Space and time in models of speech rhythm

Sam Tilsen[1]

[1]Cornell University, Department of Linguistics

Sam Tilsen
203 Morrill Hall
Ithaca, NY 14853
tilsen@cornell.edu

Short title: Space and time in speech rhythm

Keywords: Speech rhythm, metrical theory, rhythmic typology, prosody, dynamical models

Abstract

How do rhythmic patterns in speech arise? Many representations and models incorporate a mechanism whose purpose is to generate a rhythmic pattern. Here an alternative is explored: rhythmic patterns arise indirectly, from spatial mechanisms which govern the organization of articulatory gestures. In pursuing this alternative, the roles of time and space in symbolic phonological representations are analyzed in detail, and conventional understandings of stress and accent are called into question. One aspect of rhythmic patterns in particular—the directionality of stress assignment—is examined closely. A novel dynamical model is developed, which proposes a reinterpretation of directionality and various other temporal phenomena.

Introduction

In many languages, the majority of words conform to a pattern in which some syllables can occur with an accent while others cannot. These accents—often a change of pitch or loudness or duration—may have the effect of grabbing the attention of a listener, facilitating word identification, and potentially creating a rhythm, i.e. a pattern that repeats in time. The curious thing about this is that the pattern is typically predictable only from the beginning or only from the end of the word. In conventional terms: there is a directionality parameter for stress assignment, and stress is assigned “from the left edge” or “from the right edge” of the word. Schematic examples of stress patterns with left-to-right (L→R) and right-to-left (R→L) directionality are contrasted in Table 1 below.

Table 1. Schematic comparison of stress patterns
with L→R and R→L directionality

# of σ	L→R	R→L
1	**σ**	**σ**
2	**σ** σ	σ **σ**
3	**σ** σ **σ**	**σ** σ **σ**
4	**σ** σ **σ** σ	σ **σ** σ **σ**
5	**σ** σ **σ** σ **σ**	**σ** σ **σ** σ **σ**

σ: stressed syllable, σ: unstressed syllable

The spatial vocabulary (i.e. *left*, *right*, *edge*) may be a little unsettling to some readers, because words do not really have edges, and because the mapping of left/right to earlier/later is arbitrary. When we say that words have “edges,” we are using metaphors in which time is a linear space and syllables are objects arranged in that space. This makes a lot of sense to us because graphemes are spatially arranged in our writing systems, always in a way that corresponds to their temporal order in production. Hence we can say that there are syllables which are “at the left edge” or “at the right edge” or “in the middle” of a word. Such vocabulary would certainly be useful if, as a physical description, there is a spatial mapping of the components of words to a physical space in the brain. But is there really a space of this sort? This article explores the idea that such a space indeed exists.

A curious aspect of directionality is that the distribution of L→R and R→L patterns across languages is fairly balanced.[1] Moreover, specific L→R patterns observed in one language have symmetric R→L counterparts in some other language, in most cases. These symmetries might be unexpected, because time is asymmetric: causes precede effects, and entropy always increases. Indeed, there are plenty of morphological, phonological, and phonetic patterns which do reflect an “arrow of time”. Suffixation is more prevalent than affixation and suffixes tend to be more tightly bound to roots than prefixes.[2] Word-initial strengthening and word-final weakening are more common than their counterparts.[3–6] Lexical access/retrieval appears to privilege earlier sounds over later ones,[7] and in tip of the tongue states speakers are sometimes aware of just the first sound or first few sounds in a word.[8,9] Aerodynamic effects lead to the decrease of fundamental frequency over the course of an utterance, so that pitch tends to be lower later on in utterances,[10,11] and the initiations of articulatory movements precede the achievements of movement targets, an obvious but nonetheless consequential fact.[12] Since many speech-related phenomena exhibit temporal asymmetries, one might wonder why there are not similar asymmetries in the possible directionality of stress assignment. As we will eventually see, if accentuation is understood to originate from spatial patterns, rather than a temporal mechanism, the symmetry of directionality is fair less puzzling.

The main aims of this article are (i) to analyze the role of spatio-temporal reasoning in phonological representations of rhythmic structure, and (ii) to describe a novel dynamical model in which temporal patterns emerge indirectly from spatial patterns which govern the organization of articulatory movements. We adopt a primarily motoric and developmental perspective on rhythm in this context; consequently, issues related to the perception of rhythm and phenomena emerging from interactions between agents are not discussed. This restriction of scope is useful because a detailed model of the system which gives rise to rhythmic patterns on short timescales for one speaker may be a prerequisite to understanding the perceptual, social, and historical forces which influence rhythmic patterns on longer timescales for multiple speakers. The important implication of the new model is that representations that explicitly describe temporal patterns, or models that directly create a temporal pattern, are misguided: rhythm in spontaneous conversational speech is an epiphenomenon of mechanisms which organize articulatory movements, not the direct product of a rhythmic mechanism.

What is accent and what is stress?

There are a number of different ways of conceptualizing the phenomena of stress and accent. Currently a common view is that stress is a property of syllables which derives from a structural organization, and accents are articulatory gestures—often effecting a change in pitch, intensity, or phonation quality—which may be associated with stressed syllables. This view is consistent with the structural representation in Fig. 1, where syllables are labelled as strong (σ_s) or weak (σ_w), and a H* accent is associated with a stressed syllable.

From an empirical perspective, the relevant question is how stress and accent are manifested in articulatory movements and the acoustic signal of speech. Accents are often observed to effect changes of pitch, and hence are commonly referred to as *pitch accents*. For example, the H* in Fig. 1 is observable in the form of an increase in F0 relative to other syllables in the word. A more general notion of accent is possible in which other phonetic parameters such as acoustic intensity are controlled via accents.

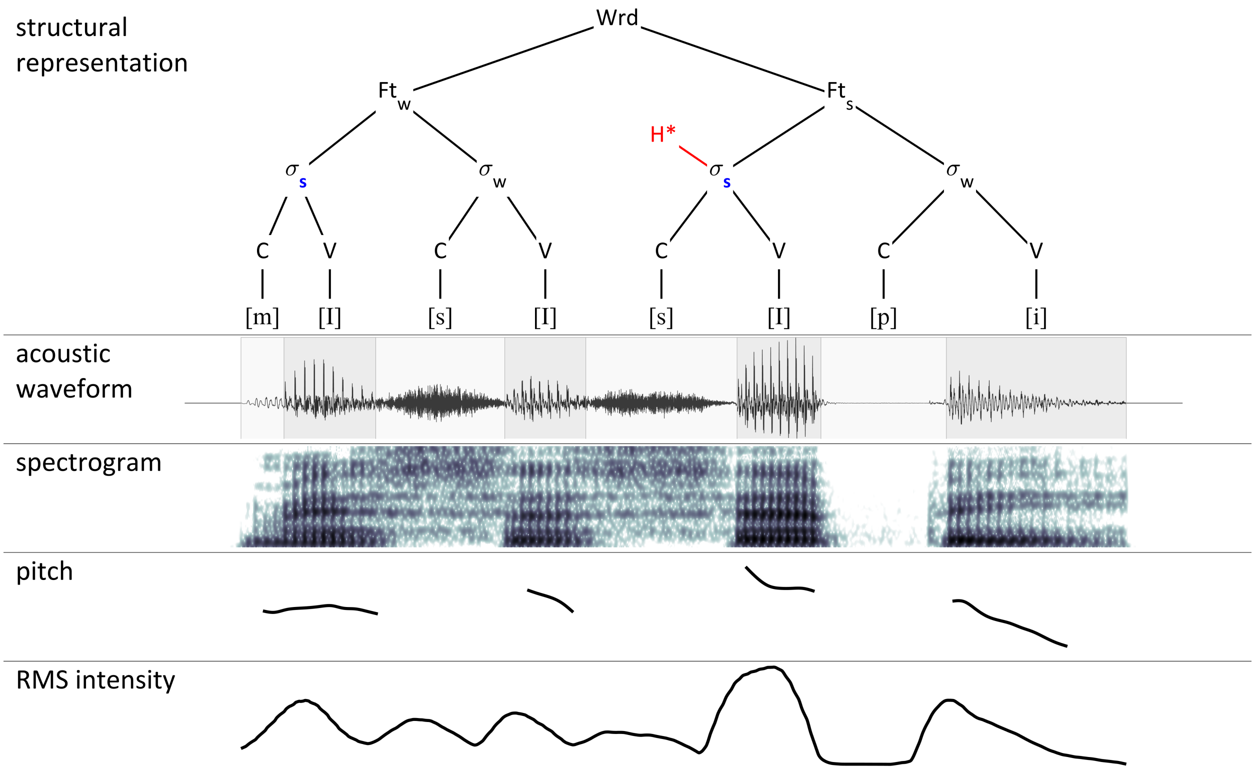


Fig. 1. Structural representation of stress and accent in the word *Mississippi*, along with various forms of acoustic information. From top to bottom: structural representation of a prosodic word with two feet and an accent associated with the penultimate syllable; acoustic waveform with segmentation; spectrogram (from 0-5000 Hz), pitch; root-mean-square intensity.

Whereas accents uncontroversially have a relatively direct influence on articulation, stress is not clearly a phenomenon of this sort. There are a number of problems with the view that stress is directly manifested via articulation. First, research on the phonetic correlates of stress have found that there are no universal acoustic effects of stress:[1] pitch, intensity, segmental duration, spectral tilt (reflecting vocal fold configuration), and articulatory kinematic variables (i.e. movement ranges/velocities/targets) appear to vary with stress in a language-specific and contextually contingent manner. If there were direct articulatory manifestations of stress, one would expect them to be fairly uniform across languages. Second, early empirical studies which purportedly found phonetic effects of stress[13,14] failed to deconfound accent from stress. Because some stressed syllables will be produced with an accent, phonetic measurements of stressed syllables will confound potential effects of stress with those of accent. Third, to avoid the aforementioned confound, a number of studies have compared unstressed syllables to stressed, unaccented syllables.[15,16] Problematically, such approaches must assume a particular phenomenology of accent in which accents are categorically present or absent in association with stressed syllables, and in which it is possible to determine whether an accent is present or not from more abstract considerations (e.g. from semantic and/or pragmatic information, or from phonological patterns). Without such assumptions, it is not possible to control for accent in an investigation of stress.

The alternative, simpler view adopted here is that all phonetic “effects” of stress are really effects of accentuation, where accentuation is understood as a gradient phenomenon with a variety of articulatory and acoustic manifestations. Those manifestations are potentially changes in F0, intensity, voice quality (i.e. spectral tilt), and articulatory velocities, targets, and durations. Stress in this simpler view is *purely* structural: there are statistical correlates of stress only because stressed syllables may be produced with accents, which in turn induce gradient effects on a variety of phonetic parameters. This is in line with operationalized determinations of stress such as in Hayes (1995),[1] where all four of the proposed diagnostics of stress are reducible to the potential for a syllable to be produced an accent.

Another complication in the phenomenology of stress and accent is that there appear to be two types of stress, in the structural sense. In many languages it is possible to distinguish between primary stress and secondary stress. Syllables with primary stress are typically produced with more extreme accentuation than syllables with secondary stress; accents on syllables with secondary stress have weaker phonetic effects which may not be statistically distinguishable from unstressed syllables.[17] In the example of *Mississippi* in Fig. 1, the third syllable has primary stress, and the first syllable may have secondary stress. While primary stress intuitions are robust and can be readily verified by tapping experiments,[18] secondary stress intuitions are not always robust across speakers of a language.

One representational approach to distinguishing primary from secondary stress involves positing that syllables are grouped into feet, feet are grouped into a prosodic word, and one of the feet in the prosodic word is the strongest. As shown in Fig. 1, the syllable with primary stress in *Mississippi* is the one that is associated with a strong foot, conceptualized as the "head" of the prosodic word. Applying the same logic within each foot, syllables with stress (whether primary or secondary) are the heads of feet. However, there are alternative representational approaches which employ different metaphors.

Conceptual metaphors in symbolic and dynamic representations of stress and accent

Formal, symbolic phonological representations of the sort in Fig. 2A-D are grounded in the metaphor that linguistic units are physical objects. The use of the metaphor does *not* presuppose that units (e.g. syllables) *are* physical objects; rather, the metaphor provides a set of mappings from our experiences with the concrete domain of physical objects to the abstract, constructed domain of linguistic units.[19,20] We use these mappings to reason analogically about linguistic units, by drawing inferences from our experience with physical objects. For example, there is no a priori reason why units in representations might not overlap, as shown in Fig. 2E. Indeed, it is well established that the articulatory movements within and between syllables typically *do* overlap. Why have no formal representational models ever been developed in which symbols overlap? Such representations have not even been discussed as a possibility. The reason is that in our typical experiences with physical objects, two distinct objects *do not occupy the same space*. This characteristic of our experience in the physical domain is transferred to how we represent and reason about the abstract domain, i.e. linguistic units.

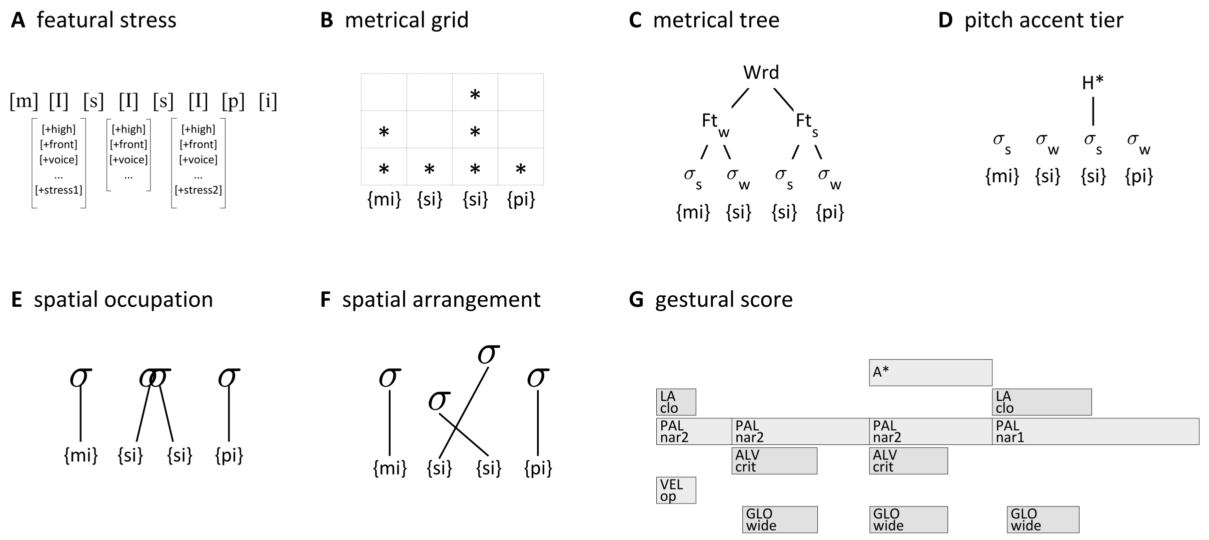


Fig. 2. Representations of stress and accent, and examples of violations of metaphoric inferences. (A) Featural representation in which primary and secondary stress features are distinguished. (B) Metrical grid in which prominence marks are associated with syllables. (C) Metrical tree where stress is represented via a hierarchical grouping structure. (D) Representation of pitch accent on an accentual tier. (E, F) Examples of violations of spatial occupation and spatial arrangement mappings. (G) Gestural score, where gestures are periods of time during which forces drive changes in the state of the vocal tract.

Symbolic representations also universally employ the metaphor that temporal order is spatial arrangement. In the conventional application of this metaphor, units are arranged horizontally and events which occur later in time are arranged to the right of events which occur earlier in time. There is no *a priori* reason why symbolic representations might not be constructed with non-linear spatial arrangements as in Fig. 2F. Indeed, when units of different "types" are considered, non-linear arrangement is used extensively.[21–23] Non-linear arrangements of units of the same type are generally avoided because such depictions violate the conventional mapping of temporal order to a linear spatial arrangement. Furthermore, because units are conceptualized as discrete objects, it is natural to infer a discretization of time in such representations, i.e. a temporal order.

Unlike symbolic representations, the gestural scores of Articulatory Phonology,[24,25] which are based on the computational model of Task Dynamics,[26,27] do not evoke the object metaphor. Gestural scores are schematic representations of dynamics; an example score for *Mississippi* shown in Fig. 2G. An articulatory gesture is a period of time in which there are forces acting upon a parameter of the vocal tract. These forces drive the state of the vocal tract toward new target state. For example, the "LA clo" gesture in Fig. 2G specifies a target value of the tract variable *lip aperture*, and activation of the gesture results in a bilabial closure (for the [m] sound in *Mississippi*). The empirical correlates of gestures are trajectories in the state space of the vocal tract, i.e. movements. Early work in the Articulatory Phonology framework focused on oral articulatory gestures; more recently gestural models of tones and intonational pitch accents have been developed.[28–31] Following this trend, we can think of accents as accentual gestures which drive the state of the vocal tract toward pitch, intensity, and/or phonation quality targets.

Unlike the aforementioned parameters, which can be readily associated with target states, durational effects of accent must be conceptualized differently, because durations are not states of the vocal tract. We will see later on that durational effects of accent have a natural re-interpretation in the current approach and are mechanistically distinct from target-state parameters. The reader should note that time is conceptualized linearly in the gestural score, but crucially, gestural scores are not conceptualized as objects, so there is no intuition that gestures cannot occupy the same space. In other words, gestures can overlap in time. Furthermore, it is not sensible to refer to gestures as ordered in time, because there is not always a unique order of overlapping events. No temporal discretization is imposed by the score.

In addition to spatial occupation and linear ordering, some representations such as the metrical tree (Fig. 1 and Fig. 2C) employ an object connection schema to evoke grouping or containment relations. By convention, an object which is connected to another object which is vertically higher in the tree is *contained by* the higher-level object. Containment schemas are implicit in metrical trees but are also depicted explicitly in bracketed grids.[32] These schemas have been used to conceptualize accentual patterns as the product of a "foot construction" algorithm, in which syllables are grouped into feet of some type (i.e. trochaic, iambic), beginning at a word edge. The reader should note that while containment is fundamental to phrase structure models of syntax, it is somewhat more contested in theoretical approaches in phonology: debates have arisen regarding whether an object can be connected (i.e. contained) by two distinct higher-level objects (e.g. ambisyllabicity[33]), whether objects on a given level must be necessarily contained by objects on the next highest level[34,35] (exhaustivity), and whether an object can contain an object of the same type (recursivity). The metrical grid (Fig. 2B) and the gestural score (Fig. 2G) are examples of representations which lack containment/grouping relations altogether.

The above analysis of conceptual models of stress and accent reveals that there are two incompatible sets of metaphors. On one hand, the traditional symbolic conception views speech as spatially arranged linguistic objects, provides notions of temporal order (i.e. discretized time), and in many cases imposes grouping/containment of objects. On the other hand, the articulatory phonology conception views speech as a state space trajectory driven by forces, and lacks temporal discretization and grouping. Below we consider how these two different sets of conceptual metaphors fare in their ability to provide a basis for classification of accentual systems across languages, aspects of which are reviewed in the next section.

Classification of quantity-insensitive accentual patterns

In some languages, accentual patterns are predictable entirely from the position (i.e. temporal order) of syllables relative to the edges (i.e. beginning and/or end) of a word—these are called *quantity insensitive* patterns. In other languages, accentuation is partly predictable from the composition of the syllables in a word (e.g. the presence of a long vowel or coda consonant in a syllable), or is unpredictable and must be determined from long-term (i.e. lexical) memories. We will address quantity sensitive and lexical patterns later; in this section we review the quantity insensitive patterns.

A classification scheme for quantity insensitive patterns is shown in Fig. 3, derived from typologies in several sources.[1,36,37] Note that our concern here is *classification* of logical possibilities, rather than *typology*, i.e. the statistical distributions of attested patterns—it is unknown to what extent the distribution results from universal forces on language evolution or from chance historical factors. As mentioned in the introduction, there is a puzzling form of cross-linguistic variation such that a given pattern must be understood in relation to a particular directionality of stress assignment. This is reflected by the two sides of the vertical division of the table, and can be considered a parameter of the typology. Another parameter is the directionality-relative location of the primary accent, which is generally the first, second, or third syllable from the edge associated with the directionality parameter. In uni-directional systems, primary accent and secondary accent (if present) are predictable from the same edge of the word; in contrast, in bi-directional systems, primary accent and secondary accent are predictable from different edges of the word. Another parameter is whether secondary accent locations are periodic or aperiodic. In periodic patterns, secondary accents occur at regular intervals, either every other syllable (binary) or every third syllable (ternary). In aperiodic patterns, there is no secondary accent (uni-directional systems) or a single secondary accent (bi-directional systems). Note that Fig. 3 also lists codes for each pattern used in the remainder of this paper, where "B" refer to the beginning of the word and "E" to the end of the word.

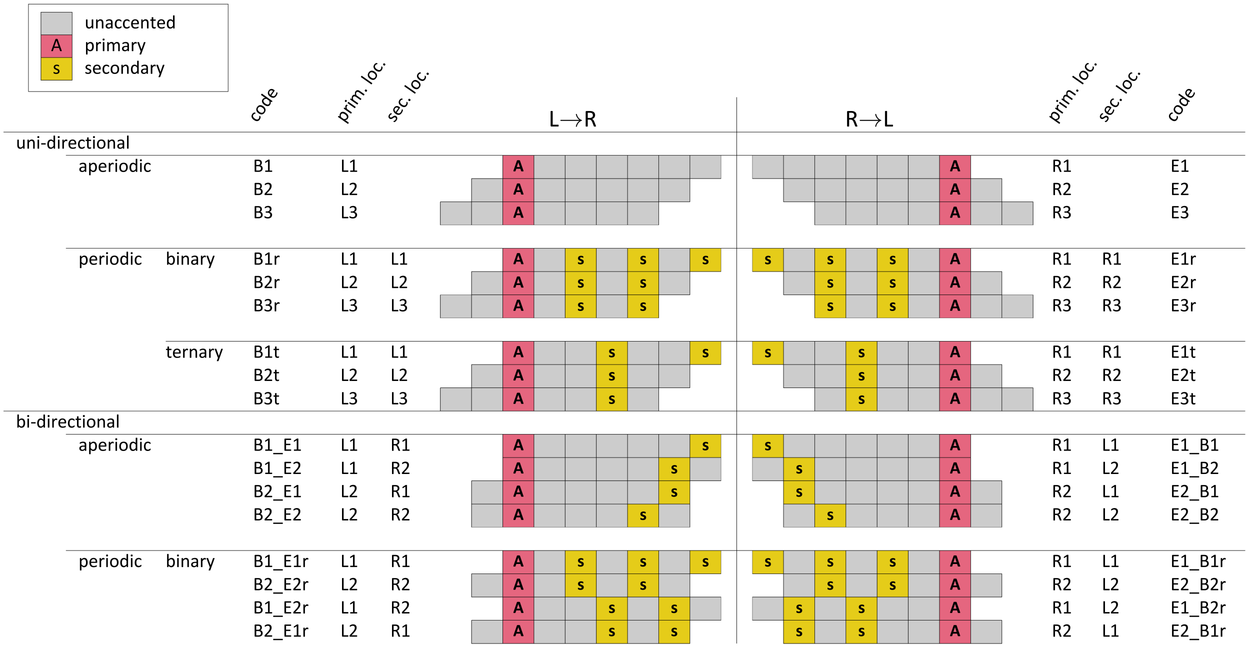

unaccented
A primary
s secondary
code
prim. loc.
sec. loc.
L→R
R→L
prim. loc.
sec. loc.
code
uni-directional
aperiodic
B1 L1 A R1 E1
B2 L2 A A R2 E2
B3 L3 A A R3 E3
periodic binary
B1r L1 L1 A s s s s s s A R1 R1 E1r
B2r L2 L2 A s s s s A R2 R2 E2r
B3r L3 L3 A s s s s A R3 R3 E3r
ternary
B1t L1 L1 A s s s s A R1 R1 E1t
B2t L2 L2 A s s A R2 R2 E2t
B3t L3 L3 A s s A R3 R3 E3t
bi-directional
aperiodic
B1_E1 L1 R1 A s s A R1 L1 E1_B1
B1_E2 L1 R2 A s s A R1 L2 E1_B2
B2_E1 L2 R1 A s s A R2 L1 E2_B1
B2_E2 L2 R2 A s s A R2 L2 E2_B2
periodic binary
B1_E1r L1 R1 A s s s s s s A R1 L1 E1_B1r
B2_E2r L2 R2 A s s s s A R2 L2 E2_B2r
B1_E2r L1 R2 A s s s s A R1 L2 E1_B2r
B2_E1r L2 R1 A s s s s A R2 L1 E2_B1r

Fig. 3. Classification of accentual systems. Directionality is represented by the vertical division of the table. Words are aligned according to the location of primary accent. R*x*/L*x* indicates syllable positions counting from the right/left edge of the word. Classification codes where B/E indicates the beginning/end of the word are included. Note that some logically possible patterns are omitted for brevity.

Now lets consider how symbolic vs. dynamic representations fare in describing the logically possible accentual patterns. Symbolic representations, which allow for notions of grouping and discretized time, provide a natural basis for understanding accentuation patterns. Accents are uncontroversially associated with syllables, rather than individual segments, and symbolic representations readily allow for the grouping of segments into syllables (through connection and/or containment schemas). In contrast, the dynamic representations of gestural scores cannot achieve the same natural description of accentuation patterns because gestures are not grouped into syllables. The association of accents with syllables might be reinterpreted in a gestural framework as a constraint that accentual gestures can be coupled only to vocalic gestures. However, vocalic gestures cannot be substituted wholesale for syllables because (i) syllables may contain multiple vocalic gestures (as in diphthongs), and (ii) in some languages there are appear to be syllables which lack vowels.[38,39] Because gestural scores do not represent discretized time or grouping, there is no natural basis for counting syllables from the edge of the score. Symbolic representations thus have a considerable advantage over gestural scores when it comes to classification of accentual patterns.

The selection-coordination framework and grouping of gestural selection

The selection-coordination framework,[12,40,41] which is an extension of Articulatory Phonology[24,25] and Task Dynamics,[26,42] imposes grouping on gestures of the score, and hence allows for a dynamic conception of speech that is more suitable for understanding accentual patterns. The selection-coordination (henceforth "s/c") framework accomplishes this by integrating gestural scores with a competitive queuing mechanism.[43–46] The empirical motivation and details of the s/c framework have been discussed extensively in earlier work;[12,40,41] here a brief overview is provided.

In the s/c model, prior to the production of a word, premotor systems associated with articulatory gestures in the word are organized into competitively selected sets $\{g\}_1 \ldots \{g\}_n$, which conform to a stable pattern of relative activation, as shown in the top panel of Fig. 4. When production of the word is initiated, there is a competition process in which the activations of the sets increase until one of them exceeds a selection threshold. At this point the gestures in the above-threshold set are executed. Note that the precise timing of the execution of co-selected gestures is governed by phasing mechanisms hypothesized in the AP framework, where a system of coupled oscillators determines a pattern of relative phasing.[12,47] During the epoch in which $\{g\}_1$ is selected, competing sets $\{g\}_2$ and $\{g\}_3$ are *gated*, i.e. their activation is prevented from increasing. Eventually feedback is received regarding the achievement of targets associated with the gestures in $\{g\}_1$. The feedback induces the suppression of this set and de-gates the competitors, allowing for the competition process to resume until the next most highly active set, $\{g\}_2$, is selected. This cycle of selection and feedback-induced suppression iterates until all sets have been selected and suppressed.

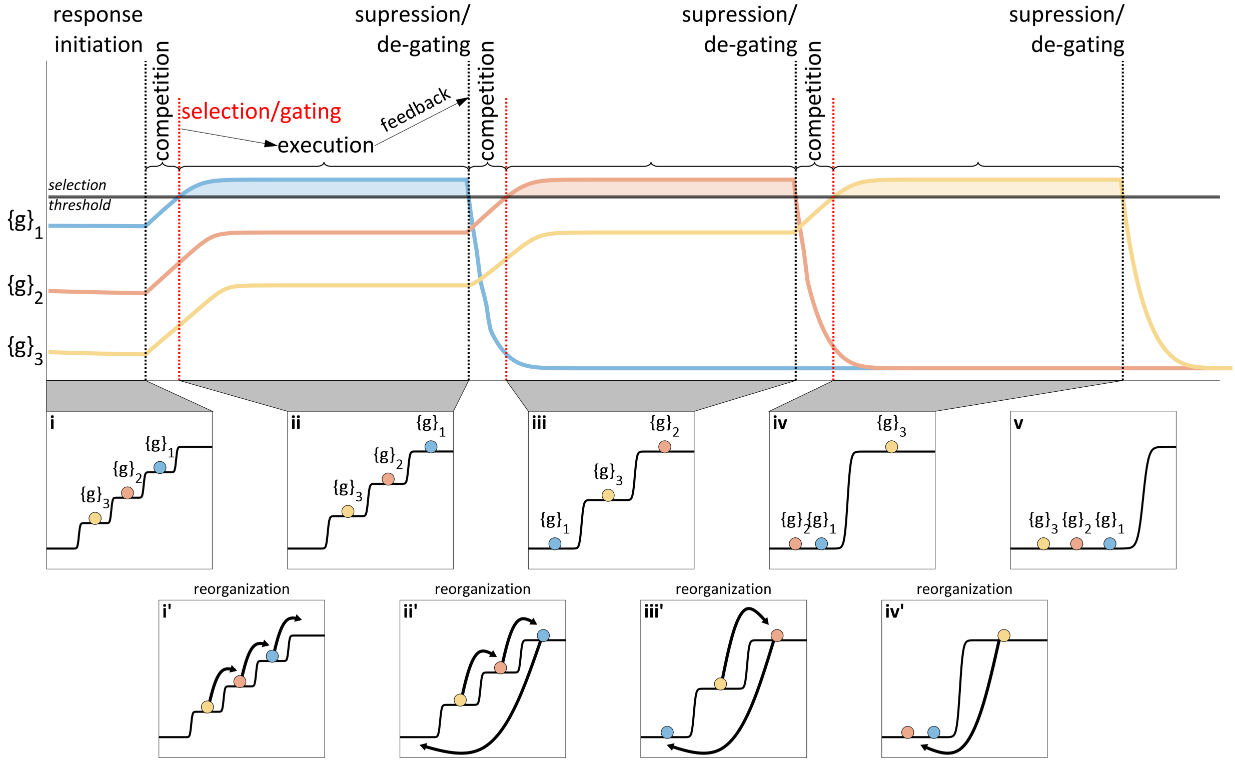


Fig. 4. Competitive queuing model of sequencing, and quantal potential functions describing steady state relative activation patterns and reorganizations. Three sets of articulatory gestures, $\{g\}_1$, $\{g\}_2$, and $\{g\}_3$ are initially organized in a stable pattern of relative activation (i). When the sequence is initiated, a rapid competition process occurs, corresponding to an abrupt reorganization of the potential (i'). The gestures in the first set to reach the selection threshold are selected (ii), while the competing sets are suppressed. Subsequently feedback drives the suppression of the selected set and the competition process resumes (ii'). The selection-feedback-suppression cycle iterates (iii, iii', …) until all sets have been suppressed.

In order to conceptualize the stability of activation patterns, the s/c framework employs a quantal potential function,[12,48] in which energy barriers maintain the relative activation pattern prior to production and during steady state epochs of production. As shown in the bottom panels of Fig. 4, the competitive queuing dynamics can be described more phenomenologically as a sequence of relatively steady state epochs (i-v) and intermittent, abrupt reorganizations (i'-iv'). In the initial organization (i), the highest level of the potential (the selection level) is unoccupied. When the response is initiated a fast-timescale reorganization of the potential occurs in which each set is promoted one level (i'). Subsequently a new stable pattern emerges (ii) in which the selection level is occupied, inducing execution. Feedback regarding target achievement eventually induces another abrupt reorganization (ii'), in which the selected system is demoted and the competitors are promoted. Alternating steady states and abrupt reorganizations continue until all systems have been demoted to the ground level.

The reader should note that in the s/c framework the association of accentual gestures with syllables is reinterpreted as the co-selection of accentual gestures with a set of oral articulatory gestures. This conception is only possible because the framework incorporates a selection mechanism which requires that gestures be organized into competitively selected sets. Such a mechanism is not available in the standard model of Articulatory Phonology. Importantly, the s/c framework does not require that there exists a spatial mapping of systems to their order of selection; the order of selection may be determined solely by an initial relative activation pattern. However, to account for directionality in accentual systems, it is useful to impose a spatio-temporal correspondence between sets of gestures and their order of selection. Below we extend the s/c model to accomplish this, but first, we consider a unique connectionist approach developed in Goldsmith (1994),[49] which to a large extent inspired the current one.

The Goldsmith model: a dynamical computational theory of accentual systems

The Goldsmith (1994) model is a connectionist network in which each node corresponds to a syllable in a word. The nodes are linearly arranged in a manner that corresponds to the order of syllables in the word. Fig. 5A shows a network for an eight-syllable word form. Each node has a real valued activation state, and in each time step the nodes transmit a portion of their activation to their nearest neighbors. The leftward and rightward transmission coefficients are α and β. The initial and final nodes can receive an external source of activation, and an external source can be uniformly applied to influence the internal activation of all nodes. All external sources are held constant throughout a simulation.

To understand the temporal evolution of the model, consider the time course of node activation shown in Fig. 5C. In the initial condition, all nodes have 0 activation. At the first time step, positional activation causes the final syllable node to become activated. In the second step, the final syllable node transmits a portion of its activation to the penultimate syllable—in this case negative activation (or perhaps, inhibition), according to the leftward transmission coefficient (here α = -0.8). At each time step, activation is transmitted further leftward, and the final node continues to receive 1 unit of positional action. After some number of iterations, the activation function over nodes stabilizes, exhibiting a pattern of activation peaks and valleys (Fig. 5B). As shown in Fig. 5D, the location of primary accent is the highest peak, and all other peaks are potential locations for secondary accents. In this example, the stable activation pattern corresponds to pattern E1r, i.e. R→L iambs.

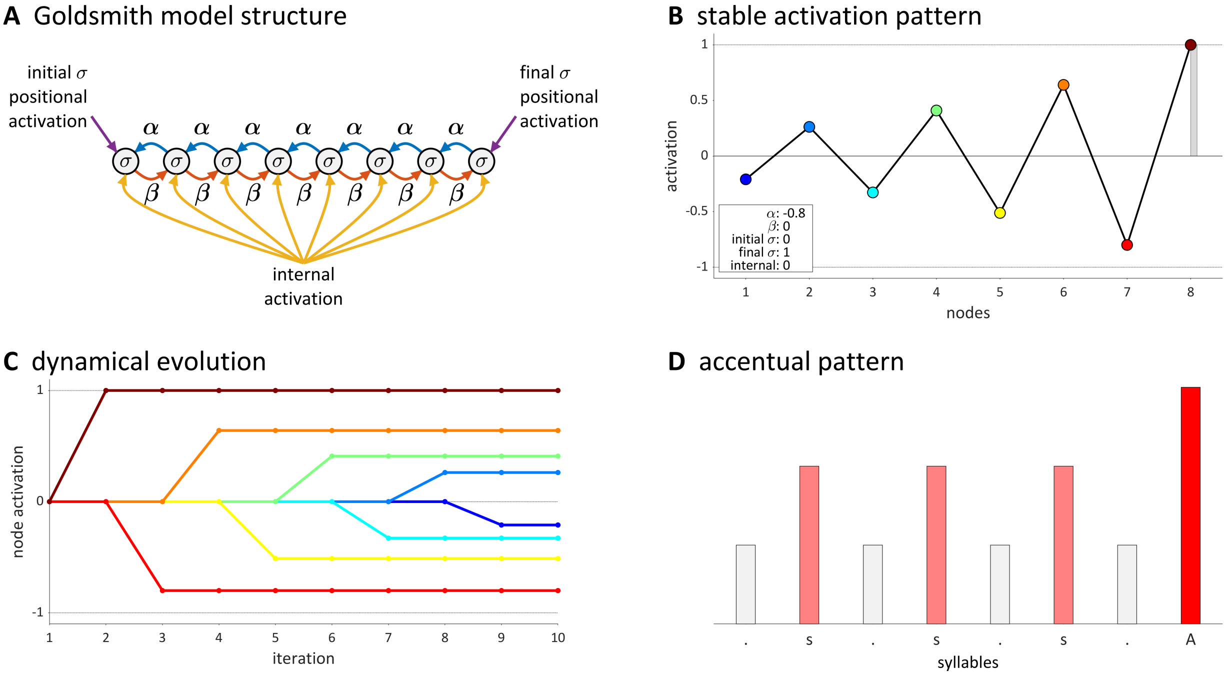


Fig. 5. Illustration of the Goldsmith connectionist model, showing a periodic E1r pattern. (A) Model structure for a word form with eight syllables. (B) Stable pattern of activation that emerges after iteration of the model with parameters α=-0.8, β=0, and final-σ activation of 1. (C) Dynamical evolution of the model. (D) Accentual pattern: 'A' primary accent; 's' secondary accent; '.' unaccented.

Fig. 6 shows examples of accentual patterns generated by various parameterizations of the model and their corresponding locations in α-β space. When α > β, the system is referred to as left-dominant and the directionality is R→L, as in Fig. 6A,B. Depending on whether the positional activation is positive (Fig. 6A) or negative (Fig. 6B), an E1r or E2r pattern is generated. Patterns with L→R directionality can be generated by imposing rightward dominance with initial positional activation, as in Fig. 6A′,B′. When the dominant transmission coefficient is positive rather than negative, the pattern will have only a single peak anchored to a word edge, thereby generating an aperiodic accentuation pattern as in Fig. 6C. Bi-directional patterns can be generated by combining initial and final activation with rightward or leftward dominance (Fig. 6D), and lexical or quantity sensitive patterns—which we address later on—can be generated by imposing node-specific (non-uniform) internal activation (Fig. 6E).

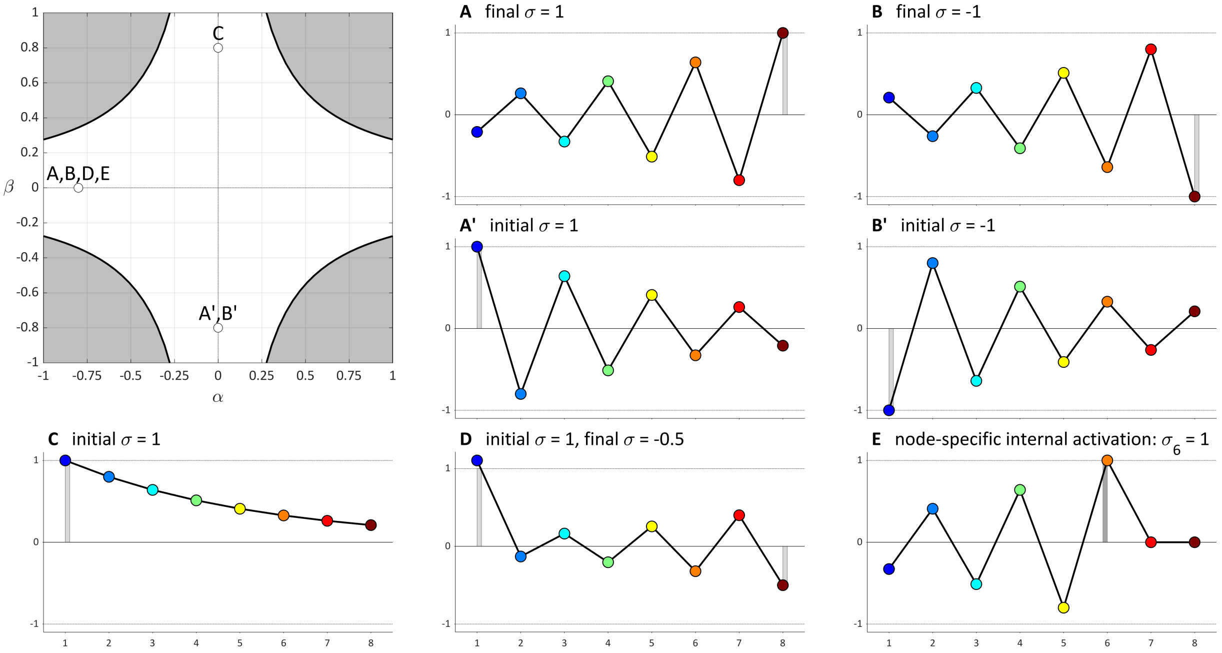


Fig. 6. Examples of accentual patterns generated by the Goldsmith model and their locations in α-β parameter space. (A, B) final excitation/inhibition with leftward dominance produces periodic R1/R2 patterns. (A', B') initial excitation/inhibition with rightward dominance produces periodic L1/L2 patterns. (C) initial excitation with right-dominance β>0 produces an aperiodic R1 system. (D) combining initial and final positional activation generates a bi-directional system. (E) non-uniform (node-specific) internal activation generates a pattern with primary stress on an arbitrary syllable.

When α or β is dominant and negative, the model can produce stable patterns which are reminiscent of standing waves, but the model does not necessarily converge to a stable pattern. The region of convergence in α-β parameter space is bounded by hyperbolas of the form y = ±A/x, where A decays exponentially as the number of nodes increases. More generally, the stabilization criterion is a detailed balance in which the total magnitude of the input to the network is matched by the total magnitude of the input dissipated from the network. Dissipation occurs whenever $|\alpha|+|\beta| < 1$, i.e. nodes transmit less absolute activation than they receive as input. Because the initial and final nodes implicitly have 0 leftward and rightward transmission coefficients, respectively, there is dissipation as at edge of the network opposite from the dominant transmission direction.

An important feature of the Goldsmith model is that, just like object-based symbolic representations, a spatial arrangement of units is imposed. This arrangement provides a basis for the nearest-neighbor constraint on interactions and for differentiating leftward and rightward transmission of activation. If we take the spatial arrangement and object-metaphors somewhat literally, there are a number of problems that arise, which are illustrated schematically in Fig. 7.

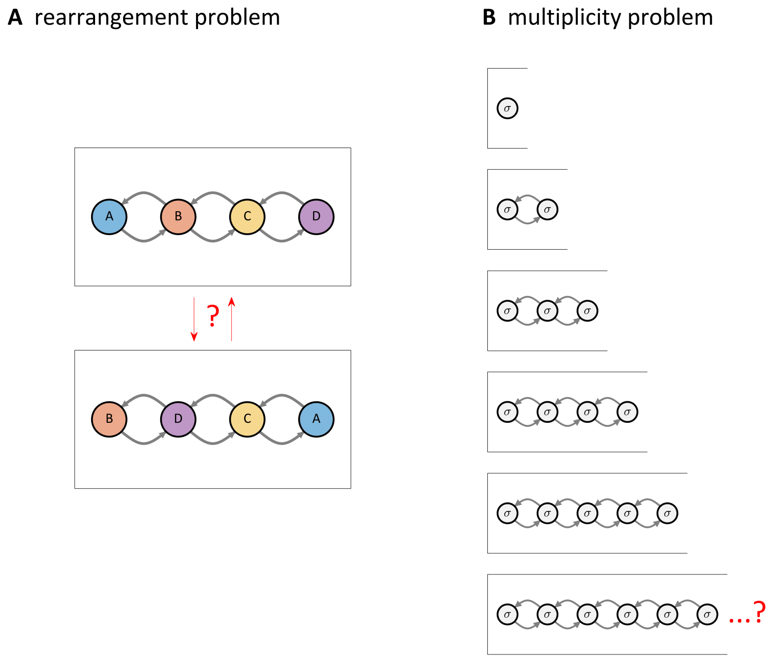


Fig. 7. Conceptual problems arising from the spatial occupation of units. (A) The rearrangement problem: how are nodes spatially arranged in a word-specific manner? (B) The multiplicity problem: can an arbitrarily large number of nodes be arranged? (C) The void space problem: what happens in unused space?

One problem is rearrangement (Fig. 7A): for different words, different spatial arrangements of nodes are required. How is this variation in spatial arrangement accomplished? Another problem is multiplicity (Fig. 7B): without any further constraints, words could be associated with an arbitrarily large number of nodes and hence an arbitrarily large space. On the other hand, if the space is constrained to be finite (so that there is a maximum node capacity) a *void space* problem arises (Fig. 7C): for words whose number of syllables is less than the maximum capacity, some of the space is "unused," assuming that unit "size" (i.e. how much space a unit occupies) is constant. Rearrangement, multiplicity, and void space may not seem problematic from an abstract perspective, but if we are to really embrace the idea that rhythmic patterns emerge from interactions in a *physical space*, such issues should be addressed.

The motor sequencing field and sets of coupled articulatory gestures

Here we conjecture that there is a physical space, in the brain, in which there is a spatial arrangement of systems that organize articulatory gestures into sets. To ground this conjecture, lets imagine that the space contains a large population of interacting microscopic units (e.g. a network of excitatory and inhibitory neurons, or perhaps cortical microcircuits). This population is labelled as the *set organization population* in Fig. 8. We assume, on the basis of empirical and theoretical studies,[50–55] that the microscopic units can enter into a regime of collective oscillation. We then posit that the full population has the ability to self-organize into subpopulations, and that these subpopulations are spatially arranged in a manner that corresponds to the *initial* organization of sets of gestures in the s/c potential. Hence one subpopulation occupies a region of the space that is associated with a set of gestures that will be selected first/earliest in time, a different subpopulation in a neighboring region of the space is associated with gestures that will be selected next, and so on. In the example in Fig. 8 there are four subpopulations of the set organization field, corresponding to the four syllables in *Mississippi*. Thus each syllable corresponds to a set of gestures which is competitively selected relative to other sets of gestures.

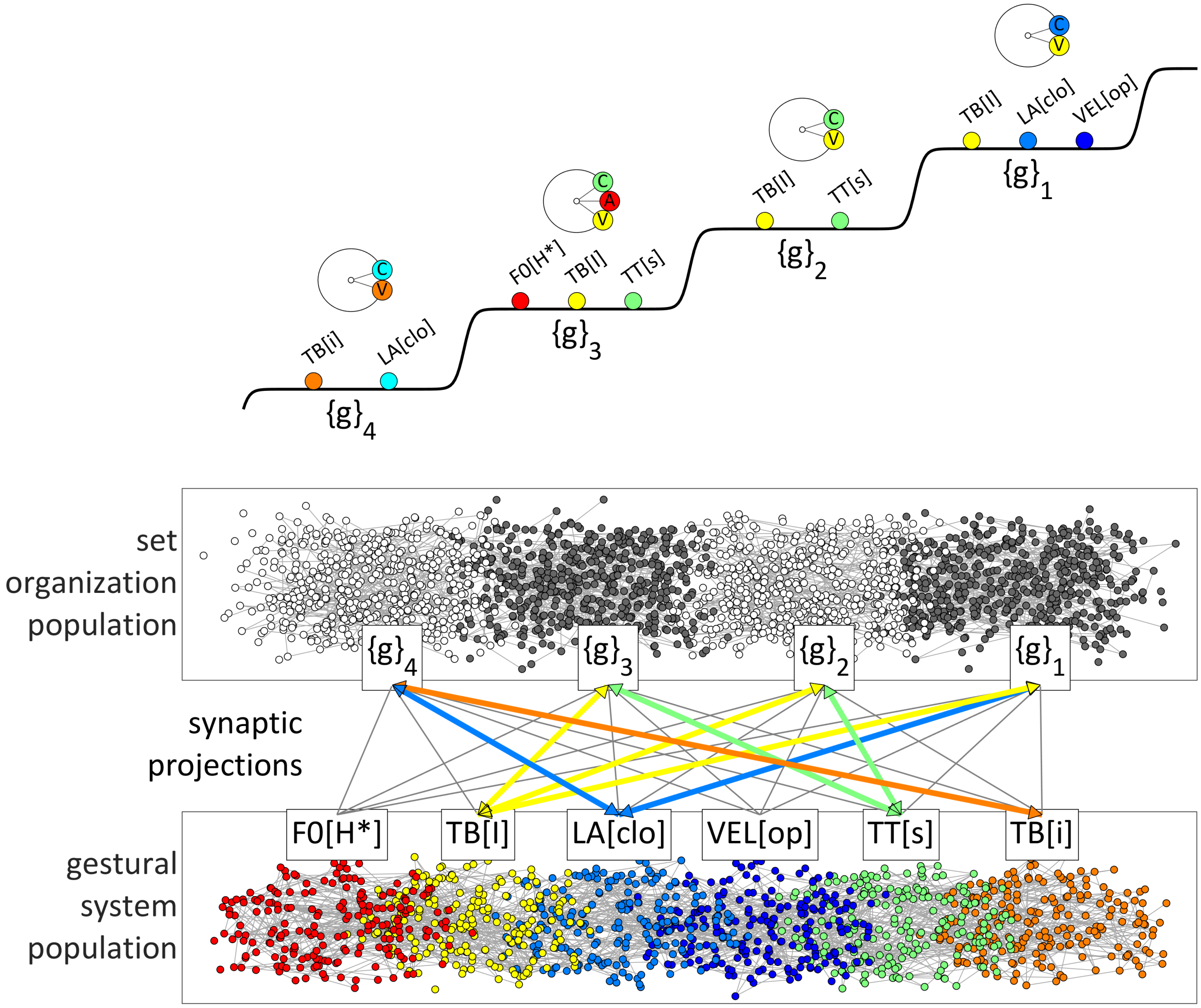
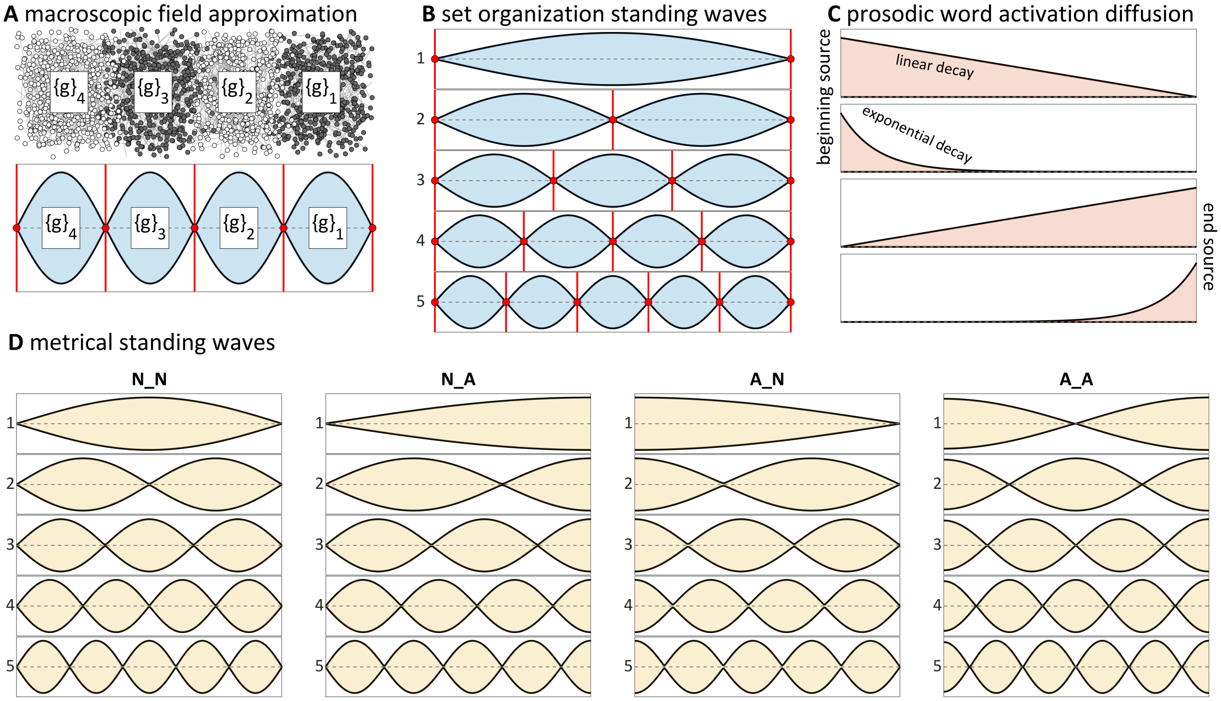


Fig. 8. Relation between macroscale and microscale conception of set organizing systems and gestural systems. A set organization population differentiates into subpopulations which encode sets of gestures. These sets arise from transient coupling of gestural populations to the set organization population. Each subpopulation of the set organization population corresponds to a different set of gestures whose initial activation is organized in the sequencing potential.

In addition to the spatially arranged population of microscopic units that encodes set organization, we posit a second population which is comprised of subpopulations that encode articulatory gestures. The spatial organization of the gestural population does not reflect a spatio-temporal relation; instead, its topology relates to a somatotopic organization based on the targets of articulatory gestures in relevant sensorimotor coordinates. The microscopic units in the gestural population and the set organization population interact bidirectionally via synaptic projections. Via a positive feedback/resonance mechanism, gestural and set organization subpopulations are able to transiently couple when they are in the collective oscillation regime. This mechanism has the effect of temporarily "binding" gestural subpopulations into selection sets. In a sense, this picture is a mechanistic, microscale elaboration of more abstract slot-filler models[56] which describe the organization of segments into syllables.

Starting from the microscale picture, we zoom out to a more macroscopic perspective and refer to a collectively oscillating subpopulation of microscopic units as a *system*. These systems are labelled $\{g\}_1 \ldots \{g\}_n$ in Fig. 8. Each system has a time-varying activation state which is derived from a short-time integration of a function of all of the states of the microscopic units in the corresponding subpopulation. Furthermore, we think of the entire population of microscopic units as a field, so that the systems are associated with distinct regions of a *motor sequencing field*.

Next, we envision that the organization of contemporaneously active sets of gestures is accomplished via a *set organization standing wave* in the motor sequencing field, which leads to picture in Fig. 9A. This particular standing wave pattern is generated by imposing zero amplitude variation (i.e. node) boundary conditions on the spatial edges of the field, which receives a periodic external input. The set organization standing wave self-organizes such that there will be one antinode (local maximum in amplitude variation) for each set of co-selected gestures (cf. the vertical axis labels in Fig. 9B).



A macroscopic field approximation
{g}4
{g}3
{g}2
{g}1
B set organization standing waves
1
2
3
4
5
C prosodic word activation diffusion
beginning source
linear decay
exponential decay
end source
D metrical standing waves
N_N
N_A
A_N
A_A
1
2
3
4
5

Fig. 9. Dynamics in the wave/field model. (A) Macroscopic conceptualization of the set organization population as a field. (B) Set organization standing waves for words of 1-5 syllables. (C) Prosodic word standing waves. (D) Metrical standing waves: each combination of wavenumber and boundary condition is a *mode* of the metrical subfield.

In order to classify accentual systems, two additional dynamical mechanisms are attributed to the motor sequencing field. One involves a *metrical* standing wave which may have symmetric boundary conditions (node-node; antinode-antinode) or asymmetric boundary conditions (node-antinode; antinode-node), and a wavenumber that corresponds to a half-integer multiple of the number of sets (in the case of symmetric b.c.) or a quarter-integer multiple of the number of sets (in the case of asymmetric b.c.). We refer to a combination of boundary conditions and wavenumber as a *mode*; the collection of all possible metrical wave modes for up to five sets is shown in Fig. 9D.

The set organization and metrical standing waves are excited by a periodic source that is located at the beginning or end (i.e. left or right edge) of the field. The frequency of the source is varied in order to excite different modes of the field. All simulations of standing waves were conducted numerically using a finite difference method applied to the 1-dimensional damped wave equation (see Appendix A for details).

The other dynamical mechanism in the motor sequencing field is *prosodic word activation diffusion*. This is modeled by the 1-dimensional diffusion equation (see Appendix A), where a source of excitation is implemented as a non-zero activation boundary condition. Depending on the value of the diffusion coefficient in the equation, the prosodic word activation diffusion pattern exhibits either a linear change in activation density or an exponential decay, as contrasted in Fig. 9C.

The activation of the motor sequencing field is derived from the interactions of the three subfields described above: (i) the set organization subfield, (ii) the metrical subfield, and (iii) the prosodic word subfield. There are a number of ways in which these interactions could be modeled; for current purposes we adopt a relatively simple approach in which motor sequencing field activation is the product of the set organization subfield with a weighted sum of the metrical and prosodic word subfields. By integrating the motor sequencing field activation over the regions of space associated with each partition/set of gestures, activation values are obtained for each set. As in the Goldsmith model, peaks in the activation pattern are associated with accents. In other words, accentual gestures can be co-selected with sets that are associated with a region of the space where there is a peak in the activation pattern. The strongest accent is assumed to couple with the most highly active set of gestures, hence primary accent is the highest peak. Mechanistically, these assumptions are sensible if the activation of a set influences its propensity to couple with accentual gestures; on the micro-scale this implies that if more neurons are spiking in a subpopulation, its interactions with other subpopulations are stronger.

Generating quantity-insensitive patterns in the wave/field model

Given the above constructs, all of the periodic quantity insensitive patterns can be generated by choice of metrical field modes, prosodic word diffusion pattern, and excitation source locations. A full list of model parameters for a range of quantity insensitive patterns is provided in Appendix A. Some examples are shown in Fig. 10A-E, in each case for words comprised of 2-5 syllables. Fig. 10A shows pattern B1r (L→R trochees) and Fig. 10B shows B2r (L→R iambs). The reader should observe that within a given pattern, different metrical modes are used, depending on the number of sets which are organized in a word (or equivalently, the number of field partitions, which often corresponds to the number of syllables). Taken together, we refer to the modes employed for a given pattern as a *progression* of modes, because the wavenumber of the mode increases with the number of organized sets. The reader should also observe that a different progression of metrical modes is used for B2r than for B1r. Ternary periodic patterns as in Fig. 10D are similar to binary ones, except that a different progression of metrical modes is chosen.

Variation in directionality is modeled by varying the location of excitation sources, which can be at the beginning or end of the field. Patterns E1r and E2r, which are the R→L counterparts of B1r and B2r, can be generated with a excitation source at the end of the field; E1r and E2r employ different progressions of metrical modes than B1r and B2r. Because prosodic word activation is strongest at the edge where the source is located, the primary accent (i.e. highest activation peak) will be at the peak closest to this edge.

For generation of aperiodic patterns, there are two reasonable approaches. One is to impose zero weight on the metrical field, as shown for pattern B1 in Fig. 10C. In this case, patterns B2 and E2 require an additional mechanism, a "clamp" which inhibits the edge of the field associated with the prosodic word source (see Appendix A). The clamping mechanism may be useful for generating patterns in which edge-units are "extrametrical". An alternative is to posit an accentual gesture competition mechanism—specific to aperiodic systems—which allows only one accentual gesture to be selected with a group of co-organized sets. In that case, B1 can be derived from B1r, B2 from B2r, etc. The accentual gesture competition mechanism has the advantage that fewer model parameters are needed to generate the full range of quantity insensitive systems.

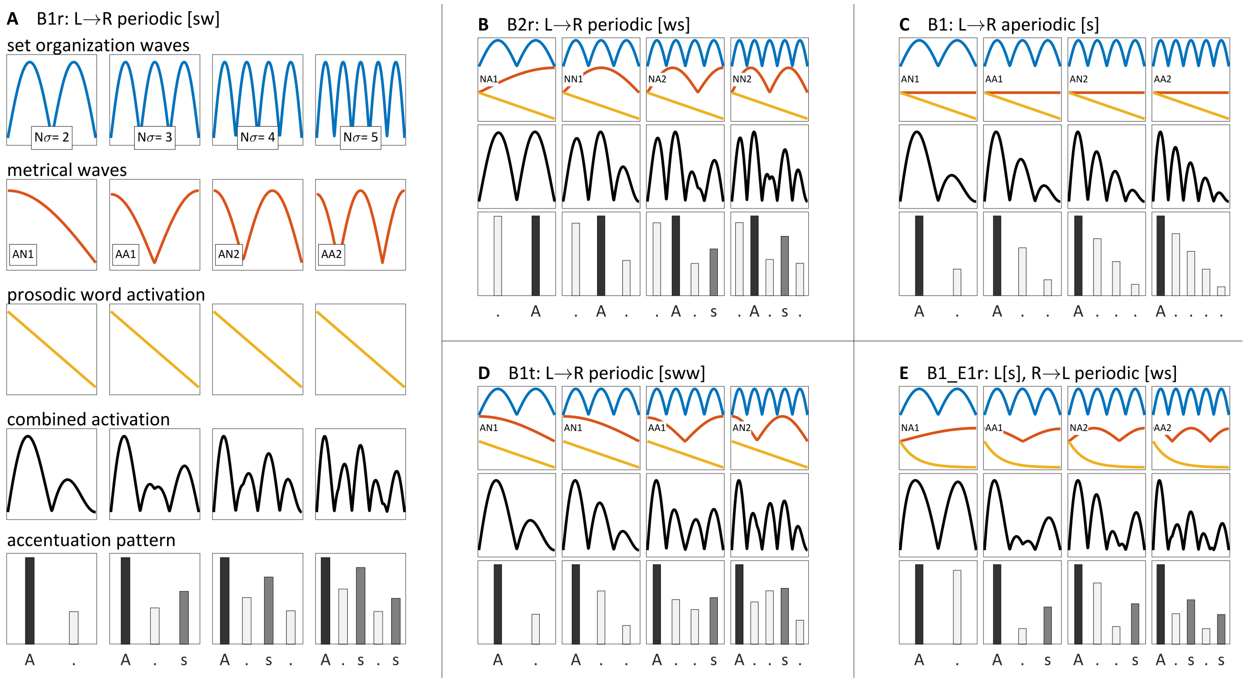


Fig. 10. Examples of quantity insensitive systems generated by the model. (A) B1r, i.e. a L→R periodic binary pattern; (B) B2r is the same as B1r except a different set of metrical modes is selected. (C) An aperiodic pattern generated with zero-weighted metrical field; (D) ternary pattern; (E) bi-directional pattern, where metrical and prosodic word sources are at different edges.

For uni-directional patterns, the excitation source location is the same for the metrical and prosodic word fields. To generate bidirectional patterns, metrical and prosodic word source locations differ. One particular example is shown in Fig. 10E, where the prosodic word subfield has a beginning source, while the metrical field has an end source. Explorations of the model indicate that substantial differences in the relative weighting of metrical and prosodic fields are required for generating bi-directional patterns (see Appendix A). This may be related to the fact that such patterns are relatively rare, and many of the logically possible bi-directional patterns in Fig. 3 are not attested in any languages.

In general, periodic accentual patterns differ with regard to the progression of metrical modes which are employed. It is reasonable to ask if there is any systematicity within or between these progressions. To address this, recall that different modes require different frequencies of source excitation. On the basis of the fact that physical energy of an emitted wave is proportional to the square of frequency, we can arrange all of the possible metrical modes in a hierarchy, according to the square of the resonant source frequency. As shown in Fig. 11A, there is a hierarchy of distinct energy levels, each occupied by two modes. The levels alternate between pairs of asymmetric and symmetric boundary conditions, and increase as wavenumber increases.

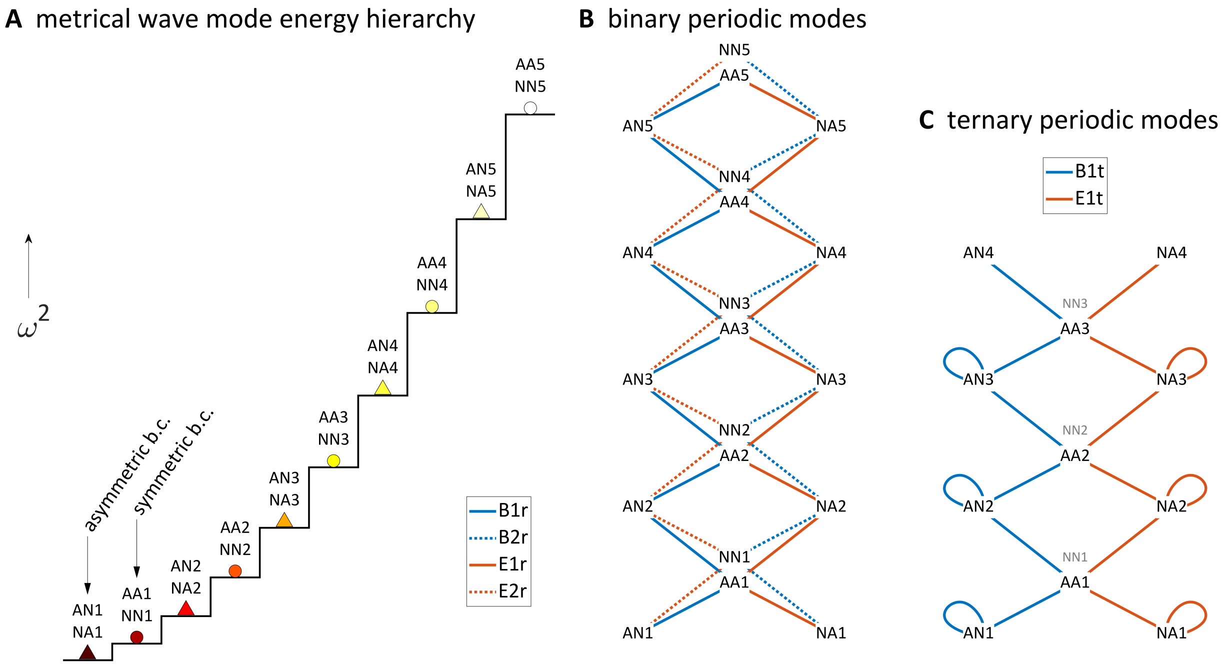


Fig. 11. Energy hierarchy of metrical standing wave modes and progressions of modes used for binary and ternary periodic systems. (A) Energy hierarchy of metrical modes based on the squared frequency of the source excitation required to excite a given mode. (B) Mode progressions for binary periodic patterns. (C) Mode progressions for ternary periodic patterns; loops indicate that a mode is used twice consecutively in the progression.

In Fig. 11B and C, the progressions of modes required for a given periodic accentual pattern are plotted, for binary and ternary patterns respectively. These progressions show that there are systematic relations within and between the progressions that generate a given accentual pattern. The mode progressions for the four periodic binary patterns (B1r, B2r, E1r, E2r) are related through a small set of symmetries involving the field boundary conditions. The modes employed for ternary patterns (B1t, E1t) are the same as those of binary pattern modes, except that each asymmetric mode in the progression is used twice, as indicated by the loops in Fig. 11C. Although it is an open question how speakers of a language learn to employ one progression of modes and not others, the fact that the progressions are systematic makes the learning problem potentially more tractable.

Importantly, the wave/field model incorporates a physical space which maps *indirectly* to temporal order: sets are mapped to space such that the most highly active set in the initial organization of a word form is located at the beginning of the field, with successively less active sets located farther toward the end of the field. Because the space is finite, the multiplicity problem is avoided; indeed, the model holds that as the number of contemporaneously organized sets increases, the space devoted to each becomes smaller. With further elaboration of the model, this could be used predict instability that gives rise to a bound on cardinality (e.g. 7±2 sets[57]). There is also no void space problem: in all circumstances the entirety of the space in the wave/field model is "used" for the purpose of organizing sets of articulatory and accentual gestures; hence there is no issue with what happens in "unused" space.

Quantity-sensitive patterns

In quantity sensitive accentual patterns, the locations of accents are influenced by syllable "weight." Weight is a phenomenon in which syllables can be classified as "heavy" or "light", according to their composition. In some quantity-sensitive languages, only syllables which contain a diphthong or long/tense vowel are heavy; in others, syllables which contain a coda consonant are also heavy.[58,59] Such patterns may also be regular, i.e. fully predictable from syllable composition, or irregular, i.e. derived from lexical long-term memory. English is an example of the latter class. The typology of quantity sensitive accentual patterns is substantially more complicated than that of quantity insensitive ones, and it is important to note that morphological structure can play a role in determining patterns. Because of this, it is beyond the scope of the current article to provide a comprehensive analysis. Our focus here is on the conditions which motivate such patterns, which are predicted by applying the wave/field model to hypothesized developmental changes in organization.

Lets consider geographical names of Native American origin for examples. (This semantic class is useful because such forms are morphologically opaque to speakers). First, the majority of such names conform to a periodic quantity-insensitive pattern (E2r), such as *Mississippi*, *Tallahassee*, and *Massachusetts*. However, some words in this class exhibit primary accent on the final syllable, as in *Kalamazoo, Manitowoc*, *Mattamuskeet*, and *Saxapahaw*. In these forms the final syllable is heavy: it contains a long/tense vowel, or a rime with a coda consonant.

The selection-coordination framework provides a new way of reasoning about how quantity-sensitive patterns of this sort emerge. The developmental hypothesis of s/c theory holds that speakers transition from competitive to coordinative control regimes in early development. Specifically, in early development gestures associated with post-vocalic consonants or the second vocalic gesture in diphthongs are organized into separate competitively selected sets, as shown under the *prototypical competitive control* endpoint of the continuum in Fig. 12A. Subsequently, via increasing reliance on internal feedback for de-gating gestures,[40] children transition to a coordinative regime in which gestures are organized into the same set, labeled as *prototypical coordinative control* in Fig. 12A. This distinction makes the use of the term “syllable” inappropriate for a general model of articulatory organization: for adults, gestures may typically be selected in syllable-sized sets, but for children in the early word stage (1-2.5 y.o.), the sets correspond more closely to moras.

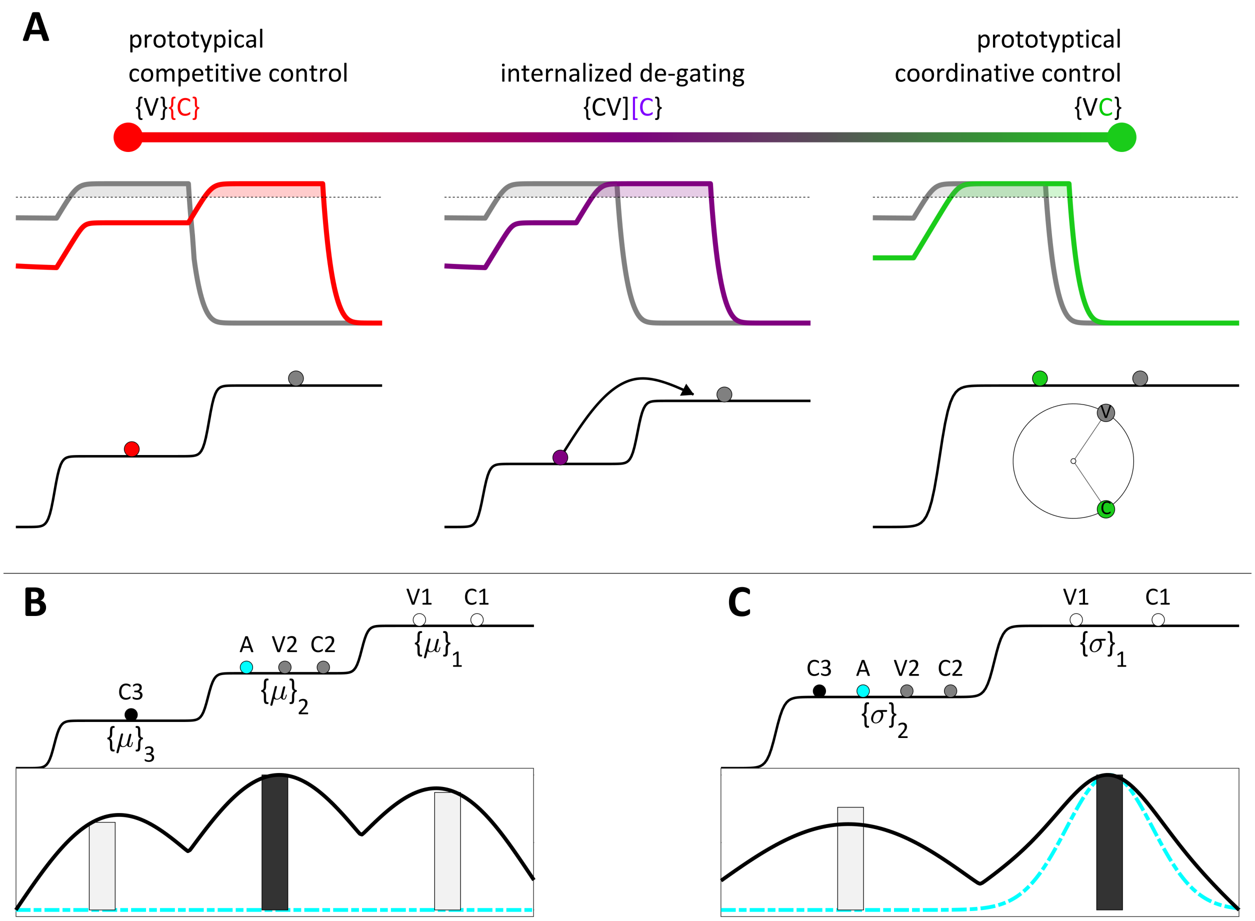


Fig. 12. Application of developmental changes in gestural organization to understanding quantity sensitivity. (A) Hypothesized developmental trajectory from competitive to coordinative control: children learn to co-select gestures which were previously organized into separate sets. (B) Left: moraic organization in which post-vocalic gesture is a separate partition of the motor sequencing field. Right: syllabic organization with lexicalization of accentual gesture selections.

The hypothesized developmental transition is useful for understanding quantity sensitivity because it predicts that in some circumstances, there is an early stage in which the organization system generates a pattern that appears to be quantity sensitive. Specifically, lets consider a CV.CVC word form. Fig. 12B shows the developmentally earlier, moraic organization where the coda consonant of the final syllable is organized as a distinct set of gestures. In this case, an E2r pattern generates accent on second-to-last set, which is the final syllable—this is consistent with quantity-sensitive accentuation. The developmentally later, syllabic organization should—according to the E2r pattern—have accent on the initial syllable, but in quantity-sensitive systems it may exhibit the deviant pattern shown in Fig. 12C. This can readily be attributed to a lexicalization of the selection of the accentual gesture: in the earlier stage, children learn to co-select an accentual gesture with some specific set of gestures in a word form, and this bias on co-selection becomes part of the long-term memory of the word form. In the wave/field model, this can be implemented by introducing a set-specific excitation source, which is directly analogous to the non-uniform internal activation used by Goldsmith (1994)[49] to generate quantity-sensitive patterns. The "lexicalization" mechanism employed here is presumably very general, and can be applied to generating accentuation patterns in so-called "free stress" languages (like English), where in some word forms learned patterns of accentual gesture co-selection override effects of the metrical field.

Additional phenomena: durational lengthening and the rhythm rule

Duration is different from other "correlates" of stress/accent, such as pitch, intensity, and phonation quality. The latter can be readily understood to involve control of the state of the vocal tract (e.g. F0), and are well suited to being modeled as gestures in the Task Dynamic framework. In contrast, accent-related durational variation must be understood differently, because durations are not state variables of the vocal tract (by definition, state variables evolve in time). It is hypothesized here that accentual gestures induce lengthening because they increase attention to external sensory feedback. As represented in Fig. 13A, this increased attention delays the timecourse of feedback-induced suppression, thereby prolonging the period of time during which selected gestures exert forces on the vocal tract. This concords with the s/c analysis of duration in early development: young children tend to produce words that are longer in duration, because they rely to a greater degree than adults on external sensory feedback. An even more general prediction of the above hypothesis is that in the absence of accentual gestures, the degree of regularity in the timing of syllables in adult speech is determined by the regularity in the timecourse of feedback-induced suppression.

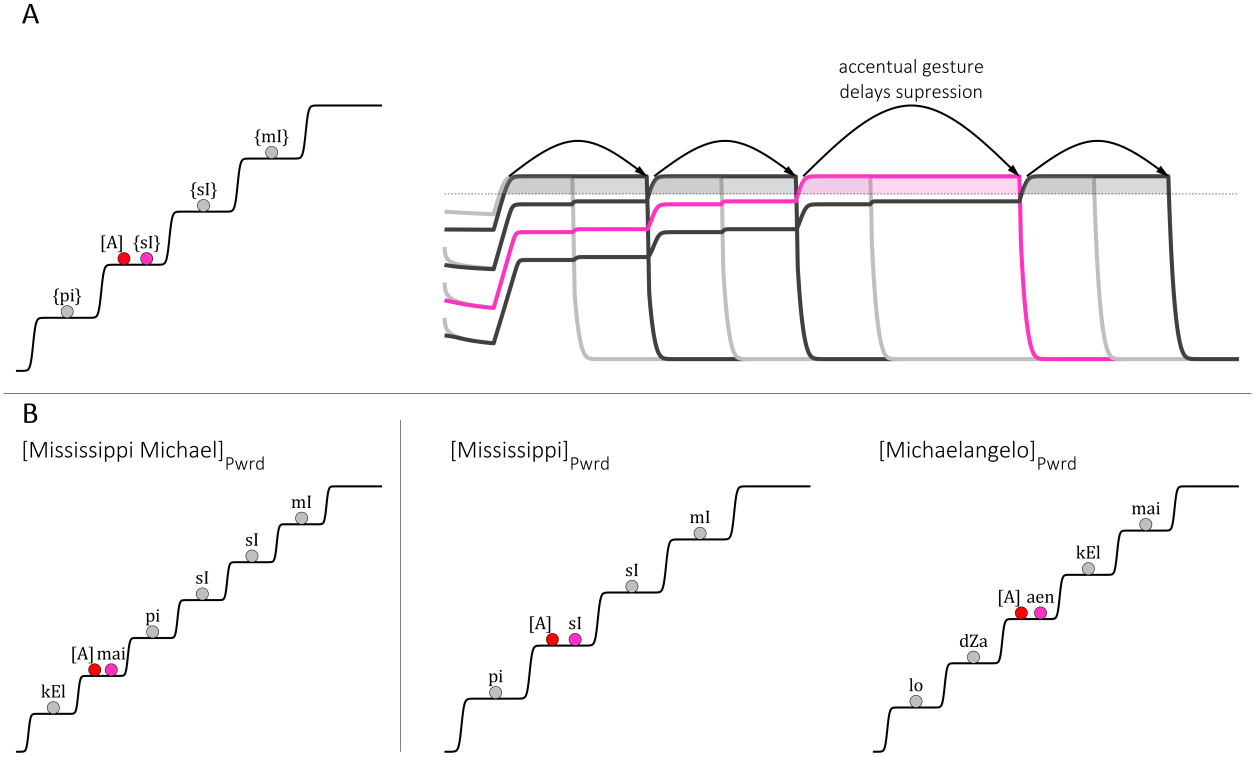


Fig. 13. Hypotheses regarding accentual influences on duration and the rhythm rule. (A) Accentual gestures increase duration by delaying suppression of selected gestures. (B) The rhythm rule as a consequence of whether there is one contemporaneous organization of gestures or a sequence of two organizations.

An important constraint on the wave/field model is that metrical/prosodic word influences on accentuation emerge only for systems which are contemporaneously organized, i.e. organized at the same time in the motor sequencing field. By hypothesis, the scope of this domain tends to correspond to the prosodic word (Pwrd). Crucially, the imposition of the constraint does not entail that there is a one-to-one mapping of utterances to prosodic words. Consider an analysis of the so-called "rhythm-rule" pattern in

Fig. 13B: in a noun-noun compound the primary accent on the first member of the compound is reduced, as in *Mississippi Michael*. The "deletion" of the accent on Mississippi is predicted when the sets of gestures in *Mississippi* and *Michael* are contemporaneously organized, i.e. produced as a single prosodic word. As the cardinality of the second member of the compound increases, e.g. *Mississippi Mikaela, Mississippi Michaelina, Mississippi Michaelangelo*, etc., the likelihood that the forms will be organized as a single prosodic word decreases. When the forms are produced as a sequence of two prosodic words, each of the two organizations is predicted to have one primary accent, and hence no "deletion" of primary accent is expected. Note that some mechanism is required for resetting the motor sequencing field when a sequence of prosodic words is produced, and this likely involves syntactic-conceptual mechanisms of the sort described in Tilsen (2018).[48] Rather than viewing the rhythm rule as a consequence of proximity of primary accents, as has been the traditional approach, the phenomenon is thus reinterpreted as a consequence of whether sets of gestures associated with a pair of words are organized at the same time or in a sequence.

Conclusion

The s/c wave/field model of speech rhythm is a radical departure from previous approaches. At issue is whether we conceptualize the generation of temporal patterns in speech as the product of mechanisms which serve the purpose of creating a rhythmic pattern, or whether temporal patterns are an indirect consequence of articulatory-accentual gesture organization. Whereas traditional approaches presuppose a representation which generates a rhythmic pattern, the current model, as well as its inspiration, the Goldsmith (1994) model, hold that rhythmic patterns arise indirectly, as a consequence of a spatial organization. By integrating a spatial model with the selection-coordination framework, we can see that regularity in the timing of execution of sets of gestures (i.e. syllables or moras) arises from regularity in the timecourse of feedback-induced suppression, and that regularity in the timing of accentual gestures arises because co-selection of accentual gestures with articulatory gestures is biased by standing waves in a motor sequencing field.

The wave/field model can be viewed in part as a reinterpretation or elaboration of the Goldsmith model. There are a number of similarities: the positional activation parameters parallel the location of source excitation, a mapping of units to a physical space is utilized, and in this space there are spatial waves (although the mechanisms which give rise to the waves differ). Also, the positive/negative sign of positional activation in the Goldsmith model corresponds to the anti-node/node boundary conditions in the wave/field model. However, the wave/field model is not simply an alternative vocabulary. By integrating the model into the selection-coordination framework, the model allows for the partitioning of the organizing space to vary in the course of development. This provides a natural basis for understanding quantity sensitive patterns through lexicalization of patterns which arose in earlier stages of development.

The focus of this paper has been on accentuation in spontaneous conversational speech, but it is evident that periodicity of accentuation, i.e. the rhythmicity of speech, may be substantially enhanced in certain contexts or genres such as poetry, chant, lyrical music, and even prepared speech. A sensible account of periodicity enhancement in such contexts involves the entrainment of selection and suppression events to an external periodic signal (as in lyrical music) or an internally generated signal (as in composition and production of poetry). However, it is also evident that the scope of organization can be adjusted to promote rhythmicity. For example, a spontaneous conversational production of the phrase *twinkle twinkle little star* is presumably organized quite differently from a lyrical production, where each syllable may become a separate prosodic word and have a primary accent.

Finally, we can re-interpret the directionality parameter of accentual patterns: it is no longer necessary to impose the *temporal order is spatial arrangement* metaphor on our conceptualization of rhythm, because we have posited a real space in which the subsystems of a word form (i.e. sets of gestures) are arranged. There are no "temporal edges" in this view. Instead, there is a field whose dynamics include a partitioning of space. This partitioning corresponds to the count of competitively selected sets in a word form, i.e. cardinality, and cardinality determines which metrical standing wave mode is dominant in periodic systems. The claim that stress is "purely structural" thus gains a more detailed meaning: stress is the side-effect of diffusive prosodic activation and metrical standing waves, which interact to create biases on the selection of accentual gestures. More careful attention to the use of spatiotemporal metaphors in our theories is what makes this new understanding possible.

Acknowledgements
I would like to thank John Coleman and Lucian Wong for helpful discussions in the course of developing this manuscript.

References

1. Hayes B. 1995. "*Metrical stress theory: principles and case studies*." University of Chicago Press.
2. Hyman L.M. 2008. Directional asymmetries in the morphology and phonology of words, with special reference to Bantu. *Linguistics* **46**: 309–350.
3. Beckman J.N. 2013. "*Positional faithfulness: an Optimality Theoretic treatment of phonological asymmetries*." Routledge.
4. Lavoie L.M. 2001. "*Consonant strength: Phonological patterns and phonetic manifestations*." Routledge.
5. Barnes J. 2008. "*Strength and weakness at the interface: Positional neutralization in phonetics and phonology*." Walter de Gruyter.
6. Keating P.A. 2006. Phonetic encoding of prosodic structure. *Speech Prod. Models Phon. Process. Tech.* 167–186.
7. Nooteboom S.G. 1981. Lexical retrieval from fragments of spoken words: beginnings vs. endings. *J. Phon.* **9**: 407–424.
8. Schwartz B.L. 2001. "*Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*." Psychology Press.
9. James L.E. & D.M. Burke. 2000. Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *J. Exp. Psychol. Learn. Mem. Cogn.* **26**: 1378.
10. Ladd D.R. 2008. "*Intonational phonology*." Cambridge: Cambridge University Press.
11. Gussenhoven C. 2004. "*The phonology of tone and intonation*." Cambridge: Cambridge University Press.
12. Tilsen S. 2018. "*Three mechanisms for modeling articulation: selection, coordination, and intention*." Cornell Working Papers in Phonetics and Phonology 2018.
13. Fry D.B. 1955. Duration and intensity as physical correlates of linguistic stress. *J. Acoust. Soc. Am.* **27**: 765.
14. Fant G., A. Kruckenberg & L. Nord. 1991. Durational correlates of stress in Swedish, French, and English. *J. Phon.*
15. Sluijter A.M. & V.J. Van Heuven. 1996. Spectral balance as an acoustic correlate of linguistic stress. *J. Acoust. Soc. Am.* **100**: 2471–2485.
16. Ortega-Llebaria M. & P. Prieto. 2011. Acoustic correlates of stress in Central Catalan and Castilian Spanish. *Lang. Speech* **54**: 73–97.
17. Campbell N. & M. Beckman. 1997. Stress, prominence, and spectral tilt. In *In: Botinis, A., Kouroutpetroglou, G., Carayiannis, G. (Eds.), Intonation: Theory, models and applications* 67–70. Athens.
18. Allen G.D. 1972. The location of rhythmic stress beats in English: An experimental study I. *Lang. Speech* **15**: 72–100.
19. Lakoff G. 2008. "*Women, fire, and dangerous things*." University of Chicago press.

20. Lakoff G. & M. Johnson. 1980. The metaphorical structure of the human conceptual system. *Cogn. Sci.* **4**: 195–208.
21. Goldsmith J.A. 1979. Autosegmental phonology. .
22. Goldsmith J.A. 1990. “*Autosegmental and metrical phonology*.” Basil Blackwell.
23. Coleman J. & J. Local. 1991. The “no crossing constraint” in autosegmental phonology. *Linguist. Philos.* **14**: 295–338.
24. Browman C. & L. Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* **49**: 155–180.
25. Goldstein L. & C.A. Fowler. 2003. Articulatory phonology: A phonology for public language use. *Phon. Phonol. Lang. Comprehension Prod. Differ. Similarities* 159–207.
26. Saltzman E. & K. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1**: 333–382.
27. Kelso J.A.S. & E. Saltzman. 1987. Skilled actions: A task dynamic approach. *Psychol. Rev.* **94**: 84–106.
28. Gao M. 2008. “*Tonal Alignment in Mandarin Chinese: An Articulatory Phonology Account*.” Doctoral Dissertation, Yale University, New Haven, CT.
29. Yi H. 2017. Lexical tone gestures. .
30. Niemann H., D. Mücke, H. Nam, *et al.* 2011. Tones as Gestures: the Case of Italian and German. *Proc. ICPhS XVII* 1486–1489.
31. Tilsen S., D. Burgess & E. Lantz. 2013. Imitation of intonational gestures: a preliminary report. *Cornell Work. Pap. Phon. Phonol. 2013* 1–17.
32. Halle M. & J.-R. Vergnaud. 1987. “*An essay on stress*.” MIT press.
33. Jensen J.T. 2000. Against ambisyllabicity. *Phonology* **17**: 187–235.
34. Peperkamp S.A. 1997. “*Prosodic words*.” Holland Academic Graphics The Hague.
35. Nespor M. & I. Vogel. 1986. Prosodic phonology. *Prosodic Phonol.*
36. Kager R.W.J. 1995. The metrical theory of word stress. *Blackwell Handb. Linguist.* **1**: 367–402.
37. Gordon M. 2002. A factorial typology of quantity-insensitive stress. *Nat. Lang. Linguist. Theory* **20**: 491–552.
38. Ridouane R. 2008. Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* **25**: 321–359.
39. Dell F. & M. Elmedlaoui. 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *J. OfAfrican Lang. Linguist.* **7**: 105–130.
40. Tilsen S. 2016. Selection and coordination: The articulatory basis for the emergence of phonological structure. *J. Phon.* **55**: 53–77.
41. Tilsen S. 2013. A Dynamical Model of Hierarchical Selection and Coordination in Speech Planning. *PloS One* **8**: e62800.
42. Saltzman E., L. Goldstein, C. Browman, *et al.* 1988. Modeling speech production using dynamic gestural structures. *J. Acoust. Soc. Am.* **84**: S146–S146.
43. Grossberg S. 1978. A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. *Prog. Theor. Biol.* **5**: 233–374.
44. Grossberg S. 1987. The adaptive self-organization of serial order in behavior: Speech, language, and motor control. *Adv. Psychol.* **43**: 313–400.
45. Bullock D. & B. Rhodes. 2002. Competitive queuing for planning and serial performance. *CASCNS Tech. Rep. Ser.* **3**: 1–9.
46. Bullock D. 2004. Adaptive neural models of queuing and timing in fluent action. *Trends Cogn. Sci.* **8**: 426–433.
47. Saltzman E., H. Nam, J. Krivokapic, *et al.* 2008. A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In *Proceedings of the 4th international conference on speech prosody* 175–184. Brazil: Campinas.

48. Tilsen S. 2018. "*Syntax with oscillators and energy levels*." Cornell Working Papers in Phonetics and Phonology 2018.
49. Goldsmith J. 1994. A dynamic computational theory of accent systems. *Perspect. Phonol.* 1–28.
50. Acebrón J.A., L.L. Bonilla, C.J.P. Vicente, *et al.* 2005. The Kuramoto model: A simple paradigm for synchronization phenomena. *Rev Mod Phys* **77**: 137–185.
51. Breakspear M., S. Heitmann & A. Daffertshofer. 2010. Generative models of cortical oscillations: neurobiological implications of the Kuramoto model. *Front. Hum. Neurosci.* **4**: 190.
52. Hong H. & S.H. Strogatz. 2011. Kuramoto model of coupled oscillators with positive and negative coupling parameters: an example of conformist and contrarian oscillators. *Phys. Rev. Lett.* **106**: 054102.
53. Kelso J.A.S. 1997. "*Dynamic patterns: The self-organization of brain and behavior*." MIT press.
54. Strogatz S.H. 2000. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Phys. Nonlinear Phenom.* **143**: 1–20.
55. Buzsáki G. & A. Draguhn. 2004. Neuronal oscillations in cortical networks. *science* **304**: 1926–1929.
56. Shattuck-Hufnagel S. 1979. Speech errors as evidence for a serial-ordering mechanism in sentence production. In *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* 295–342.
57. Miller G.A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**: 81.
58. Zec D. 2007. The syllable. *Camb. Handb. Phonol.* 161–194.
59. Hyman L.M. 1985. "*A theory of phonological weight*." Dortrecht: Foris Publications.

Appendix: model details

The set organization and metrical subfield standing waves in the model were generated from numerical simulations of the damped 1D scalar wave equation, shown in Eq. 1:

Eq. 1 $$\frac{\partial^2 u(x,t)}{\partial t^2} + \gamma \frac{\partial u}{\partial t} = v^2 \frac{\partial^2 u(x,t)}{\partial x^2}$$

Here $u(x,t)$ is the wavefunction, v is the wave speed, and γ the damping coefficient. The damping term is included for generality but no damping was imposed in the simulations reported in this manuscript. Numerical solutions were obtained with a finite difference method, using the central difference approximations in Eqs. 2-4:

Eq. 2 $$\frac{\partial^2 u(x_i,t)}{\partial x^2} = \frac{u(x_{i-1},t) - 2u(x_i,t) + u(x_{i+1},t)}{\Delta x^2}$$

Eq. 3 $$\frac{\partial^2 u(x_i,t)}{\partial t^2} = \frac{u(x_i,t-1) - 2u(x_i,t) + u(x_i,t+1)}{\Delta t^2}$$

Eq. 4 $$\gamma \frac{\partial u}{\partial t} = \frac{\gamma}{2\Delta t}[u(x_i, t+1) - u(x_i, t-1)]$$

Substituting Eqs. 2-4 into the wave equation gives the following updating rule Eq. 5:

Eq. 5 $$\left[\frac{1}{\Delta t^2} + \frac{\gamma}{2\Delta t}\right] u(x, t+1) = \frac{2}{\Delta t^2} u(x,t) - \left[\frac{1}{\Delta t^2} - \frac{\gamma}{2\Delta t}\right] u(x, t-1) + \left[\frac{v^2}{\Delta x^2}\right][u(x_{i-1}, t) - 2u(x_i, t) + u(x_{i+1}, t)] + s(x, t+1)$$

In all simulations, the temporal boundary condition u(x,t=0) = 0 was imposed: this represents the simplifying assumption that the motor sequencing field is quiescent when gesture sets are initially organized. Standing waves associated with all combinations of Dirichlet and Neumann boundary conditions and wavenumbers from 1 to 10 were generated. The Dirichlet (node) condition fixes *u* at a boundary to 0 at all times. The Neumann (antinode) condition sets *u* at a boundary to be equal to *u* at the adjacent point in space for each time. In all simulations the time step Δt was set to 0.0001 s, the transmission velocity *v* = 40, and the field length L = 1. To facilitate numerical simulation, the spatial resolution Δx was chosen such that the Courant number u_S = *v* Δt/Δx = 1. Input source frequencies *f* were chosen to produce resonant standing waves for each combination of wavenumber and boundary conditions according to *f* = *v*/λ, where λ is the wavelength. For asymmetric boundary conditions, λ = 4L/(2*n*-1), where *n* is the wavenumber. For symmetric boundary conditions, λ = 2L/*n*. For beginning (B) and end (E) sources, a sinusoidal input of frequency *f* and unit amplitude was added in each time step, to the point in the field adjacent to the relevant boundary.

The prosodic word field dynamics were generated from numerical simulations of the 1D diffusion equation, shown in Eq. 6. The spatial central difference approximation from Eq. 2 was again used, leading to the updating rule in Eq. 7, where *D* is a diffusion parameter. Beginning (B) and end (E) excitation sources are modeled as spatial boundary conditions u(x=0,t) = 1 and u(x=L,t) = 1, respectively. Two values of D were used, *D*=0.5 which generates a stationary linearly decaying activation pattern, and *D*=0.005 which generates a stationary exponentially decaying activation pattern.

Eq. 6 $$\frac{\partial u}{\partial t} = D \frac{\partial^2 u(x,t)}{\partial x^2}$$

Eq. 7 $$u(x, t+1) = u(x, t-1) + D[u(x_{i-1}, t) - 2u(x_i, t) + u(x_{i+1}, t)]$$

Activation functions were calculated as follows. For set organization and metrical subfield standing waves, the positive envelope of *u* is extracted over time by taking the maximum of u over time at each point in space. The envelope is then normalized to range from 0 to 1. The normalized metrical subfield standing wave and the prosodic word diffusion function are added together according to the weights specified in Table A.1 below. The result is multiplied by the normalized set organization standing wave. Note that the linearly decaying and exponentially decaying diffusion patterns in the prosodic word subfield are labelled LL1 and ZZ1. For accents which are aperiodic and which are associated with a non-edge partition (i.e. B2, E2), it is necessary to impose a "clamp", i.e. inhibition of the field in the adjacent edge partition. This is accomplished by adding -1 activation at each spatial position in the edge partition. An alternative approach is to employ the periodic patterns B1r, B2r, E1r, and E2r to generate the aperiodic patterns B1, B2, E1, and E2. This is done by incorporating an independent mechanism which allows only one accentual gesture to be selected for each group of co-organized sets.

Table A.1 Wave/field model parameters for quantity insensitive patterns

	metrical subfield			prosodic word subfield			
pattern	modes (σ=2…8)	source	weight	mode	source	weight	clamp
B1	AN1,AA1,AN2,AA2,AN3,AA3,AN4	B	0	LL1	B	1	
B2	NA1,NN1,NA2,NN2,NA3,NN3,NA4	B	0	LL1	B	1	B
B1r	AN1,AA1,AN2,AA2,AN3,AA3,AN4	B	0.95	LL1	B	1	
B2r	NA1,NN1,NA2,NN2,NA3,NN3,NA4	B	0.95	LL1	B	1	
B1t	AN1,AN1,AA1,AN2,AN2,AA2,AN3	B	0.95	LL1	B	1	
E1	NA1,AA1,NA2,AA2,NA3,AA3,NA4	E	0	LL1	E	1	
E2	AN1,NN1,AN2,NN2,AN3,NN3,AN4	E	0	LL1	E	1	E
E1r	NA1,AA1,NA2,AA2,NA3,AA3,NA4	E	0.95	LL1	E	1	
E2r	AN1,NN1,AN2,NN2,AN3,NN3,AN4	E	0.95	LL1	E	1	
E1t	NA1,NA1,AA1,NA2,NA2,AA2,NA3	E	0.95	LL1	E	1	
B1_E1	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	E	0.5	ZZ1	B	1	
B1_E1r	NA1,AA1,NA2,AA2,NA3,AA3,NA4	E	0.5	ZZ1	B	1	
B1_E2	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	E	0.5	ZZ1	B	1	E
B1_E2r	AN1,NN1,AN2,NN2,AN3,NN3,AN4	E	0.5	ZZ1	B	1	
E1_B1	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	B	0.5	ZZ1	E	1	
E1_B1r	AN1,AA1,AN2,AA2,AN3,AA3,AN4	B	0.5	ZZ1	E	1	
E1_B2	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	B	0.5	ZZ1	E	1	B
E1_B2r	NA1,NN1,NA2,NN2,NA3,NN3,NA4	B	0.5	ZZ1	E	1	
B2_E1	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	E	0.05	ZZ1	B	2	B
B2_E1r	NA1,AA1,NA2,AA2,NA3,AA3,NA4	E	0.1	ZZ1	B	1.75	B
B2_E2	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	E	0.25	ZZ1	B	1	B
B2_E2r	AN1,NN1,AN2,NN2,AN3,NN3,AN4	E	0.25	ZZ1	B	1.75	B
E2_B1	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	B	0.05	ZZ1	E	2	E
E2_B1r	AN1,AA1,AN2,AA2,AN3,AA3,AN4	B	0.1	ZZ1	E	1.75	E
E2_B2	ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1,ZZ1	B	0.25	ZZ1	E	1	E
E2_B2r	NA1,NN1,NA2,NN2,NA3,NN3,NA4	B	0.25	ZZ1	E	1.75	E