[1] = \varepsilon^2. \quad (12)$$

The (squared) loss in the distance to the top-left by

$$\Delta dbl^2 = [1+1]-[(1-\varepsilon)^2+1] = [1+1]-[1-2\varepsilon+\varepsilon^2+1] = 2\varepsilon-\varepsilon^2. \quad (13)$$

The summed distance will increase only if (12) is larger than (13), which is never the case if $0 \leqslant \varepsilon \leqslant 1$.

In contrast, in Fig. 5(b) a system of nine points is shown that has been optimized for distinctiveness at a relatively low noise level ($\delta=0.1$). Here maximizing distinctiveness is not equivalent to maximizing summed distance, because of the relatively low noise level. To see why the noise level determines whether it is equivalent, consider a small change to the configuration, for instance moving the central point a bit to the left. Such a change will decrease the distance to some points, and increase the distance to some other

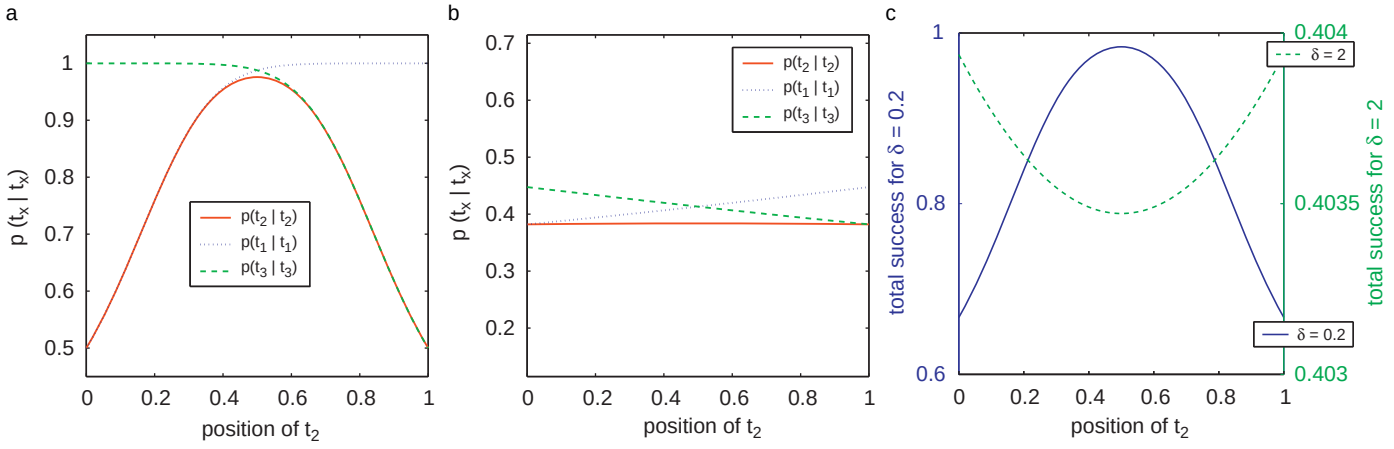


Fig. 4. Distinctiveness as dependent on distance and noise, one-dimensional example. Panel (a) shows confusion probabilities in a low-noise environment, panel (b) shows them in a high-noise environment. Panel (c) compares the probabilities of correct recognition for the low-noise and the high-noise conditions. Note that the high-noise condition has a minimum at maximal distance, whereas the low-noise condition has a maximum.

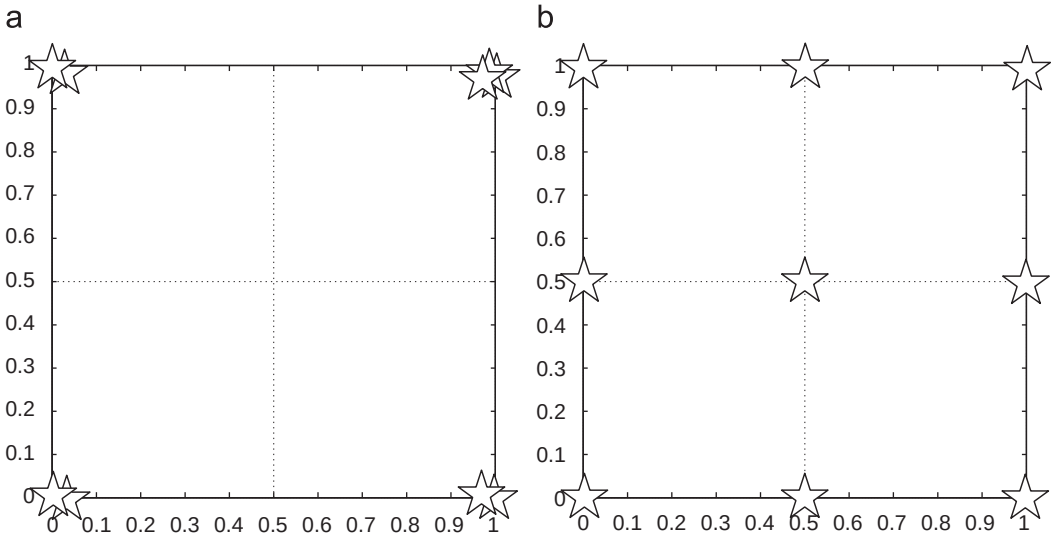


Fig. 5. The noise level determines how many signals can be kept distinct.

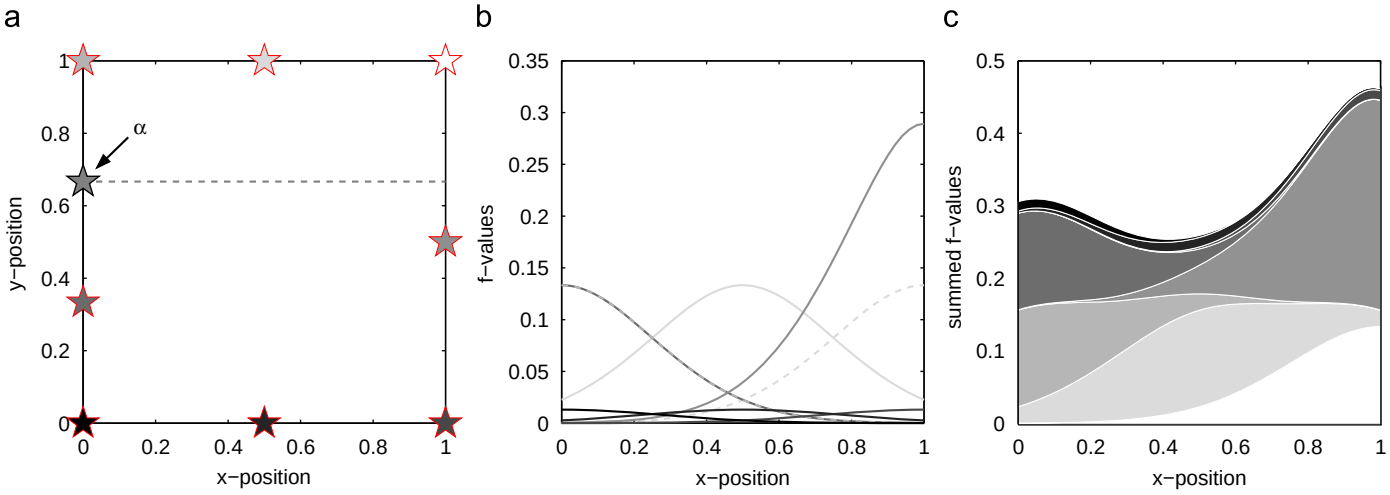


Fig. 6. Figure (a) shows a local optimum of a nine-signal repertoire optimized for distinctiveness. What would happen if we move the signal at the right end of the interval in (a) horizontally to left? The probability of correct recognition of that signal, α, is inversely proportional to the sum of the f-scores of all other signals (see Eq. (5)). Figures (b) and (c) show why this probability is in a local optimum with α at its current location. Parameters: $\delta = 0.3$.

points. Now, note that the distance-to-confusion function is approximately linear for relatively small distances (see Fig. 1). Therefore, maximizing distinctiveness corresponds approximately to maximizing average distances only if distances are small *relative to the noise level*, or equivalently, if the noise level is high *relative to the distances*.

4.3. Distinctiveness is a non-linear function of distances

Fig. 6 shows another two-dimensional, nine signal system. It has, after running the hill-climbing algorithm, converged to a local optimum (a). Why is this configuration stable? Consider moving the signal α at the left-most end of the interval, along that same interval. For each alternative x-coordinate of that signal, we can calculate the estimated probability of confusion with other signals. The f-values for all the other signals are plotted in figure (b). For instance, the f-value of the central-right signal (its contribution to the confusion about α) goes from very low if α is at the left-most end of the interval to very high (0.3) if α is at the right-most end of the interval.

The probability of correct recognition of α, and hence its contribution to the total distinctiveness, is inversely proportional to the sum of all f-values. In Fig. 6(c) we therefore give a plot of the sum of all these values (with the contribution of each signal indicated in different colors). That sum is in a local minimum at the actual location of α, which suggests that—at least initially—distinctiveness will not improve by shifting α to the left. This is not the whole story, though, because the probability of correct recognition of the other signals will also change. Nevertheless, the distinctiveness of a repertoire is a non-linear combination

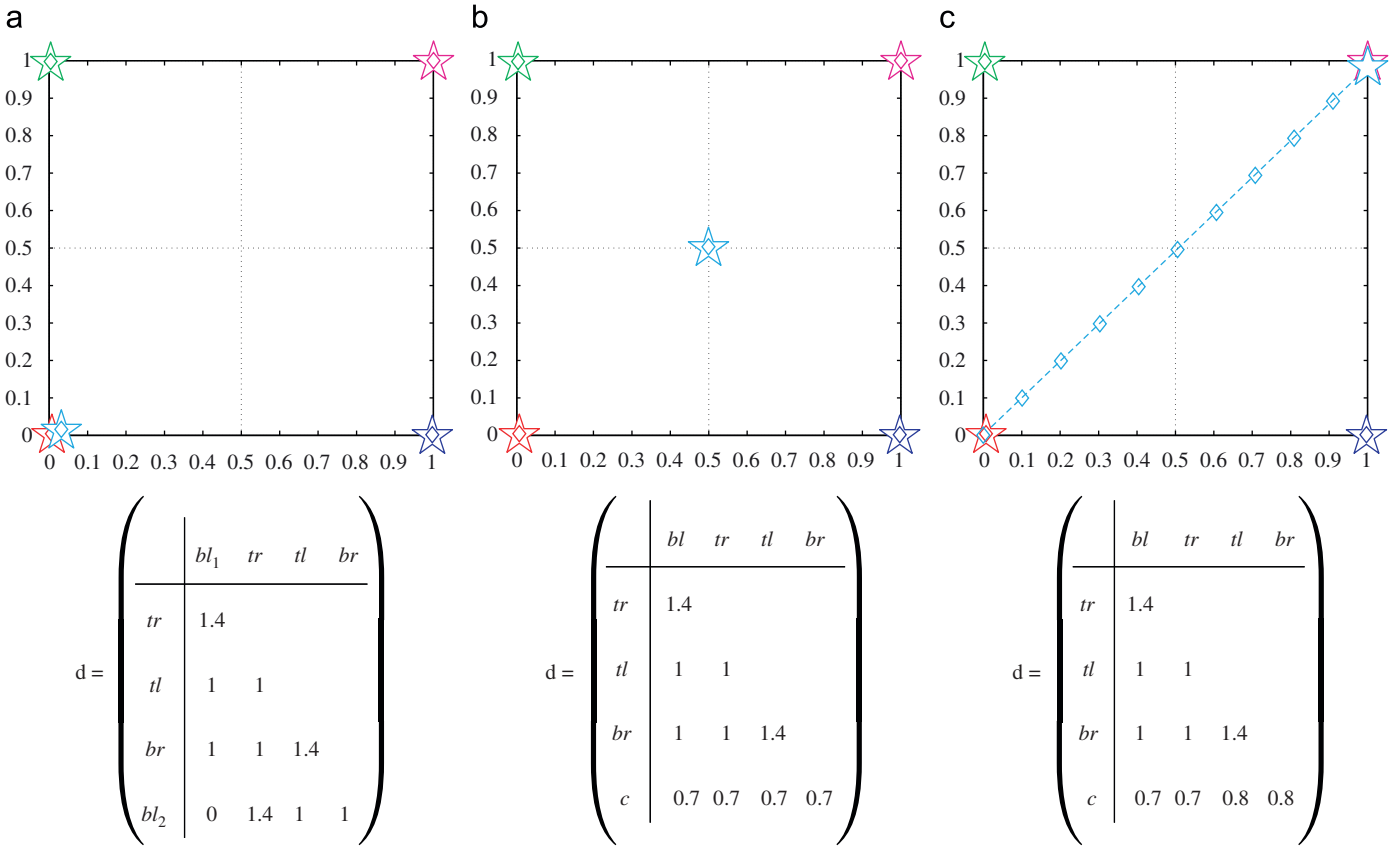


Fig. 7. Why do trajectories stretch out? Three configurations and their distance matrices. Abbreviations: bl, bottom-left; tr, top-right; tl, top-left; br, bottom-right; c, center.

of the distances between the signals. Due to this non-linearity, the resulting stable configurations are sometimes counter-intuitive.

4.4. Why trajectories stretch out

Finally, in Fig. 7 we explore the question of why many trajectories in our simulations stretch out. In figure (a) we show five signals (in the bottom-left corner there are two signals on top of each other). The signals are points in the acoustic space, which we will here interpret as *stationary* trajectories of some arbitrary length. The graph shows the configuration that maximizes the summed distance between the signals. The figure also gives the distance matrix that gives the distance between every pair of signals. The values are equal to the Euclidean distance (the distance as measured with a ruler), $\sqrt{2} \approx 1.4$ (across the diagonals), 1 (horizontally or vertically) and 0 (for the pair in the bottom-left corner). The average distance is $\overline{d} = 10.2/10 = 1.02$.

Fig. 7(b) shows an alternative configuration, with the fifth signal in the center. The distance matrix shows that the distance of the fifth signal to the bottom-left corner has increased, but at the expense of the distances to the three other corners. As a result, the average distance has actually gone down to $\overline{d} = 0.96$. The reason is that this configuration does not make optimal use of the longest available distances over the diagonal. Importantly, however, at low noise levels, the distinctiveness of this configuration is in fact higher than of the configuration in (a). The reason is that with relatively little noise and long distances, the distinctiveness-distance function flattens out. Hence, there is more to be gained from avoiding confusion between the fifth and the bottom-left signal, then there is from maintaining the excessive "safety margin" with the other signals. In other words, the configuration in (b) sacrifices some average distance, to gain a lower average confusion probability.

Fig. 7(c) shows yet another configuration, now with the fifth trajectory stretched out over the whole diagonal. As is clear from the given distance matrix, this configuration yields larger distances than in (b). To go from (b) to (c) there is no trade-off. The distances from the central, fifth signal to the top-left and bottom-left corners can be increased without decreasing the distances to the other two signals. The reason is that the distance between a stationary trajectory t and a stretched out trajectory t' is equal to the distance between t and the centroid of t' when t is on a line through all the points of t', but larger when it is not.

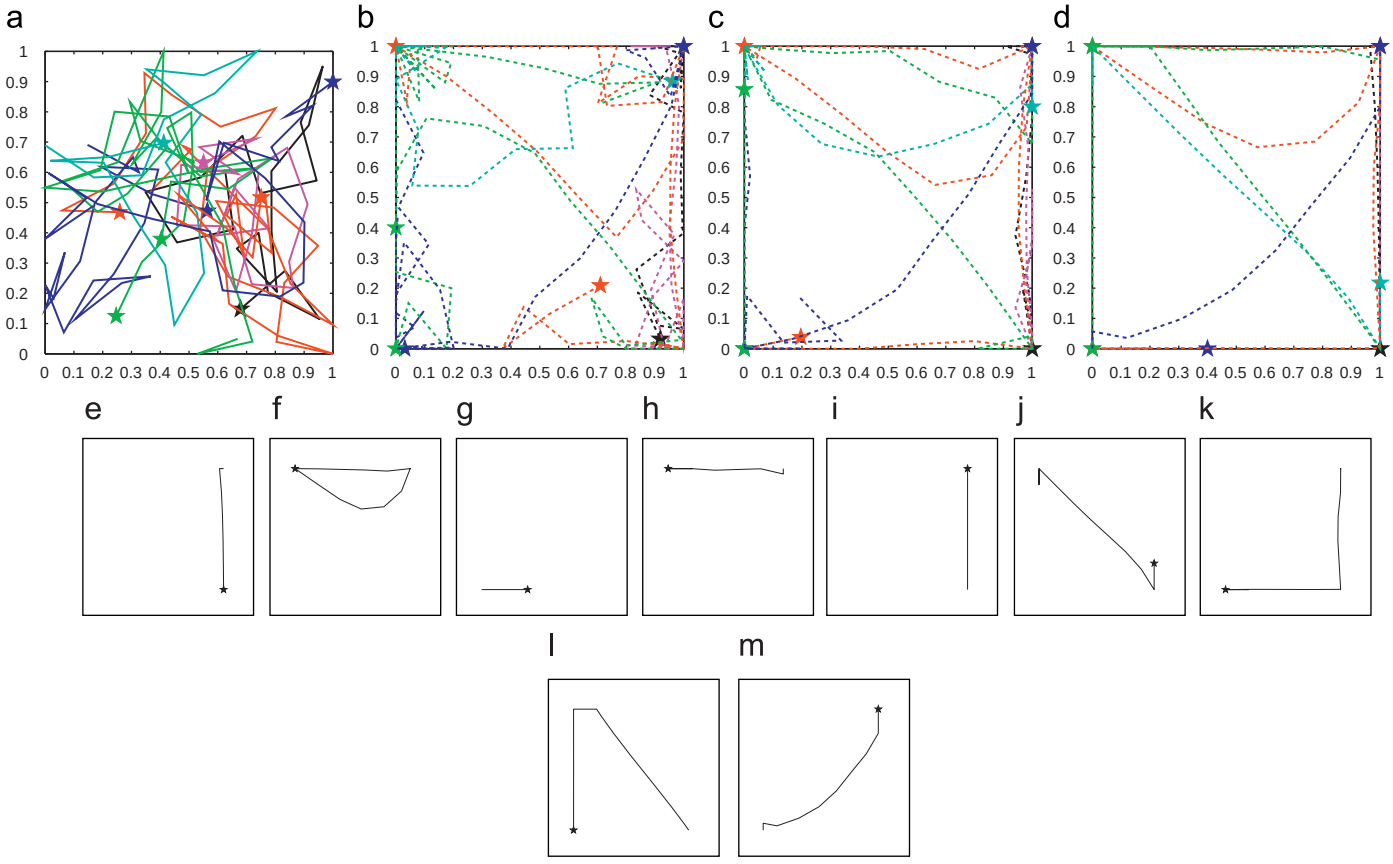


Fig. 8. A repertoire in a two-dimensional acoustic space optimized for distinctiveness. Common parameters: $T = 9, P = 20, 2d, S = 0.2, \rho = 0.2, \delta = 0.25$. Figures (a–d) show the configurations at various stages of the hill-climbing process. Figures (e–m) show each of the individual trajectories in figure (d).

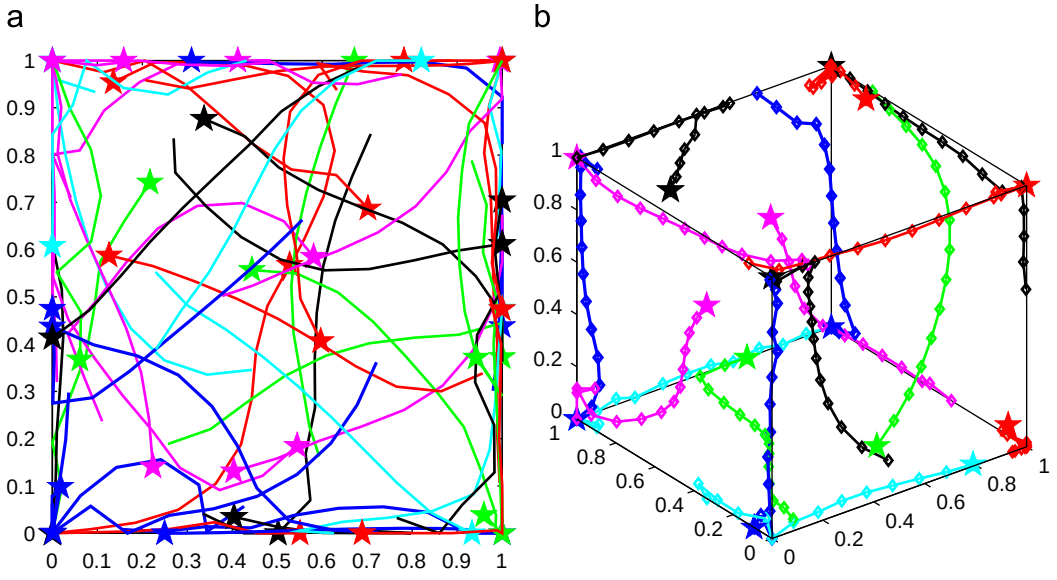


Fig. 9. Repertoire in a two-dimensional and three-dimensional acoustic space optimized for distinctiveness.

In Figs. 8 and 9 we show results from running the basic model under various parameter settings, including with repertoires with many trajectories and with three-dimensional acoustic spaces. These results show that the observations made in the simple systems above, generalize to a wide range of conditions.

4.5. Locally optimal repertoires are ESSs

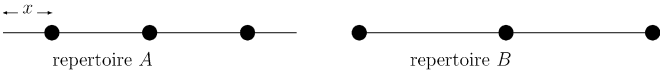
So far, we have seen that repertoires of signals with a temporal structure will, when optimized for distinctiveness, stretch out. Rather than staying away as far as possible from other trajectories along its whole length, each trajectory will be close to some trajectories for some of its length, and close to other trajectories elsewhere. In qualitative terms, these systems show superficially combinatorial structure.

We have not, however, dealt with the question whether an innovation is able to invade and become established in a population where it is very infrequent. To investigate these questions, we change the definition of distinctiveness to tell us something about pairs of languages. This way we can ask the question: how well will a repertoire R' (with T trajectories) do when communicating with a repertoire R? Pairwise distinctiveness $\mathscr{D}$ is defined as follows:

$$\mathscr{D}(R,R') = \sum_{t=1}^{T} \frac{f(d(R_t, R'_l))}{\sum_{t'=1}^{T} f(d(R_t, R'_{t'}))}. \tag{14}$$

The quantity $\mathscr{D}(R,R')$ can be interpreted as the estimated probability of a signal uttered by a speaker with repertoire R, to be correctly interpreted by a hearer with repertoire R'.

When we now consider the invasion of a *mutant* repertoire R' into a population with *resident* repertoire R, four measures are of interest: $\mathscr{D}(R,R)$, $\mathscr{D}(R,R')$, $\mathscr{D}(R',R)$ and $\mathscr{D}(R',R')$. That is, how well does each of the repertoires fare when communicating with itself or with the other repertoire, in the role of speaker or of hearer? Specifically, for the invasion of R', it is necessary that $\mathscr{D}(R',R) > \mathscr{D}(R,R)$ or $\mathscr{D}(R,R') > \mathscr{D}(R,R)$, or some weighted combination of these requirements. That is, a successful mutant must do better against the resident language, than the resident language does against itself.



This situation turns out to be very common. Consider the above one-dimensional example: The configuration on the right (B) is better on all accounts. Obviously, there will be less confusion between its signals because they are further apart (when $x = 0.1$ and $\delta = 0.1$, $D(A) = \mathscr{D}(A,A) = 0.70$ vs. $\mathscr{D}(B,B) = 0.84$). But configuration B will even do better when communicating with A, both as a hearer ($\mathscr{D}(A,B) = 0.78$) and as a speaker ($\mathscr{D}(B,A) = 0.76$).

The HILLCLIMBING-CRITERION(R,R') is redefined as follows in each of the conditions "hearer benefits" (HB), "speaker benefits" (SB) or "equal benefits" (EB):

$$\text{HB:}\quad \mathscr{D}(R,R') \geqslant \mathscr{D}(R,R), \tag{15}$$

$$\text{SB:}\quad \mathscr{D}(R',R) \geqslant \mathscr{D}(R,R), \tag{16}$$

$$\text{EB:}\quad \tfrac{1}{2}(\mathscr{D}(R',R) + \mathscr{D}(R,R')) \geqslant \mathscr{D}(R,R). \tag{17}$$

It turns out that all the stable configurations we found in simulations with the optimization criterion (OP, Eq. (8)), are also stable under criteria HB, SB and EB. Thus, locally optimal repertoires are evolutionary stable strategies.

4.6. Not all ESSs are locally optimal

ESSs are strategies that cannot be invaded by any other strategy. In evolutionary game theory, ESSs are therefore considered likely outcomes of an evolutionary process. However, if there are many ESSs in a given system, the initial conditions will determine which ESS will emerge ("equilibrium selection"). In our simulations with the HB, SB and EB conditions, we also observe ESSs that do not correspond to the locally optimal configurations that we found with the OP condition.

Fig. 10(a–d) show the configuration of the repertoire at different numbers of iterations of the hill-climbing algorithm under the HB condition. Fig. 10(i) gives the pairwise distinctiveness measures for each combination of these four configurations. At the diagonal of this matrix are the distinctiveness scores of each configuration. As is clear from this matrix, each next configuration can invade a population with the previous repertoire. In boldface we see the approximate evolutionary trajectory (the actual steps in the simulation are much smaller). Figure (d) is an ESS. However, figures (e–f) show that this configuration is not stable when the OP criterion is used. Fig. 10(j) gives the pairwise distinctiveness measures for each combination of these five configurations. The diagonal elements in this matrix are the highest values in their row and column, which shows that none of these configurations could have invaded a population using (d) under the HB (or SB, or EB) condition. Nevertheless, once adopted, communication is more successful with every next configuration (as the diagonal elements show). The locally optimal configuration in (f), however, is an ESS under all four conditions.

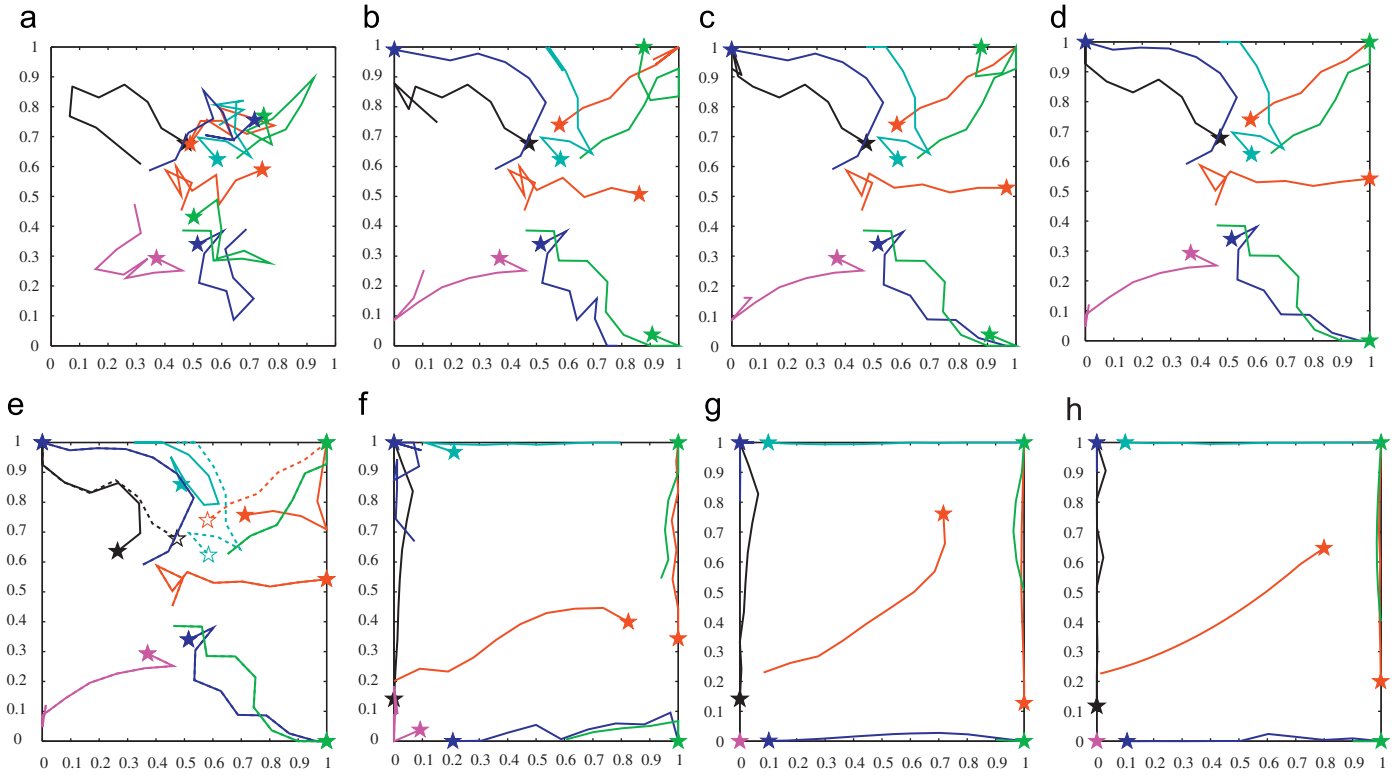
We find suboptimal ESSs in simulations with the SB and EB conditions as well. Fig. 11 shows stable configurations that emerge in each of the four conditions. Interestingly, these suboptimal ESSs disappear when a different distance-to-confusion function is used. We used

$$f(d) = \frac{1}{1 + e^{(1/\delta d^2)}}, \tag{18}$$

instead of Eq. (4). In this, and other simulations (two-dimensional and three-dimensional) with that same function, all ESSs observed show the same type of superficially combinatorial phonology that we found in the OP condition. This shows among other things that the exact behavior of systems as complex as this one can depend on the details of their implementation.

4.7. Individual-based model

As a final test of the appropriateness of the basic model, we studied an individual-based simulation of a *population*



i

D* =

	a	b	c	d
a	**0.250**	**0.255**	0.256	0.257
b	0.257	**0.411**	**0.412**	0.414
c	0.260	0.412	**0.443**	**0.445**
d	0.262	0.415	0.449	**0.458**

j

D* =

	d	e	f	g	h
d	**0.458**	0.445	0.374	0.353	0.354
e	0.445	**0.465**	0.391	0.368	0.370
f	0.4030	0.417	**0.599**	0.563	0.569
g	0.387	0.400	0.564	**0.629**	0.614
h	0.389	0.402	0.570	0.615	**0.634**

Fig. 10. Locally optimal repertoires are ESSs, but not all ESSs are locally optimal. (a–d) show configurations in an evolutionary simulation with the hearer benefit condition (HB, $D(R, R') > D(R, R)$) at various time steps; (d) is an ESS in the HB condition; (e–h) show results from a simulation in the optimization condition (OP, $D(R', R') > D(R, R)$) that used (d) as its initial condition. (h) is an ESS in all conditions (OP, HB, SB, EB) considered. (i) shows a matrix that gives the pairwise distinctiveness scores for every combination of configurations in (a–d); (j) the matrix that gives the pairwise distinctiveness scores for every combination of configurations in (d–h). The approximate evolutionary trajectory is indicated with boldface in these matrices. Parameters are: $T = 9$, $P = 10$, $D = 2$, $N = 0.05$, $S = 0.1$.

of agents that each try to imitate each other in noisy conditions. This simulation is similar to the model described above, but now each agent in the population has its own repertoire, and it tries to optimize its own success in imitating and being imitated by other agents of the population.

This version of the model is similar to the imitation games of de Boer (2000). That paper only modeled point-like signals (vowels) and did not investigate combinatorial phonology. The game implemented here is a slight simplification of the original imitation game. First, all agents in the population are initialized with a random set of a fixed number of trajectories, using the "elaborate" initialization scheme from Section 3.5. Then for each game, a speaker is randomly selected from the population. This speaker selects a trajectory, and makes a random modification to it. Then it plays a number of imitation games (50 in all simulations reported here) with all other agents in the population. In these games, the *initiator* utters the modified trajectory with additional noise. The *imitator* finds the closest trajectory in its repertoire and utters it with noise. Games are successful if the imitator's signal is closest to the modified trajectory in the initiator's repertoire. If it turns out that the modified trajectory has better imitation success than the original trajectory, the modified trajectory is kept, otherwise the original one is restored.

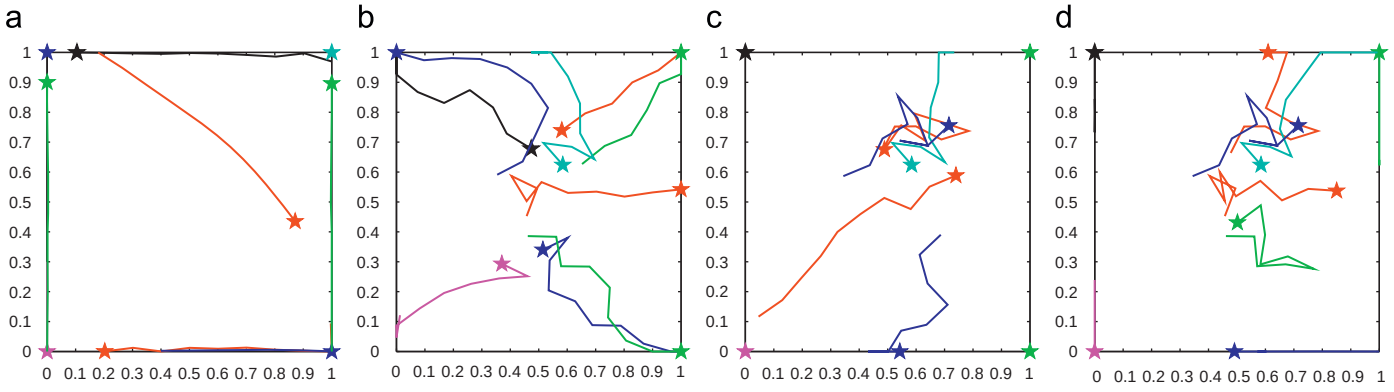


Fig. 11. Not all ESSs are locally optimal. Results from four simulations, each with the initial condition as in Fig. 10(a). Different payoff functions lead to different ESSs, although for all payoff functions considered, locally optimal configurations as in (a) are stable. Parameters are: $T = 9, P = 10, D = 2, \rho = 0.2, \delta = 0.2, S = 0.1$.

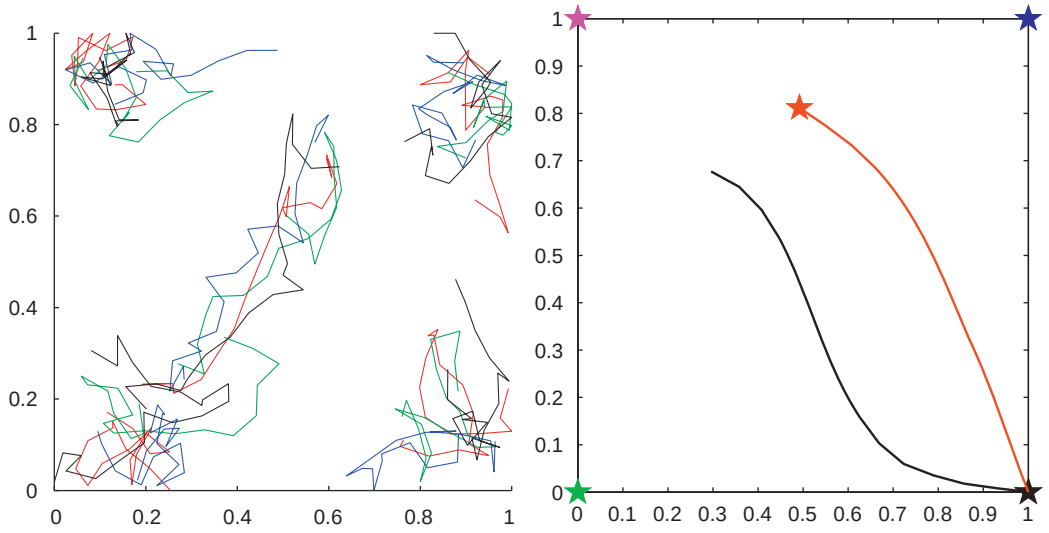


Fig. 12. Comparison of population-based models with the optimization model. Frame (a) shows the (five) trajectories of four agents (from a population of ten). Notice the similarities with the optimized trajectories in frame (b).

For vowel systems, it has been shown that optimizing a single repertoire leads to similar systems as a population-optimization system (compare de Boer, 2000; Liljencrants & Lindblom, 1972). It can be shown that for trajectories the same is true. This is illustrated in Fig. 12.

In this figure the left frame shows the system of five trajectories that resulted from playing imitation games in a population. The right frame, for reference, shows a system of five trajectories that resulted from optimizing distinctiveness as in the basic model. It can be observed that in both cases, the corners are populated by four trajectories, which are bunched up. The fifth trajectory, in contrast, follows the diagonal. As before, an analysis in terms of phonemes suggests itself: the four corners are basic phonemes, while the fifth trajectory uses one of the corners as a starting phoneme and the opposite corner as the ending phoneme. Although less clean and not fully conclusive, the results from the individual-based model seem to be consistent with the observations from the basic model.

4.8. Measuring combinatorial phonology

So far, we have relied on an intuitive notion of what it means for a repertoire of trajectories to show combinatorial phonology. As the paper investigates the emergence of categorical and combinatorial coding on a qualitative level, this has not been an obstacle. Fig. 8(e–f), by showing all individual trajectories, give perhaps the most convincing example: every trajectory in the final repertoire (Fig. 8(d)) shares approximate begin or end points with some other trajectory (often the corners of the acoustic space), and one can uniquely characterize each trajectory by giving the sequence of corner points it visits (that is, for *recognition*; for *reproducing* the sounds we would need more information). This was still impossible in the earliest stages of the simulation (Fig. 8(a)).

However, for a more quantitative understanding of the emergence of combinatorial phonology, a numerical measure of the degree of recombination would be useful. When the basic building blocks of the repertoire are

known, it is straightforward to give a measure of the degree of combination. An example would be the number of times a given building block is being reused: $\varphi = N/k$, where φ is the measure of recombination (phonemicity), N is the number of words in the repertoire and k the number of building blocks.

A difficult problem, however, is finding the basic building blocks, given a repertoire of signals. The traditional linguistic procedure for identifying phonemes in an unknown language relies on the notion of "minimal pair": a pair of words that have different meanings and differ only in one sound. This analysis might seem straightforward, but when put to use in an automatic phoneme discovery procedure, a number of pitfalls emerge. First of all, it already makes use of an existing set of basic building blocks (usually the phonetic categories as defined by the International Phonetic Association). Secondly, it needs to decide what is meaningful variation and what is variation caused by automatic influence from neighboring sounds. However, this can only be deduced from the articulation, and the use of these sounds in a language, not from the signals themselves. Although infants do learn which sounds are phonemes and which sounds are allophones, they do not do this on the basis of the sounds alone, they also make use of the meanings of the words in which they occur and of proprioception of their own articulations. Such information is not available to an automated phoneme-detection procedure.

A final problem in the analysis is how to decide that two building blocks are equal. In many cases of allophony, it can be argued that the allophones are instances of the same building block, because there are no minimal pairs and their articulation is very similar. However, the case is not always so clear-cut, and it is therefore not uncommon to find disagreement amongst linguists about the exact set of building blocks that are used in a language.

In the case of artificially generated sets of trajectories, these problems are even more apparent. One needs to make assumptions about what kind of building blocks are possible, what constitutes natural behavior for a trajectory between building blocks, and how to define similarity between building blocks. In the analyses that we have presented, we have implicitly assumed that building blocks are points, that trajectories tend to follow straight lines from point to point and that points that are close together are the same building block.

Appendix A presents an approach to calculating the degree of phonemicity of a repertoire of trajectories, based on such simplifying assumptions. The quantitative results from applying this approach to random, optimized and hand-designed systems of trajectories support the qualitative observations we have relied on in the main text.

5. Discussion and conclusion

When optimizing a repertoire of temporally extended trajectories in an abstract acoustic space, the trajectories tend either to occupy the corners of the available space or to stretch out from corner to corner. It appears as if trajectories become *far apart where possible* and *close together where necessary*. A repertoire with this structure can be analyzed as reusing certain points as building blocks of its trajectories and thus to have combinatorial structure. Such a system also has discrete coding, because the building blocks are clearly separate and the trajectories between them stretched out and predictable from the position of the building blocks. As there is nothing in the trajectory that explicitly codes discreteness or combinatorial structure, and as agents that would use these repertoires of trajectories need not be aware of their structure, their combinatorial nature is purely *superficial*.

Most of the results presented here were obtained through direct optimization of repertoires of trajectories. However, we have also shown that similar repertoires of trajectories can emerge in a population of agents that try to imitate each other as well as possible. Apparently agents that strive for maximum success in imitation in noisy conditions, using information from simple interactions (imitation games) alone, converge towards repertoires that are similar to repertoires that are optimized directly.

Finally, we have shown that repertoires of trajectories that are optimized for acoustic distinctiveness (and thus combinatorial) are evolutionary stable. Agents that have a repertoire of trajectories that is more distinctive can invade a population of agents that have less distinctive (but otherwise similar) repertoires, at least if the only fitness criterion is the robustness to acoustic noise of their repertoires. Conversely, a population of agents with an optimal repertoire cannot be invaded by agents with less optimal repertoires. We have also shown that there is a path of ever increasing fitness towards the optimal (and combinatorial) repertoire.

Our model differs from other attempts to explain combinatorial speech in several ways. First of all, both holistic and (superficially) combinatorial signals have temporal structure. All signals in the model are of the same duration. Secondly, our model does not use articulatory targets. The resulting structure is purely emergent and therefore called "superficial". In fact, no distinction is made between holistic and combinatorial signals in the model; the difference only becomes apparent when analyzing the structure of the repertoires.

We argue that agents can make use of this structure to evolve towards productive use of recombination. When the structure becomes available in the population, it becomes advantageous for agents to make use of it. They can use it to store the repertoire of trajectories more compactly, to perceive and produce trajectories in a more robust way and eventually to more easily create new trajectories. In this way, agents that use combinatorial structure productively can invade a population of agents that do not. This is only possible when there already exists a repertoire that is superficially combinatorial. Only then is there a path of continuously increasing fitness towards productive

combinatorial coding, and eventually, to phonemic speech. We have shown that optimization for acoustic distinctiveness can result in such a repertoire.

Natural language phonology is categorical and combinatorial. What we have shown in this paper is that these properties have functional significance: they aid the reliable recognition of signals by the hearer. We have also shown that there is a path that leads from a signal system without these properties, to one that can be viewed as having those properties. Crucially, we have shown that each step on this path represents an improvement, both when it first appears in a population and when it is already common.

It turns out that a system that shows categorical and superficially combinatorial structure is advantageous even for a population of speakers and listeners that is not aware of this structure. We note that these results are consistent with several different scenarios on the origins of combinatorial speech. In particular, it is not necessary for combinatorial phonology to have emerged purely through genetic evolution. Rather, we see natural selection as a force that has shaped the parameters of the self-organizing process, and cultural self-organization as a process that determined which genetic adaptations would be beneficial. Hence, *self-organization is the substrate of evolution* (Boerlijst & Hogeweg, 1991; Kirby & Hurford, 1997; Smith, 2004; Waddington, 1939).

Acknowledgments

W. Z. is funded by the Netherlands Organisation for Scientific Research (Exacte Wetenschappen), project number 612.066.405. B. d B. is funded by the Netherlands Organisation for Scientific Research project number 276-75-007. Part of this research was performed whilst W. Z. was at the Language Evolution and Computation research unit and the Institute of Evolutionary Biology of the University of Edinburgh; an early version of this paper appeared as chapter 4 of a dissertation written at that institution (Zuidema, 2005). We like to thank Nick Barton, Jim Hurford, Simon Kirby, Matina Donaldson, Tecumseh Fitch, Mark Steedman, Björn Lindblom, Michael Studdert-Kennedy and three anonymous reviewers for their comments on earlier drafts of this paper.

Appendix A

In this paper we have only shown graphically and qualitatively that the repertoires emerging from our simulations show combinatorial phonology. To allow more quantitative statements, and evaluate the statistical significance, we need a numerical value that expresses the degree to which building blocks are reused in the system. Such a measure will be called the system's *phonemicity*. Unfortunately, defining such a measure raises many conceptual and technical issues, and no prior work exists on which we can base ourselves.

In linguistics, researchers have never found it necessary to measure the phonemicity of a language directly from the speech signals. All languages use recombination of a small set of signals to create words. It is not always possible to determine the exact number of contrastive building blocks unambiguously and there is even some controversy about what exactly make up the building blocks of speech. However, it is never necessary to determine the building blocks and thus the degree of phonemicity from the acoustic properties of the signal alone, as word meaning is always used in this procedure through the use of minimal pairs.

In animal behavior research, it would be useful to be able to have a measure of phonemicity in the study of the structure and the complexity of animal vocalizations, but it is clear that the minimal pair procedure is of little use here. A certain degree of recombination appears to be used in many animal signalling systems, but it is highly unlikely that these systems are as complex as human speech. All studies that investigate the complexity of animal call systems that we are aware of use a set of heuristics based on the ability of human observers to detect patterns in spectrograms. Hence, a well-defined measure of phonemicity also lacks in the biological literature.

Probably the best conceptual basis for such a measure is that of the compressibility of the underlying set of signals. A purely holistic set of signals cannot be compressed much, while a completely combinatorial set of signals should be most compressible. Of course, the set of signals should still contain enough diversity such that different signals can be distinguished from each other.

Measuring compressibility in real signals is problematic, however. Like the trajectories in our model, the features of real signals vary on continuous scales (amplitude, frequency, phase, etc.). Traditional (lossless) compression algorithms based on redundancy in strings of discrete symbols are therefore not usable. Nevertheless, there is redundancy in continuous data—points on a smooth trajectory can for instance be predicted from the previous points—and an algorithm that measures phonemicity should make use of this.

When measuring phonemicity it is thus most useful to think of *control points* and *predictable trajectories* between these control points. In the measure that will be defined below, we will assume, for simplicity, that there are only two control points per trajectory (the start and endpoints). We further assume that a trajectory moves more or less along a straight line between two control points. This is not strictly true for the kinds of system that emerge from the optimization process, as sometimes trajectories appear to move towards one point at first, and then towards another. Our measure therefore systematically underestimates the true phonemicity of emerged systems.

If a system shows combinatorial structure, it is expected that start and endpoints of trajectories tend to cluster together. Intermediary points on a trajectory are expected to be more spread out through the available space. Using

Liljencrants and Lindblom's (1972) measure of dispersion, we can then define a measure of phonemicity. The average dispersion of all start and endpoints, E, can be calculated as follows:

$$E = \frac{1}{2N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} [D(i_1,j_1) + D(i_1,j_L) + D(i_L,j_1) + D(i_L,j_L)].$$

The average dispersion of all other points on the trajectories, P can be calculated analogously

$$C = \frac{1}{N(N-1)(L-2)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=2}^{L-1} [D(i_k,j_k) + D(i_{L-k+1},j_k)].$$

The dispersion of two points, $D(\mathbf{p}_1, \mathbf{p}_2)$ is defined as follows:

$$D(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{\varepsilon + \|\mathbf{p}_1 - \mathbf{p}_2\|^2},$$

where $\mathbf{p}_1$ and $\mathbf{p}_2$ are the points between which the measure is calculated, and ε is a small value to prevent infinite values for overlapping points. A value of 0.01 has been used for the measurements presented here. A measure of phonemicity is then given by

$$P = \log_{10} \frac{E}{C}.$$

This measure gives a higher value when a system of trajectories shows stronger combinatorial structure.

In order to investigate whether the optimization procedure described in the paper results in more combinatorial structure, the phonemicity values of the initial, random, trajectories are compared with those of optimized trajectories. The results are given in Fig. 13. For systems of 4, 5, 9 and 30 trajectories, the random data set consisted of 1000 sets of signals. The optimized data set consisted of 100 sets of trajectories that had been optimized for 100,000 steps.

As the figure shows, for systems containing more than four signals the phonemicity measure gives a higher value for optimized systems than for random systems. This is significant with $p<0.05$ using the Kolmogorov–Smirnov test. For random trajectories, the phonemicity has a peak around 0 for all system sizes. This indicates approximately equal average dispersion of start- and endpoints and of intermediary points on a trajectory. For optimized systems with four trajectories, there is a large peak near the value of zero, corresponding to systems that have all four trajectories bunched up in the four corners of the space. However, smaller peaks are observed for higher values of phonemicity, corresponding to systems where one or more trajectories go from one corner of the space to another. For reference, completely combinatorial systems of 4, 9 and 16 trajectories (using 2, 3 and 4 fully connected start- and endpoints, respectively) have phonemicity values of 0.50, 0.83 and 0.93, respectively. These results show that optimization of distinctiveness of trajectories does reliably result in systems that show (superficial) combinatorial structure.

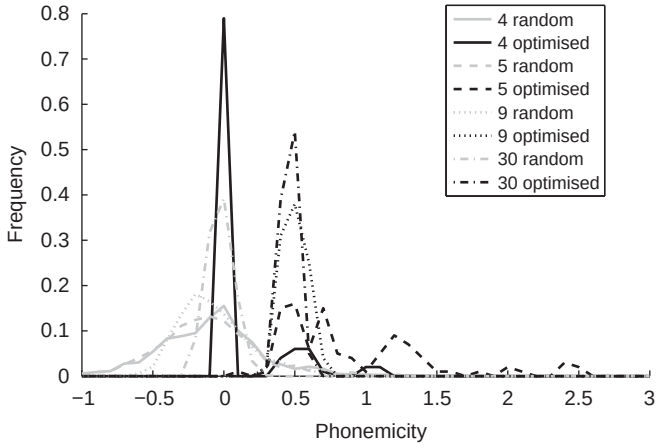


Fig. 13. Optimized repertoires always show a significantly higher degree of phonemicity (as defined in the appendix) than random repertoires. Shown are the frequencies (y-axis) with which a certain degree of phonemicity (x-axis) is obtained in 100 runs for each condition. Conditions varied in the number of trajectories per repertoire (4, 5, 9 and 30) and whether or they were randomly chosen or optimized (the OP criterion from Section 3.5). For all repertoire sizes larger than four, the differences between the random and optimized conditions are highly significant.

References

Andrew, R. J. (1976). Use of formants in the grunts of baboons and other nonhuman primates. *Annals of the New York Academy of Sciences*, *280*, 673–693.

Arcadi, A. (1996). Phrase structure of wild chimpanzee pant hoots: Patterns of production and interpopulation variability. *American Journal of Primatology*, *39*, 159–178.

Barton, N., & Zuidema, W. (2003). Evolution: The erratic path towards complexity. *Current Biology*, *13*(16), 649–651.

Benz, A., Jäger, G., & Van Rooij, R. (Eds.) (2005). *Game theory and pragmatics*. Palgrave.

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge: Cambridge University Press.

Boerlijst, M., & Hogeweg, P. (1991). Self-structuring and selection: Spiral waves as a substrate for prebiotic evolution. In C. Langton, C. Tayler, J. Farmer, & S. Rasmussen (Eds.), *Artificial life II* (pp. 255–276).

Bogert, B. P., Healy, M. J., & Tukey, J. W. (1963). The frequency analysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and shape cracking. In M. Rosenblatt (Ed.), *Time series analysis* (pp. 209–243). New York, NY: Wiley.

Butcher, A. (1994). On the phonetics of small vowel systems: Evidence from Australian languages. In *Proceedings of the 5th Australian conference on speech science and technology*, Vol. I (pp. 28–33). Australian Speech Science and Technology Association.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York, NY: Harper & Row.

de Boer, B. (1999). *Self organisation in vowel systems*. Ph.D. thesis, Artificial Intelligence Lab, Vrije Universiteit Brussel.

de Boer, B. (2000). Self organization in vowel systems. *Journal of Phonetics*, *28*, 441–465.

de Boer, B. (2001). *The origins of vowel systems*. Oxford, UK: Oxford University Press.

Deacon, T. (2000). Evolutionary perspectives on language and brain plasticity. *Journal of Communication Disorders*, *33*(4), 273–290.

Deuchar, M. (1996). Spoken language and sign language. In A. Lock, & C. R. Peters (Eds.), *Handbook of human symbolic evolution*. Oxford, UK: Clarendon Press.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, *22*, 567–631.

Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Science*, *4*(7), 258–267.

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition*. Cambridge, UK: Cambridge University Press.

Hockett, C. (1960). The origin of speech. *Scientific American*, *203*, 88–111.

Jackendoff, R. (2002). *Foundations of language*. Oxford, UK: Oxford University Press.

Ke, J., Ogura, M., & Wang, W. S.-Y. (2003). Modeling evolution of sound systems with genetic algorithm. *Computational Linguistics*, *29*(1), 1–18.

Kirby, S., & Hurford, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands, & I. Harvey (Eds.), *Proceedings of the 4th European conference on artificial life* (pp. 493–502). Cambridge, MA: MIT Press.

Komarova, N. L., & Nowak, M. A. (2003). Language, learning and evolution. In M. H. Christiansen, & S. Kirby (Eds.), *Language evolution* (pp. 317–337). Oxford, UK: Oxford University Press.

Kuhl, P., Williams, K., Lacerda, F., Stevens, K., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 month of age. *Science*, *255*, 606–608.

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*(1–3), 239–269.

Lieberman, P. (1984). *The biology and evolution of language*. Cambridge, MA: Harvard University Press.

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

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1984). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for language universals* (pp. 181–203). Berlin: Mouton.

MacNeilage, P. F., & Davis, B. L. (2000). On the origin of internal structure of word forms. *Science*, *288*, 527–531.

Masataka, N. (1987). The perception of sex-specificity in the long calls of the tamarin (saguinnes labiatus labiatus). *Ethology*, *76*, 56–64.

Maynard Smith, J. (1982). *Evolution and the theory of games*. Cambridge, England: Cambridge University Press.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *246*, 15–18.

Mitani, J. C., & Marler, P. (1989). A phonological analysis of male gibbon singing behavior. *Behaviour*, *109*, 20–45.

Nowak, M. A., Krakauer, D., & Dress, A. (1999). An error limit for the evolution of language. *Proceedings of the Royal Society of London, Series B, Biological Sciences*, *266*(1433), 2131–2136.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences of the USA*, *96*, 8028–8033.

Ohala, J. J. (1981). The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, & M. F. Miller (Eds.), *Papers from the parasession on language and behavior* (pp. 178–203). Chicago: Chicago Linguistic Society.

Oudeyer, P.-Y. (2001). Coupled neural maps for the origins of vowel systems. In G. Dorffner, H. Bischof, K. Hornik (Eds.), *Proceedings of the international conference on artificial neural networks. Lecture notes in computer science* (Vol. 2130, pp. 1171–1176). Berlin: Springer.

Oudeyer, P.-Y. (2002). Phonemic coding might be a result of sensory-motor coupling dynamics. In B. Hallam, D. Floreano, J. Hallam, G. Hayes, & J.-A. Meyer (Eds.), *Proceedings of the 7th international conference on the simulation of adaptive behavior* (pp. 406–416). Cambridge, MA: MIT Press.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*(3), 435–449.

Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford: Oxford University Press.

Parker, G. A., & Maynard Smith, J. (1990). Optimality theory in evolutionary biology. *Nature*, *348*, 27–33.

Payne, R. S., & McVay, S. (1971). Songs of humpback whales. *Science*, *173*(3997), 585–597.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 137–157). Amsterdam, The Netherlands: John Benjamins.

Plotkin, J. B., & Nowak, M. A. (2000). Language evolution and information theory. *Journal of Theoretical Biology*, pp. 147–159.

Redford, M. A., Chen, C. C., & Miikkulainen, R. (2001). Constrained emergence of universals and variation in syllable systems. *Language and Speech*, *44*, 27–56.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell Systems Technical Journal*, 27(July, October), 379–423, 623–656.

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, *228*(1), 127–142.

Steels, L., & Oudeyer, P.-Y. (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)*. Cambridge, MA: MIT Press.

Stokoe, W. C. (1960). *Sign language structure: An outline of the visual communication systems of the American deaf*. Studies in linguistics: Occasional papers 8. Department of Anthropology and Linguistics, University of Buffalo.

Studdert-Kennedy, M. (1998). The particulate origins of language generativity: From syllable to gesture. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive bases*. Cambridge, UK: Cambridge University Press.

Studdert-Kennedy, M. (2000). Evolutionary implications of the particulate principle: Imitation and the dissociation of phonetic form from semantic function. In C. Knight, J. R. Hurford, & M. Studdert-Kennedy (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge: Cambridge University Press.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.

Ujhelyi, M. (1996). Is there any intermediate stage between animal communication and language? *Journal of Theoretical Biology*, *180*, 71–76.

Waddington, C. H. (1939). *An introduction to modern genetics*. London: Allen Unwin.

Westermann, G. (2001). A model of perceptual change by domain integration. In *Proceedings of the 23rd annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum.

Zuidema, W. (2005). *The major transitions in the evolution of language*. Ph.D. thesis, Theoretical and Applied Linguistics, University of Edinburgh.

Zuidema, W., & de Boer, B. (2003). How did we get from there to here in the evolution of language? *Behavioral and Brain Sciences*, *26*(6), 694–695.