Hierarchical spatiotemporal dynamics of speech rhythm and articulation

by

Samuel Edward Tilsen

B.A. (Northwestern University) 2001
M.A. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Linguistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Keith Johnson, Chair
Professor Sharon Inkelas
Professor Richard Ivry

Spring 2009

Chapter 1: Introduction

The similarity between representations of structure across linguistic subdomains is remarkable. Branching, hierarchical, tree-like structures are used as descriptive and theoretic tools in feature geometry, syllable constituency, metrical and prosodic theories, and various morphological and syntactic theories. Why is hierarchical structure observed across linguistic subdomains? Is it because scientists find such structures conceptually appealing, or is it because general cognitive principles give rise to hierarchies? To view linguistic units as hierarchically arranged often goes hand-in-hand with thinking of them as connected to, contained within, and specially related to other units. The hierarchical conceptualization of articulatory and prosodic units necessarily evokes metaphors that discretize time, and many representations of the prosodic hierarchy make use of a conventional image-schema in which prosodic categories are objects, and relations between those categories are connections. These object- and connection- based representations of hierarchical structure are useful for many purposes, but for understanding temporal patterning in speech, new models are needed.

This dissertation explores the idea that general principles governing self-organizing dynamical systems can explain how hierarchical patterns emerge in language. The focus is on several levels of phonological patterns: articulatory gestures, segments, syllables, feet, and prosodic phrases. The theory proposed herein is that linguistic systems can be conceptualized as idealized waves interacting in space and time. This "wave and field" theory is founded in part on a theoretical approach to understanding self-organizing systems developed by Herman Haken, and on the application of this approach to cognitive systems. Fundamental to these endeavors are the conceptual tools of dynamical systems theory, which offer general, powerful ways of geometrically visualizing pattern formation processes. The theory presented herein is primarily integrative, combining ideas from linguistics, motor coordination, and cognitive neuroscience. The gist of the wave and field theory is that all speech units are instantiated physically as neural ensembles (or groups of such ensembles) whose activation is embedded in multi-dimensional fields, and that statistically the activation of these ensembles exhibits approximately wave-like dynamics in the course of speech planning and production. Thus, relations between prosodic units, and linguistic structure itself, can be conceptualized as emerging from interactions between wave-like systems, and the spatial organization of these waves can be understood with dynamical fields. These concepts generate new predictions about what sorts of patterns may be observed in speech, and some of which are tested in this dissertation.

The conceptual tools of the wave theory allow for branching, hierarchical structure to be reconceptualized in a new way. To give the reader a very basic idea of how hierarchical structure can be related to wave interaction, Figure 1 (a) shows an abstract hierarchical structure, and (b) and (c) show how that structure can be interpreted as interacting waves. The "units" X and Y are reconceptualized as oscillatory systems, which are pictured as points moving around a circle, as in (b). The unit "Z"—typically thought of as the *combination* of X and Y, or as a unit *connected* to X and Y, which may also *contain* X and Y—is likewise conceptualized as a point moving around a circle. However, an important general constraint holds upon how quickly X, Y, and Z move around the circle: X and Y move at approximately the same speed, while Z moves at approximately an integer ratio of that speed, typically 1/2 or 1/3, but generally $1/n$, where n is an integer. Hence for every revolution that Z makes around the circle, X and Y make approximately n revolutions. Figure 1 (c) shows the oscillatory amplitude variation of X, Y, and Z over time.

These oscillations are amplitude modulated because they represent the transient *activation* of the systems.

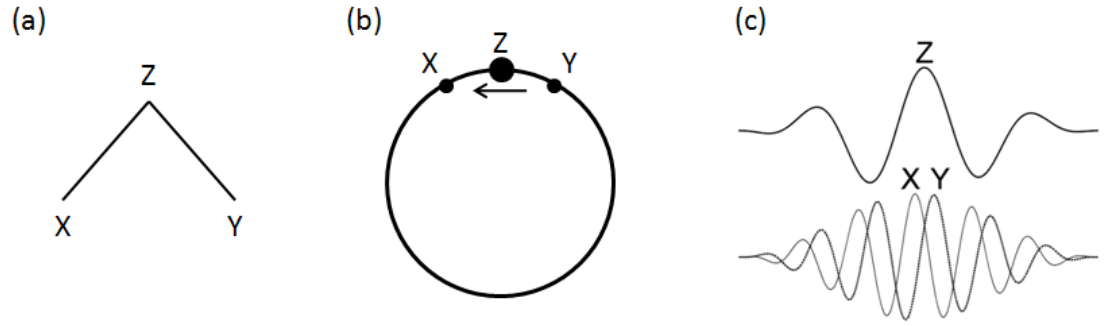


Figure 1. Illustration of the relation between hierarchical structure, oscillatory motion, and interacting waves.

Figure 1 expresses the insight that relatively fast lower-level oscillatory systems are coupled to relatively slower, higher-level oscillatory systems. By taking this *principle of harmonic frequency-locking* as a guiding principle of the wave theory, the prosodic hierarchy can be reconceptualized as in Figure 2:

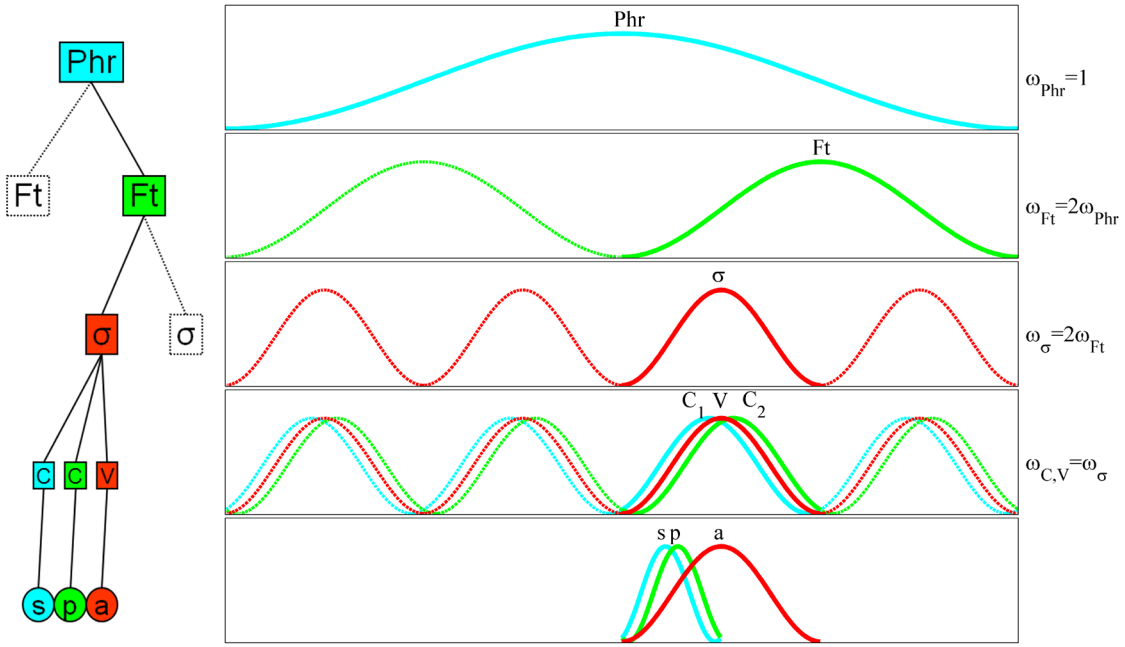


Figure 2. Relations between the representation of prosodic hierarchy and a multiscale dynamical model. The word "spa" is used as an example. Prosodic units are conceptualized as oscillatory systems; prosodic and segmental levels in the hierarchy are conceptualized as timescales, which correspond to the inherent frequencies of their associated oscillators, and which are related by integer ratios. Articulatory gestures exhibit overdamped mass-spring dynamics.

The figure shows that the levels of the prosodic hierarchy associated with segments, syllables, feet, and phrases can be conceptualized as distinct timescales. These timescales characterize the oscillatory frequencies of the systems associated with each level. There is much more to this theory, particularly with regard to asymmetries in the strength and directionality of coupling between wave systems—these considerations will be introduced in subsequent chapters.

The field aspect of the theory presented in this dissertation is schematized in Figure 3. Dynamical fields allow for wavelike planning activity to represent activation in multiple spatial dimensions, which is useful for conceptualizing how articulatory targets are planned. Field activity is derived from excitatory (gestural activation) and inhibitory (intergestural inhibition) components. In Figure 3, several stages of planning field activity are depicted: first an [i] is planned, then there is an intermediate stage with both [i] and [a] active, and then the plans for [i] are inhibited as the [a] becomes fully active. The symbol (+) in the bottom right panel corresponds to a target for the vowel [a] after an [i] has been planned, and the symbol (□) corresponds to an [a] target in the absence of any prior planning. Experimental evidence for spatial patterns of this sort is presented in Chapter 6.

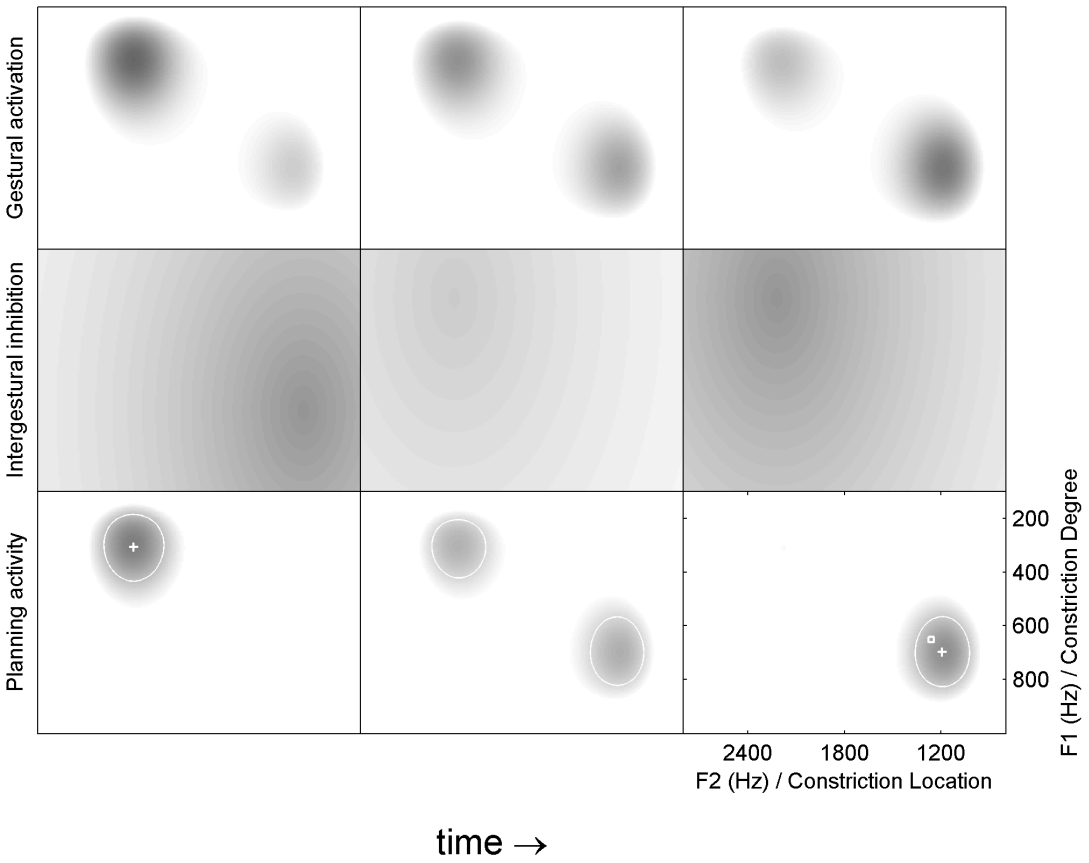


Figure 3. Schematic representation of a dynamical planning field at three stages of time in the transition from planning an [i] to planning an [a]. The top row depicts gestural activation, the middle row depicts intergestural inhibition, and the bottom row depicts planning activity, which is the integration of the gestural activation and inhibition. Chapter 7 describes these concepts in more detail. The (+) in the bottom right panel is an [a] target after an [i] has been planned, and the (□) is an [a] target in the absence of any prior gestural planning.

This dissertation tests two important predictions of the wave and field theory. First, the dissertation tests the prediction that there may be dynamical interactions of a temporal nature between the systems that govern articulatory gestures and higher-level prosodic systems, such as feet and phrases. This means that regularities in how foot intervals are timed relative to a phrase may influence aspects of how articulatory gestures are timed. For example, the timing of tongue and lip movements for the [s] and [p] in the word *spa* may be influenced by timing between foot and phrase waves (cf. Figure 2). Second, the dissertation tests the prediction that there may be dynamical interactions of a spatiotemporal nature between contemporaneously planned gestures. This means that the articulatory context in which an articulatory gesture is planned will have an influence on the *target* of that gesture. For example, the target of the vowel [a] may be different when an [i] is contemporaneously planned vs. when an [a] is contemporaneously planned.

Before proceeding to the next chapter, a roadmap for the dissertation is provided here. Chapter 2 examines the concept of HIERARCHY in general, and in application to linguistics. Section 2.1 shows that the concept of hierarchy is often used with a network of related concepts, including connection, containment, and relation. Although basic concepts of rank and order are more central to hierarchy, the more interesting uses of hierarchy involve some form of connection and/or containment. Section 2.2 examines how hierarchical structure in language is represented with node and connection schemas, and explores how the concept of "activation" has been integrated into representations of hierarchical structure. Section 2.3 considers several different perspectives on the philosophical question of *why* we observe hierarchical structure—is hierarchical structure imposed by scientists on language, or does it emerge from general cognitive principles? Section 2.4 explains why current representations of hierarchical structure do not make sufficiently detailed temporal predictions about interactions between speech rhythm and articulatory gestures. It also suggests that patterns that are understood as constraints upon, or operations upon, hierarchical structure, can be reconceptualized to arise from principles governing wave interactions.

Chapter 3 reviews some of the experimental and theoretical work that provides a foundation for the wave theory. Section 3.1 presents some of the basic concepts of coordination dynamics, such as relative phase coupling, generalized relative phase coupling, and potential functions. Section 3.2 examines in detail the task dynamic model of articulatory gestures. Section 3.3 shows how the task dynamic model of gestures was extended to understand the relative timing of gestures. Section 3.4 describes a coupled-oscillators model of cross-linguistic variation in the durations of feet. Section 3.5 describes experimental and theoretical work on the relative timing of feet within phrases. Section 3.6 proposes that the dynamical approaches to modeling articulatory gestures, syllables, feet, and phrases can be integrated into a single model that makes testable predictions about how prosodic systems may interact with gestural systems. Section 3.7 concludes the chapter by presenting a brief argument for why the wave theory is neuro-cognitively plausible.

Chapter 4 presents experimental evidence that temporal patterns on rhythmic and gestural timescales interact in a way that is not amenable to treatment within non-dynamical frameworks. An experiment was conducted to test whether variability in rhythmic timing is correlated with variability in intergestural timing. To that end, gestural kinematics were recorded with electromagnetic articulometry in a phrase repetition task, where metronome tones were used to guide the relative timing of feet within a phrase. In previous studies employing this paradigm, Cummins & Port (1998) found that variability in rhythmic timing depends upon the target defined by the metronome tones. Using this observation, the experiment reported in this chapter shows that articulatory timing is also more variable when rhythmic timing is more variable. These results strongly suggest that there are dynamical interactions between higher-level prosodic systems and gestural systems.

Chapter 5 reports on a corpus study that investigates how the likelihood of segmental deletion in speech is influenced by the rhythm of speech. Phonetic deletions of consonants and vowels were identified in the Buckeye corpus of conversational speech by comparison of the citation form of a word and its phonetic transcription. 2-3 s stretches of phonetically labeled speech, centered around deletions, were extracted from the corpus. Another dataset of 2-3 s stretches of speech that did not contain a deletion were extracted from the corpus. The stretches were subjected to low-frequency spectral analysis using a method of rhythmic analysis described

in Tilsen & Johnson (2008). By comparing the distributions of rhythm spectrum peaks from deletion and non-deletion datasets, and controlling for speech rate, the chapter shows that deletions are associated with more rhythmic speech. Furthermore, vowel and consonant deletions are associated with rhythmicity at different frequencies. These results lend further credence to the idea that aspects of higher-level prosodic timing interact with lower-level articulatory timing in a complex way.

Chapter 6 presents experimental evidence for the field theory—the idea that the systems responsible for articulatory planning are distributed in a multidimensional space, and interact in the course of planning and producing speech. A primed vowel shadowing task was conducted, in which the effects of suballophonic and cross-phonemic priming were tested. The results indicate that subtle within-vowel category acoustic information is perceived and rapidly integrated into motor planning. This sort of effect can only be understood if the cognitive representation of a vowel target includes more than just the phonemic category of the vowel. The results also showed that between-vowel priming was dissimilatory. This effect is suggestive of analogous effects observed in nonspeech motor control, which are hypothesized to arise from inhibition between competing movement targets.

Chapter 7 explores in greater detail how levels of the prosodic hierarchy can be reconceptualized as a multifrequency system of coupled oscillators, and shows how this system accounts for the experimental and corpus findings in chapters 4 and 5. Model simulations with rhythmic-gestural coupling demonstrate that as the strength of coupling between rhythmic systems is increased, the effect of noise-induced perturbations on the relative timing of gestural systems is decreased. Furthermore, this chapter explains how the wave theory can be integrated with a field theory to account for the experimental findings in chapter 6. The chapter discusses a number of outstanding issues in the wave and field theory, and presents a speculative extension of the idea to syntactic systems.

Chapter 2: Hierarchy and Connectionism

It is important to keep in mind that most theoretical constructs are likely to mean different things to different people. Because this dissertation explores the idea that some linguistic units in the prosodic hierarchy can be usefully conceptualized as interacting waves, we should be aware that "the prosodic hierarchy" is not an uncontested concept. To better understand it, this chapter begins by examining hierarchical structure more generally, and then goes on to consider how hierarchical structure is represented in linguistic theories, why hierarchies are observed, and what our representations of them can predict about temporal patterns in speech. Each section of this chapter considers one of several basic questions that underlie this dissertation:

- What is meant by the use of the word "hierarchical"?
- What sorts of hierarchical structure occurs in language, and how do we represent hierarchical structure?
- *Why* do we observe hierarchical structures in language?
- What sorts of predictions do our representations of hierarchical structure make, and what sorts of predictions *cant* they make?

2.1 The concept of HIERARCHY

To begin, we consider what is meant by "hierarchical." More precisely, we aim to disassemble the concept of HIERARCHY into more basic concepts, or to better understand the network of concepts that give rise to HIERARCHY. The Oxford English Dictionary defines a "hierarchy" as (in addition to a division of angels): "in *Natural Science* and *Logic*, a system or series of terms of successive rank (as *classes*, *orders*, *genera*, *species*, etc.), used in classification." This definition begs further questions: what are "terms of successive rank," what is the "system" or "series," and what is the "classification"? The problem with attempting to define the concept is that any definition presupposes that there is a catch-all notion of hierarchy, rather than a multitude of uses. Deeper analysis of the concept leads one to the conclusion that HIERARCHY involves a conglomeration of related concepts: RANK, RELATION, CONNECTION, and CONTAINMENT.

RANK is the simplest and most central aspect of a hierarchy. One way to understand RANK is as a blend of SIZE and ORDER, whereby SIZE is commonly metaphorically mapped to another domain, and ORDER evokes position in a space with some orientation. For example, feet belong to a higher level than syllables, and phrases are larger than heads. The use of words like "higher" and "larger" imply that notions of size and position are used to understand the relations between members of this hierarchy. For another example, Sternberg, Knoll, Monsell, and Wright (1988) cite the use of "hierarchical" to describe the partitioning of a movement sequence into disjoint subsequences, and furthermore characterize a notion of the "depth" of a hierarchy based on the number of partitionings. In both cases, the quantity of time associated with something (a linguistic unit or a movement sequence) is metaphorically mapped to a size and/or a distance in space, and according to this metaphoric size, the quantity of time is positioned in space relative to other quantities of time, such that larger size units are above smaller size units. Thus a TIME is QUANTITY/SPACE mapping underlies both of these examples of hierarchy.

Non-temporal hierarchies are common too. For example, cherubim are of higher order than seraphim, etc., because of a mapping between CELESTIAL INFLUENCE and SIZE. Non-temporal hierarchies are used in linguistics as well. The cross-linguistic person/number marking hierarchy is based upon a mapping between size and the likelihood of a grammatical person and number being morphologically expressed. This does not rely on a metaphor to quantify time. In contrast, the hierarchies that concern this dissertation are ones that *are* temporal, in that the rankings that construct these hierarchies are based on a metaphor that quantifies time. Importantly, there is no HIERARCHY without RANK: the concept of RANK is central to HIERARCHY.

HIERARCHY may or may not be used in association with concepts of CONTAINMENT, RELATION, and CONNECTION, but without these concepts, it is somewhat less interesting. For example, consider a hierarchy of bodies of water: puddles, ponds, and lakes. These are ranked by their actual size, but we do not say that lakes contain ponds; specific ponds are not related in any special way to other specific ponds, puddles, or lakes, and these are normally not connected to each other. In contrast, a hierarchy of bodies of moving water—brook, stream, river—may exhibit regularity in relational and connectional properties. Many brooks may be connected to a stream, many of which may in turn be connected to a river. These connections are interesting because they predict that creeks will interact with rivers in certain ways (for example, if a brook is polluted, the stream to which it is connected will be polluted as well. Important relations arise too. Streams are "related" to other streams feeding the same river; for example, a particular species of fish found in the river will swim upstream to related streams to breed, but may not be found in streams of a different river. Hence the streams are related by virtue of their shared relation to the river, in the same way that sister nodes are related because they share a mother node.

In contrast, it makes less sense to say that streams are "contained" in the rivers they connect to. CONTAINMENT is perhaps the most interesting and theoretically troublesome associate of the concept of HIERARCHY. Using another example, a primacy contains provinces, which contain dioceses, which contain archdeaconries, which contain deaneries, which contain parishes. These containment relations are both physical and functional, in that the smaller units are geographically "inside" the larger ones, and that power is distributed such that the person in charge of the larger unit has direct authority *only* over the contained units, and not over other units which, though lower on the hierarchy, are not contained.

In the context of HIERARCHY, CONTAINMENT usually entails RELATION and CONNECTION, but not vice versa. For example, there is an important division between theories of prosodic structure that utilize CONTAINMENT and those that do not. Basic metrical grid theories (e.g. Selkirk, 1984), in which there is a hierarchy of levels on which beats occur (i.e. beats associated with syllables, feet, word stress, phrase stress), do not imply that each syllable is or is not contained *within* a foot. In contrast, containment is an important implication of arboreal metrical theories, which unambiguously require that a syllable be either in a foot or not in a foot, and that each foot be either in a phonological word or not. Figure 4 illustrates this distinction: the grid-based representation in (b) does not impose any groupings, or containment relations, between elements; despite this, the grid retains a notion of rank, represented by the number of marks associated with each syllable. (a) makes the positive claim that the initial syllable of banana is *NOT* integrated into the metrical structure (i.e. it is extrametrical), because containment relations are necessarily implied by the pattern of connections. Furthermore, the arboreal structure in (a) posits a relation between the syllables *pu* and *dding*, whereas the grid only shows that these syllables happen to be adjacent. Note that containment relations in (a) are indirectly expressed

through connection patterns; although these connections are spatial in representation, they imply temporal containment.

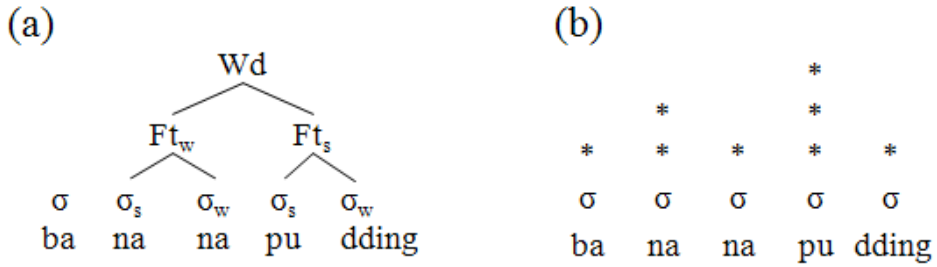


Figure 4. (a) Arboreal representation of the hierarchical metrical structure of the phrase banana pudding. (b) Metrical grid representation.

In sum, the above considerations highlight the futility of attempting to define "hierarchy" as if it were a coherent concept, rather than a contextually shaped nexus of related concepts. RANK is at the core, but RELATION, CONNECTION, and CONTAINMENT open up HIERARCHY to more diverse usage. There are theoretical consequences of whether a given system for representing hierarchy includes connection.

2.2 Representation of hierarchical structure

The geometric similarities in representational schemas used to describe patterns in syntax, morphology, and phonology are striking. Figure 5 provides some examples of such schemas. In syntax (a), relations between phrases and constituents within phrases are commonly represented with nodes and connections, where the nodes are "units" of some sort, and the connections are lines drawn between them. In morphology (b), several prominent representational approaches use more or less the same image-schematic constructs. Likewise in phonology, nodes and connections are widely used in metrical theory to represent relations between syllables within feet (cf. Figure 4), segments within syllables (c), and features within segments (d).

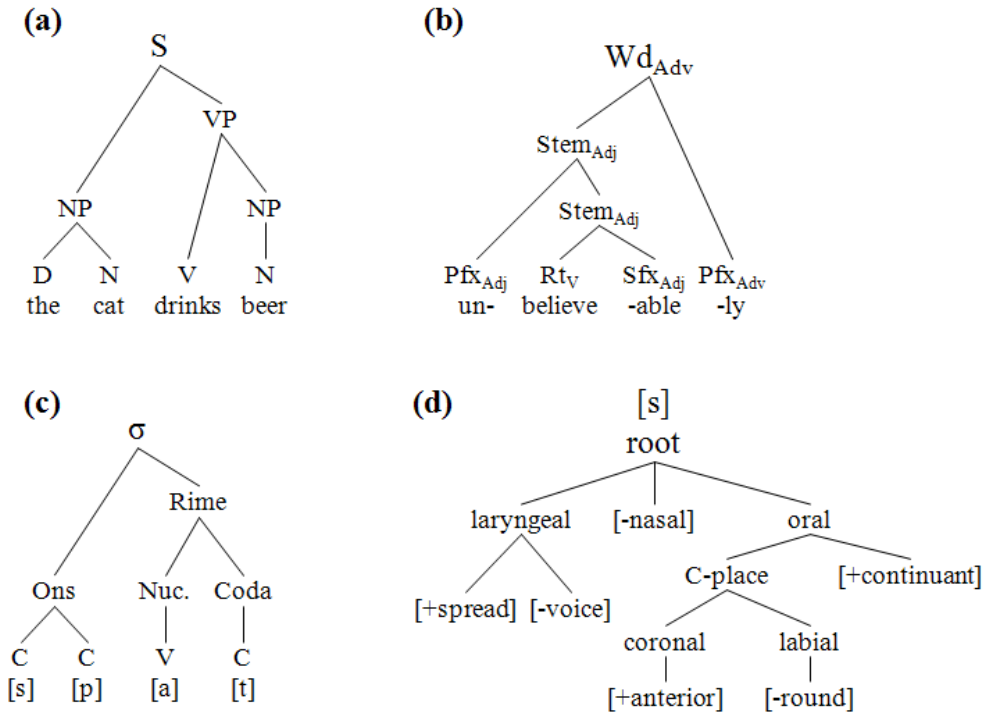


Figure 5. Some examples of hierarchical structures.

It is useful to have general terminology to describe these representational tools. To that end, we will use "connections" to refer to the lines that relate objects/units/features, and we will used "nodes" to refer to the objects/units/features that are related by connections. Taken together, *nodes* and *connections* form *schemas*, which are not only the structures constructed with nodes and connections, but also are conventional sets of assumptions about how nodes and connections can be related. For example, one convention that is generally observed in representation is *non-intersection*: connections do not intersect. Another common convention is *connective uniqueness*: for any two nodes, there at most one connection between them.

Generally speaking, nodes and connections, as schematic, metaphoric, representational devices, are many things to many people. To mathematicians, or more specifically to graph theorists, they are edges and vertices, and one can easily devote an entire career to studying their general properties within some set of arbitrary constraints. To biologists, they are branching structures, such as observed in trees, lung alveoli, some coral colonies, etc., and they arise from growth processes. To neuroscientists, they are neuron cell bodies and axonal and dendritic projections, which also arise from growth processes, and of particular interest is how they interact or "communicate" with each other. To computer scientists, they are connectionist networks. To complexity theorists, they are geometric objects that are statistically self-similar on multiple scales, perhaps fractal. For linguists, they are used to reason about linguistic structure in a variety of linguistic domains. That node and connection schemas are *used* to reason, are *used* conceptualize, and are *used*—as tools—to think about various domains, is a very important point.

An important augmentation of the node and connection schemas is the addition of *spreading activation*, which is a dynamical form of communication between nodes. This conceptual addition follows from viewing the connections as three-dimensional tubes and the nodes as containers. Spreading activation can thus be conceptualized metaphorically as the flow

of a liquid from one container to another, or more abstractly as a transfer of information between nodes. The nodes are containers, and some energy-like substance can move from one container to another. A virtue of this conceptual blend is that it allows for real-time dynamical modeling of the distribution of energy across a network.

An early example of a spreading activation network is the model applied to speech production described in Dell (1986). This model was designed to accomplish phonological encoding, "the processes by which the speech sounds that compose a morpheme or string of morphemes are retrieved, ordered, and organized for articulation," and to replicate certain speech-error patterns. Figure 6 shows an example of the model structure.

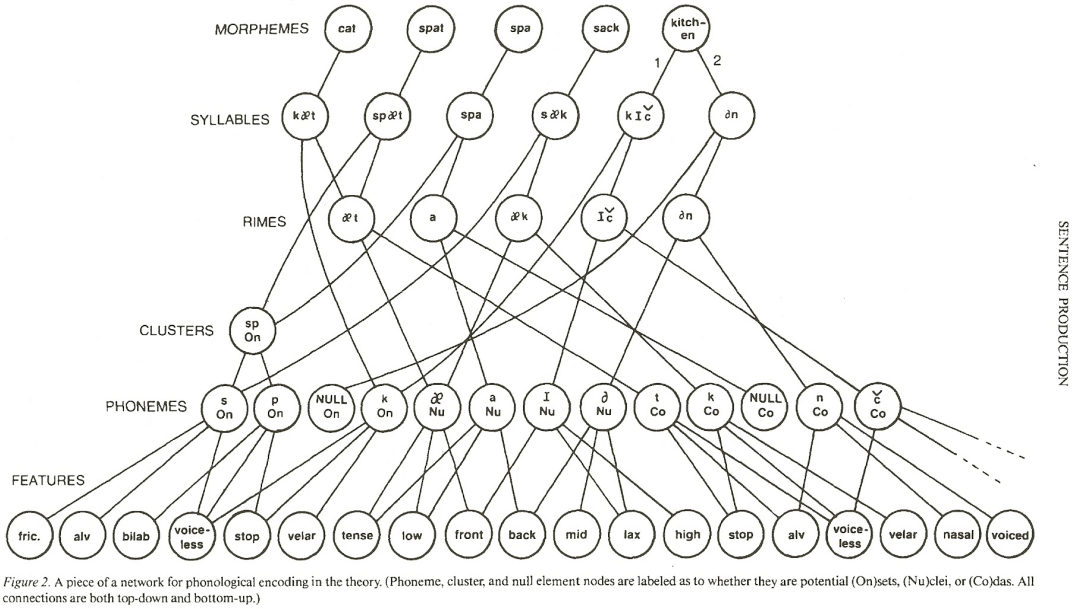


Figure 2. A piece of a network for phonological encoding in the theory. (Phoneme, cluster, and null element nodes are labeled as to whether they are potential (On)sets, (Nu)clei, or (Co)das. All connections are both top-down and bottom-up.)

Figure 6. Spreading activation network from Dell (1986). Figure used without permission.

Not surprisingly, levels of structure play an important role in the model, and the connections between levels are of paramount importance. Dell (1986) showed that, by parameterizing speech rate, the rate of decay of activation within nodes, and the rate of flow of activation between nodes, this system could model various speech errors. More specifically, Dell made several simplifying assumptions: 1) morphemes in the morphological representation of a planned utterance are activated fully in sequence and activated partly in anticipation, 2) all downward activation flow has a constant spreading rate: *p*, and 3) all upward activation flow is a percentage of *p*. Distinguishing between upward and downward flow rates in this way suggests that, conceptually speaking, each connection in the figure is actually two one-way connections between activation containers, through which activation/information flows in opposite directions. The relative transmission rates of these bidirectional connections are determined by the relative rank of the nodes, which is essentially a hierarchical property. This is a natural inference to make, but it is only readily imaginable in the context of a node and connection schema. Note that transmission of activation between units is not a natural outgrowth of a model in which features are listed in association with segments, as in Chomsky & Halle (1968), for example.

A subtle and somewhat awkward metaphoric mapping in the Dell (1986) model is SELECTION OF AN ACTION is *choosing the fullest container*. The time course of motor execution (i.e. productions of a syllable) is parameterized as *r*. Every *r* time steps, the syllable onset, nucleus, and coda with the highest activation levels are executed. This is another way of saying that every so often, the syllable subconstituents most full with activation are chosen for production. The model emphasizes the dynamics of activation flow between nodes, which determine the content of speech events, rather than the timing of those events. In other words, the model uses a network-dynamical speech planning system to predict which linguistic units will appear in an utterance (or a sequence of syllables), but not precisely when their articulatory realizations will occur.

This observation raises a crucial distinction: it is important to differentiate between the temporal dynamics of planning and temporal dynamics of execution. Traditional generative phonological models provide neither, although they do loosely reflect in an abstract, discretized way the timing of execution; they do not address the timecourse of planning processes whatsoever. Note that in the Dell (1986) model the order of the morphemes for a particular utterance, which influences the order of their associated syllables and related systems, must be predetermined by specifying that activation of some units precedes the activation of others. The arbitrary specification of sequence is a general problem to which we will return later.

A final point worth discussing in the context of connectionist representations of hierarchical structure is that node-connection schematization is more general than hierarchical structure. A network system in which all nodes are randomly connected has no semblance of a hierarchy. What makes nodes and connections hierarchical is the structure that we impose upon the network, which presumably reflects our observations. The Dell (1986) model illustrates how a node-connection system can be imbued with hierarchical structure. What exactly are the criteria for considering this network hierarchical? Notice that a spatial orientation has been associated with the network. Strictly speaking this is not necessary, as the nodes could be spatially distributed in a random fashion; importantly, however, the spatial orientation in this representation implies several forms of asymmetry in the network. First, there is an asymmetry in connectivity, whereby "higher" level nodes are generally connected to a greater number of "lower" level nodes than relatively lower nodes are (when taking into account indirect connections, but restricting the count to downward paths). Second, there is a intraconnectional asymmetry built into the model, whereby upward flow of activation differs from downward flow of activation. Thirdly, the spatial arrangement of the representation to some extent corresponds to the temporal size of the units, with syllables above rimes and onset clusters, both of which are above phonemes.

2.3 The origin of hierarchical structure?

The use of the same connectionist representational devices across domains begs for an explanation. It speaks to the powerfulness and utility of node and connection schemas. It does not matter so much that there are domain-specific differences in the "rules" and/or "constraints" that govern how these nodes can be connected, what exactly the nodes refer to, what they contain, how their contents change over time, and how they can be organized and reorganized. What is striking is that metaphors of connection, relation, and containment, which structure the systems under investigation, are used in a variety of domains—there is no logical necessity for this, no a priori reason to suppose that patterns in different linguistic domains are structured in this way. It is thus important to question *why* this is so. There are several perspectives on the answer:

Representational utility: node and connection schemas are useful representational tools.
Observer imposition: scientists impose structure on an unstructured world.
Cognitive reality: node and connection schemas reflect cognitive processes/mechanisms

Representational utility. The answer could be that hierarchical structures are representational *tools*. It just so happens that they help us understand various linguistic patterns. They do not necessarily indicate any sort of cognitive organization, nor do they directly represent brain processes. They are, perhaps, the sorts of things that could arise indirectly from cognitive processes.

Observer imposition. A skeptical answer is that hierarchical structure is *not* observed across linguistic domains; instead we impose hierarchical structure upon our observations. We do this, presumably, because we see nodes and connections, or branching structures, in nature. In the most linguistically-relevant domain of the physical world, i.e. neurons in the brain, this is certainly true, and our experience with such structures extends to a wide variety of domains. In other words, our node and connection representational schemas, with which we posit hierarchical structures, are a form of hallucination. We see structure that is not there. Of course, we can also knowingly choose to use such structures, because they help us understand patterns. We can do this in spite of an awareness that the structures are not cognitively real. Yet this view is overly simplistic. If hierarchical structure were an entirely arbitrary dissection of a kaleidoscopic flux of impressions, then the past several decades would constitute one of the largest scale mass hallucinations in a scientific community in history. It is preferable to think that something is *sort of* correct about connectionist representational schemas, in the way that it is *sort of* right to say that the sun goes around the Earth. The geocentric model is perfectly adequate to explain a circumscribed set of phenomena, and moreover is intuitively obvious and easy to comprehend.

Cognitive reality. A more constructive answer is that hierarchical structure is observed in speech because similar cognitive mechanisms give rise to similar patterns in different linguistic domains. This answer is constructive, perhaps, not by virtue of being correct, but by its provocation of experimentation to test the idea. This dissertation tests one very specific prediction of a theory that posits a unifying explanation for the emergence of hierarchical structure. The specific prediction is that the temporal patterning of articulation and the temporal

patterning of higher-level prosodic units can interact. The next section explains why node and connection schemas do not readily lend to making predictions about temporal patterns.

An alternative perspective is that the question of why we observe hierarchical structure presupposes a false ontology—namely, that we are "observing" "something" in "the external world". Allowing that we have constructed linguistic phenomena rather than described something may or may not lead one to reject the validity of the approach that has been taken, as well as the scientific method more generally. In that case, a legitimate approach to studying language and cognition would be to study the theories of linguistics as narratives written by linguistic theorists. Ultimately, we should remain skeptical of all representational systems—we should be in a state of perpetual skepticism.

2.4 Predictions of hierarchical representation

This last section of the chapter makes two points: (1) that node and connection-schemas prevalent in metrical/prosodic theories do not make predictions about interactions between speech rhythm and articulatory gestures; and (2) that many patterns understood to arise from operations or constraints on node and connection schemas may potentially be understood to arise from principles governing wave interactions. This dissertation will not explore in depth how principles of wave interaction may account for various linguistic patterns, but it is my assertion that the theory can do so parsimoniously, and a couple of examples are provided.

Metrical/prosodic theories generally use one of two image schemas: grids or trees. In the grid approach (Liberman 1975), each syllable constitutes a slot in a grid, and some number of prominence marks are assigned to each slot. In the tree approach (Liberman & Prince 1977; Halle & Vernaud 1978; Selkirk 1980), syllables are grouped into constituents, i.e. feet, which are labeled either strong or weak; feet can then be grouped into larger constituents (prosodic words or phrases), which are also labeled strong or weak. In both cases, the organization of segments within syllables is represented independently with some version of syllable subconstituency employing onsets, nuclei, codas, and perhaps rimes and moras.

Grid and tree approaches are mutually compatible. The primary differences between them are what they emphasize representationally. Trees emphasize the hierarchical nature of prominence, but make it more difficult to see the variation in prominence from one syllable to the next. In contrast, grids very directly depict prominence variation and rank, but eliminate the representation of connection and containment relations. Not only are grids and trees compatible representations, but grids can be derived from trees with mapping rules, given some arbitrary assumptions. However, Prince (1983) and Selkirk (1984) argued against the need for trees in metrical theory, on the basis that the parameters necessary for tree approaches can be captured with constraints on grid patterns. Several versions of hybrid tree-grid representations were later developed as well (cf. Kager, 1995).

Crucially, none of these approaches makes any predictions about temporal patterns arising from interaction between rhythmic units and gestures. Even more basically, metrical theories makes very few predictions about the dynamics of prosodic units such as syllables and feet, and have next to nothing to say about the temporal coordination of articulations. These are not necessarily indictments of these theories. It is clear, however, that to make testable

predictions about temporal patterning in speech, one needs to look beyond node and connection schemas, or their impoverished cousins, metrical grids.

Lets consider an example that is particularly relevant to one of the experiments reported in this dissertation. Node and connection representations predict (or through the absence of the ability to predict, they imply) that an [sp] onset cluster should be articulated the same way regardless of the metrical context in which it is produced. One reason for this is that the representation discretizes time severely. This discretization discourages the endeavor of using the representation to make quantitative predictions about observables such as syllable durations or relative timing of syllable onsets, as well as predictions about segmental timing. It similarly makes it difficult to understand how other factors, such as changes in speech rate, prosodic phrasing, and segmental content might affect these temporal patterns. It is not impossible to generate temporal predictions from tree and grid structures, but the absence of a working model that can do this speaks to an important point: making temporal predictions is not one of the goals of node and connection representations.

A second issue of concern with metrical theories is that the lowest level units in the prosodic hierarchy—segments—are generally connected only to the syllable subconstituents (or in some approaches to syllables themselves). This follows from strict layering (Selkirk 1984), although some less traditional models allow for direct connections to feet, or no connections whatsoever. Even if these connections are viewed as dynamical conduits of activation (and they are generally not viewed as such), this predicts that the higher level units can have only indirect effects on the timing of articulations associated with segments. The indirectness of these effects follows from the absence of a direct connections from feet to syllables, and this absence follows from the desire for representational parsimony and strict layering (or perhaps, exhaustive containment) of segments within syllables. Ultimately, the representations used in metrical theories do not lend themselves to testable predictions about the effects of metrical structure on the 10-100 ms timescale relevant to articulation, because they do not allow for simultaneous dynamical interaction between segmental systems and syllable/foot/higher-level metrical systems.

A third point to make about the scope of node and connection representations is that there is a large amount of rhythmic variation in speech that they cannot capture. Two utterances of the same phrase, having identical metrical structure, can sound fairly similar but differ greatly in the timing of rhythmic and articulatory units. This suggests that metrical representations are biased to represent highly abstracted rhythmic percepts; they do not aim to address important questions about how temporal variation arises in the course of production, and how it is filtered out in the course of perception.

Lastly, we should note that the development of node and connection schematic representation of metrical structure has been influenced to some extent by aesthetic considerations, which favor exhaustive parsing and strict layering. It is always possible for people to disagree on what constitutes an aesthetic representation.

It is noteworthy that one of the motivations for node and connection-based representations is to allow processes involving nonadjacent elements to be formalized as local operations. For example, rules or abstract constraints can be used to account for phenomena such as rhythm rules (Nespor & Vogel, 1989; Hayes, 1995). Such patterns are commonly understood to arise from the avoidance of clash (adjacent prominences) or a preference for alternation between more and less prominent syllables. In English, a rhythm rule is observed in primary

stress ~ secondary stress alternations such as *àntîque àrmóir* vs. *ántîque sófa* and *háll fòurtéen* vs. *fóurtèen hálls*. This alternation does not occur obligatorily, nor does it occur in all phrases with metrical structures which would allow for it.

Rule and constraint-based formulations of such patterns, which utilize node and connection schemas, do not give us insight into the deeper cognitive mechanisms from which the patterns arise, nor do they address the variability in such patterns. *Why* is "clash" something that speakers avoid (or, *why* is euphony something that speakers prefer)? How is avoidance accomplished? Why are some clashes avoided, while others are not? Why do some languages avoid some clashes, while others do not? Should speech rate condition variability in the rhythm rule? Is the rhythm rule more or less likely to apply in spontaneous speech compared to rehearsed speech?

It seems plausible that the avoidance of stress clashes arises from the way that successive feet systems are coordinated: their associated planning waves may be coupled in an anti-phase manner, so as to minimize their overlap. The forces responsible for this can in effect alter the coupling between feet and syllables, resulting in a new pattern of rhythmic coordination. This view allows that there can be other factors—prosodic, segmental, lexical, etc., that affect the dynamics of stressed syllable activation waves. The wave interaction model implicates a cognitive mechanism: neural ensemble coupling dynamics; in contrast, the constraints or rules used in metrical theories are at best descriptions of the effects of these cognitive mechanisms. Furthermore, the wave theory allows for the change in metrical wave dynamics induced by interaction with other waves to be a function of the rate of speech and the extent to which the speech has been rehearsed. These possibilities do not follow in any straightforward way from node and connection representations.

Furthermore, a tantalizing possibility is that many of the parameters of foot construction or constraints upon foot construction can be understood dynamically. Examples of such parameters are: *boundedness* (a size restriction on feet to a maximum of two syllables), *foot dominance* (the sequential location of the stressed syllable in foot), and *quantity sensitivity* (the relation between light and heavy syllables to foot structure). Other metrical parameters, such as *directionality* of foot construction (earlier to later or vice versa), and *iterativity* of foot construction, may be unnecessary because they are an artifact of the way that node and connection representations structure temporal relations between units. To understand how all of these parameters can be associated with wave interactions is a major undertaking. If it can be done, then the result will be a theory that unifies our understanding of variation in metrical patterns across languages with our understanding of the rhythmic-articulatory interactions shown in the following chapters.

Thus, while the node and connection representation of metrical/prosodic structure has been extremely successful in helping us identify linguistic patterns and contrast them between languages and utterances, it seems that in some respects—particularly with respect to understanding phonetic variation, the raw temporal patterning of speech—nodes and connections are approaching the limits of their utility. In the following chapter we consider the background for a wave theory that can more readily account for temporal patterns.

Chapter 3: Dynamical models of articulation and prosody

This chapter reviews some of the experimental and theoretical work which has successfully employed concepts from dynamical systems theory to understand phonological and prosodic patterns. Section 3.1 presents some of the basic concepts of coordination dynamics, such as relative phase coupling, generalized relative phase coupling, and potential functions. Section 3.2 examines in detail the task dynamic model of articulatory gestures. Section 3.3 shows how the task dynamic model of gestures was extended to understand the relative timing of gestures. Section 3.4 describes a coupled-oscillators model of cross-linguistic variation in the durations of feet. Section 3.5 describes experimental and theoretical work on the relative timing of feet within phrases. Section 3.6 proposes that the dynamical approaches to modeling articulatory gestures, syllables, feet, and phrases can be integrated into a single model that makes testable predictions about how prosodic systems may interact with gestural systems. Section 3.7 concludes the chapter by presenting a brief argument for why the wave theory is cognitively plausible.

This dissertation explores the idea that linguistic units can be usefully conceptualized as systems that oscillate like waves. One advantage of this approach is that there is an extensively developed mathematical framework for understanding how waves interact. From this perspective, linguistic patterns (synchronic, diachronic, typological) emerge more or less directly from principles that govern interactions between waves. These principles are derived (1) from the forms of the equations which describe the waves, (2) in constraints upon the parameters for those equations, and (3) in the way various symmetries are broken by those constraints.

The use of the word "waves" here is by itself almost meaningless, and can only be given meaning through specific instantiations of the idea. The "wave" is metaphoric here, in the sense that we can use our observations of how physical waves behave to understand the mechanisms responsible for hierarchical linguistic patterns. More generally, our knowledge of waves and their logically possible interactions, physical or mathematical, can be used to reason about the origins of linguistic patterns.

The reader should not take the use of the word "wave" too literally. Our source domain for constructing an alternative conceptualization of hierarchical structure revolves around an "ideal wave," which is more general than any physical wave could be, and which interacts with other waves in a more general manner than any physical waves could interact. The ideal wave is imaginary, but may also have measurable physical correlates (later in this chapter the neurophysiological bases for the wave approach are discussed). Strictly speaking, the objects of investigation here are the systems that give rise to wave-like patterns. These systems can be modeled as "oscillators" in dynamical systems theory, and they give rise to the waves/oscillations.

Most of the theory and modeling that follows in this dissertation—both my own and that of others—is guided by dynamical systems theory. Dynamical systems theory is not really a "theory" per se, nor or even a "theory of" something, but rather, a set of geometrical and mathematical concepts—"thinking tools"—that help scientists understand how diverse phenomena are related, and help them describe how observables change in time. Dynamical systems theory can be considered an area of applied mathematics, in which the behaviors of complex systems are described with differential equations. Dynamical systems help us visualize how patterns emerge from the behavior of complex systems, and how changes in the parameters

of systems influence observable patterns. The equations are not observable, but the patterns are—the equations are a useful way of understanding how and why patterns emerge.

I will often use the word "dynamics" in two different senses: first, to refer to the motions of, or trajectories of, or patterns of change in some variable, and second, to refer to the laws or forces governing the changes of those variables. Accordingly, I will sometimes describe a pattern of change as a "dynamic," or a set of laws/forces dictating the change of a variable as a "dynamic".

3.1 Coordination dynamics

A major innovation in the past several decades has been the application of concepts from theories of self-organization in physical systems to the understanding of various cognitive and behavioral patterns involving the coordinated movement of multiple effectors. Kelso (1981, 1984), to better conceptualize rhythmic interlimb coordination, employed the concept of an *order parameter* or *collective variable* that had been developed by physicist Hermann Haken. An order parameter is a low-dimensional variable that describes the collective behavior of a self-organized system composed of many individual systems driven far from thermal equilibrium (Haken, 1975, 1983; cf. Kelso (1995) for a historical introduction to dynamical modeling of pattern formation processes and movement behavior). An oft-cited example is the Rayleigh-Benard experiment, in which oil heated in a pan forms convection rolls. The individual molecules in the system are enslaved to an orderly, coherent pattern of motion. Rather than needing to describe the motions of all these individual molecules, a collective variable can be used to describe the rolling motion, thereby using fewer degrees of freedom in the analysis of the system.

Another important concept is that of a synergy, or coordinative structure (Bernstein, 1967), which describes the cooperation of multiple effectors to accomplish a task goal. Perturbation of one effector in a synergy is compensated for by others. For example, Abbs & Gracco (1984) found that when unexpected loads were applied to the lower lip during a speech gesture involving both lower and upper lip movement, compensatory muscle responses were observed in the upper lip in under 75 ms. Many other studies employing similar paradigms have found evidence for various synergies in speech and non-speech movements.

A central concept used in modeling the rhythmic coordination of effectors is that of *relative phase*. In general, relative phase is a collective variable that corresponds to the phase difference between oscillatory systems, each of which has its own well defined individual phase that can typically be formulated as an angle in polar coordinates. In normal walking, the relative phase of the legs is approximately 180 degrees (or π radians, or 0.5 unit phase), which is a mode known as anti-phase coordination. In hopping, where both feet leave and return to the ground at the same time, the relative phase is 0, which is called in-phase coordination. If systems are coupled so as to mutually influence each other, and exhibit some degree of stability in relative phase, the systems can be referred to as *synchronized*. We will avoid the use of the word "synchronization," because it is commonly used with varying degrees of generality.

One of the major insights in studies of coordinated movement has been that some relative phase relations in coupled systems are more stable than others. In a classic paper, Haken, Kelso, & Bunz (1985) reported that as movement frequency (a *control parameter*) is increased, anti-phase finger wagging movements become unstable; the relative phase of the system then undergoes a transition to the more stable in-phase mode of finger wagging. Figure 7 (a)

illustrates what a typical pattern of motion in this experiment looks like, and (b) shows how relative phase (φ) undergoes a transition as movement frequency increases.

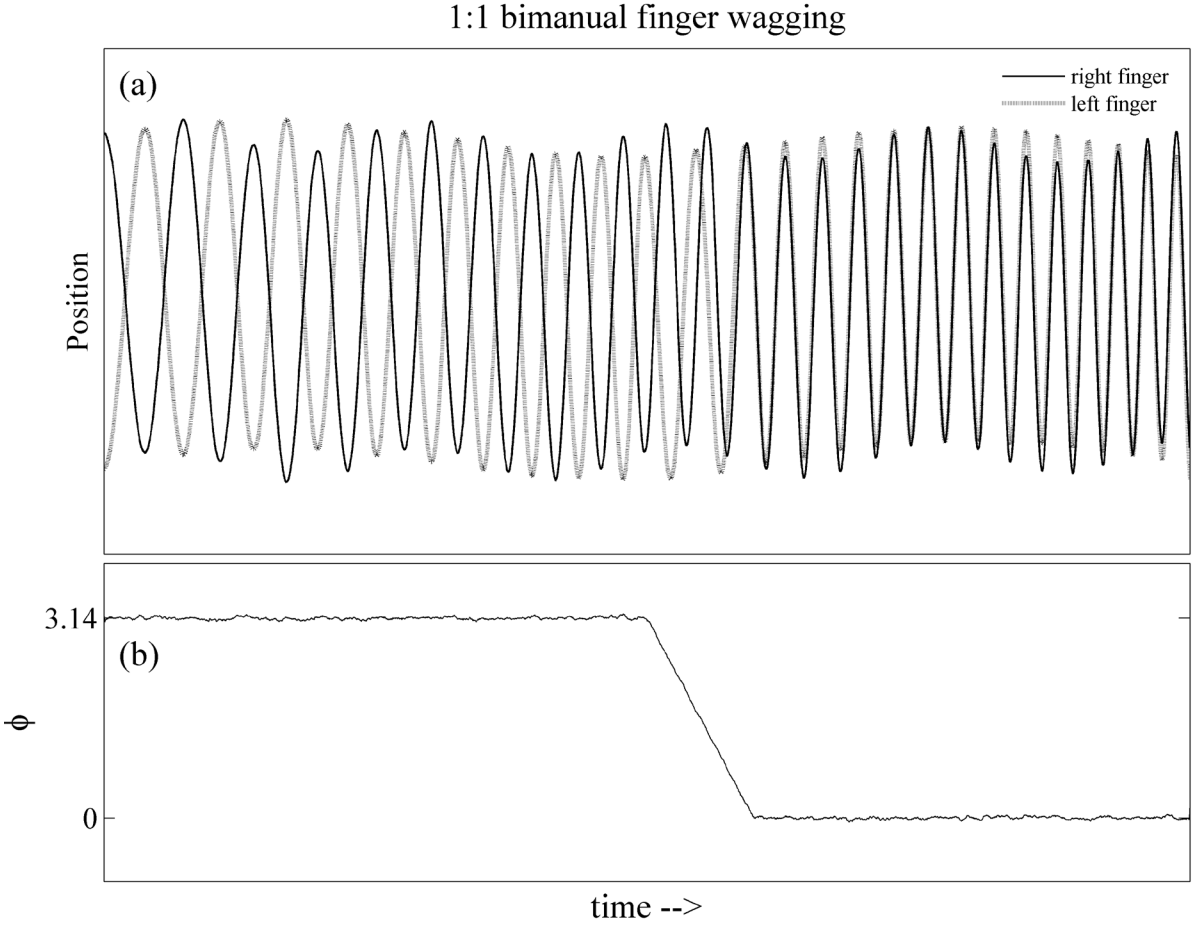


Figure 7. Schematized phase transition from anti-phase to in-phase coordination in a bimanual finger wagging task. (a) displacements of the right and left index fingers from the midline. (b) relative phase of the fingers.

To represent the differential stability of coordinative modes of relative phase, the concept of a *potential* has also been borrowed from physics. In this context, the potential is a function that describes the force acting upon a collective variable, and its minima and maxima are the stable and unstable equilibria of the system. Haken, Kelso, & Bunz modeled the stability of relative phase in finger wagging using a potential function that is the superposition of two cosine waves that have a 1:2 frequency ratio [i.e. $V(\varphi) = -a \cos(\varphi) - b \cos(2\varphi)$]. They hypothesized that the control parameter of movement frequency corresponds to the ratio of amplitudes of the components of the potential function (b/a), so that as frequency increases the ratio changes and the stable equilibrium corresponding to anti-phase coordination becomes unstable. Figure 8 shows potential functions corresponding to several values of this ratio. The reader should observe how the anti-phase mode becomes unstable at a critical ratio of 0.25. The potential can be thought of as a force field that directs the change of relative phase, such that the relative phase velocity is the negative of the derivative of the potential with respect to relative phase [$d\varphi/dt = -dV/dt$]. This means that the force advances or delays the phases of component oscillators so that

the relative phase moves toward the nearest local minimum in the potential function. An oft-used analogy to the relative phase potential is the motion of a ball (relative phase) rolling down a sticky hill (potential), which stops when it reaches the bottom (a local minimum).

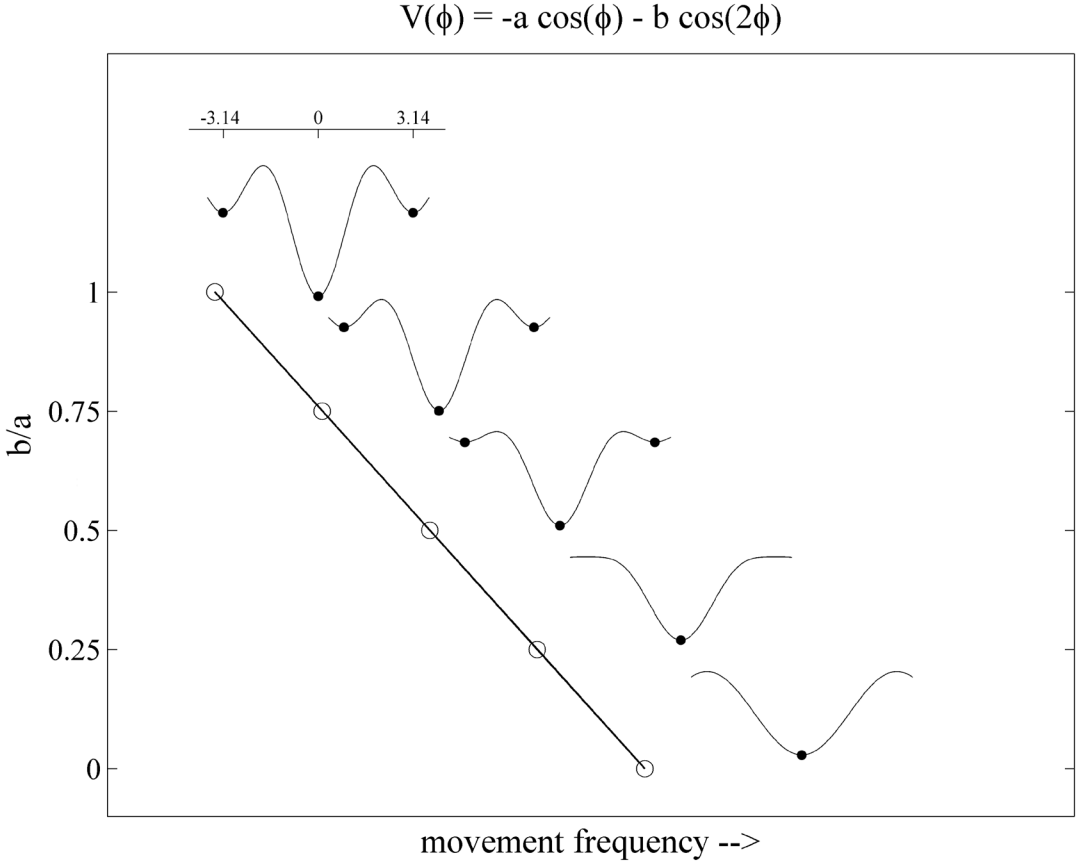


Figure 8. Changes in the relative phase potential as movement frequency increases. Potential functions for five values of the amplitude ratio b/a are shown next to their corresponding ratio values.

Another conceptual advance in this framework has been the notion of *generalized relative phase*, which allows for a stable relative phase to be defined between systems of differing frequencies (cf. Keith & Rand, 1984; Sternad, Turvey, & Saltzman, 1999; Saltzman & Byrd, 2000; Pikovsky et al., 2001). To see why this is necessary, consider Figure 9 (a), which shows the motions of two hypothetical effectors in a 1:2 frequency-locked system. If no phase transformation is performed, then the relative phase is periodic, as shown in Figure 9 (b). It is desirable for theoretic purposes to identify an order parameter which is constant, and which can therefore attain an energy minimum. For two frequency-locked oscillatory systems with individual phases θ_1 and θ_2, whose inherent frequencies ω_1 and ω_2 approximate an *m:n* ratio, the generalized relative phase $\varphi_G = \omega_1\theta_2 - \omega_2\theta_1$ is constant over time, as shown in Figure 9 (c). For example, in the case of 1:2 frequency locking, the phase of the slower oscillator is multiplied by the frequency of the faster oscillator prior to taking the phase difference.

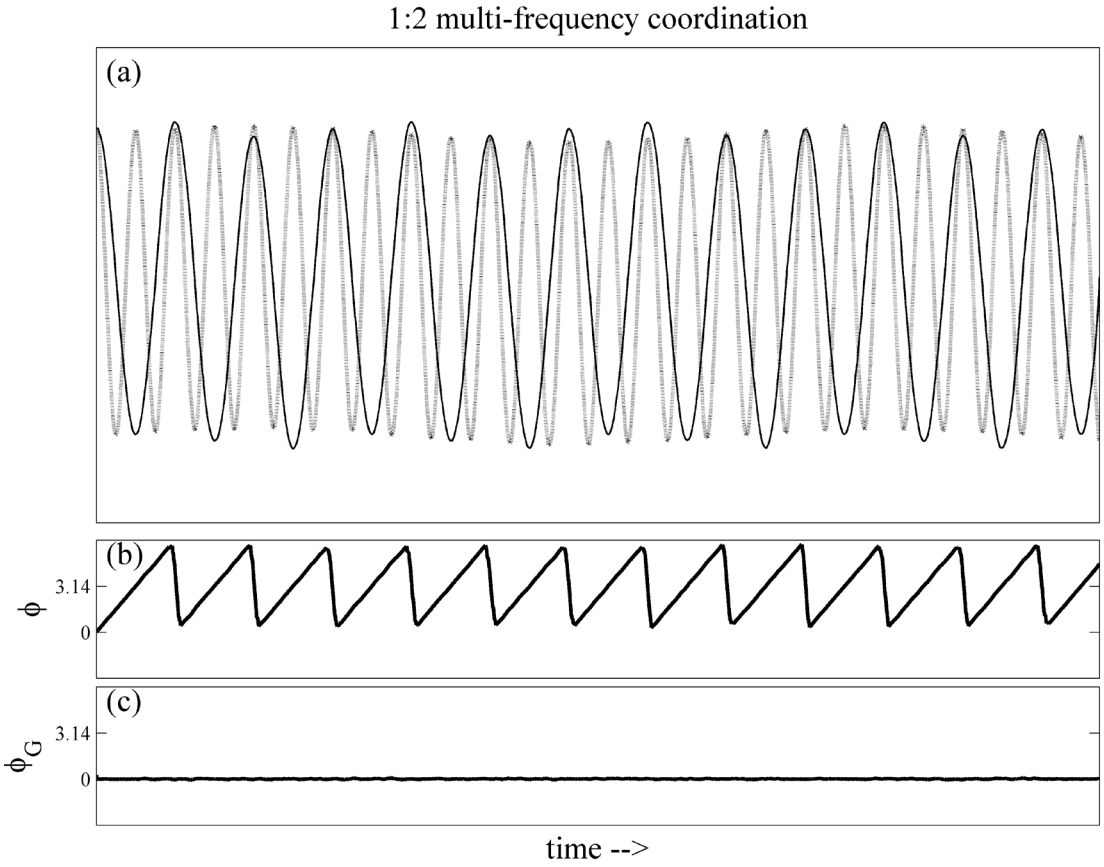


Figure 9. An example of 1:2 frequency-locking. (a) movement amplitudes of hypothetical effectors. (b) relative phase, which is periodic. (c) generalized relative phase, which is constant.

With generalized relative phase, stability of various modes of frequency-locking can be described with potential functions of generalized relative phase variables (Saltzman & Byrd, 2000). A more general approach, in which the potential function is a superposition of different modes of frequency locking with differing amplitudes, is described by Haken, Peper, Beek, & Daffertshofer (1996), who used this idea to model phase transitions between frequency-locked modes of drumming performed by skilled musicians. They observed that relatively high-order modes of frequency locking such as 2:5 and 3:5 became unstable, resulting in phase transitions to nearby lower-order ratios such as 2:3 and 1:2. Their account posits that the amplitudes of higher-order frequency-locking terms in the potential function decrease as movement frequency increases.

Experimental and theoretical work in the dynamical systems approach has further investigated these and other ideas, particularly in the domain of rhythmic interlimb coordination. Stochastic influences on coordination were introduced by Schöner, Haken, & Kelso (1986). Handedness asymmetry has been shown to cause a slight phase-lead of the dominant limb in bimanual finger wagging (de Poel, Peper, & Beek, 2007), as well as asymmetry in coupling strength (Treffner & Turvey, 1995). The Haken-Kelso-Bunz (1985) model predicted that coupling strength should increase with movement amplitude. This has been confirmed through perturbations of rhythmic movements with varying amplitudes (Peper, Boer, de Poel, & Beek, 2008), and effects of movement amplitude have been proposed to arise from neurophysiological

time delays (Peper & Beek, 1998). Additionally, researchers have observed stability of anti-phase coupling and low-order frequency locking in studying rhythmic coordination between nonhomologous limbs (Carson, Goodman, Kelso, & Elliot, 1995) and between people (Schmidt, Carello, & Turvey, 1990).

Speech coordination is related to such phenomena because it involves interacting systems associated with a hierarchy of timescales. What exactly are these timescales? Lets begin by considering only articulatory and prosodic domains. The shortest timescale encompasses speech gestures, and informative work has been done in the task dynamics framework, in which target relative phases of gestures are derived from coupled systems (Browman & Goldstein, 1990; Saltzman, 1986; Saltzman & Munhall, 1989; Saltzman & Byrd, 2000; Nam & Saltzman, 2003). Dynamical treatments of rhythmic units, such as moras, syllables, feet, and phrases, have also been proposed by several researchers (Port, 1986; Port, Cummins, & Gasser, 1995; Cummins & Port, 1998; O'Dell & Nieminen, 1999; Barbosa, 2002). The following sections will review these dynamical models.

3.2 Gestural dynamics

The most well developed and influential dynamical framework for modeling the spatial and temporal variation in articulation is *task dynamics* (Saltzman & Kelso, 1983; Saltzman, 1986). The development of task dynamics was motivated in part by the desire to understand contextual variability in articulatory kinematics, i.e. how the position, velocity, and acceleration of skilled movements (in speech, "gestures") change over time and across contexts. Loosely speaking, in the way that articulatory or phonetic features are the indivisible units/objects of speech in feature-geometric representations, gestures are the irreducible, "atomic" units in task dynamics. However, gestures are not objects, they are much more than containers, and perhaps they are not even the type of thing that should be conceptualized as reducible or irreducible. Instead, gestures should be thought of as abstract movements resulting from idealized forces. Some examples include decreasing labial aperture, adducting the vocal folds for vibration, lowering the velum for nasality, etc., and also combinations (synergies) of such movements. What does it mean to say that a gesture IS a motion? In a way, it means that a gesture is a system of equations.

On a metaphoric level, task-dynamics conceptualizes speech movements by analogy to mechanical principles, using the same equation Sir Isaac Newton developed to describe motion, i.e. Eq. 3.1. This equation states a rule for the change in time of the velocity of a particle of mass m under the action of forces F_0. In familiar terms, *force equals mass times acceleration* (acceleration is the change over time of velocity, which is the change over time of position). In many cases one can distinguish between several forces acting upon the particle. For example, the motion might be well described with a friction or damping force proportional to the velocity and a driving force (F), as in Eq. 3.2. Sometimes the driving force F is a function of the position of the particle (x), as in Eq. 3.3. For example, in a linear mass-spring system, which is a type of "harmonic oscillator," the driving force $F(x)$ is proportional to the elongation/compression of the spring (x). The constant of proportionality is called Hooke's constant (k), this parameter represents the elasticity of the spring. The minus sign in Eq. 3.3 represents the observation that an elongated spring exerts a compressing force (here elongation is defined to be positive,

compression negative). Using the relations $v = x'$ and $dv/dt = x''$, and putting Eqs. 3.2 and 3.3 together, we have the equation of motion for the mass, Eq. 3.4.

$$m\,dv/dt = F_0 \quad (3.1)$$
$$m\,dv/dt = F - \gamma v \quad (3.2)$$
$$F(x) = -kx \quad (3.3)$$
$$mx'' + \gamma x' + kx = 0 \quad (3.4)$$

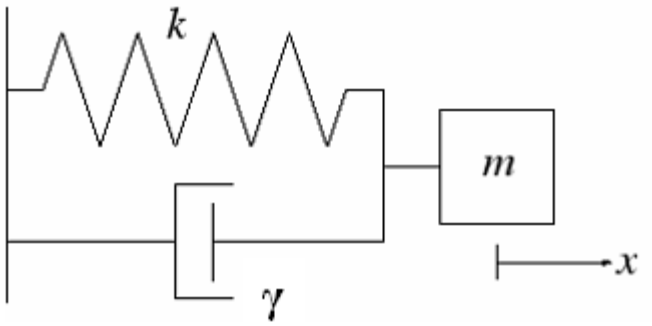


Figure 10. *Horizontal mass-spring system with damping*. Schematic of mass-spring system; k = spring stiffness, γ = damping, m = mass, x = position.

Saltzman & Kelso (1983a) and Saltzman (1986) used mass-spring dynamics to model the coordination of effectors in skilled movements, including speech. Here we will discuss an elaboration of this approach in Saltzman & Munhall (1989)—henceforth SM89. In the SM89 approach, individual gestures are dynamical systems in two coordinate systems: a context-independent *tract variable* space and a context-dependent *model articulator* space. The tract variable space parameterizes the real spatial dimensions of the vocal tract in such a way that these dimensions correspond to gestural "intents," which are associated with *goals*, *targets*, or *tasks*. The mapping from tract-variable space to goal space is assumed to be simple enough in speech to neglect the distinction; in contrast, for reaching movements, Saltzman & Kelso (1983) make use of a three-way distinction between goal space, body space (analogous to tract variable space), and effector space (analogous to articulator space).

Tract variable coordinate space can be transformed into kinematic model articulator coordinate space, but only by assuming an arbitrary weighting of various articulators; this weighting is necessary because of the one-to-many mappings between tract variables and model articulator variables. Later we will see how this mapping allows for modeling the phenomenon of compensatory articulation (or more generally, motor equivalence). The tract variable dynamical systems for each gesture are uncoupled (independent of each other), and are nearly identical in form to the equations of motion for the damped mass-spring system described above. To understand the model, one must understand the meanings of the parameters, i.e. the metaphoric mappings between mass-spring systems and gestures. We now examine these mappings in more detail.

In SM89 the tract-variable equations are presented in matrix form, meaning that the equations describe the dynamics of all tract-variables. Since the tract variables are uncoupled, however, we can consider just one to illustrate some key points. Take for example the classic syllable [ba]. The task-dynamic model associates the labial gesture for the consonant [b] with several tract variables, such as lip protrusion (LP), lip aperture (LA), and glottal opening (GLO).

The dynamics of each variable are defined by a differential equation. For example, lip aperture (LA = z), is defined as:

$$m z'' = -b z' - k \Delta z, \quad \Delta z = z - z_0 \tag{3.5}$$
$$m z'' = - b z' - k z + k z_0 \tag{3.6}$$
$$k z = k z0 \tag{3.7}$$

Substituting for Δz in Eq. 3.5 gives Eq. 3.6, which shows more clearly how z_0—the tract-variable target—together with the stiffness parameter k, defines a point attractor for LA. Observe the similarity between Eq. 3.6 and Eq. 3.2, the only difference being that in Eq. 3.6 the motion of the lip aperture variable is determined by three forces. The first term on the r.h.s. of Eq. 3.6 represents a damping force, and the second term represents an elastic (stiffness) force, analogous to the elastic force of a spring. Both the damping and elastic forces act in opposition to the desired direction of motion (evident from their signs), and are proportional to the velocity and position of LA, respectively. The third term on the r.h.s., however, is positive and constant. This term can be thought of as a driving force, and can be visualized, in the mass-spring system, as the application of a constant force to the mass. The parameter z_0 is a gesture-specific target or goal. It defines a point attractor for the system, i.e. an equilibrium where the acceleration and velocity of the system is zero.

To visualize the effect of changing the target parameter z_0, examine the plots in Figure 11 below. In panel (a), the target z_0 is set to a value of 1 at t=1, and the position of z moves to that value. In panel (b), the target z_0 is set to a value of 2 at t=1, and again, the system moves to that value. Before t=1, the systems are in equilibrium, there are no driving forces acting upon the systems and their positions and velocities are zero. At t=1, when a constant driving force is suddenly applied, the systems are set in motion.

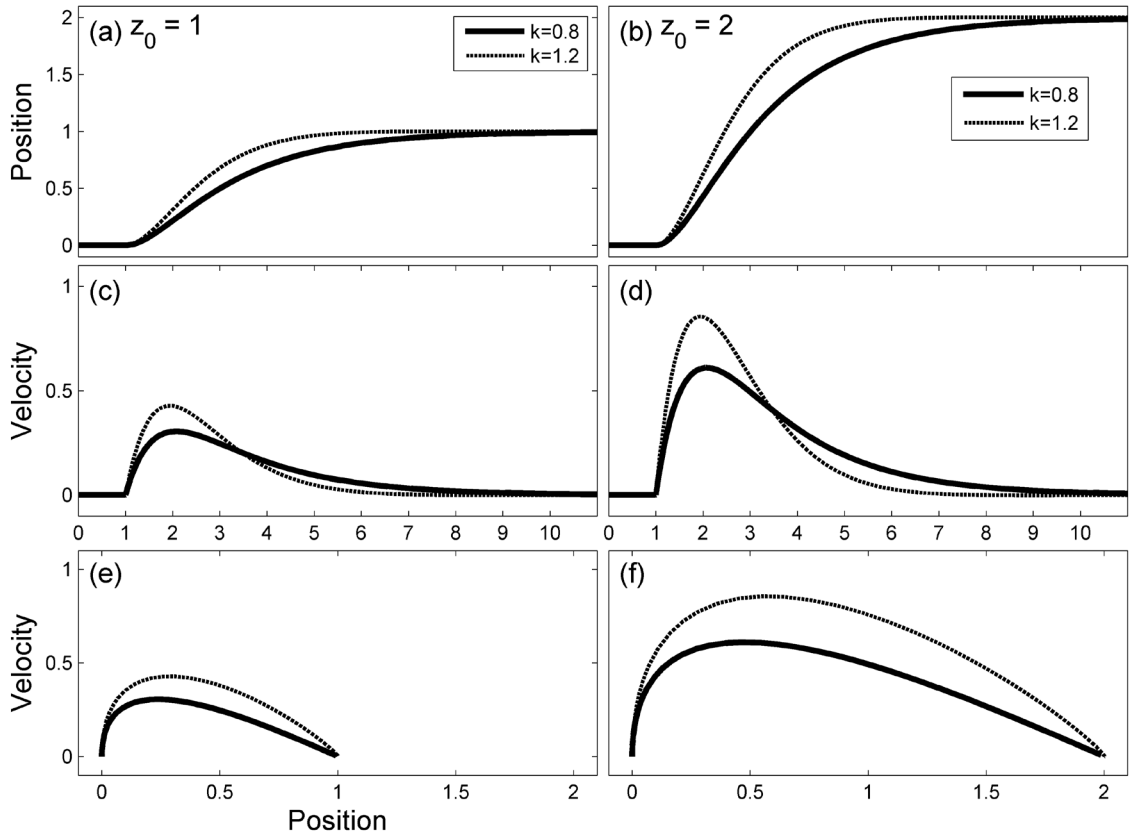


Figure 11. *Comparison of mass-spring systems with different stiffness and target parameters*. (a, b) position over time; (c, d) velocity over time; (e, f) position-velocity phase space. In both cases, the target z_0 is changed from 0 to z_0 at t=1. For a given target, the stiffer system reaches equilibrium more quickly. The equilibrium is a function of the target value only.

The target equilibrium (z_0) alone determines the new equilibrium position that is reached; the strength of the constant driving force upon the mass is $k\, z_0$, and equilibrium is achieved when this force is exactly balanced by the elastic force of the spring, $k\, z$. This is shown in Eq. **Error! Bookmark not defined.**3.7, which follows from Eq. 3.6 when the acceleration and velocity terms are zero, i.e. when the system is in equilibrium. One can also see this in the phase-space plots in Figure 11: the velocity of the mass returns to zero at the target/equilibrium position.

It is the application of a constant driving force or "activation" that Saltzman & Munhall associate with a gesture. Hence a gesture specifies a set of driving forces on a set of tract variables (although above we only considered one variable). The driving forces are distinct from the tract variables, yet the two belong to the same conceptual system—the latter represent current locations in tract variable space, while the former correspond to target locations. The operative metaphor here maps conceptual structure in the domain of mass-spring systems to conceptual structure in the domain of gestures. Table 3.1 lists a number of specific mappings in this metaphor that deserve further attention. It is important to keep in mind that these are metaphors, not identities; hence the degree of opening between the lips is *conceptualized as* the position of a mass on a spring—it is not believed in actuality to BE the position of a mass.

Table 3.1. *Mappings in the* SPEECH GESTURES are FORCES UPON A MASS-SPRING SYSTEM *metaphor*.

SPEECH GESTURE DOMAIN			MASS-SPRING SYSTEM DOMAIN
tract variable (e.g. lip aperture)	z	→	position of mass
first derivative of lip aperture	z'	→	velocity of mass
second derivative of lip aperture	z''	→	acceleration of mass
no opening between lips	0	→	an arbitrary position of mass
target lip aperture / neutral lip aperture	z_0	→	equilibrium position of mass
gestural activation	$k\ z_0$	→	driving force exerted on mass
"massless" (or, of unit mass)	$m=1$	→	inertial parameter
damping	b	→	damping
"speed" of gesture	k	→	stiffness

For the mappings to cohere, the tract variable space and mass-spring space are arbitrarily endowed with coordinates, i.e. an origin and directions. In the mass-spring system the origin of the 1-dimensional variable z is by convention defined to be the resting position of the (unforced) mass; positive and negative directions of motion are likewise arbitrarily defined. In the metaphoric blend, this amounts to mapping a neutral degree of lip aperture to the resting position of the mass, and increases of lip aperture to increases in the position of the mass (and vice versa for decreases).

The parameter m (mass) is normally set to 1, which reflects the assumption that tract variables are "massless". (Technically speaking, this actually means that all systems are all of equal, unit mass). The mass parameter is thus relatively unimportant. In contrast, the parameters b (damping), and k (stiffness) are of utmost importance. Different ranges of k are correspond to differences in the speed of a gesture, which is hypothesized to reflect a fundamental difference between consonantal and vocalic gestures. To see this, consider the mass-spring system in Eq. 3.6. If the spring stiffness constant is relatively large, then the gestural activation force $k\ z_0$ is also relatively large and the system will move toward equilibrium more quickly (Figure 11 a, b). Vice versa, a more elastic spring will experience a weaker driving force. So, the difference in speed of consonantal and vocalic gestures can be treated as a difference in the stiffness parameter.

This line of reasoning is an example of how inferences about model parameters are derived from the image-schematic structure of the metaphor behind the model. In this case, the image-schematic structure involves the motion of a mass-spring system; because the differential equations describing the forces acting upon such a system allow for parameterization of the strengths of these forces, it is intuitively natural to use the parameters of the conceptual model to account for observed phenomena in the target domain of speech gestures.

The damping parameter is no less important than the stiffness parameter. The SM89 tract-variable systems are critically damped, which means that the damping parameters are large enough to prevent any oscillatory behavior in the system. Figure 12 (a),(c), and (e) show the position, velocity, and phase space trajectory of an underdamped system, where $b = 0.5$. Figure 12 (b), (d), and (f) show an undamped system, where $b = 0$.

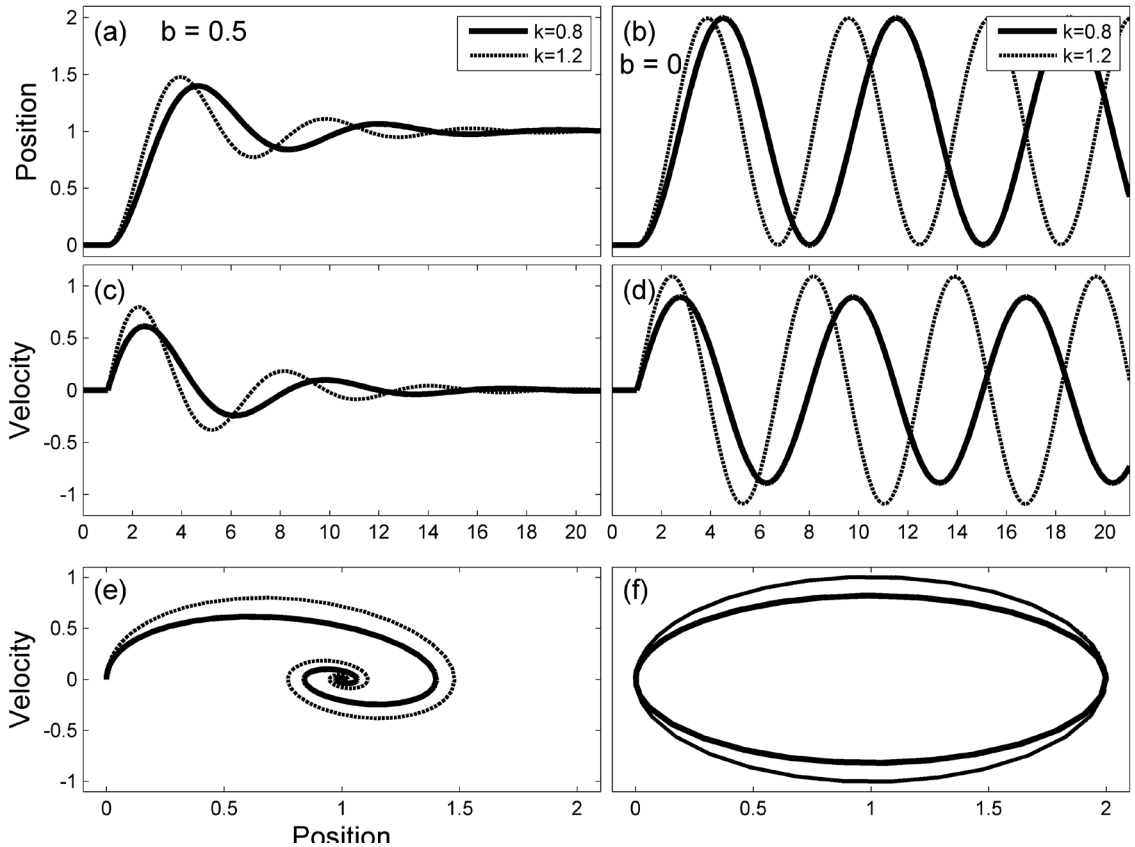


Figure 12. *Comparison of underdamped and undamped mass-spring systems*. In both cases, simulation time = 21 sec. and the target z_0 is changed from 0 to 1 at t=1.

Note that the underdamped system shown in Figure 12 (a) overshoots the target, and then oscillates around the target with progressively smaller amplitude. The phase space reveals a point attractor, but in this case the phase space trajectory of the system is spiral. In contrast, the attractor in phase space of the undamped system shown in Figure 12 (b) is periodic; this is called a *limit cycle*: the system exhibits periodic motion around the target value. Note that the frequency of this oscillatory system is related to the stiffness parameter.

So far we have only considered how SM89 models the first phase of a gesture, which constitutes the "turning on" of driving forces upon a set of tract variables. But gestures that are initiated are also terminated eventually—if the lips open, they should close when the gesture is over. To accomplish this, the driving force is "turned off".

One problem with this abrupt activation and deactivation is that the onset and offset trajectories are mirror images, since the same parameters describe the system. To avoid this, SM89 posited two separate sets of forces exerted on tract variables: an active gestural control force and neutral, default forces. The total forces acting upon a tract variable are the sum of gestural forces and neutral (or, resting, default) forces. Both gestural and neutral forces include damping, elasticity, and driving components. The neutral driving force is weighted so that the total normalized activation of gestural and neutral driving forces sums to 1. The advantage to separating neutral and gestural forces is that the velocity and duration of the onset of a gesture can be modeled independently of the velocity and duration of the offset. When the stiffness of

the neutral forces is lower than the gestural stiffness ($k_N < k$), the result is that, upon deactivation of the gesture, relaxation to equilibrium occurs more slowly.

For reference, a schematic of tract-variables from Saltzman & Munhall (1989) is shown in Figure 13. Since the tract variables are independent (i.e. uncoupled), their movements do not affect each other. Their motions are entirely determined by when and the extent to which driving forces are activated and deactivated by gestures and by their associated neutral attractors. What then controls the time course of activation and deactivation? We will next consider the SM89 approach below, and more recent approaches in subsequent sections.

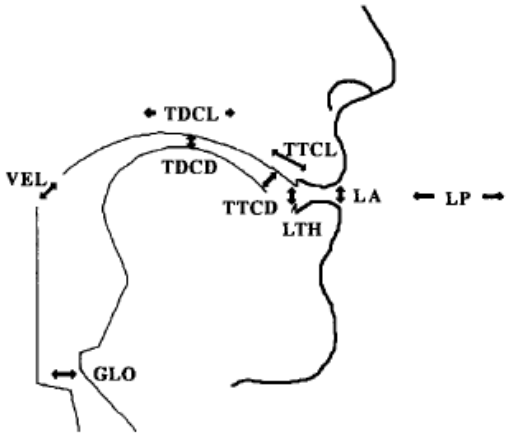


FIGURE 3 Schematic mid-sagittal vocal-tract outline, with tract-variable degrees of freedom indicated by arrows. (See text for definitions of tract-variable abbreviations used.)

Figure 13. Tract variables from Saltzman & Munhall (1989). *Figure used without permission.*

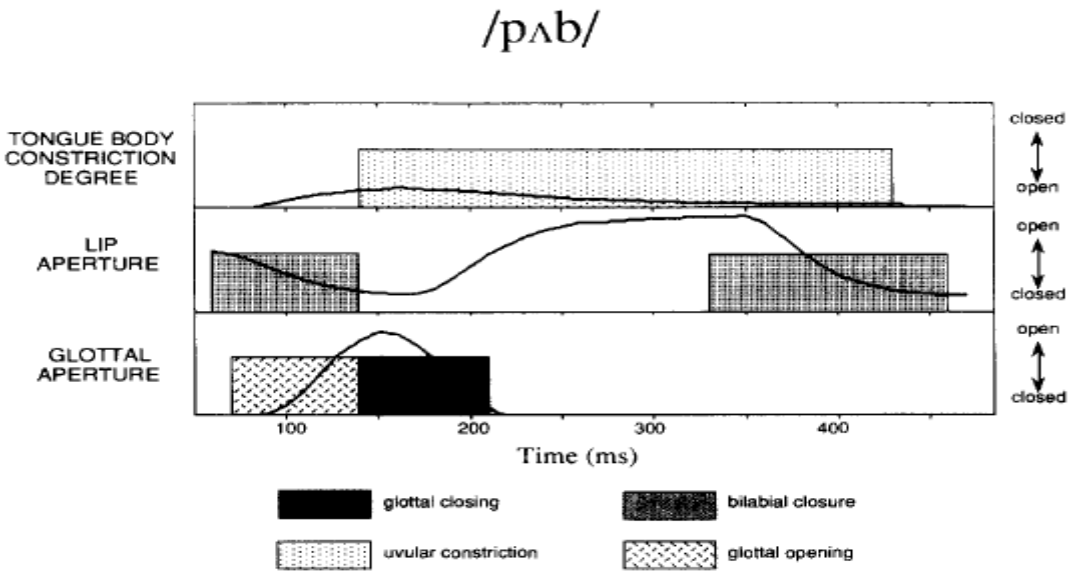


FIGURE 6 Gestural score used to synthesize the sequence /pʌb/. Filled boxes denote intervals of gestural activation. Box heights are uniformly either 0 (no activation) or 1 (full activation). The waveform lines denote tract-variable trajectories generated during the simulation.

Figure 14. Gestural score and tract-variable trajectories from Saltzman & Munhall (1989). *Figure used without permission.*

In SM89, the activations and deactivations of driving forces (and corresponding changes in tract variable equilibria) are specified as intervals on the *intergestural level.* Tract-variable targets are thus changed in the course of an utterance, and these changes are activation and deactivation of driving forces on tract-variable systems. SM89 represented the application of these driving forces non-dynamically, in the form of a so-called "gestural score". In this approach, every possible linguistic gesture is associated with a vector specifying the target, damping, and stiffness parameters for the active force components of the specific tract variables involved in the gesture. When the gestural score (Figure 14) reaches a point in time when a particular gesture becomes active, those parameters govern the associated tract variables.

One important thing to note about this system is that it is intended to model the directly observable interactions of articulators in speech. Coproduction patterns can be captured as overlap in gestural activations. This requires a system for determining the values of tract variable parameters when multiple gestures simultaneously exert an influence upon the same tract variable. SM89 accomplished this blending through *parameter tuning*. Parameter tuning integrates parameter values into an averaged weighted sum, and the weights are determined by dynamical equations representing lateral inhibition in a neural network. The details of this tuning are too complicated to consider here, but the specifics are not essential to understanding the concept.

One important question to ask about the task-dynamic model, or any model for that matter, is: what evidence is there to support it? The evidence comes from patterns and phenomena that the model can describe which were not previously well-described in a coherent framework. There are a number of advantages to the task-dynamic approach, which we will consider in turn:

1. Realistic movement trajectories and temporal dynamics:
    - the model offers parameters for conceptualizing differences in kinematic trajectories observed in speech.
    - the model allows for temporally realistic simulation of articulatory kinematics

2. Compensatory articulation:
    - the model offers an account of the immediate compensation observed in some effectors when other effectors in a synergy are perturbed; this is accomplished by the partitioning of gestural driving forces in the mapping from a tract-variable to multiple articulator variables.

3. Integrated accounts of phonological and prosodic patterns:
    - the model opens up the possibility of providing coherent explanations for previously unrelated phenomena; for example, Beckman, Edwards and Fletcher (1992) use this framework to propose that both schwa-deletion in English and German and high-vowel devoicing in Japanese and Korean can be understood to arise from gestural overlap.

The original motivation for applying task dynamics to speech movements (Saltzman 1986) was the desire to model articulatory variation, that is, details of articulator movement trajectories (kinematics) such as movement velocity, amplitude, and duration. The problem with modeling articulator movements directly is that there are too many degrees of freedom—because

in many cases, articulators interact with each other. For example, the motion of the lower lip, if the goal is to form a bilabial closure, will depend upon the motions of the jaw and upper lip; likewise the motions of the jaw and upper lip will depend upon the motion of the lower lip.

The task-dynamic approach was designed to handle this degrees-of-freedom problem. Task-dynamics allows the problem to be reformulated with fewer degrees of freedom, by hypothesizing that the *goals* or *tasks* of speech movements are cognitively instantiated in a goal/task-space (or, tract-variable space). Each goal or task is thus a one-dimensional attractor, rather than multidimensional.

If the task-dynamic model is based upon tract variables, then how does it accomplish the realistic modeling of movement trajectories? The answer lies in 1) learned mappings (through coordinate transformations) between tract variables and articulator variables, and 2) in articulator weightings that determine the extent to which various articulators contribute to the accomplishment of goals in tract variable space. Through these mappings and weightings, one is able to derive realistic articulator movement trajectories. We will consider them in more detail when we discuss compensatory articulation. Besides realistically modeling the kinematics of articulators and dynamics of vocal tract geometry, the task dynamic model can also describes such variables on realistic time scales. This is advantageous because it allows for more detailed comparison between model predictions and empirical data.

Another nice quality of the task-dynamic approach is that it can model compensatory articulation, which is a change in one articulation to compensate for perturbation of another. For example, Folkins & Abbs (1975) and Abbs and Gracco (1983) showed that if the jaw is unexpectedly loaded (i.e. a downward force is applied) during the closure movement for a /p/, lip closure is still accomplished, thanks to increased lip movements. In other words, if the jaw is unable to raise the normal amount for a /p/, the lips compensate. Kelso et. al. (1984) showed that this compensation is task-specific rather than reflexive: jaw perturbation during a /z/ induces compensation in the tongue, rather than then lips. Furthermore, the compensation occurs very quickly, within 20-30 ms of the onset of jaw perturbation. These findings suggest that compensatory articulation is "automatic", rather than arising from a normal feedback process involving cortical control ("intentional control"). Though automatic, the compensation pattern is not fixed like a reflex; it is specific to utterance goals, not stereotyped.

The task dynamic model of Saltzman & Kelso (1983) can simulate such compensatory movements because it distinguishes between tract variables and model articulator variables. As explained above, the tract-variables and an articulatory weighting matrix determine the extent to which tract-variable (task-space) goals are accomplished by the relevant model articulators. So in the example of lip closing in /ba/, while there is one lip aperture variable in tract-space, there are three relevant variables in model articulator space: the jaw, lower lip, and upper lip. The key idea behind the tract-variable-to-articulators mapping is that accomplishing the task of closing the lips (i.e. moving the tract variable lip aperture to its point attractor) involves a division of labor between three articulators. This division of labor is defined in the articulatory weighting matrix (defined either arbitrarily or from empirical observations).

In the mapping from tract variable space to model articulator space, the driving force upon the tract-variable LA is "activated" in the gestural score and partitioned by the articulatory weighting matrix into driving forces exerted upon associated model articulator variables. The tract variable lip aperture will continue to change until reaching equilibrium, which is determined by its target, and thus the net driving forces upon the articulators will also continue to be non-zero until LA is in equilibrium.

If one of the model articulators, for example the jaw, is suddenly prevented from moving, this can be thought of, in model terms, as the absence of a driving force on that system (or alternatively, as an additional opposing force—either way, this effect is the articulator remains "frozen"). Because some of the gestural activation force upon the tract-variable is lost in the mapping to articulator movements, lip aperture changes more slowly. (The changes in lip aperture are calculated from the model articulator positions). This model predicts that in the presence of perturbation, gestures will take longer to reach their targets, which gels with the empirical observations.

3.3 Intergestural timing

While the SM89 approach to intergestural timing employed a non-dynamical specification of activation intervals in a gestural score, a more recent approach utilizes systems of coupled limit-cycle oscillators to determine intergestural phasing relations. This approach, described in Saltzman & Byrd (2000)—henceforth SB00, begins with the definition of a task-space relative phase variable (Eq. 3.8), which represents the phase difference between two gesture activations. This model was developed to: "lend computational plausibility to the hypothesis that gestures belonging to single segments are coupled so as to yield narrow [relative] phase windows, while those belong to different segments are coupled to yield wider [relative] phase windows (Byrd 1996)." (2000:6).

$$\psi = \varphi_1 - \varphi_2 \tag{3.8}$$

$$\varphi_i = -\tan^{-1}\left((x'_i/\omega_{0i}) / x_i \right) \tag{3.9}$$

$$A_i = \left(x_{i2} + (x'_i/\omega_{0i})^2 \right)^{1/2} \tag{3.10}$$

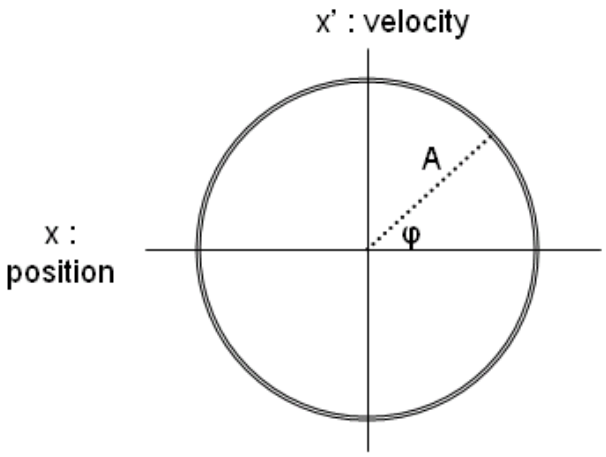


Figure 15. Relation between Cartesian phase space coordinates and polar phase space coordinates.

The relative phase variable ψ belongs to the task-space level, which is to say that the state of this variable corresponds to some goal or target. As is made explicit in later work (Goldstein, Byrd, & Saltzman 2006), the steady-state reached by this task-space relative phase variable determines intergestural timing in the gestural score. In other words, the difference between the

onsets of two gestural activations in the gestural score is a goal determined by a planning process. As we will see, this planning process involves the stabilization of relative phase between systems which are called *planning oscillators* in Goldstein, Byrd, & Saltzman (2006). Although in the SB00 discussion of their model, the limit cycle φ_i systems are called "model articulators," they cannot serve the same functions as the model articulators in the SM89 task-dynamic model, particularly with respect to simulating realistic movement trajectories. The planning oscillators are distinct from tract variable or model articulator variables, exhibiting limit cycle dynamics, and they correspond to higher-level, premotor planning processes.

The φ_i in Eq. 3.8 correspond to the phase angle of the *i*th planning oscillator. These phase angles are obtained by transforming Cartesian position x_i and velocity x'_i phase space coordinates to polar coordinates (phase angle φ, radial amplitude A). The equations for the transformation are shown in Eq. 3.9 and Eq. 3.10. ω_{0i} is the constant natural frequency of the *i*th oscillator. This transformation can be easily visualized by considering the phase space of an oscillator. Since the trajectory of the system is periodic (approximately circular or elliptical), the phase angle can be conceptualized as the angle between the Cartesian state vector and the *x*-axis (Figure 15).

Rather than using harmonic oscillators (i.e. systems with linear mass-spring dynamics), SB00 adopt a hybrid oscillatory dynamics (Kay, Kelso, Saltzman & Schoner, 1987). The hybrid systems are oscillators with linear damping and two sources of nonlinear damping (Eq. 3.11). α_i is the nonlinear van der Pol damping coefficient of the *i*th oscillator, and β_i is the nonlinear Rayleigh damping coefficient. If the nonlinear damping coefficients are zero, then we have the form of the harmonic oscillator (Eq. 3.12). For current purposes, it is not necessary to understand how manipulation of the nonlinear damping parameters changes the behavior of the oscillators; one need only be aware that these terms represent additional damping forces which make the oscillator motions follow more biologically realistic movement trajectories.

$$x''_i = -\alpha_i x'_i - \beta_i x_i^2 x'_i - \gamma_i x'^3_i - \omega_{0i}^2 x_i \quad (3.11)$$

$$x''_i = -\alpha_i x'_i - \omega_{0i}^2 x_i \quad (3.12)$$

Eq. 3.11 contains no terms that couple the systems. When two systems are coupled, the motion of each oscillator (if the coupling is bidirectional) is in some way influenced by the other one. To accomplish this, a term involving a state variable from the other oscillator must be added to the Eq. 3.11. This term can be thought of as describing a coupling *force*. This coupling force is like the gestural activation driving force, except that it is not constant. Rather, the coupling force is a function of the current state of the other oscillator.

SB00 use a potential function to model the effects of coupling between any two oscillators. The potential is a force that governs the change of a relative phase variable, and the relative phase of any two oscillators is the difference between their current phases, as in Eq. 3.8. Below I present in greater detail the conceptual basis for this maneuver, which was described in Haken (1983) and subsequently used by Haken, Kelso, and Bunz (1985) to model bimanual coordination. To understand the procedure, we return to the motion of a particle.

Haken (1983), treating harmonic oscillators (Eq. 3.13), considers a special case in which *m* is very small (relatively massless) and the damping constant γ very large, which corresponds to over-damped motion. In this special case, the first term of the l.h.s. of Eq. 3.13 can be neglected. Furthermore, the damping coefficient can be eliminated by using an appropriate time scale—this leads to Eq. 3.14. After these simplifications are made, the next step is to employ the notion of *work*. Work equals force times distance (Eq. 3.15). For example, lifting a body to a certain height

requires that some work be done, i.e. that some force is applied over some distance. Work is quantified as the product of the force applied times the distance the body is lifted. Because the force normally depends upon the position, an equation for work must be formulated for an infinitesimally small distance. By integrating Eq. 3.15, one obtains the total work over a finite distance (Eq. 3.16). The potential, *V*, is defined as the negative of *W*. From Eq. 3.15, this gives the relation in Eq. 3.17. In other words, Eq. 3.17 is an expression for the force acting upon a system that makes reference to the derivative of a potential function. For the harmonic oscillator $x' = -kx$, the potential is found by integrating, and has the form in Eq. 3.18.

$$mx'' + \gamma x' = F(x) \quad (3.13)$$

$$x' = F(x) \quad (3.14)$$

$$dW = F(x)\,dx \quad (3.15)$$

$$-V = W = \int_{x0\text{-}x1} F(x)\,dx \quad (3.16)$$

$$F(x) = -\,dV / dx \quad (3.17)$$

$$V(x) = \tfrac{1}{2} k x^2 , \qquad x' = F(x) = -kx \quad (3.18)$$

Haken (1983) then compares *V* with the work done in lifting a weight. The comparison suggests interpreting the curve defined by *V* as the slope of a hill, as in Figure 16. Doing work to raise a particle up the hill endows the particle with the potential to fall back down and return to the equilibrium point.

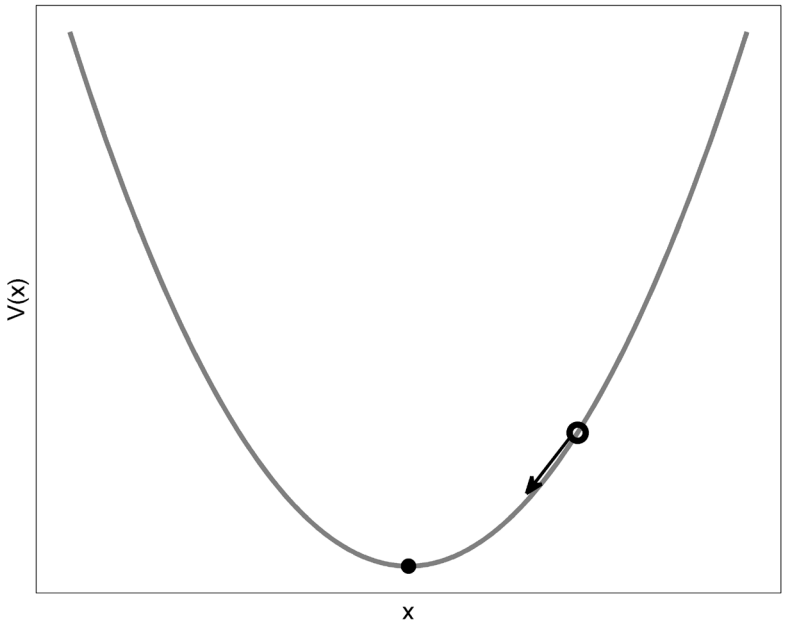


Figure 16. Potential function and motion of particle.

The crucial idea here is that the change of a variable in some system under investigation, given the assumption of massless-ness and an appropriate time scale, will be like the motion of a particle in a potential function. Not all mappings in this metaphoric blend hold, because the

particle, having negligible mass and thus no momentum, does not overshoot the equilibrium and oscillate until settling down to the stable equilibrium point. Instead, when the horizontal slope at x=0 is reached, F(0) = 0 and x' = 0, thus the particle stops. If one wishes, this can be attributed to the "stickiness" of the hill. The potential function thus describes the change of a variable, and the rate and direction of this change are conceptualized as motion down the slope of a valley.

Potential functions are used to describe the behaviors of much more general physical systems than just massless particles rolling down hills. Other forces can also be described with potential functions. There is something intuitively appealing about using force field potentials to describe coupling forces of nonphysical oscillatory systems. One thing that is particularly appealing about the metaphor is that between the source domain of physical forces and target domain of cognitive systems there are neurophysiological oscillations that may also be subject to similar forces.

We return now to how SB00 derive the equation of motion for the task-space relative phase from a potential function. They start by defining a potential function (Eq. 3.19); this potential function represents a hypothesis about how the task-space relative phase variable changes over time. That is, they hypothesize that the relative phase ψ changes over time as if it is a particle rolling down the slope of the potential function in Eq. 3.19, in an overdamped manner, with one minimum (ψ_0) defined as 0, which represents synchronization of the oscillators. The motion of the particle rolling down the hill is the negative of the derivative of the potential function, thus giving Eq. 3.20 for the change of relative phase.

$$V(\psi) = -a\cos(\psi - \psi_0), \qquad \psi_0 = 0 \tag{3.19}$$

$$\psi' = -dV/d\psi = -a\sin(\psi - \psi_0) \tag{3.20}$$

$$\psi = \varphi_2 - \varphi_1 \quad , \quad \psi' = \varphi_2' - \varphi_1' \tag{3.21}$$

$$\varphi_i' = S_i\,\psi' \tag{3.22}$$

In principle, any continuous function could be chosen for the potential. The cosine function in (Eq. 3.19) was chosen because it meets several important conditions. First, the cosine is an even function, symmetric about the x origin, i.e. $\cos(x) = \cos(-x)$, which is important because the model should hold whether relative phase is defined as φ_1-φ_2 or φ_2-φ_1. Second, it is periodic, so that $\cos(x) = \cos(x + 2\pi)$, which also means that the potential function can be represented with a Fourier series. Third, it has one local minimum per cycle.

The final step is to transform the task-space state velocity (ψ') into coupling terms to add to Eq. 3.11. To do this, an expression is needed for the ith planning oscillator velocity in terms of relative phase, as in Eq. 3.21 and Eq. 3.22. S_i is the ith component of an unweighted pseudoinverse of the Jacobian matrix of Eq. 3.21 (i.e. a 1 x 2 matrix of partial derivatives of ψ with respect to φ_2 and φ_1). This component is the least squares solution for Eq. 3.22, and in this case $S = [-\frac{1}{2}, \frac{1}{2}]^T$. The S_i represent the strength and direction of the effect of the potential force on the ith oscillator. The phase velocities are then transformed into coupling terms in planning oscillator Cartesian coordinate space (x', x). After these transformations, the coupling term for the ith oscillator is Eq. 3.23. Thus the equation of motion for the ith oscillator is thus Eq. 3.24.

$$x_{i\text{-}couple}'' = (\omega_{0i} A_i \sin \varphi_i)\, S_i\, (a \sin(\psi - \psi_0)) \quad (3.23)$$

$$x''_i = \boxed{-\alpha_i x'_i} - \boxed{\beta_i x_i^2 x'_i - \gamma_i x_i'^3} - \boxed{\omega_{0i}^2 x_i} + \boxed{(\omega_{0i} A_i \sin \varphi_i)\, S_i\, (a \sin(\psi - \psi_0))} \quad (3.24)$$

LINEAR DAMPING FORCE + NON-LINEAR DAMPING FORCES - ELASTIC/ SPRING FORCE + COUPLING FORCE

The first term on the r.h.s contains the linear damping, the second and third terms the nonlinear damping, the fourth term the linear elastic driving force, and the last is the force from coupling. This coupling term is a function of both oscillators, because $\psi = \varphi_2 - \varphi_1$. SB00 further generalized this coupled limit-cycle oscillators approach so that target relative phases can be specified as "windows", rather than points. This is accomplished by choosing a potential function with a much flatter slope in the vicinity of the target. The major difference between the flattened potential (window approach) and the sinusoidal potential (punctate approach) is how quickly relative phase settles down to the target. In 30 sec. simulations reported in SB00, the "final" relative phase, i.e. the relative phase after 30 sec. of simulation, differed between the two approaches. More specifically, initial relative phase and oscillator velocities influenced final relative phase for the window approach, but not for the punctate approach. The authors argue that compared to the punctate approach, the window approach to relative phasing more faithfully represents variability in relative timing.

It is noteworthy here that the choice of when to terminate the simulation is entirely arbitrary, and because the difference between the relative phase outputs of punctate and window approaches depends heavily upon when the simulation is terminated, the assumption made regarding an appropriate amount of time to run the system should be justified in some way. Both the window and punctate approaches are punctate approaches in the limit of infinite time, and their outputs will reflect this if the system is given enough time for relative phase to settle.

An important empirical observation that lends credence to a relative phase-based analysis is the c-center effect, identified by Browman & Goldstein (1988). The *c-center effect* is a generalization about the relative timing of onset-consonantal gestures and vocalic gestures. In syllables with simplex onsets, vocalic and consonantal gestural initiations are approximately synchronized. However in syllables with complex onsets, the initiations of consonantal gestural are equally displaced in opposite directions from the beginning of the vocalic gesture (alternatively, the vocalic gesture starts about midway between the consonantal ones). In other words, CCV and CV are influenced by the same phasing relation, but in CCV the relation applies to the center of the consonantal gesture onsets, i.e. the point midway between their onsets (here, 'C' and 'V' will stand for associated gestures, not segments).

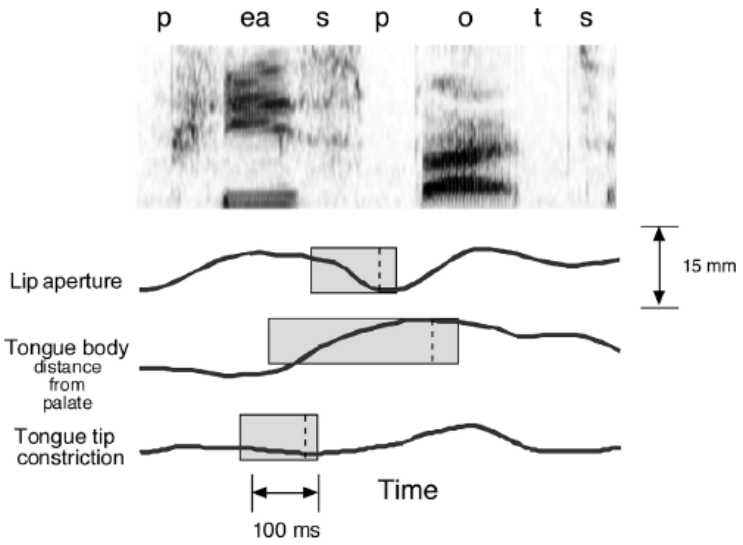


Figure 17. Example of c-center effect from Goldstein, Saltzman, & Byrd (2006). *Figure used without permission.*

Browman & Goldstein (2000) proposed that the c-center effect could be modeled with competing relative phase specifications between pairs of gestures (i.e. lexical specifications). They hypothesized that in complex onsets (C_1C_2V), both consonants are influenced by the same C-V phase relation to the vowel gesture. This is represented by Figure 18 (a). If this were the only specified phasing relationship influencing the timing of the gestures, then the consonantal gestures would be highly overlapped, and that would make perception of the gestures difficult. However, if an additional C_1-C_2 anti-phase relation influences the timing of the gestures, as shown in Figure 18 (b), and if this competes with the C-V phase relation, then the outcome is partly overlapped but still perceptible consonantal gestures with some preservation of the C-V relation, shown in Figure 18 (c). This compromise between the two phasing specifications is hypothesized to cause the c-center effect. The moment that ends up being in-phase synchronized with the beginning of the vowel gesture is the *c-center*, which is the point halfway between the beginnings of the C gestures. Generalizing to 3 and 4 consonant onsets, the c-center is the point halfway between the beginnings of gestures associated with the first and last consonants.

The C_1-C_2 anti-phase relation has a functional motivation in *recoverability*: the gestures must not be entirely overlapped, otherwise one might completely obscure the other. For example, if two oral consonantal gestures were phased synchronously, a hearer would have difficulty recovering cues for the constriction located further back in the vocal tract. In #CCV, if a labial and dorsal constriction are made synchronously, then the primary acoustic cue for perceiving the dorsal gesture, its release burst, would be obfuscated by labial constriction. The C-V phasing relation also has a functional motivation: it allows for advantageous *parallel transmission* of information.

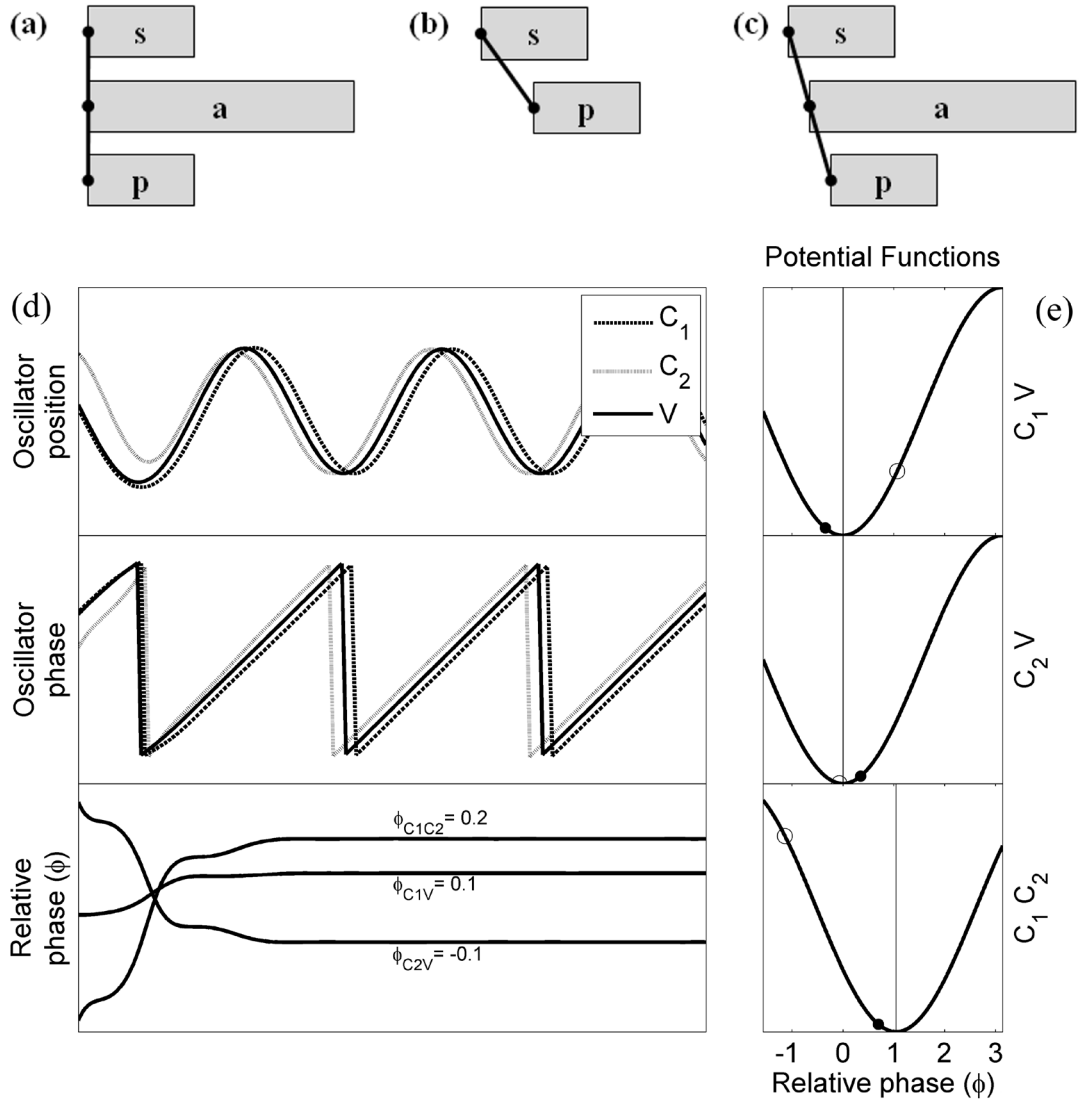


Figure 18. Coupling graphs and simulation of c-center effect through competitive coupling. Potential functions depict relative phase targets with vertical lines, initial relative phases (randomly chosen) with open circles, and final relative phases with filled dots. Evolutions of oscillator positions, phases, and relative phases from random initial conditions are shown on the left. Note that initial transients die off quickly, and that once the system has stabilized, C_1 and C_2 are equally displaced from V in opposite directions.

Nam & Saltzman (2003) simulated the competitive coupling proposed by Browman & Goldstein (2000) using a system of coupled oscillators. Rather than treating the relative phases of gestural onsets as arbitrarily specified lexical information (as in Saltzman & Munhall 1989), the newer model incorporates the idea that target relative phases are derived from the stable relative phases of a system of coupled oscillators. These oscillators are *gestural planning systems*, whose properties are developed partly in Saltzman & Byrd (2000), and were described above in detail.

The NS03 equations of motion for one of the consonants and the vowel are shown in Eqs. 3.25 and 3.26.

$$\begin{aligned} x''_{C1} = & -\alpha_{C1} x'_{C1} - \beta_{C1} x_{C1}^2 x'_{C1} - \gamma_{C1} x_{C1}'^3 - \omega_{0\,C1}^2 x_{C1} && \text{[uncoup. osc. motion]} \\ & + (\omega_{0\,C1} A_{C1} \sin \varphi_{C1})\, S_{C1}\, (a_{C1V1} \sin(\psi - \psi_{C1V1})) && \text{[coupling of C1 to V1]} \\ & + (\omega_{0\,C1} A_{C1} \sin \varphi_{C1})\, S_{C1}\, (a_{C1C2} \sin(\psi - \psi_{C1C2})) && \text{[coupling of C1 to C2.]} \end{aligned} \quad (3.25)$$

$$\begin{aligned} x''_{V1} = & -\alpha_{V1} x'_{V1} - \beta_{V1} x_{V1}^2 x'_{V1} - \gamma_{V1} x_{V1}'^3 - \omega_{0\,V1}^2 x_{V1} && \text{[uncoup. osc. motion]} \\ & + (\omega_{0\,V1} A_{V1} \sin \varphi_{V1})\, S_{V1}\, (a_{C1V1} \sin(\psi - \psi_{C1V1})) && \text{[coupling of V1 to C1]} \\ & + (\omega_{0\,V1} A_{V1} \sin \varphi_{V1})\, S_{V1}\, (a_{C2V1} \sin(\psi - \psi_{C2V1})) && \text{[coupling of V1 to C2]} \end{aligned} \quad (3.26)$$

One can see from these equations of motion that in addition to parameters for the uncoupled motions of the oscillators, there are now three potential functions from which three different coupling terms are derived, one for each coupling relation. Note also that each potential function has its own amplitude parameter, a_{xy}, representing the strength of the coupling force.

There is an important distinction between gestural systems and gestural planning systems. In the task dynamic framework, gestures are modeled as overdamped mass-spring systems that have fixed-point attractors. With parameter modifications, gestures can vary in speed, amplitude, and duration. The gestural planning oscillators are very different. They exhibit limit cycle dynamics and their relative phases are governed by potential functions. The planning oscillators bear an indirect relation to observable articulator movements. It is the stabilized relative phases of the planning oscillators which in the NS03 model are hypothesized to determine the relative phases of the damped mass-spring gestural systems. It remains an open question how long it normally takes for these systems to stabilize in the course of speech planning. Another possibility is that the initial conditions are not random, and that lexical memory biases the initial conditions. In any case, if stabilization does not occur in a trivially fast manner, then the model predicts that relatively prepared speech will exhibit less articulatory variability than relatively unprepared speech.

Figure 18 (d) shows a simulation of the c-center effect in planning oscillators based upon the NS03 model. For each pair of coupled oscillators (and there are three such pairs in a C_1C_2V system), the figure shows the potential function governing the evolution of the relative phase between the oscillators. Regardless of initial relative phase, the system settles into a stable configuration that corresponds to a balance between the competing forces associated with the potential functions. Following NS03, the target relative phase of the consonants (Φ_{C1C2}) was approximately 1 radian (60°), rather than true anti-phasing. The minimization of potential energy is accomplished by equal displacement of C gestural phases from the phase of the V gesture, hence producing a c-center effect. (Note that as an alternative to lexical specification of Φ_{C1C2} = 1 radian, the same sort of C-center effect is also produced if the potential governing a $\Phi_{C1C2} = \pi$ target relative phase is relatively low-amplitude compared to the potential governing C-V relative phase. This preserves a parsimonious version of the model in which only in-phase and anti-phase lexical specification is possible.)

In contrast, for coda (VC_1C_2) gestural timing, Browman & Goldstein (2000) hypothesized that only a V-C_1 phasing relation is specified. Subsequent consonants are phased only to each other, as represented in Figure 19. The idea here is that there is no competition between phase specifications and thus no c-center effect.

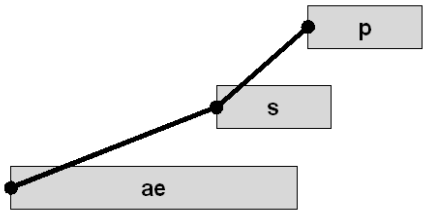


Figure 19. Schematization of VCC phase relations.

The NS03 coupled oscillators model of gestural overlap in CCV and VCC can also account for the relatively greater variability in gestural overlap of coda consonant gestures compared to onset gestures. To do this, Nam & Saltzman added noise to the system by randomly perturbing the frequencies (ω_0) of the oscillators. This "detuning" of frequency was implemented in the simulations by adding a noisy term to the potential function (Sternad, Amazeen, Turvey, 1996). In Eq. 3.27, the parameter *b* represents the amount of detuning. This parameter was chosen randomly from a Gaussian distribution on successive simulations. The results showed that CCV phasing was less variable than VCC phasing. This finding can be understood to follow from the larger number and thus greater strength of coupling forces in CCV than in VCC.

$$V(\psi) = -a \cos(\psi - \psi_0) + b(\psi - \psi_0) \tag{3.27}$$

Goldstein, Byrd, & Saltzman (2006)—henceforth GBS06—clarified the relation between limit cycle planning and the critically damped oscillators with point attractor targets in Saltzman & Munhall (1989). The planning oscillators endow the intergestural level with its own dynamics in the following way: the steady state relative phase values of a system of planning oscillators determines the relative phases of activations in the gestural score. The target relative phases of the planning oscillators are assumed to be specified lexically. These lexical relative phase specifications are represented in couplings graphs, like the one in Figure 20.

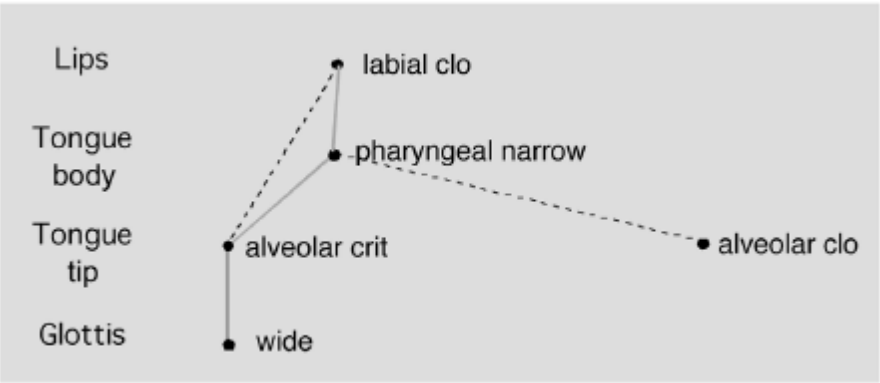


Figure 20. *Coupling graph from Goldstein, Byrd, & Saltzman (2006).* Coupling relations are represented by solid and dotted lines. *Figure used without permission.*

In the coupling graph, coupling relations (represented by solid and dotted lines) may or may not exist between any pair of gestures. Goldstein, Byrd, & Saltzman (2006) further

hypothesize that only two target relative phases are necessary: in-phase (synchronous) and anti-phase (syncopated). In coupling graphs, in-phase coupling is represented with a solid line and anti-phase coupling with a dashed line. The justification for needing only these two phase relations is that in-phase and anti-phase coupling have been shown to be the two most stable modes of coordination of repeated movements (e.g. Haken, Kelso, Bunz, 1985).

GBS06 theorize that syllable-initial consonants and following vowels are coordinated in-phase with one another (called the *onset relation*), and that the coda relation (V-C) is anti-phase. C-C target phasing is proposed to be anti-phase. Doubly articulated consonants such as k͡p can be hypothesized to be in-phase synchronized. GBS06 also make the intriguing suggestion that target phase relations and phonological combinatoriality are related:

> "free combination occurs just where articulatory gestures are coordinated in the most stable, in-phase mode. Consequently, onset gestures combine freely with vowel gestures because of the stability and availability of the in-phase mode. Coda gestures are in a less stable mode (anti-phase) with vowels and therefore there is an increased dependency between vowels and their following consonants; though individual combinations may be made more stable by learning. Within onsets and within codas, modes of coordination may be employed that are either anti-phase or possibly not intrinsically stable at all. These coordinative patterns of specific coupling are learned, acquired late, and typically involve a restricted number of combinations" (2006).

GBS06 cite a number of forms of evidence for this model of syllable structure. The first involves phonological asymmetries in onset and coda patterning. The distribution of coda segments is often relatively restricted compared to onset segment distribution. Combinatorial restrictions on coda-nucleus pairs are more prevalent than restrictions on onset-rime pairs. Further, codas potentially have weight, while onsets rarely (arguably) have weight. The second form of evidence involves statistical biases toward more easily in-phase produced CV pairs. It is known that in CV sequences, coronals and front vowels, dorsals and back vowels, labials and central vowels tend to co-occur more frequently than segments that do not share the same place of articulation. Moreover, these statistical biases are absent in VC structures. Third, there are different segment-internal gestural timing relations for C gestures in CV than VC, as evidenced in the variability patterns described above. Fourth, resyllabification has been understood experimentally as phase transition to a more stable mode of coordination. Tuller and Kelso (1990, 1991) argued that phase transitions from VC → CV suggest that resyllabification is a phase transition to a more stable mode of coordination.

The dynamical approach to understanding intergestural timing and syllable-internal structure has thus been highly successful in accounting for temporal and distributional patterns on the gestural timescale. It is noteworthy that this approach has not been extensively applied to perceptual patterns, nor fully integrated with a dynamical account of perception. There are, however, several extensions of these ideas to understanding prosodic systems, which we consider next.

3.4 Rhythmic timing: syllables and feet

The next highest level or largest timescale of linguistic organization above the syllable is the foot. Generally speaking, there are two sorts of approaches to studying feet and their relations to syllables (cf. Liberman, 1975; Liberman & Prince, 1977; Selkirk, 1984). On the one hand, there are approaches in which the foot is conceptualized as a container, and in which relations between syllables and feet are represented with node and connection schemas. On the other hand, there are approaches in which the foot is thought of as an inter-stress interval, a so-called i.e. stress-foot. This latter approach is more consistent with the idea of a metrical grid, and is exclusively used in dynamical approaches to rhythmic timing.

The phonetic tradition on rhythm has primarily been concerned with typological classes. Pike (1945) cited metaphors of "Morse code rhythm" and "machine-gun rhythm" from Lloyd James (1940), in describing a typological distinction between languages like English and Dutch vs. Spanish and Italian. Pike (1945) named these types stress(foot)-timed and syllable-timed languages. Abercrombie (1967) proposed that the distinction boils down to the primary unit of isochrony in a language. Thus in English and Dutch, the timing of stressed syllables is purported to be more regular (or, less variable) than the timing of syllables. In syllable-timed languages, the reverse is supposed to hold. Mora-timing has also been proposed for some languages, such as Japanese and Tamil (Port, Dalby, & O'Dell, 1987). These distinctions have not been well supported in empirical studies (O'Dell & Nieminen, 1999; Dauer, 1983; Bolinger, 1968; Lehiste, 1977).

There is, however, evidence that across languages the durations of inter-stress-intervals (ISIs) are well-described by linear models where the independent variable is the number of syllables contained in the intervals. Erickson (1991) analyzed ISI duration data (mean interval durations) from Dauer (1983) with linear regression (see Table 3.2). Erickson used a linear model (i.e. $I_n = a + bn$) to predict the mean duration of ISIs in a given language from the number of syllables n within those intervals.

Table 3.2 Erickson (1991) analysis of ISI duration from Dauer (1983)

Rhythm class	Language	Linear Model	
stress-timed:	English	$I = 201 + 102n$	$r = 0.996$
	Thai	$I = 220 + 97n$	$r = 0.973$
syllable-timed:	Spanish	$I = 76 + 119n$	$r = 0.997$
	Greek	$I = 107 + 104n$	$r = 1.000$
	Italian	$I = 110 + 105n$	$r = 1.000$

The nice thing about this analysis is that the distinction between stress-timing and syllable-timing is reflected in the size of the intercept coefficient. Erickson (1991) interpreted the slope coefficients as representing the normal contribution of a syllable to ISI duration, and the intercept coefficients as the extra duration associated with the stressed syllable in the interval. Under this interpretation, the distinction between stress- and syllable-timing is mostly the amount of extra duration contributed by a stressed syllable. That interpretation, however, is not the only one possible. The extra duration might be distributed across the syllables in each group in more complicated ways. Moreover, these languages differ substantially in the amount of reduction in unstressed syllables (Dauer, 1983); indeed, a number of researchers have argued that

the distinction between stress and syllable-timing can be understood with respect to statistics describing vocalic and consonantal intervals in speech, or voiced and voiceless intervals (Ramus, Nespor, & Mehler, 1999; Grabe & Low, 2002). It is a chicken and the egg sort of question whether the reduction patterns give rise to the rhythm, or vice versa, and this calls into question the meaningfulness of regression analyses, which do not speak to an underlying principle of rhythmic organization the way that the isochrony hypothesis does.

O'Dell & Nieminen (1999) proposed to model the increased duration of stressed syllables in stress-timed language with a coupled oscillators model. In their approach θ_1 and θ_2 are the phases, and ω_1 and ω_2 the frequencies, of a stress group oscillator and syllable oscillator. The stress group oscillator corresponds to the stress-foot, i.e. a foot defined as the interval between stressed syllables. The stress foot has been used by researchers to address the isochrony hypothesis (e.g. Lehiste, 1977; Ohala, 1975), but it is not equivalent to the foot as it is treated in metrical theory. However, metrical feet and stress-feet are related in that they are defined with respect to stress.

The oscillating systems in this model are simple harmonic oscillators in polar coordinates, where the angular velocities and radial amplitudes of the system are constant (Eq. 3.28). In other words, this system describes a point moving in a circular orbit. This type of oscillator is somewhat simpler than those employed in task-dynamic approaches described above, but nonetheless can exhibit similar behaviors when coupled to other oscillators. As in the Saltzman & Byrd (2000) approach, they use a generalized relative phase (Eqs. 3.29 and 3.30), which allows for the inherent frequencies of the oscillators to differ. By hypothesis, if there are *n* syllables in a foot, then the ratio ω_1: ω_2 is 1:*n*.

$$\theta'_n = \omega_n \tag{3.28}$$

$$\varphi_{12} = m\theta_2 - n\theta_1 \qquad n = \#\ of\ \sigma\ in\ Ft,\ m = 1 \tag{3.29}$$

$$\varphi_{mn} = \omega_m\theta_n - \omega_n\theta_m \tag{3.30}$$

$$H_n(\varphi) = \varphi / (r + n) \tag{3.31}$$

$$\theta'_1 = \omega_1 + H_1(\varphi_n) \tag{3.32}$$

$$\theta'_2 = \omega_2 - rH_2(\varphi_n)$$

$$T_1(n) = \frac{1}{\omega_1 + H(\phi_n)} = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2}n \tag{3.33}$$

The ON99 generalized relative phase coupling function (Eq. 3.31) is parameterized by *r*, which corresponds to the strength of the stress oscillator relative to the syllable oscillator. The functions H_1 and H_2 are somewhat analogous to the potential functions used by Saltzman and Byrd (2000), but do not employ the metaphor of a particle moving in a force field. Eq. 3.32 shows the equations of motion for the stress and syllable oscillators.

O'Dell & Nieminen get explanatory mileage out of the idea that the coupling between stress and syllable oscillators can be asymmetric. The syllable oscillator forces the stress group oscillator to a greater extent if $r < 1$ and vice versa if $r > 1$. Note that the equilibrium of the coupling function is always relative phase 0. The period *T(n)* of the stress group oscillator at equilibrium is a function of the number of syllables in the group (Eq. 3.33). Thus the period/duration of the steady-state (equilibrium) interstress interval in the coupled oscillators system is a linear function of the number of syllables. Associating *r* with *a/b* and rearranging the linear function into the form $I = b(r + n)$ allows one to use the coupled oscillators model to capture the duration data by parameters *r*, the relative strength of the syllable and stress group

oscillators, and *b*, the slope of the linear model $I = b(r + n)$. Figure 21 shows a graph of *r* vs. *b* for different languages from O'Dell & Nieminen (1999).

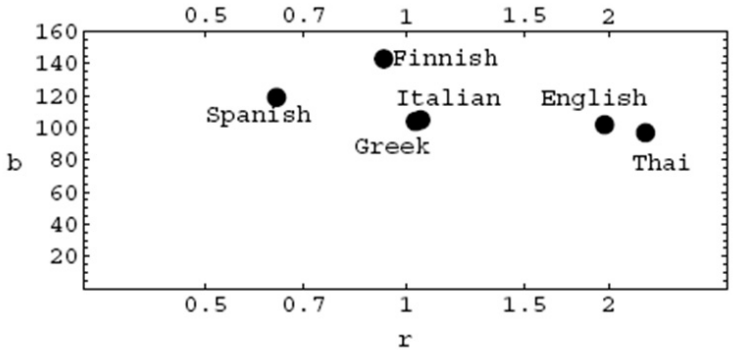


Figure 1. Relative coupling strength (*r*) vs. regression slope (*b*) for various languages.

Figure 21. Coupled oscillators model parameterization of rhythmic classes from O'Dell & Nieminen (1999). *Figure used without permission.*

A noteworthy difference between the ON99 model and Nam & Saltzman (2003) gestural planning oscillators model is the form of the coupling function. The ON99 coupling function is a linear function of generalized relative phase, and implicitly requires a target generalized relative phase of 0. This means that the force exerted on relative phase does not grow larger as it gets further from the target, nor become smaller when it gets closer. The basic insight offered by this model is that patterns of interstress interval duration can be understood to arise from differences in coupling strength between oscillatory systems. This contrasts with the NS03 model of intergestural timing, which accounts for linguistic patterns with differences in relative phase targets.

3.5 Rhythmic timing: feet and phrases

The idea of the foot as a rhythmic unit naturally opens up the possibility of higher level, larger timescale units that involve multiple feet—which, for lack of a better term, are called *phrases*. Cummins & Port (1996) and (1998) used a *speech cycling task* to probe the relative timing between phrases and feet, and revealed two very interesting patterns. In this task, subjects hear a high-low two tone metronome pattern and repeat a phrase (e.g. *dig for a duck*) so that the first and last stressed syllables of the phrase (i.e. *dig* and *duck*) align with the metronome beats, as depicted in Figure 22 (a). After a number of metronome pattern repetitions, the metronome fades out while subjects continue repeating the phrase, trying to maintain the metronome rhythm. The control parameter in this design is the target phase (Φ) of the second (L) metronome tone relative to the phrasal period defined by successive first (H) tones. In Cummins & Port (1998) the H to L tone interval was fixed at 700 ms and the phrasal periods were altered such that target phases were varied randomly on successive trials from 0.30 to 0.70 of the phrasal period.

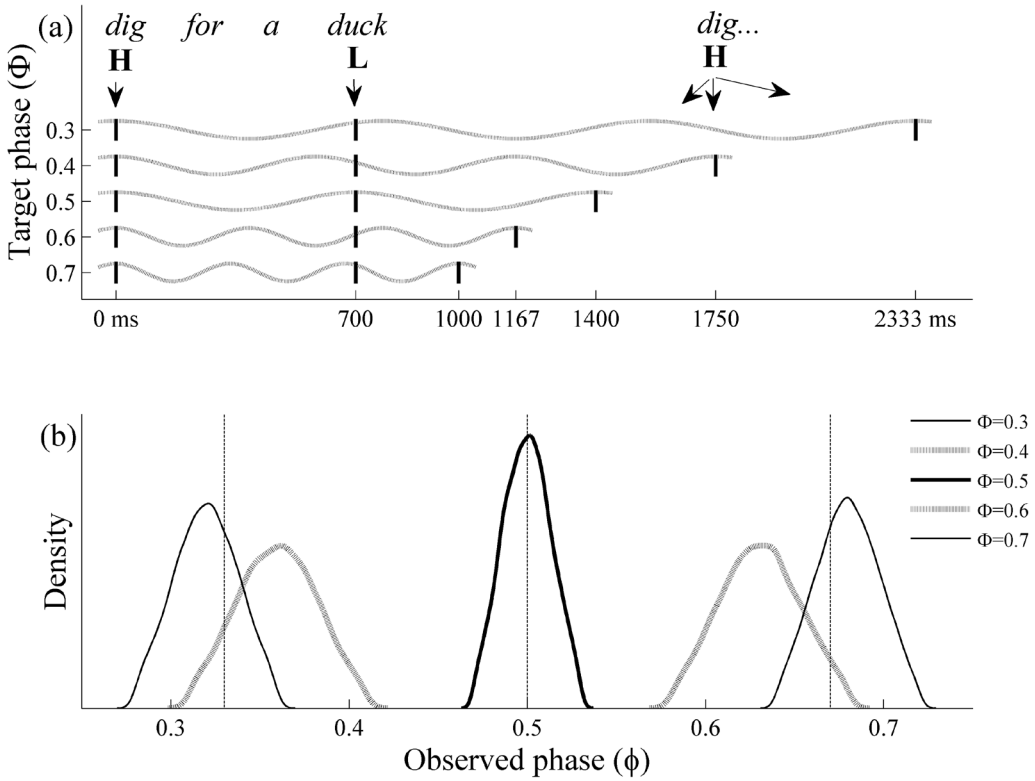


Figure 22. Schematic illustrations of speech cycling task design and harmonic timing effect with target phases 0.30, 0.40, 0.50, 0.60, and 0.70. (a) intertone interval durations in each of the five target phase conditions, along with hypothesized foot system oscillations. (b) relative phase distributions for each target phase condition (Φ). Shifts in observed phases toward harmonic ratios (vertical lines) are evident in the distributions, as well as increased variability in the higher-order target phase conditions. The same pattern is observed with or without metronome tones present. See text for further details.

In the speech cycling task, Cummins and Port discovered a *harmonic timing effect* whereby productions of the second stressed syllable were shifted toward phases of the phrasal period that were close to the low-order integer ratios 1/3, 1/2, and 2/3. Observed phases were measured by approximating the locations of p-centers, which are salient points in time believed to correspond to the "beats" of syllables (Allen, 1972, 1975; Morton, Marcus, & Frankish, 1976; Howell, 1988; Pompino-Marschall, 1989). In Figure 22, the observed relative phase φ is the ratio of the interval between p-centers of *big* and *duck* to the interval between successive p-centers of *big*. Cummins and Port (1998) found that the distributions of observed phases were shifted toward the nearest low-order harmonic ratio, as shown in Figure 22 (b). The same pattern was observed with and without the metronome tones present.

Another important observation was that variance in produced phase was lower for target phases nearer to the low-order harmonic ratios; hence for the five target phase conditions shown above variability in produced phase was highest in the $\Phi_{0.4}$ and $\Phi_{0.6}$ conditions, moderate in the $\Phi_{0.3}$ and $\Phi_{0.7}$ conditions, and lowest in the $\Phi_{0.5}$ condition. This suggests that the rhythm of the $\Phi_{0.5}$ condition is easiest to produce, while the other target phases require more difficult rhythms.

Cummins & Port (1998) and Port (2003) suggested that the results of the speech cycling task can be modeled with phase-locked pulses from a multifrequency system of coupled oscillators, where harmonic phrase and foot oscillators are phase-locked and either 1:2 or 1:3 frequency-locked. Figure 22 (a) shows the oscillations of the foot oscillators corresponding to each target rhythm. Observe that the phrase oscillator is entrained to the H metronome tones. The foot oscillator, in turn, is frequency-locked to the phrase oscillator, but not entrained to any external perceptual stimulus. The foot oscillator produces pulses at phase 0, and stressed syllable beats are attracted to the pulses. The motivation for this approach derives from the finding that low-order harmonic frequency ratios are more stable than higher-order ones (Haken, Peper, Beek & Daffertshofer, 1996). Correlations between target phase and variability may arise due to the relatively lesser stability of 1:3 frequency-locking compared to 1:2 locking.

The results of the speech-cycling task are interesting because they support the notion that there exist multi-timescale dynamical interactions between linguistic systems, in this case feet and phrases. The Cummins & Port analysis conceptualizes hierarchically-related metrical structures as frequency-locked coupled dynamical systems that are responsible for rhythmic coordination of speech. An important contribution of this approach is that it allows readily for predictions to be made regarding the relative timing of higher-level prosodic constituents such as syllables, feet, and phrases. Furthermore, it provides for predictions about how articulation interacts with prosodic structure, which we turn to below.

3.6 Coupling between gestural and rhythmic systems

In preceding sections, we examined how dynamical systems have been used to model patterns of intergestural timing and patterns of prosodic timing between syllables and feet, and between feet and phrases. We now discuss the predictions that arise from integrating gestural and prosodic dynamical models. These predictions are tested by experimental and corpus studies reported in subsequent chapters.

If short timescale gestural systems and longer timescale rhythmic systems do not interact with each other, then their dynamics can be understood independently and there should be

nothing particularly interesting about the behavior of the system as whole that cannot be learned from studying its parts. There is no a priori reason for such interactions to occur, although it seems intuitively obvious that some form of interaction must take place. Absence of an interaction should entail, for example, that the relative timing of tongue and lip gestures in an [sp] cluster in the word "spa" will be unaffected by the rhythmic context in which these gestures are articulated. To the contrary, if rhythmic and gestural systems do interact, then intergestural timing should be influenced by the rhythmic context in which the gestures are produced. Experimental perturbations of speech rhythm should influence the relative timing of movements associated with nearby gestures.

One reason to suspect that rhythmic and gestural systems do interact is that there are correlations between patterns of deletion/reduction of speech gestures and the typological classes of stress- and syllable-timed languages (Ramus et al., 1999; Dauer, 1983). However, to date there has been no conclusive demonstration that rhythmic and gestural systems interact within a given utterance.

Because dynamical coupling can usefully model empirically observed temporal patterns involving feet and phrases, as well as temporal patterns involving gestures, an obvious question to ask is whether evidence can be found for a more general model which accounts for patterns on both timescales. A synthesis of the two models requires bridging the gap between gestural timing and rhythmic timing. An obvious way to connect these disparate timescales is through generalized relative phase coupling. In addition to phrase-foot coupling and intergestural coupling, such a system presumably includes interactions between feet and syllables and/or moras. One would expect such a model to cover the gamut of relevant timescales, from phrases (groups of feet), to feet (intervals between stressed syllables), to syllables, moras, and on the lowest level, gestures and effectors. Figure 23 illustrates how prosodic units in the hierarchical branching model of prosodic and segmental structure can be related to oscillatory systems associated with a range of timescales.

Of course, not all languages exhibit evidence for systematic patterning on all of these scales; for example, languages that lack closed syllables and long vowels generally exhibit no evidence for a moraic level of prosodic structure distinct from the syllabic level (e.g. Hua, Cayuvava; cf. Blevins, 1995), and many languages have no stressed syllables (e.g. Cantonese, Yoruba; cf. Hyman, 2006). Any satisfactory model should have parameters available to accommodate cross-linguistic variation, although such variation will not be our primary concern here. Further, the values of some model parameters may vary from speaker to speaker and from utterance to utterance. A long-term goal of any modeling enterprise should be the identification of intrinsic constraints on model parameters by thorough investigation of cross-linguistic, inter-speaker, and intra-speaker variation.

It should also be observed that higher-level prosodic constituents, such as the phonological word or phonological phrase, as currently conceptualized, are not straightforwardly integrated into the hierarchical model presented here. This is because these units are determined by syntactic and semantic structure, unlike prototypical feet and syllables. This dissertation will not leave for future research the question of how higher-level prosodic units fit into the picture.

The multi-timescale dynamical model reconceptualizes the units of the traditional model as oscillators, which can be described with differential equations. The "levels" (or types of units) in the traditional model become timescales, which correspond to the inherent frequencies of their associated oscillators. These frequencies are related by low-order ratios of integers, and connections between units are represented by coupling terms that are potential functions. Hence

the two most fundamental concepts structuring the node and connection schemas are preserved: CONTAINMENT (of feet within phrases, syllables within feet, etc.) is implicit in the frequency relations between timescales, and CONNECTION (between systems) is a coupling interaction between oscillators, parameterized by a potential function. Note that gestural systems (as opposed to gestural planning systems) are not oscillatory; rather, these systems exhibit overdamped mass-spring dynamics. The relative phases of the gestural onsets are determined by the gestural planning systems, either as input to a gestural score as in Nam & Saltzman (2003), or preferably through coupling, which may obviate the need for a gestural score entirely. Figure 23 partly illustrates the relations between a traditional node and connection model and a multiscale wave dynamical model of the utterance *take on a spa*, focusing specifically on the syllable "spa".

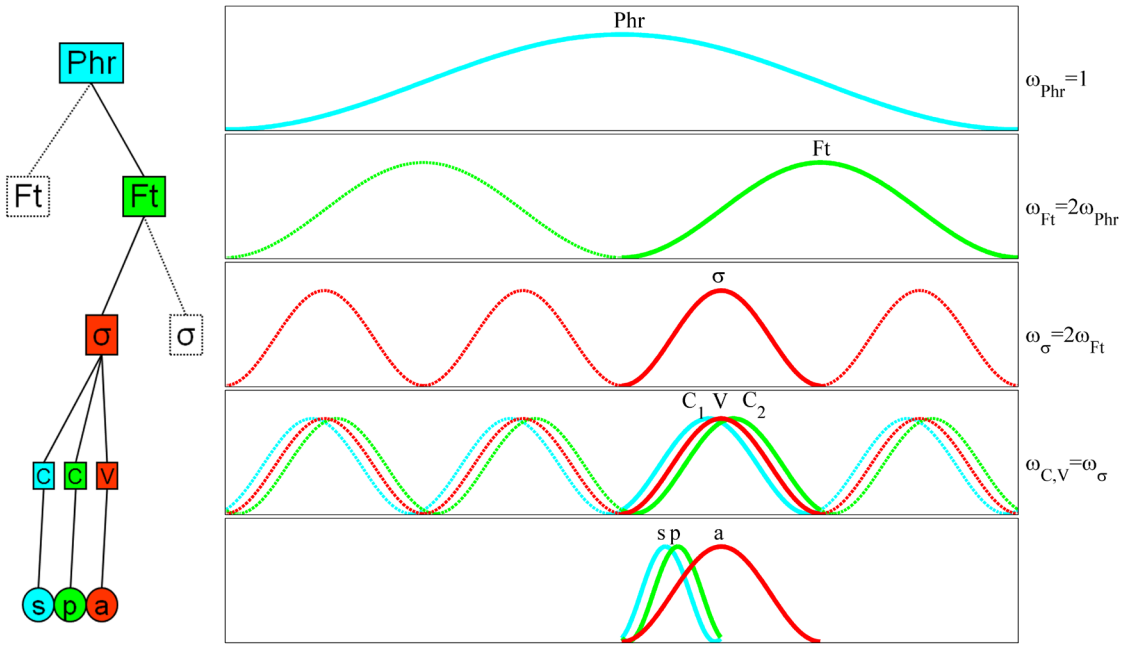


Figure 23. Relations between the hierarchical model of rhythmic/gestural units and a multiscale dynamical model. Units are conceptualized as oscillatory systems; prosodic and segmental levels in the hierarchy are conceptualized as timescales, which correspond to the inherent frequencies of their associated oscillators, and which are related by integer ratios. Gestures exhibit overdamped mass-spring dynamics.

There are many possibilities for exactly what form the equations describing a multiscale model can take, and in the end these should be resolved empirically. The issue of how the coupling relations between oscillators should be structured will be considered in Chapter 7. Crucially, regardless of how the multiscale dynamical model is formulated, the model predicts the possibility of interactions between disparate timescales. For example, consider repeated productions of the phrase *take on a spa*, in which the words *take* and *spa* are stressed syllables. In addition to rhythmic temporal patterns between the phrase and feet (inter-stress intervals), there are intergestural patterns that involve the tongue movement associated with [s] and the

lower lip movement associated with [p]. The multiscale dynamical model allows for the possibility of interactions between gestural timing and lower frequency foot and phrase timing. Chapter 4 presents the results of an experimental test of this idea. The reasoning behind it is as follows.

The correlation between variability and produce phase observed by Cummins & Port (1998) suggests that higher-order phrase-foot rhythms are more difficult to produce. A likely explanation for this difficulty is that foot-phrase coupling and syllable-foot coupling are weaker when governed by oscillatory systems with higher order frequency-locking. This accords with the finding that higher-order frequency-locking is relatively less stable, and that such relations follow from the relatively low amplitude of terms in the potential function that correspond to higher-order locks (Haken et al., 1996). Now assume that there are random noise-induced perturbations of oscillator phases at all timescales. It follows that if gestural systems are coupled to rhythmic systems, then intergestural timing should be more variable when rhythmic coupling is weaker. Equivalently, increased variability in rhythmic timing will be associated with increased intergestural variability. In the case of *take on a spa*, the relative timing of kinematic landmarks associated respectively with the tongue and lip gestures in [s] and [p] will be least variable when the phrase is spoken to a metronome target of $\Phi_{0.5}$, and will be more variable when the metronome target is $\Phi_{0.3}$, $\Phi_{0.6}$, etc. Chapter 7 will explain in more detail how this *rhythmic-gestural covariability* arises from multiscale coupling, but the gist is that more strongly coupled rhythmic systems exert a more coherent, stronger overall effect on gestural systems, thereby counteracting noise-induced perturbations of intergestural relative phase.

3.7 Neurological basis of the wave approach

One of the main technological hurdles to finding neurological evidence for the wave theory appears to be the difficulty of recording simultaneous intracellular potentials from a large number of neurons. Even in animals, where reduced safety concerns enable more extensive intracellular recording, the feat of recording even 1,000 neurons simultaneously remains to be accomplished—and it is likely that 10,000 neurons is a better approximation of the size of a local functional ensemble. Moreover, even if a sufficient number of neurons could be recorded, knowledge of their interconnections, including the precise axonal and dendritic conduction delays would be crucial to understanding their collective behavior. It is also important, depending upon how much detail one wants to consider, to understand dendritic currents separately from cell bodies. Even more challenging, no ensemble exists in a vacuum, nor is completely distinct from other ones, local and more distant, which means that knowledge of many other neurons would be probably be necessary. The difficulty of measuring all of this information (not to mention analyzing it) is currently prohibitive of any direct full-scale test of neural ensemble dynamical interactions.

The currently insurmountable obstacles to observing neural ensembles in vivo has led to a lot of interesting mathematical modeling of the interactions within and between neural groups of varying sizes, cell types, and organization. One outstanding approach is found in the work of Izhikevich (2005), who presents a model of 100,000 cortical spiking neurons, with an 80-20% ratio of excitatory to inhibitory neurons. Connectivity in the model is sparse, with the probability of connection between any two neurons only 10%. Each neuron is described by the simple spiking model (Izhikevich, 2003). This model of a neuron is intended to be as biologically plausible as the Hodgkin-Huxley model, but more computationally efficient. Different types of neurons, e.g. regular-spiking, intrinsically bursting, fast spiking, etc. are characterized by the settings of four parameters in the two-dimensional system of ordinary differential equations shown below.

$$v\,dv/dt = .04v^2 + 5v + 140 - u + I \qquad (3.34)$$
$$u\,du/dt = a(bv - u)$$
$$\text{if } v >= 30\ mV, \text{ then } v = c,\ u = u + d$$

In Eq. 3.34, v represents neuron membrane potential, u represents membrane recovery, and I represents input current (Izhikevich, 2003). The after-spike resetting rule sets v to c and u to u + d when v exceeds 30 mV. These equations are derived from experimental observations and bifurcation theory, allowing one to fit the spiking dynamics of a cortical neuron to a quadratic equation, such that time and mV scales are reasonably accurate. The variable u "accounts for the activation of K+ ionic currents and inactivation of Na+ ionic currents, and it provides negative feedback to v. After the spike reaches its apex (+30 mV), the membrane voltage and the recovery variable are reset according to [the resetting rule]. Synaptic currents or injected dc-currents are delivered via the variable I" (Izhikevich, 2003). The parameters a and b describe the time scale and sensitivity of the recovery variable u to membrane potential v, and the parameters c and d describe the after-spike resetting of u and v.

With the appropriate parameter settings, a wide variety of neuronal dynamics can be reproduced using these equations (Izhikevich, 2004). In the large model, Izhikevich (2005) only employs two different types of neurons: excitatory neurons and inhibitory neurons, and biologically plausible membrane voltage dynamics are represented. The excitatory neurons are "regular-spiking," which is reported to be the most common type of neuron in the cortex; they exhibit spike-frequency adaptation, such that in response to an step injection of direct-current, they fire a few spikes with high frequency and then their spiking frequency decreases. The other type of neurons in the model are fast-spiking inhibitory neurons, which fire spikes at high frequency and have very little adaptation (Izhikevich, 2003). The model has 1 ms. resolution, which is argued to be sufficiently precise. Synaptic connections have fixed conduction delays between 1 and 20 ms. Learning occurs because the synaptic weights change according to a spike-timing dependent plasticity rule (STDP), in which the change in a weight depends upon the timing between spikes of the pre-synaptic and post-synaptic neurons. The arrival of a pre-synaptic spike before a post-synaptic spike increases the synaptic weight, reflecting the notion that the pre-synaptic neuron had something to do with the post-synaptic spike; conversely, the arrival of a pre-synaptic spike after post-synaptic firing decreases the synaptic weight, reflecting the notion that the pre-synaptic neuron was not relevant to the firing of the post-synaptic neuron. This falls within the class of Hebbian learning rules. There is also a slow decay in the synaptic weights over time. In the initial conditions, all neurons in the model are randomly connected to 10% of the other neurons, and all synaptic weights are equal. Electric current (thalamic input) is then randomly input to one neuron every millisecond.

In simulations of this model, the network eventually reaches a balance between excitation and inhibition, and furthermore, exhibits two interesting behaviors: rhythmic oscillations in the delta range of 4 Hz, and in the gamma range of 30-70 Hz. Importantly, many distinct polychronous groups emerge. Izhikevich (2005) hypothesizes that polychronous groups represent memories and experience: "Coherent external simulation builds certain groups that represent this stimulation in the sense that the groups are activated when the stimulation is present". He also mentions an important property of polychronous groups, *incomplete activation*: "When a group is activated, whether in response to a particular stimulation or spontaneously, it rarely activates entirely. Typically, neurons at the beginning of the group fire with precise spike-timing pattern imposed by group connectivity, but the precision fades away as activation propagates along the group. As a result, the connectivity in the tail of the groups does not stabilize, so the group as a whole changes".

Two of the results of the Izhikevich simulations are of critical interest to the wave theory. First, the network produces oscillations at multiple frequencies. The 4 Hz (250 ms period) oscillation itself is a welcome finding, because of its potential relevance to speech system observable durations. Perhaps with more sophisticated and realistic models, a range of delta and theta oscillations (say 1-10 Hz) may emerge, which would correspond to behavioral observables. Of course, the cognitive planning wave frequencies need not necessarily correspond to observable linguistic durations in any simple way.

The second important result of the Izhikevich (2005) model is that "polychronous" neural groups emerge. These are not groups that spike simultaneously, but in a way such that the integrated spike rate of the group approximates the positive phase of a wave-like oscillation. This is nice because it justifies continuous wave approximations to the spiking activity within an ensemble. Taken together with the potential for multifrequency oscillations, the simulation

results suggest that a wave-based statistical approach to cortical neural ensemble dynamics is not horribly implausible.

An alternative (yet more removed and less precise) neurological observable is the local field potential, which can be measured more directly with extracellular recording or less directly with EEG (electro-encephalography) and MEG (magneto-encephalography). Both EEG and MEG have good temporal resolution, but MEG has better spatial resolution because the magnetic fluctuations induced by extracellular currents are less affected by the scalp. Fluctuations in the LFP as measured indirectly by EEG or MEG are partially representative of collective neural dynamics. Along these lines, the work of György Buzsáki points to neural mechanisms that may provide the basis for a wave theory of language. Buzsaki (2006), among others, has argued that coexisting brain rhythms can correspond to specific behavioral functions, and he has placed particular emphasis on rhythms in the theta range (4-10 Hz, 250-100 ms), cf. Figure 24. Theta rhythms are probably the best candidates for carrier waves of cognitive planning systems associated with syllable-internal structure. Buzsaki and others have shown how learning of temporal and spatial sequences in the hippocampus involves the association of events and places with different phases of theta cycles.

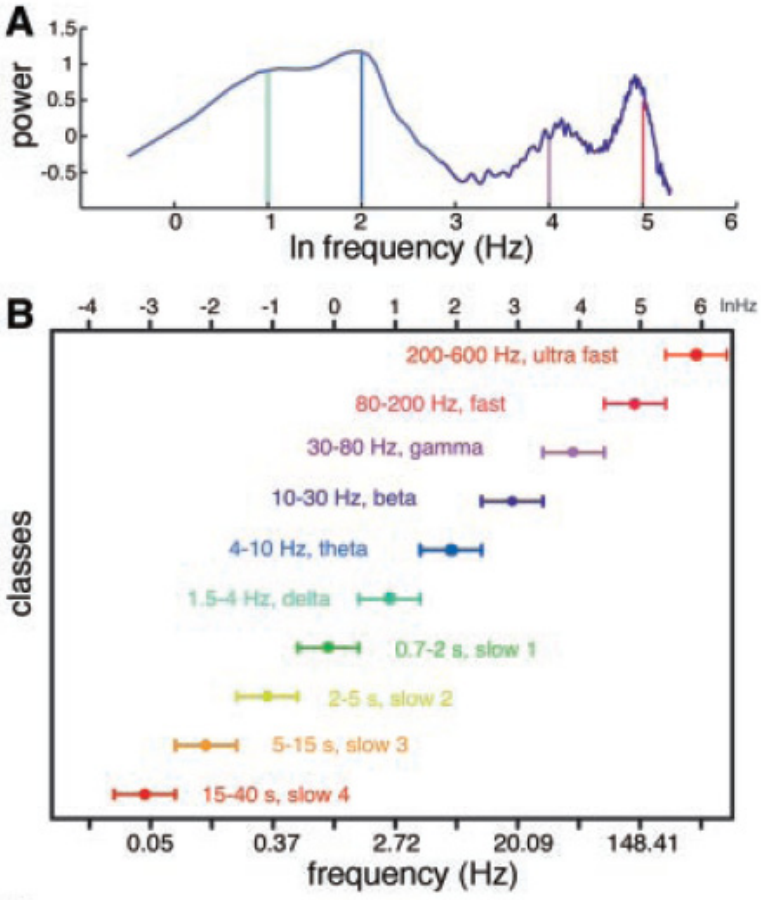


Figure 24. *From Buzsaki and Draguhn (2004).* Power spectrum of mouse hippocampal EEG (A), and commonly used rhythmic range labels (B). *Figure used without permission.*

In summary, work on realistic neural network modeling and studies of the local field potential provide suggestive neurophysiological evidence that points to the utility of thinking of neural populations as exhibiting wave-like dynamics. In this section, we have barely scratched the surface of such evidence. In the following chapters we turn to behavioral evidence to test the wave theory prediction that gestural and rhythmic systems are coupled.

Chapter 4: Covariability of speech rhythm and gesture

This chapter presents experimental evidence that rhythmic timing and gestural timing interact in an interesting way. This interaction is evidenced by rhythmic-gestural *covariability*, meaning that when rhythmic timing is more variable, intergestural timing is also more variable, and that measures of rhythmic and intergestural variability are significantly correlated. The experimental method is described in detail, and results and discussion are presented. Modeling of results is deferred to chapter 7.

To test for evidence of rhythmic-gestural interaction, a speech cycling task (as in Cummins & Port 1998) was conducted using the phrase *take on a spa*, and movements of the jaw, lower lip, and tongue blade were recorded using electromagnetic articulometry. The experiment was designed to address the following hypothesis:

> *Hyp*. *Rhythmic-gestural covariability*: in higher-order target phase conditions, (a) rhythmic timing will be more variable, (b) intergestural timing between [s] and [p] will be more variable, and (c) variability in rhythmic timing will be correlated with variability in intergestural timing.

The *target phase* is a ratio that represents the relation between the phrase duration and the duration of the foot (the inter-stress interval, i.e. the duration between *take* and *spa*). Note that "higher-order" here refers to the size of the denominator of the nearest integer ratio with a relatively small integers. Hence the lowest order target phase is 0.5, corresponding to a ratio of 1/2, and meaning that the foot duration is half of the phrase duration. The target phases 0.4 and 0.6 are relatively higher-order, because the corresponding integer ratios (2/5 and 3/5) have larger denominators. The target phases 0.3 and 0.7 are exactly 3/10 and 7/10, but there are nearby lower-order ratios 1/3 and 2/3. Hence the target phase conditions, from lowest to highest order, are: $0.5 < 0.3, 0.7 < 0.4, 0.6$. Exactly how target phase is defined in the context of the task is described below.

4.1 Method

4.1.1 Task and participants

Subjects were native speakers of English, ages 18-30. 6 subjects participated in two sessions, 1 subject participated in three sessions, and 1 in one session (due to scheduling restrictions). The phrase *take on a spa* was used for all subjects. At the beginning of their first session, the subjects were given instructions and practiced the $\Phi_{0.5}$ rhythmic condition. Sessions were organized in blocks of five trials, so that subjects performed the task with the same target rhythm for five consecutive trials, after which they switched to a different target rhythm. Each session consisted of 10 blocks, two for each of five different target phases. The order of blocks/target rhythms was restricted so that target phases in consecutive blocks always differed by more than 0.1, in order to reduce the possibility of carryover effects from one block to another. Block orders were randomly assigned to subjects. After performing each of the five

target phase blocks, subjects rested for several minutes, and then performed another five blocks in the same order. Thus each subject produced 10 trials per target condition in each session.

There was one major difference between the Cummins & Port 1998 (henceforth CP98) speech cycling task design and the one used here. CP98 continuously varied target phases from 0.3 to 0.7, meaning that the second metronome tone was located from 30% to 70% through the phrasal period. In the present design, target relative phases $\Phi = [0.3, 0.4, 0.5, 0.6, 0.7]$ were used. This allowed for a larger number of measurements at each of these data points, and was done partly due to time constraints on use of the articulometer. It is noteworthy that in an earlier version of the speech cycling task, Cummins & Port (1996) fixed the phrasal duration and varied the H to L tone duration. This is less useful way to manipulate the target phase for current purposes because it confounds articulatory timing with speech rate, and because both low and high target phases require relatively unnatural productions of the phrase. In contrast, keeping the H-L tone duration fixed better controls for speech rate, and avoids relatively unnatural rhythmic targets.

Following CP98, subjects were instructed to wait until the third repetition of the metronome pattern to begin on each trial, and to produce the phrase so that the word *take* coincides with the first (H) tone and the word *spa* with the second (L) tone. The metronome tones were 50 ms long and offset by 700 ms in all target phase conditions. The high tone was 1200 Hz, the low tone 600 Hz; both were windowed with a Tukey window ($r = 0.5$). The metronome pattern repeated 12 times from the start of each trial, and the last three pairs of tones were faded out by successive 50% decreases in amplitude. Subjects were instructed to continue repeating the phrase with the same rhythm after the metronome pattern faded out. After a duration of time corresponding to 14 cycles of the metronome pattern, they were signaled to stop. Subjects were told to speak with a normal volume and pitch, and not to tap their feet or hands, or to imagine doing so. They were instructed not to take shallow breaths between each repetition of the phrase, but rather, to take a deep breath when necessary and skip one cycle of the phrase. When taking a breath without the metronome, they were told to estimate the duration of one cycle.

To reduce the influence of preceding and subsequent context on the gestural trajectories of [s] and [p], a number of measures were taken in the design of the carrier phrase, *take on a spa*. It was deemed important to employ a four-syllable phrase, because a three-syllable phrase is produced abnormally slowly in the 700 ms intertone interval employed by Cummins & Port (1998). A guiding principle behind this design was that the tongue and lip gestures entering and exiting [sp] should be of large magnitude, in order to make landmarks associated with their articulatory trajectories easily detectable. The unstressed vowel [ə] preceding [sp] encouraged subjects to pass through a relatively neutral vocal tract configuration immediately prior to the articulations of interest. A phrase with the fewest potential coarticulatory confounds would require no tongue raising or fronting, and no large magnitude lip closure, protrusion, or retraction gestures in the two stressless syllables preceding [sp]. However, this ideal design is not achievable with the set of English function words. The function word *on* was chosen because pilot investigations revealed that both the vowel and alveolar nasal in *on* [an] tended to be assimilated in place of articulation with the preceding velar stop in *take* [teik]. Moreover, rather than a distinct tongue gesture for [n] occurring, a lengthened nasalized back vowel [ã:] was normally produced, resulting in the phonetic sequence [teˈkã:əspa]. The vowel following [sp] was chosen to be a low, central/back vowel [a] in order to maximize the speed of the releases

associated with [s] and [p]. No coda consonant followed the vowel, in order to further maximize movement amplitude and avoid confounds from anticipatory coarticulation.

4.1.2 Data collection

Kinematic data were collected with a sampling rate of 200 Hz, using a Carstens Articulograph AG200 (an electromagnetic articulometer, EMA; cf. Hoole, 1996). Four transducer coils were used, all in the midsagittal plane, in the following locations: (1) on the forehead for reference, (2) on the lowermost projection of the jaw when the subject looked straight ahead, (3) on the outermost projection of the lower lip, and (4) on the blade of the tongue, approximately 1-1.5 cm from the tongue tip. These sensor locations are shown in Figure 25 below. In EMA studies, a bite plate is commonly used to discern the angle of the occlusal plane relative to the transmitter coils. Unlike other EMA experiments, subjects were required to wear the (somewhat uncomfortable) transmitter helmet for a rather extended period of time (approximately an hour to an hour and a half during each session). This necessitated occasional readjustment of the helmet. The weight of the helmet is balanced by a line suspended from a bar above the head of the subject; this presents the dilemma that the more weight is taken off the helmet, the more susceptible to movement the helmet becomes, especially when subjects do not sit perfectly still. In addition, there is some variation in the helmet angle which subjects find most comfortable. Consequently, to estimate the occlusal plane throughout the experiment, one would have to recalibrate with a bite plate every time an adjustment is made, which is impractical.

However, the experimental hypothesis involves only within-subject comparisons of relative timing of gestures, rather than between-subject comparisons of articulator motions. For relative timing measurements, variation in the location of the occlusal plane is not problematic, and so no bite plate was used. Furthermore, because gestures are understood as synergistic organizations of effector movements, measurements are derived from lip-jaw and tongue-jaw synergies. There is thus no need to decouple tongue and lip movement from jaw movement, which is more accurately accomplished when two jaw sensors have been used (Westbury, Lindstrom, & McClean, 2002). Strictly speaking, the bilabial gesture associated with [p] is a synergy of jaw, lower lip, and upper lip movement, the last of which was not recorded. Since upper lip movement is normally of relatively lower magnitude than lower lip movement, changes in lower lip position are fairly representative of the timing of the bilabial gesture. One issue that may arise in a study which attempts to investigate the relative contributions of individual articulators to synergies in a speech cycling task is that these relative contributions are speaker-specific, task-specific, and vary within speaker and task across time (Alfonso & Van Lieshout, 1997).

It should be noted that the sensor coil on the tongue blade undoubtedly influences the observed articulatory patterns to some extent. Subjects may have produced abnormal alveolar closures to adjust for the presence of the sensor. Inspection of the audio recordings suggests that [s] was only minorly affected by the sensor, and it is unlikely that this had a profound effect upon temporal relations between [s] and [p], or that the adjustment differed significantly between rhythmic conditions.

Audio was recorded at 22050 Hz using a microphone clipped onto the shirt of the subject. A secondary audio signal was collected for synchronization with the EMA using a table

microphone. So that the metronome tone would not be recorded along with the speech signal, subjects wore earbud headphones. Inspection of kinematic data with and without the earbuds revealed that the presence of the earbuds in the magnetic field generated by the EMA helmet produced no noticeable increase in noise or signal distortion.

4.1.3 Data Analysis

To reduce signal noise, kinematic data were smoothed with an unweighted, 55 ms window moving-average filter. For each trial, the reference signal on the forehead was taken as the origin, and the data were rotated in the X-Y plane so that the mean horizontal location of the jaw was on the Y-axis. This rotation, which was generally less than 20˚, compensated for variation in helmet orientation and facilitated visual comparison of data across subjects, without significantly affecting relative timing measurements. Figure 25 shows how the coordinate axes are related to the sensor locations.

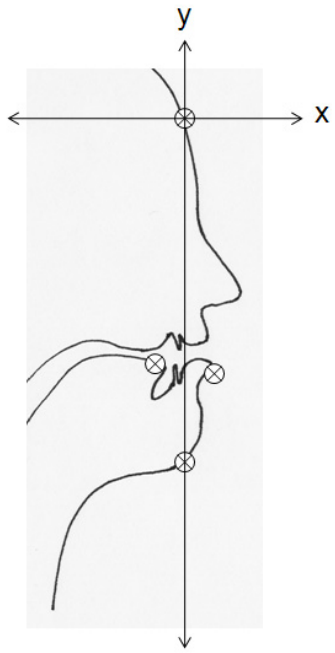


Figure 25. Sensor locations and kinematic data coordinate axes after rotation.

All horizontal and vertical minima (x, y), maxima (X, Y), velocity minima (dx, dy) and velocity maxima (dX, dY) in the neighborhoods of [sp] from *spa* and [t] from *take* were identified algorithmically in Matlab by looking for zero-crossings in difference vectors. In general, the horizontal motion (fronting) of the tongue from [ã:ə] to [s] was a higher amplitude and faster movement than the vertical motion (raising) of the tongue. For this reason, the maximum horizontal position of the tongue (Tj X) and the maximum horizontal velocity of the tongue (Tj dX) were deemed the most appropriate landmarks corresponding to the [s] gesture for all but one subject. Regarding the lower lip, both horizontal (Lj X) and vertical (Lj Y) positional maxima were found to be appropriately representative landmarks of the [p] gesture. The

appropriateness of these landmarks can be seen by qualitative inspection of 2D articulator trajectories. Figure 26 (b) shows that a rapid and relatively-high amplitude fronting of the tongue (Tj dX) occurs just prior to the second metronome tone (M2).

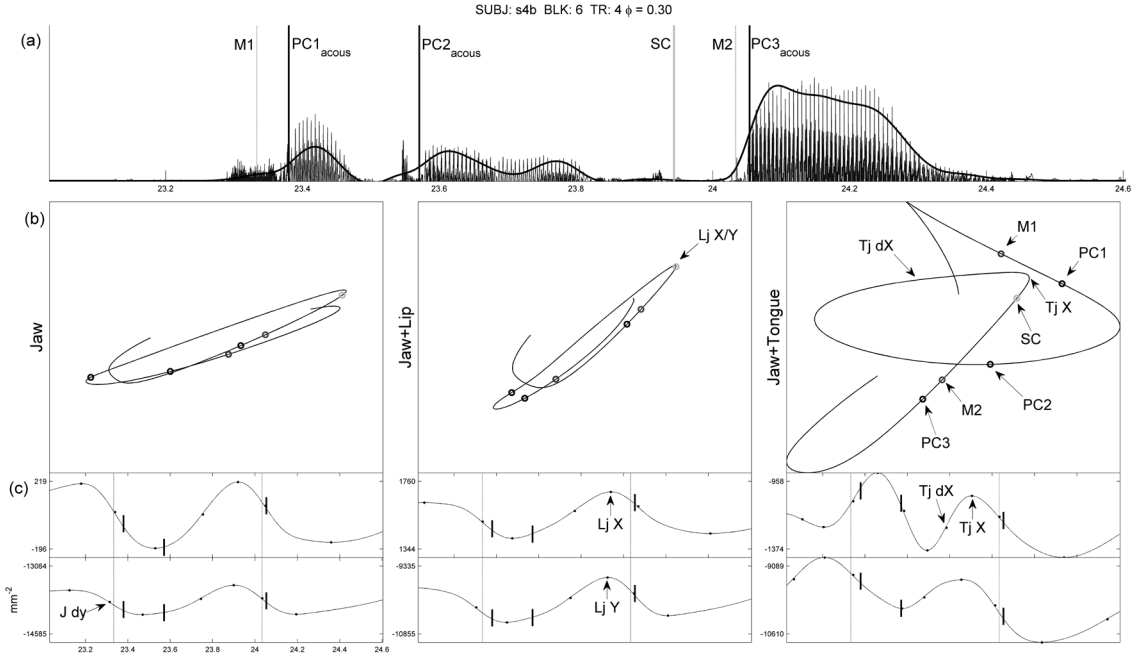


Figure 26. Example of acoustic and kinematic data, which are oriented as in Figure 25. (a) magnitude of the acoustic waveform, along with locations of metronome tones and acoustically-estimated p-centers. (b) 2D trajectories of the jaw, lower lip, and tongue sensors. (c) horizontal and vertical components of motion. See text for further details.

Interval durations between the [s] and [p] gestural landmarks described above (i.e. Tj dX, Tj X, and Lj X/Y) serve as the dependent variables in the intergestural timing analysis. The analysis is concerned with the variance of these durations, and compares within-subject variances in the higher-order Φ conditions to the variance in $\Phi_{0.5}$. These comparisons utilize ratios of variances, which are equivalent to F-tests for significantly different variance. Although relative phase measurements would be more conceptually direct objects of analysis, there are numerous challenges in defining gestural onsets and offsets—which is necessary to define phase variables—in connected speech of the sort utilized in this experiment. For this same reason, the extraction of phase angles in 2D Euclidean space is impractical; for less gesturally complex productions such as [babababa], a polar coordinate system and corresponding phase angle provides a more intuitive characterization of relative timing (cf. Kelso, Saltzman, & Tuller, 1986). Intergestural interval durations are suitable measures for variability analysis as long as there are no substantial differences in the rate of articulation of [sp] clusters across the experimental conditions. With a couple possible exceptions in $\Phi_{0.7}$ (cf. section 4.2), this assumption held true. Measures of the amplitudes of motion and durations between gestural landmarks will be used to explain several anomalous patterns associated with the $\Phi_{0.7}$ condition, but these measurements will not be of primary interest.

Estimated p-centers of the syllables *take* and *spa* served as the bases for rhythmic timing analyses. The p-center is an approximate location of the "beat" of a syllable. In early work, p-centers were equated with the temporal location of a finger tap made in time with a stressed syllable (cf. Allen, 1972, 1975). More recently the p-center has become a "perceptual-center," and its location in a single syllable can be determined using a dynamic rhythm-setting task in which the timing of a syllable relative to a repeating reference frame is adjusted with a knob (Morton, Marcus, & Frankish 1976). Numerous algorithms have been developed to estimate the locations of syllable p-centers (Howell, 1988; Pompino-Marschall, 1989; de Jong, 1994; Scott, 1993, 1998; Patel, Löfqvist, & Naito, 1999); however, there currently exists no consensus on how to predict p-center locations, nor on whether they correspond to gestural and/or acoustic events. While Cummins & Port (1998), following Scott (1993), used an estimation algorithm that treats the p-enter as a rapid energy-rise in the acoustic signal, others have argued that the p-center corresponds to a gestural event (Fowler, 1983; Tuller & Fowler, 1980), such as a maximum in jaw opening speed, or even a composite of gestural and acoustic events (de Jong 1994). Further, there are reasons to suspect that the locations of perceived and produced beats of syllables may differ (Treffner & Peter, 2002).

In the purely acoustic algorithm used by Cummins & Port (1998), the speech signal is filtered with a passband of [700-1300] Hz, using a 1^{st}-order Butterworth filter, which has gradual roll-offs. The effect of this filter is to greatly diminish spectral energy corresponding to F0 and higher-frequency fricative energy in the signal. The magnitude (absolute value) of the bandpass-filtered signal was then lowpass filtered using a 4^{th}-order Butterworth filter with a 10 Hz cutoff. The result is a smoothly-varying representation of mostly vocalic energy in the signal. P-centers are estimated to be the midpoints in time between the points when the signal amplitude is 10% above its local minimum and 10% below its local maximum. These estimates yield points that are near vocalic energy velocity maxima.

In the absence of a well-accepted algorithm that is robust to interspeaker and intersyllable variation, it was judged best to examine whether p-centers estimated from kinematic landmarks or from acoustic information were more closely-timed to metronome tones. It was apparent from visual inspection and verified quantitatively that points of minimum vertical jaw velocity (J dy), i.e. when the jaw was opening most quickly, were more closely and less variably timed with the 1st metronome tone (M1) than acoustically-estimated p-centers. This can be seen in Figure 26, where PC1 (acoustically-estimated) lags behind the metronome beat by approximately 50 ms (interestingly, this is approximately the VOT of [t^h]). In contrast, the acoustically-estimated PC3 was generally more reliably-timed with the 2nd metronome tone (M2) than any kinematic landmark. Hence, the rhythmic measure used here will take J dy as an estimate of PC1 and use the acoustic algorithm described above to locate PC3. The rhythmic φ variable for phrase *n* is $[PC3_n^{(acous)} - PC1_n^{(Jdy)}]/[PC1_{n+1}^{(Jdy)} - PC1_n^{(Jdy)}]$. Using a purely acoustic definition of this variable does not result in major changes to variability patterns, but does result in a shift of the observed phase distributions so that all distribution modes are earlier and less clearly reflect harmonic timing patterns. To illustrate φ_{PC3} distributions in section 3.1, smoothed histograms are shown (i.e. Gaussian kernel density estimates: 60 points, 0.015φ bandwidths). Also automatically detected in each phrase was the *s-center*, i.e. the sibilance center, which is the point in time corresponding to maximum sibilance energy associated with the [s] in *spa*. In this case, a passband of [5000-7500] Hz was used.

To identify locations where subjects took a breath, the acoustic data from every trial were visually and auditorily inspected; in the course of this process, spurious and missing p-centers

and s-centers were corrected. Where relevant in subsequent analyses, the data from each trial were separated into inter-breath groups, so that analyses were not confounded by the pauses associated with breaths. No φ is defined for the last phrase of each trial and inter-breath group, so these phrases were excluded from the analysis. In addition, the first trial of each session was excluded.

Each session yielded approximately 85-110 repetitions of the phrase in each target phase and metronome condition. Subject *s1* performed three sessions, all other subjects performed two sessions. One session from *s7* and one from *s8* were discarded due to equipment malfunction. In the remaining session performed by *s8*, kinematic landmarks failed to exhibit the same sort of consistency as those produced by the other subjects, making meaningful temporal analyses impossible—hence data from *s8* were not analyzed. One session from *s5* was not analyzed because the subject suffered from fatigue and exhibited a lack of attention to the task. Several subjects attempted two or three trials in $\Phi_{0.7}$ before they were capable of performing a full trial without halting--these trials were excluded.

The dependent variable of rhythmic-timing to be analyzed is observed phase (φ_{PC3}), defined in the manner described above as the timing of the p-center of *spa* (PC3) relative to the preceding and following p-centers of *take* (PC1). The dependent variables of intergestural timing are Tj dX - Lj X/Y and Tj X - Lj X/Y, corresponding to intervals between lower lip horizontal/vertical maxima and tongue horizontal positional and velocity maxima. Only intergestural results from phrases produced without the metronome are presented in section 3, although intergestural patterns were not substantially different with the metronome. Temporal patterns produced without the metronome are more theoretically appropriate tests of the covariability hypothesis, because movement patterns have been found to stabilize locally and globally with external perceptual stimuli present (Fink, Foo, Jirsa, & Kelso, 2000).

Approximately 2.0% of 12,645 eligible phrases were excluded because φ fell outside of the range 0.15 to 0.85 or was more than 2.5 s.d. from the mean phase for a given subject and target phase. Approximately 2.5% of phrases were excluded because one or more of the intergestural measures exceeded 3.0 s.d. from the mean value. Some of these outliers can be attributed to circumstances in which a subject misarticulated or hesitated abnormally. Overall more than 95% of eligible phrases were retained in the analysis.

4.2 Results

4.2.1 Rhythmic timing variability

Previous observations of increased rhythmic variability in the higher-order target rhythm conditions of the Cummins & Port (1998) speech cycling task were generally replicated both with and without the metronome present, albeit more consistently across subjects without the metronome. Harmonic timing effects were observed as well. Figure 27 shows rhythmic φ density estimates for each subject with the metronome (top), and without the metronome (bottom).

First examining target phase conditions $\Phi_{0.3}$ and $\Phi_{0.4}$, we can see that subjects *s1*-*s6* exhibited harmonic timing patterns in which one or both distributions were shifted toward 1/3. The distribution produced by *s7* is shifted toward 1/2 with the metronome, and is bimodal without the metronome, exhibiting shifts toward 1/2 and 1/3.

In target conditions $\Phi_{0.6}$ and $\Phi_{0.7}$, 3 of the subjects exhibited shifts toward 2/3, while 4 subjects exhibited shifts toward 1/2. This intersubject variation is consistent with what Cummins & Port (1998) observed: individuals in their experiment differed in regards to whether they exhibited evidence of the 1/3 and/or 2/3 phase attractors. Such intersubject variation is also consistent with findings in the domain of polyrhythmic drumming (Haken et al. 1996). Additionally, subject *s5* exhibited a large deviation of the $\Phi_{0.5}$ distribution away from the expected 0.5; several other subjects exhibited similar deviations of smaller magnitude. These deviations can most likely be attributed to inaccuracies in p-center estimation (cf. section 4.1.3), resulting from intersubject variation in the gestural/acoustic events associated with syllable beats.

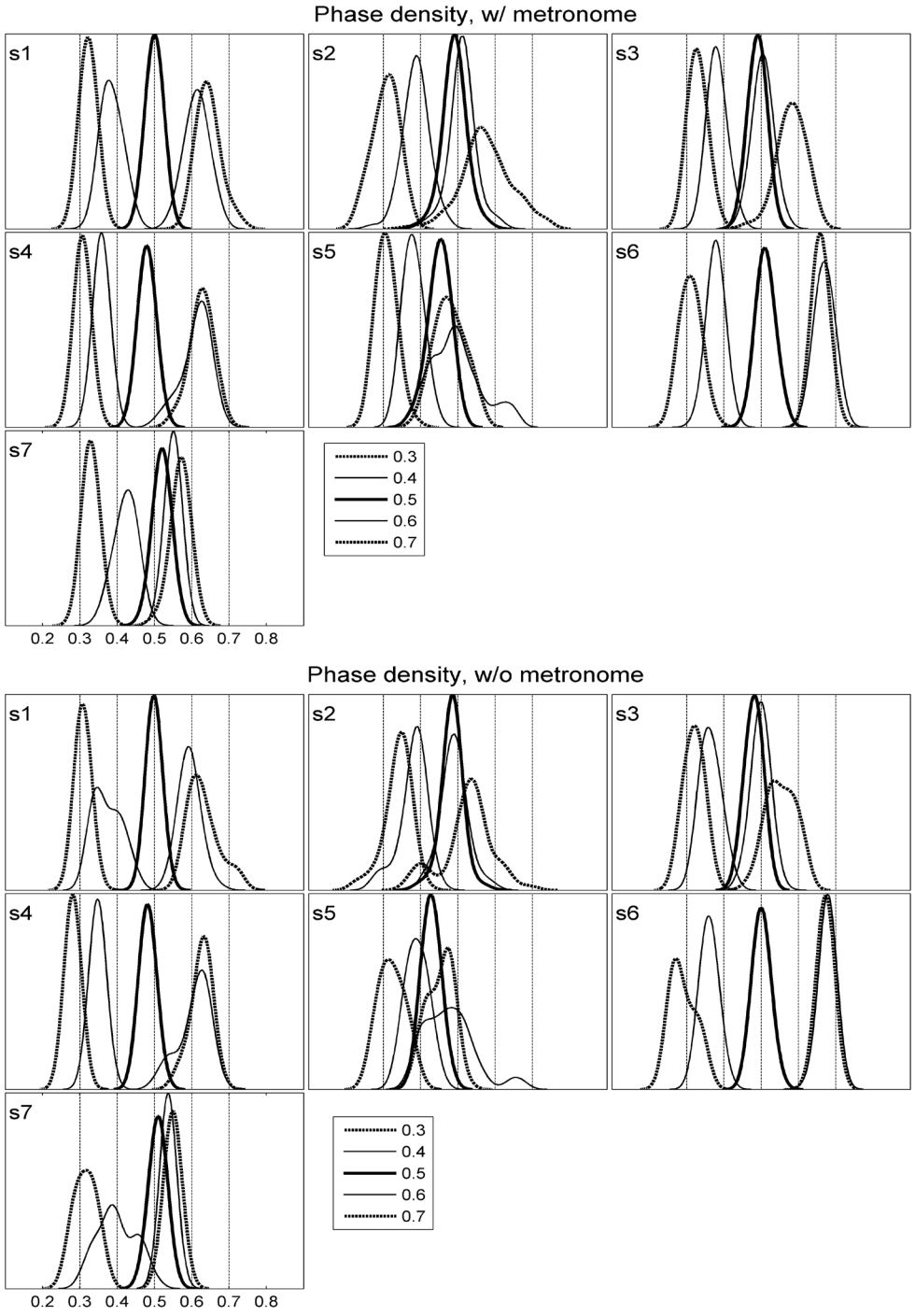
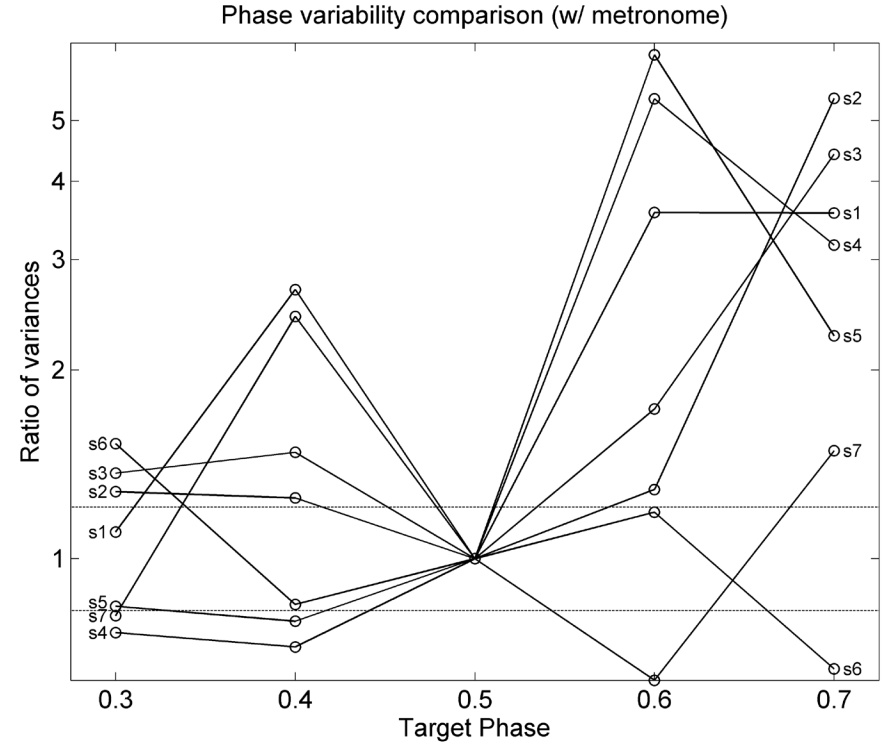


Figure 27. Observed phase densities (φ PC3) for each subject with the metronome (top) and without the metronome (bottom) show harmonic timing effects and a tendency for lower variability in Φ0.5 compared to the other target Φ.

Without the metronome, φ variability was generally lowest in $\Phi_{0.5}$. This accords with previous findings that the lowest-order harmonic mode of frequency-locking is the most stable, and is important for subsequent analyses in this study. With the metronome present, this

variability pattern was less consistently observed. Figure 28 compares phase variances within subjects across each Φ to the variance observed in $\Phi_{0.5}$, in phrases produced with the metronome present (top) and phrases without the metronome (bottom). Ratios of variances, which are plotted on the y-axis, are equivalent to F statistics used to test for significantly different variance.



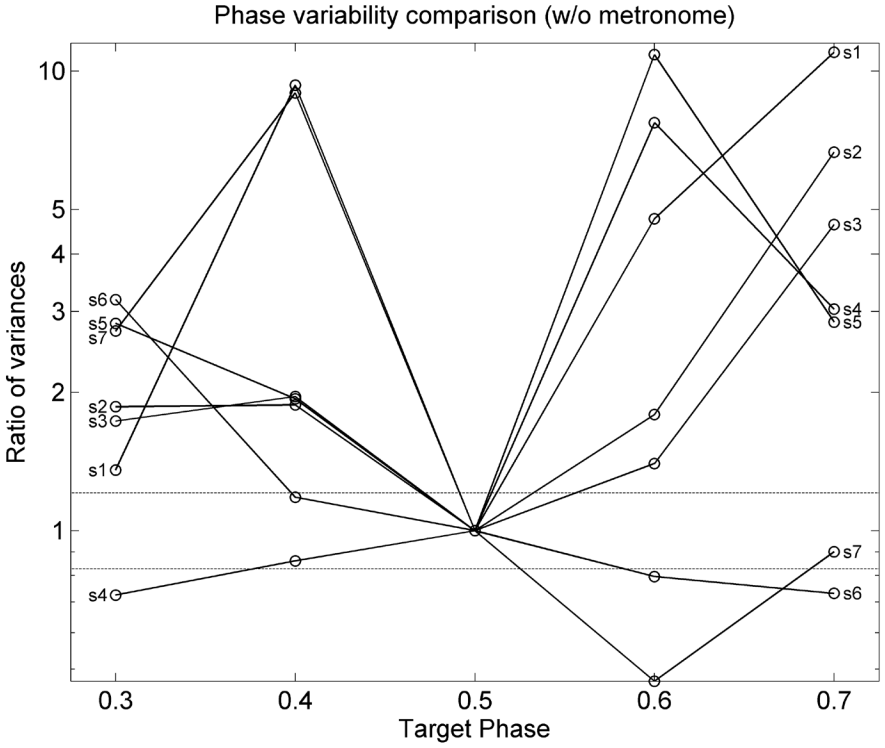


Figure 28. Comparison of rhythmic timing variability between target phase conditions. (top) with metronome; (bottom) without metronome. Ratios of variances in each Φ condition to Φ0.5 are plotted (on a logarithmic axis) for each subject. Horizontal lines show approximate confidence intervals for significantly different variance.

With the metronome present, rhythmic variability patterns in $\Phi_{0.3}$ and $\Phi_{0.4}$ were not consistent across subjects: some exhibited the expected greater variability, others were significantly less variable than in $\Phi_{0.5}$. A greater number of subjects produced the expected variability patterns with the metronome in $\Phi_{0.6}$ and $\Phi_{0.7}$, where there were only two instances of lesser variance. The cases of unexpected lower variance in higher order Φ may arise from an interaction between block order and changes in subject attention over the course of each session, although other unknown factors may be involved as well.

Without the metronome, the majority of subjects were significantly more variable in the higher Φ compared to $\Phi_{0.5}$. This finding supports the hypothesis that rhythms corresponding to the higher-order target phases are more difficult to produce reliably, and lends credence to the Cummins & Port (1998) and Port (2003) interpretation that such patterns result due to increased competition between relative phase potentials associated with 1:2 and 1:3 frequency-locked phrase and foot oscillators.

4.2.2 Intergestural timing variability

As hypothesized, intergestural timing between [s] and [p] gestures was more variable in the higher-order Φ. The gestural landmarks of interest here are the maximum horizontal position and velocity of the tongue (Tj X and Tj dX), and maximum horizontal and vertical position of the lip (Lj X and Y), which are associated with [s] and [p] respectively (cf. section 1.2 for discussion and illustration of how these measures are obtained). Figure 29 compares variances in each Φ condition to the variance in $\Phi_{0.5}$ for each subject, showing Lj X - Tj dX (top) and Lj X - Tj X (bottom).

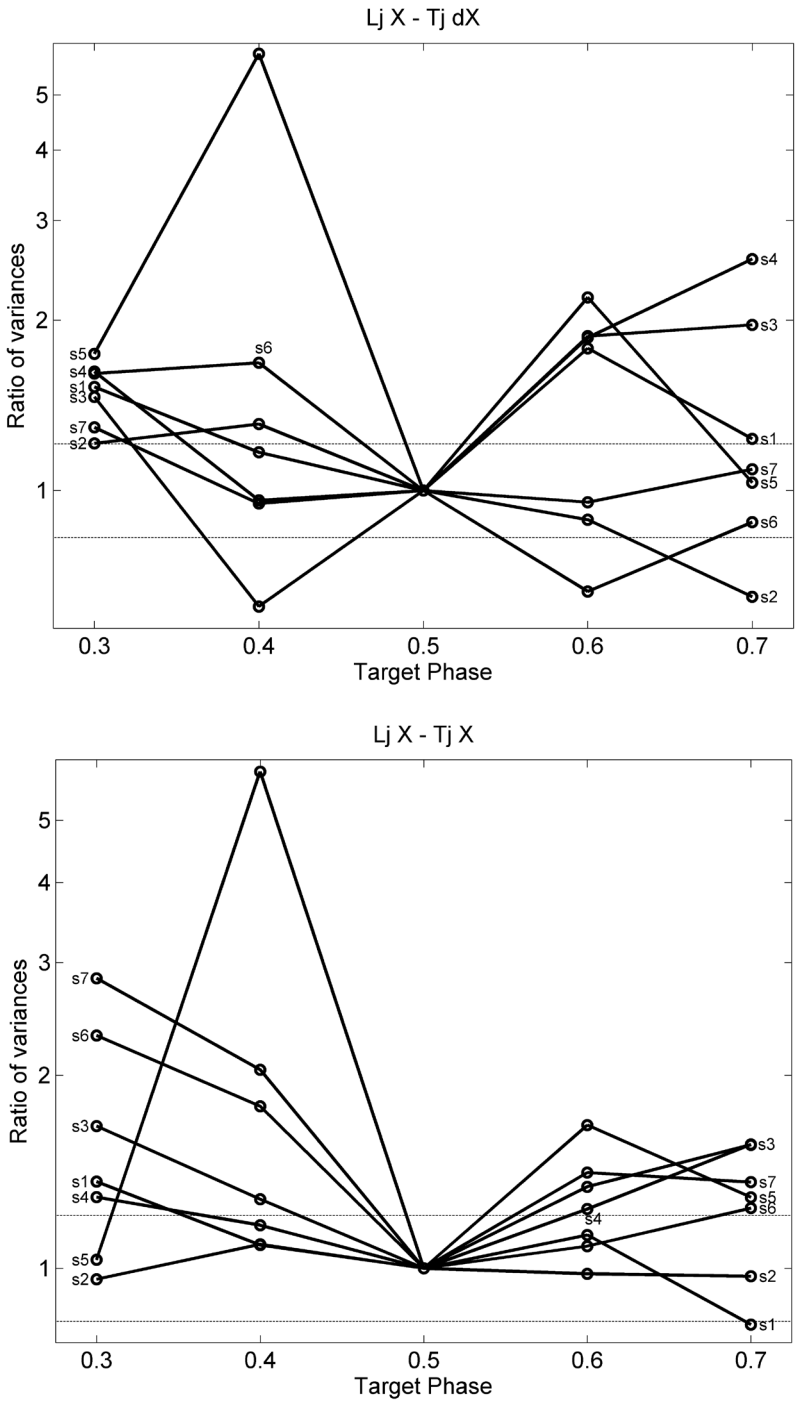


Figure 29. Intergestural variability comparison. Ratios of variances in each Φ to Φ0.5 are shown for each subject, from phrases produced without the metronome. (top) ratios of variances for Lj X - Tj X interval duration. (bottom) ratios of variances for Lj X - Tj dX interval durations. Horizontal lines indicate approximate 95% confidence intervals.

The measures in Figure 29 and their counterparts using Lj Y instead of Lj X are provided in Table A.1 in appendix A. Because subject *s6* exhibited an abnormally high standard deviation and mean interval duration in Lj X for $\Phi_{0.6}$, Lj Y was shown in the figure instead. Only kinematic measures from phrases produced without the metronome are presented here. The without-metronome condition is considered more appropriate for analysis for a several reasons. First, the expected rhythmic variability patterns were more reliably observed without the metronome (cf. section 2.1). Second, the auditory stimulus adds an additional source of temporal regulation that is not normally present in speech. As observed by Fink, Foo, Jirsa, & Kelso (2000), the metronome stimulus provides a local "anchoring" during which movements become less variable, and a doubly-anchored movement pattern (i.e. two metronome tones) can increase the stability of the entire pattern. However, analyses conducted on phrases produced with the metronome revealed qualitatively similar variance patterns across Φ to phrases produced without the metronome.

Both Lj X - Tj dX and Lj X - Tj X show greater intergestural variability for most subjects in $\Phi_{0.3}$ relative to $\Phi_{0.5}$. This finding was less consistent across subjects for $\Phi_{0.4}$ and $\Phi_{0.6}$, but nonetheless the majority exhibited either comparable or greater variability, with only two instances of lesser variability associated with the higher-order Φ in Lj X - Tj dX. In $\Phi_{0.7}$ the hypothesized variability pattern was observed in Lj X - Tj X intervals more so than in Lj X - Tj dX intervals.

One possible explanation for inconsistency in the relative variances between the higher target phase conditions $\Phi_{0.6,\ 0.7}$ and $\Phi_{0.5}$, is that some subjects may have substantially altered the amplitude of their articulatory movements because of the proximity of the upcoming onset of the next phrase. In other words, in $\Phi_{0.6,\ 0.7}$ rhythms, having to produce the syllable *take* relatively quickly after *spa* may have induced some subjects to decrease the magnitude of the [s] and [p] gestures. Subjects *s2* and *s5* showed relatively large reductions of gestural amplitude in $\Phi_{0.7}$ compared to the other conditions, which may explain why these subjects did not exhibit the expected effects (cf. Table A.2 in Appendix A for tongue gesture horizontal amplitudes).

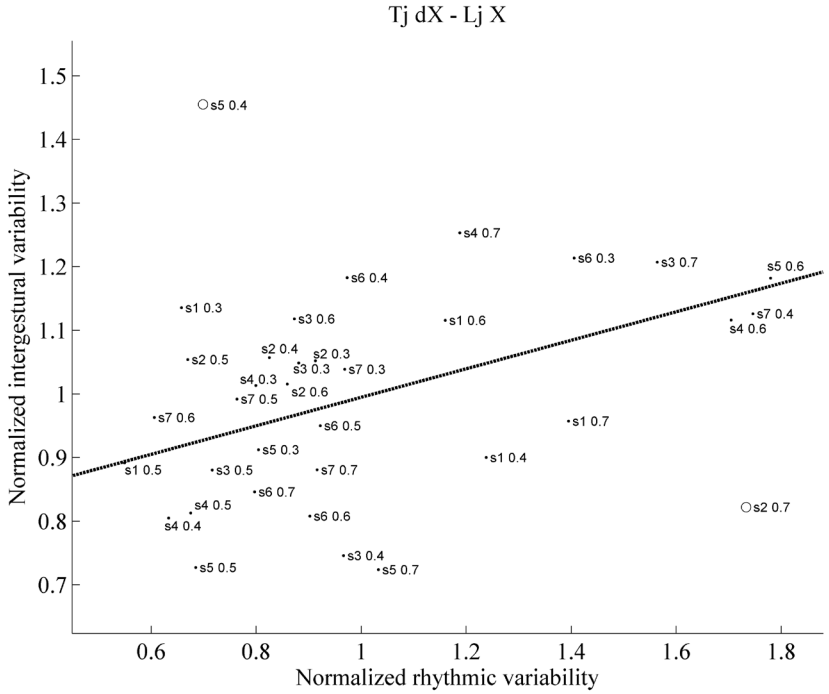
This observation raises the question of whether a relation between gestural amplitude and target phase may be the underlying factor determining intergestural variability patterns. It has been observed in interlimb coordination that as movement frequency increases, movement amplitude decreases (Haken et al., 1996; Beek, Peper, & Stegeman, 1995; Peper, de Boer, de Poel, & Beek, 2008). However, there are no obvious reasons for there to exist such amplitude differences among the lower target phase conditions $\Phi_{0.3,\ 0.4,\ 0.5}$, and by-trial ANOVA did not show $X_{Tj\ dX}$ - $X_{Tj\ X}$ amplitude to be a significant source of variance in intergestural standard deviations (Lj X - Tj dX: $F = 1.3$, $p < 0.25$; Lj X - Tj X: $F = 0.57$, $p < 0.45$) for those target phases.

Along the same lines, changes in intergestural interval duration across target rhythm conditions may play a role in influencing standard deviations. Examination of mean Lj X - Tj X and Lj X - Tj dX interval durations (cf. Table A.1) shows that several subjects (particularly, *s5* and *s7*) produced decreased interval durations in $\Phi_{0.7}$. Like decreased amplitudes, decreased durations should have the effect of reducing interval variance, and indeed analyses of variance confirmed this. The effect of duration raises the possibility that increased variability in $\Phi_{0.3,\ 0.4}$ may have resulted from increased durations relative to $\Phi_{0.5}$, but examination of durations shows that this was not the case; moreover, when analyses of variance were conducted on intergestural standard deviations in $\Phi_{0.3,\ 0.4,\ 0.5}$, duration did not have significant effects (Lj X - Tj dX: $F = 0.32$, $p < 0.57$; Lj X - Tj X: $F = 1.13$, $p < 0.29$), nor were there significant interaction effects

between duration and amplitude. Neither gestural amplitude nor duration appears to have influenced the variability patterns in $\Phi_{0.3,\ 0.4,\ 0.5}$. Furthermore, the effects of gestural amplitude and duration opposed tendencies for increased variability in $\Phi_{0.6,\ 0.7}$. Hence as hypothesized, increased intergestural variability was observed in the higher order Φ.

4.2.3 Rhythmic-gestural variability correlation

To determine whether variability in rhythmic timing and variability in intergestural timing are correlated, linear regression was performed on normalized variability measures. Normalized variability measures were computed within subjects, by dividing the standard deviations of all rhythmic phase (φ) and intergestural interval measures in each target phase condition by the mean standard deviation of those measures across target conditions. Figure 30 shows the data points and regression lines for both of the variability measures presented in the preceding section.



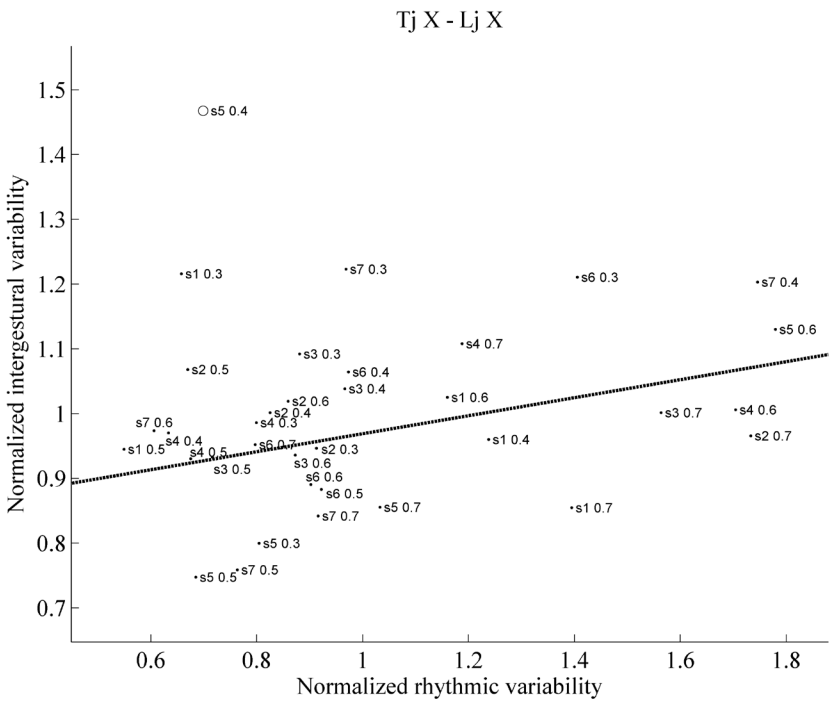


Figure 30. Rhythmic-intergestural normalized variability correlations. (top) Tj dX - Lj X intergestural variability; (bottom) Tj X - Lj X intergestural variability. Linear regression lines are shown, outliers (o) were excluded from regression.

Rhythmic and intergestural normalized variabilities were correlated significantly for Tj dX - Lj X ($R^2 = 0.26$, F = 10.85, $p < 0.005$), and for Tj X - Lj X ($R^2 = 0.20$, F = 7.71, $p < 0.001$). Three outliers (shown by open circles (o) in the figure) were removed from regression analyses; including these outliers results in marginally significant R^2. Within-subject normalization is appropriate because of absolute differences in standard deviations between subjects. These analysis establish that rhythmic and gestural temporal variation are correlated, and indicate that rhythmic and gestural systems interact in the planning and production of speech.

4.3 Discussion

The data confirm the hypothesis of rhythmic-gestural covariability: when rhythmic timing was more variable—which occurred generally in Φ conditions other than $\Phi_{0.5}$—intergestural timing between [s] and [p] was also more variable. Although there were several exceptions (some of which are attributable to changes in gestural magnitude in $\Phi_{0.7}$), the majority of subjects exhibited the hypothesized pattern. The following discussion is concerned with understanding what other factors may be relevant to the experimental data, and what sort of general mechanism is required to produce rhythmic-gestural covariability. Chapter 7 will show that a model of coupled oscillators that integrates the rhythmic and gestural dynamical models described in the introduction can simulate the observed covariability. The data speak to the existence of general principles regarding interactions between systems associated with different timescales, which may be applicable to a variety of human behaviors.

The key finding of the study was that the standard deviations of rhythmic timing and intergestural timing were correlated. These correlations were significant, but either variable only accounted for 20-25% of the variance in the other, which is not a very large amount. This begs the question of what accounts for the rest of the variance, which leads one to a laundry list of potential sources of variance. First, subject attention normally changes over the course of the experiment, such that target phase blocks performed later on may be more variable. Subject practice may have the opposite effect. If attention and practice have differing effects on rhythmic and articulatory behavior, then this noise would be magnified. The positioning of the EMA helmet is likely another source of noise; due to the natural discomfort subjects may experience wearing the helmet for extended periods of time, it must be adjusted, which induces small changes in the orientation of the helmet relative to the sensor coils, in turn introducing changes in sensor positions and the relative timing of kinematic landmarks. This source of noise is judged to be relatively small, but because it affects the intergestural measurements more so than the rhythmic ones, it may constitute part of the missing contribution to variance. Intersubject differences, particularly in rhythmic/temporal performance, presumably play a substantial role, and intersubject variation in gestural amplitude, relative gestural amplitudes across target phase conditions, and gestural durations, may contribute to variability as well. That a significant rhythmic-gestural variability correlation was observed in spite of the aforementioned sources of error speaks to the robustness of the effect.

Any explanation for the observed rhythmic-gestural covariability should be consistent with a number of related effects. For one, the harmonic timing observed between rhythmic systems in the phrase repetition task should be related in some manner to the mechanisms responsible for covariability. The relative stability of 1:2 and 1:3 frequency locking not only accounts for harmonic timing effects, but also for increased rhythmic variance in Φ conditions in which phrase and foot systems are related by 1:3 locking (or are subject to more interference between 1:2 and 1:3 frequency-locking). This is in line with the account proposed by Cummins & Port (1998) and Port (2003). A parsimonious understanding of rhythmic-gestural covariability should make use of this increased rhythmic variance in explaining intergestural variance.

For some subjects, gestural amplitude and intergestural interval duration were found to have effects on temporal variability patterns in $\Phi_{0.7}$. These effects should not be precluded by any model, but will not be of primary concern here. Because intergestural variability patterns in $\Phi_{0.3\text{-}0.5}$ were not significantly influenced by these factors, we should be able to understand the mechanism(s) behind the rhythmic-gestural covariability effect independently of amplitude and duration influences on variance. From a modeling perspective, not only would more data be required to better understand the effects of intersubject variation in intergestural duration and gestural amplitude, but changes to the data collection procedure would be necessary as well.

A difficult question here is whether phrasal tempo or frequency need be taken into account in understanding covariability. A cornerstone of the dynamical approach has been the idea that increased movement frequency corresponds to decreased coupling strength, particularly for higher-order frequency-locks. The crux of the difficulty is that—in contrast to gestural and rhythmic interval durations—the hypothesized gestural and rhythmic oscillatory systems are not directly observable. The interpretations of tempo or frequency variation in interval durations and in hypothesized oscillatory periods differ substantially. In the speech cycling task, the frequency of the phrasal period increases across the target phase conditions. But the foot durations do not change accordingly, since the first foot duration in the phrase remains close to 0.700 s in all conditions, and the second foot shortens to some extent only in the higher $\Phi_{0.6,0.7}$. Moreover, the

frequencies of hypothesized oscillatory systems governing foot timing vary, but not in proportion to the phrasal frequency: the foot oscillator is actually faster in $\Phi_{0.4}$ than in $\Phi_{0.5}$, and only a little slower in $\Phi_{0.3}$ than in $\Phi_{0.5}$ (cf. chapter 7). It is thus far from clear whether the speech-cycling task instantiates variation in tempo, and variation in frequency of hypothesized foot-oscillators is not monotonically related to phrasal period.

Likewise, observed gestural durations and hypothesized gestural planning frequencies are not simply related. Gestural durations do not increase all that much across conditions for most subjects, and are not oscillatory. In contrast, the hypothesized gestural planning oscillators have frequencies that are equivalent to the syllable oscillator frequency, which itself obtains a 2:1 or 3:1 ratio to the foot oscillator frequency. Thus, frequency varies in the systems hypothesized to govern relative timing, but generally not in the observables of foot and gestural durations. Whether frequency and tempo have effects on rhythmic and gestural timing should be investigated by simultaneous variation of both the phrasal repetition period and foot intervals, but cannot be appropriately addressed in the present context, where the period of the first foot was constant across target rhythms.

Despite the focus on coupled oscillatory systems in the modeling in Chapter 7, it should be noted that other classes of models might make similar predictions, under the right assumptions. One possibility is that limits on attentional resources can explain rhythmic-gestural covariability. If attention is a limited resource, and some attention must be allocated to perform both rhythmic and intergestural timing with accuracy, then circumstances in which one type of timing is more difficult (and hence requires more resources) should result in fewer resources being allocated to the other. If more attentional resources are allocated to rhythmic timing in the higher-order Φ, then there should be a concomitant reduction in attention to intergestural timing, leading to increased intergestural variability. One drawback to this sort of explanation is that it does not relate the phenomenon of harmonic timing to covariability.

Models based upon hierarchically-organized timekeepers (Wing & Kristofferson, 1973; Vorberg & Wing, 1996; Ohala, 1975) have been successful in explaining other sorts of variability patterns, such as negative lag-1 autocovariance, which has been argued to arise from motor delays. One might expect timekeepers to be similarly useful here. Can rhythmic-gestural covariability patterns be simulated with a group of timekeepers arranged hierarchically? The traditional assumption that a timekeeper is started by exactly one other timekeeper and then runs its course independently of that triggering event (Vorberg & Wing, 1996)--seems to prohibit such covariability effects from arising. Timekeeper models do not normally provide a way for one process to continuously influence another (or for feedback between processes), nor a way for multiple processes to simultaneously exert an influence over the timing of a single lower-level process. However, one cannot rule out that such models can be amended to accomplish such effects, and moreover, as Schoner (2003) has pointed out, timekeeper models can be reformulated as dynamical systems.

4.4 Summary and future directions

The experimental results demonstrate an interesting interaction between rhythmic and intergestural timing in speech. However, the phrase repetition task is highly unnatural, involving non-spontaneous, highly-practiced repetition and an external signal (the metronome) directing timing. This unnaturalness raises the question of whether the covariability effect obtains in

normal, spontaneous speech. Exactly how that question can even be addressed is not obvious either. To do so, one needs a metric for characterizing the "rhythmic variability" of normal, spontaneous speech. Of course, this measure would change rapidly, since the rhythm of speech, from a metrical perspective, varies fairly quickly.

One way to test for covariability in spontaneous speech is to compare intergestural timing variance between gestures occurring within a metrically constant stretch of speech with variance in timing between the same gestures occurring within a metrically less constant stretch of speech. "Metrically constant" speech could be defined as a succession of several feet with the same pattern of stressed and unstressed syllables, and non-constant speech could be defined as a succession of several feet with differing metrical structure. For example, if one were to look in a corpus (with both acoustic and kinematic data, of course), one could compare variability in trochaic speech sequences like "Sally spoke of swimming often" to sequences like "George really has spoken very often in public", where the former is metrically constant and the latter less so. However, effects from surrounding context may be hard to ignore, complicating such an endeavor. The following chapter takes a different approach to characterizing the rhythm of spontaneous speech, using a method of low-frequency spectral analysis, and correlating rhythmic measures with segmental deletions. This approach provides an empirically meaningful way to investigate how rhythmicity interacts intergestural variability in speech produced outside of the lab.

Appendix 4.A

Table 4.A.1 Intergestural interval mean durations, standard deviations, and sample sizes.

	$\Phi_{0.3}$		$\Phi_{0.4}$		$\Phi_{0.5}$		$\Phi_{0.6}$		$\Phi_{0.7}$	
Lj X - Tj dX	mean	(st. dev., N)								
s1	0.103	(0.042, 200)	0.091	(0.036, 227)	0.095	(0.034, 248)	0.107	(0.045, 265)	0.091	(0.037, 289)
s2	0.164	(0.030, 239)	0.139	(0.031, 216)	0.152	(0.027, 190)	0.114	(0.026, 186)	0.111	(0.022, 192)
s3	0.153	(0.030, 178)	0.130	(0.019, 148)	0.145	(0.024, 172)	0.133	(0.033, 173)	0.121	(0.034, 162)
s4	0.131	(0.028, 187)	0.110	(0.021, 187)	0.115	(0.022, 179)	0.157	(0.030, 158)	0.140	(0.035, 170)
s5	0.141	(0.024, 108)	0.122	(0.044, 92)	0.124	(0.018, 104)	0.100	(0.027, 107)	0.076	(0.019, 109)
s6	0.156	(0.027, 203)	0.145	(0.017, 191)	0.173	(0.028, 197)	0.222	(0.092, 168)	0.187	(0.036, 175)
s7	0.109	(0.023, 80)	0.105	(0.019, 82)	0.094	(0.020, 85)	0.065	(0.019, 99)	0.058	(0.021, 68)
Lj X - Tj X										
s1	-0.038	(0.041, 200)	-0.035	(0.037, 227)	-0.030	(0.035, 248)	-0.031	(0.038, 265)	-0.038	(0.032, 289)
s2	0.058	(0.026, 239)	0.044	(0.027, 216)	0.047	(0.026, 190)	0.029	(0.026, 186)	0.030	(0.026, 192)
s3	0.017	(0.035, 178)	0.005	(0.030, 148)	0.018	(0.027, 172)	0.017	(0.031, 173)	0.008	(0.033, 162)
s4	0.019	(0.026, 187)	0.017	(0.025, 187)	0.016	(0.023, 179)	0.046	(0.026, 158)	0.038	(0.029, 170)
s5	0.006	(0.020, 108)	-0.011	(0.047, 92)	0.000	(0.019, 104)	-0.017	(0.025, 107)	-0.030	(0.022, 109)
s6	0.018	(0.026, 203)	0.017	(0.021, 191)	0.027	(0.026, 197)	0.060	(0.091, 168)	0.043	(0.035, 175)
s7	0.007	(0.022, 80)	-0.005	(0.019, 82)	-0.017	(0.013, 85)	-0.044	(0.016, 99)	-0.052	(0.015, 68)
Lj Y - Tj dX										
s1	0.050	(0.056, 200)	0.055	(0.053, 227)	0.054	(0.049, 248)	0.038	(0.052, 265)	0.015	(0.050, 289)
s2	0.160	(0.030, 239)	0.138	(0.030, 216)	0.147	(0.030, 190)	0.118	(0.023, 186)	0.101	(0.027, 192)
s3	0.119	(0.018, 178)	0.114	(0.018, 148)	0.102	(0.014, 172)	0.092	(0.019, 173)	0.086	(0.027, 162)
s4	0.101	(0.020, 187)	0.084	(0.016, 187)	0.091	(0.019, 179)	0.113	(0.026, 158)	0.097	(0.025, 170)
s5	0.153	(0.025, 108)	0.146	(0.021, 92)	0.127	(0.017, 104)	0.107	(0.018, 107)	0.075	(0.017, 109)
s6	0.131	(0.034, 203)	0.099	(0.035, 191)	0.136	(0.027, 197)	0.178	(0.022, 168)	0.138	(0.025, 175)
s7	0.151	(0.022, 80)	0.141	(0.028, 82)	0.120	(0.017, 85)	0.101	(0.020, 99)	0.085	(0.017, 68)
Lj Y - Tj X										
s1	-0.091	(0.060, 200)	-0.070	(0.058, 227)	-0.072	(0.051, 248)	-0.100	(0.052, 265)	-0.113	(0.052, 289)
s2	0.055	(0.026, 239)	0.043	(0.026, 216)	0.042	(0.028, 190)	0.033	(0.024, 186)	0.019	(0.034, 192)
s3	-0.017	(0.028, 178)	-0.010	(0.031, 148)	-0.026	(0.018, 172)	-0.024	(0.019, 173)	-0.028	(0.024, 162)
s4	-0.011	(0.021, 187)	-0.009	(0.018, 187)	-0.007	(0.021, 179)	0.002	(0.021, 158)	-0.006	(0.020, 170)
s5	0.018	(0.018, 108)	0.013	(0.026, 92)	0.003	(0.017, 104)	-0.010	(0.018, 107)	-0.031	(0.020, 109)
s6	-0.008	(0.030, 203)	-0.028	(0.026, 191)	-0.010	(0.020, 197)	0.016	(0.020, 168)	-0.005	(0.022, 175)
s7	0.049	(0.025, 80)	0.030	(0.029, 82)	0.009	(0.017, 85)	-0.008	(0.022, 99)	-0.025	(0.019, 68)

Table 4.A.2 Intergestural movement amplitude (mm^{-2})

$X_{Tj\ dX} - X_{Tj\ X}$	mean				
s1	198	199	178	207	176
s2	227	185	219	173	151
s3	309	258	311	268	226
s4	185	133	136	161	146
s5	505	449	432	339	254
s6	558	531	598	575	530
s7	240	272	261	293	268

Chapter 5: Correlations between speech rhythmicity and segmental deletion

This chapter uses spectral analysis of speech rhythm to show that there are strong tendencies for deletion to occur more often in more rhythmic speech. This finding is important because it shows that rhythm and articulation interact. Presented here is a corpus study investigating how the rhythm of speech influences the likelihood of segmental deletion in conversational speech. As in the previous chapter, the aim is to test the hypothesis that higher-level prosodic timing interacts with lower-level articulatory timing. "Segmental deletion" here refers to *phonetic* deletion, which is the non-obligatory absence of a speech segment. It tends to arise from gestural overlap, and perhaps more indirectly from overlap of gestural planning systems. This contrasts with phonological deletion, which is generally categorical and obligatory, and which is commonly held to follow from rules and/or constraints specified by a phonological grammar. If overlap of gestural planning systems gives rise to segmental deletion, and if the relative timing of gestural planning systems is affected by the dynamics of higher-level rhythmic systems, then the prediction is that there should be correlations between speech rhythmicity and phonetic segmental deletion.

An important methodological question discussed here is how speech rhythm can be measured, quantified, and distinguished from rate of speaking. For current purposes, "rhythm" is best conceptualized as *periodicity*, and should be understood in a pretheoretical, phonetic sense, as opposed to a grouping based upon metrical categories. The quantification of rhythm in this study uses a relatively new methodology that employs low-frequency Fourier analysis of the amplitude envelope of vocalic energy in speech (Tilsen & Johnson 2008). This contrasts with previous studies of speech rhythm which generally have relied upon durational measures. A phonetically-transcribed conversational English corpus was used to identify deletions of consonants and vowels.

5.1 Methodological considerations

This section provides some background to the methodologies employed in this corpus study. Whereas most prior work on speech rhythm has used interval durations to characterize rhythmicity, the current study uses a spectral approach. This section also describes a method of artificial epenthesis that was used to investigating the rhythmic consequences of deletion.

5.1.1 Measuring rhythm: intervals

Most prior work on speech rhythm has focused on cross-linguistic differences, i.e. rhythmic typology. In contrast, my objective here is to understand how local rhythms in a given utterance influence the likelihood of deletion. However, the findings and methodologies of the cross-linguistic work provide an instructive point of departure. Rhythmic typology has long been concerned with a distinction between stress-timed and syllable-timed languages (Pike 1945). Stress-timed languages (such as English and Dutch) purportedly exhibit more regularity in intervals between stressed syllables, while syllable-timed languages exhibit more regularity

between successive syllables, regardless of stress. Abercrombie (1965) reformulated this distinction as the *rhythm class hypothesis*, proposing that interstress durations should be relatively less variable (more isochronous) in stress-timed languages, and vice versa for intersyllable durations in syllable-timed languages. However, there has been a general failure to find positive evidence for this distinction, or for any systematic cross-linguistic differences in isochrony of intersyllable and interfoot durations (Bolinger, 1965; Lehiste, 1977; Dauer, 1983).

A different sort of analysis has been used by Ramus et al., (1999) with some success. Their approach compares the ratios and variabilities of consonantal and vocalic interval durations in sentences read by speakers of various languages, and has been able to differentiate between syllable- and stress-timed languages. A plausible reason for the success of this approach may be found in the observation that syllable-timed languages tend to exhibit relatively less deletion, less vowel-reduction, and simpler syllable structure (Dauer, 1983).

Both of the approaches described above rely on the same type of measurements, namely, interval durations. These interval-based approaches measure rhythmic properties using durations between syllables, stressed syllables, moras, and sequences of consonants and vowels. In all these cases the interval endpoints are defined by pre-conceived units, such as C, V, μ, σ, and Ft. Importantly, these units are conceptualized as CONTAINERS, and the sizes (durations) of the containers represent their contribution to perceived and/or produced rhythm. So doing, these methodologies ignore the contents of the containers, or endow them with abstract labels such as "consonantal" or "vocalic". There are many phonological patterns that motivate linguistic units such as C, V, μ, σ, and Ft, but there is no *a priori* reason to believe these categories are the most relevant markers of rhythmic organization in speech. It is worth noting that in the acoustic speech signal alone, or in articulatory trajectories for that matter, these units and the intervals between them are not so well-defined. With that in mind, let us consider a very different approach to measuring rhythm.

5.1.2 Measuring rhythm: frequency

Spectral analysis of speech rhythm relies much less upon interval durations than previous methods of analyzing rhythm. There are two main steps in this approach. First, the speech signal is transformed into a slowly-varying representation of vocalic energy in the signal, called the *vocalic energy amplitude envelope*. Figure 31 illustrates this step with a 2.6 s stretch of speech in which a male speaker says "at least based on money raised it seems like…". Figure 31 (a) shows the vocalic-energy amplitude envelope superimposed over the bandpass-filtered signal, and Figure 31 (b) shows the amplitude envelope over the original signal.

The amplitude envelope is obtained by lowpass-filtering (4^{th} order Butterworth, 10 Hz cutoff) the magnitude (absolute value) of the bandpass-filtered (1^{st} order Butterworth, 700-1300 Hz) original waveform. The 700-1300 Hz passband Butterworth filter (which has gradual roll-offs in the frequency domain) provides a fairly good representation of vocalic and sonorant energy in the signal, primarily by eliminating high frequency noise due to frication and low frequency energy due to F0. Cummins & Port (1998), following Scott (1993), used this passband to approximate p-centers, which are the locations of syllable beats (Allen, 1972, 1975; Morton, Marcus, & Frankish, 1976). The fact that such energy is relatively loud and perceptually salient makes it a good source of a continuous signal corresponding more closely to rhythmic percepts.

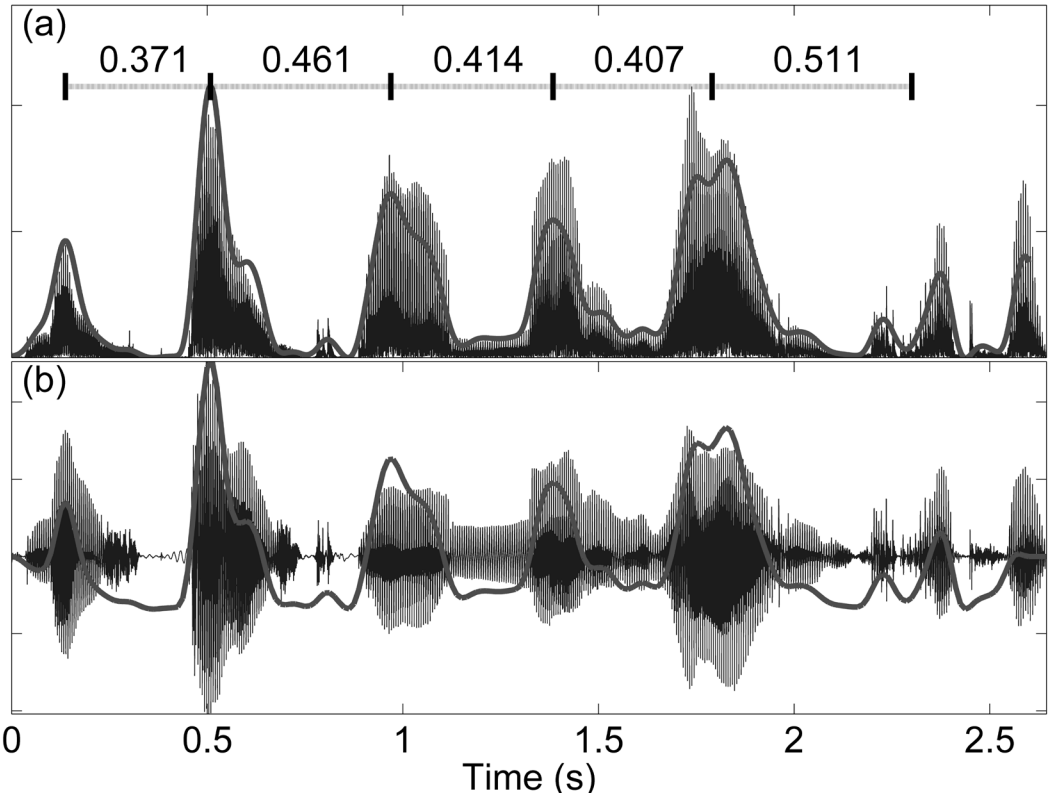


Figure 31. 2.6 s stretch of speech in which a male speaker says "at least based on money raised it looks like..." (a) lowpass-filtered magnitude of bandpass-filtered speech signal (b) amplitude envelope superimposed over original waveform.

Figure 32 compares the frequency responses of the 1st order Butterworth filter with 700-1300 Hz passband and a 4th order elliptical filter with 50 dB attenuation outside the passband and 1 dB of ripple in the passband. The Butterworth filter can be seen to have more gradual roll-offs from the passband in the frequency response. This has the effect of preserving relatively more spectral energy outside the passband in the filtered signal, preventing high and front vowels (with high F2 and low F1) from being neglected in the amplitude envelope.

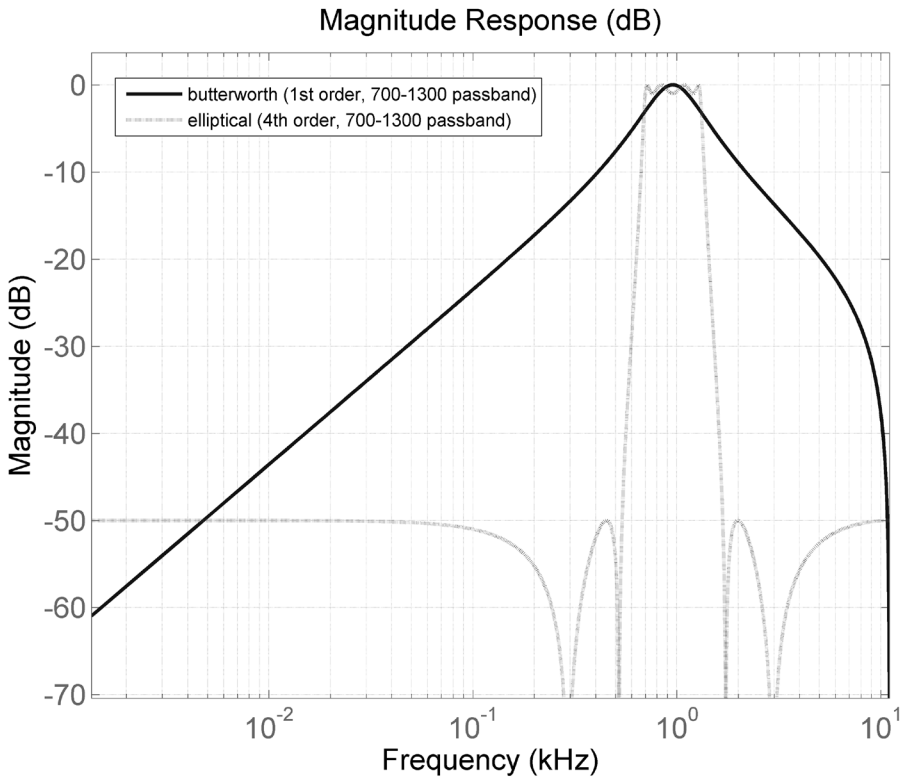


Figure 32. Comparison of Butterworth and 4-th order elliptical filters with 700-1300 Hz passbands.

To prepare the amplitude envelope for Fourier analysis, the mean is subtracted and the signal windowed (Tukey, $r = 0.1$), as can be seen in Figure 31Figure 33 (b). Then after zero-padding to N samples and normalization to unit variance, a Fourier transform is applied. The result is a frequency-domain representation of the signal, where the variance of the time series has been partitioned into components of differing amplitude at N analysis frequencies. The normalization of the amplitude envelope carries through to the sum of the magnitudes of the Fourier coefficients, which follows from Parseval's Theorem (c.f. Chatfield 1975; Jenkins 1968). The power spectrum, which is the squared magnitude of the Fourier coefficients, reveals the contributions of various frequencies to the amplitude envelope.

Figure 33 shows the power spectrum corresponding to the amplitude envelope of the utterance in Figure 31. Notice that the intervals between peaks in the amplitude envelope in Figure 33 cluster around 430 ms, and that this corresponds to a frequency of approximately 2.3 Hz in the power spectrum in Figure 33. The higher the spectral peak, the more regular (periodic) the rhythm of the stretch of speech being analyzed. The location of the peak on the frequency axis (horizontal) indicates the rhythmic components in the stretch of speech being analyzed.

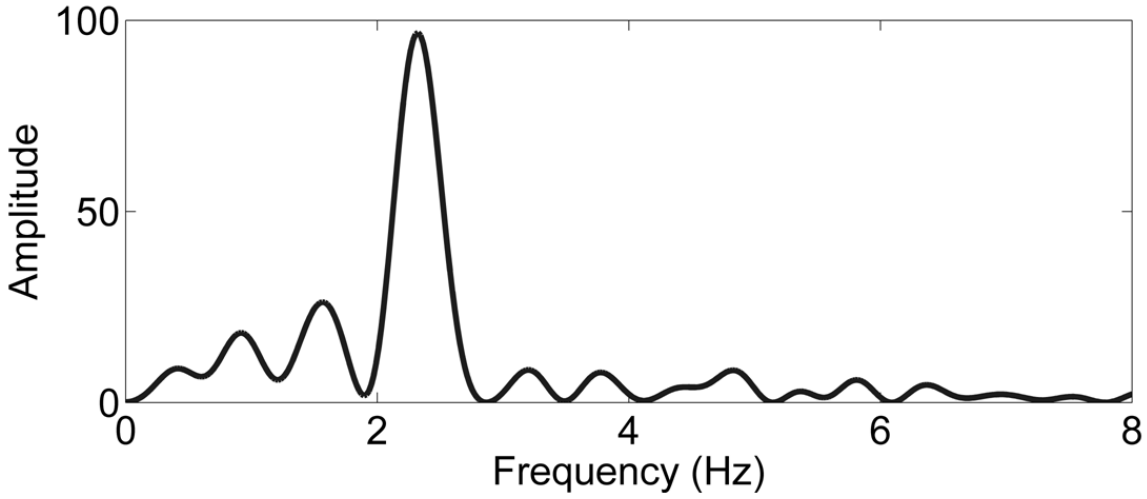


Figure 33. Power spectrum of the amplitude envelope in Figure 31.

The technique of Fourier analysis of a speech amplitude envelope is not a perfect method, one reason being that the signal is not of infinite duration or exactly periodic. This causes sidelobes (smaller bumps) to be present in the spectrum. This is mitigated to some extent by cautious windowing. Note that Hamming windows and the like more drastically alter the shape of the utterance and may not be entirely appropriate, whereas the Tukey window can be parameterized to distort less material. In general, only relatively high peaks in each spectrum can be taken to represent rhythmic characteristics of an utterance.

A fair argument can be made that the power spectrum is more appropriate for measuring rhythm than interval durations are. The spectral profile represents a sort of wisdom of the crowd: each sample in the amplitude envelope contributes to the power spectrum. This constitutes a substantial departure from interval-based approaches, since no intervals need be defined (other than the interval of speech to analyze). Further details of this method can be found in Tilsen & Johnson (2008).

5.1.3 Buckeye Corpus

Speech analyzed for this study was taken from the Buckeye corpus (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005). The Buckeye corpus contains approximately 300,000 words of conversational speech between interviewers and 40 central native Ohio English speakers from a balanced set of ages and genders. Transcribers using acoustic and spectrographic information labeled the corpus with phonetic transcription. This allows for occurrences of deletions to be identified by comparison of the phonetic transcription to citation forms taken from a pronunciation dictionary.

Stretches of speech ranging uniformly in duration from 2-3 s were extracted from the corpus. Figure 34 shows an example. The top panel shows the amplitude envelope and original signal, along with the citation, transcription, and deletions (in this case, the second vowel of “actually” and the first vowel of “Columbus” were deleted). The bottom panel shows the power spectrum (peaked line) against the average power spectrum for all 2-3 chunks (flatter line), and their 2.5 standard deviation region (filled).

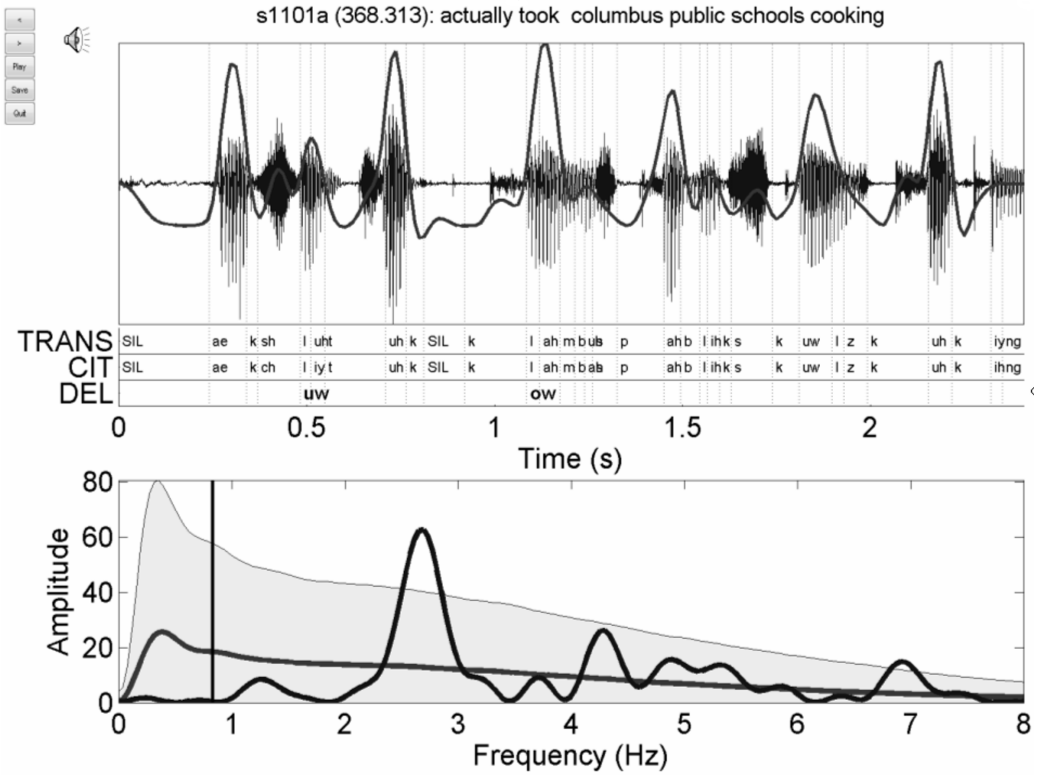


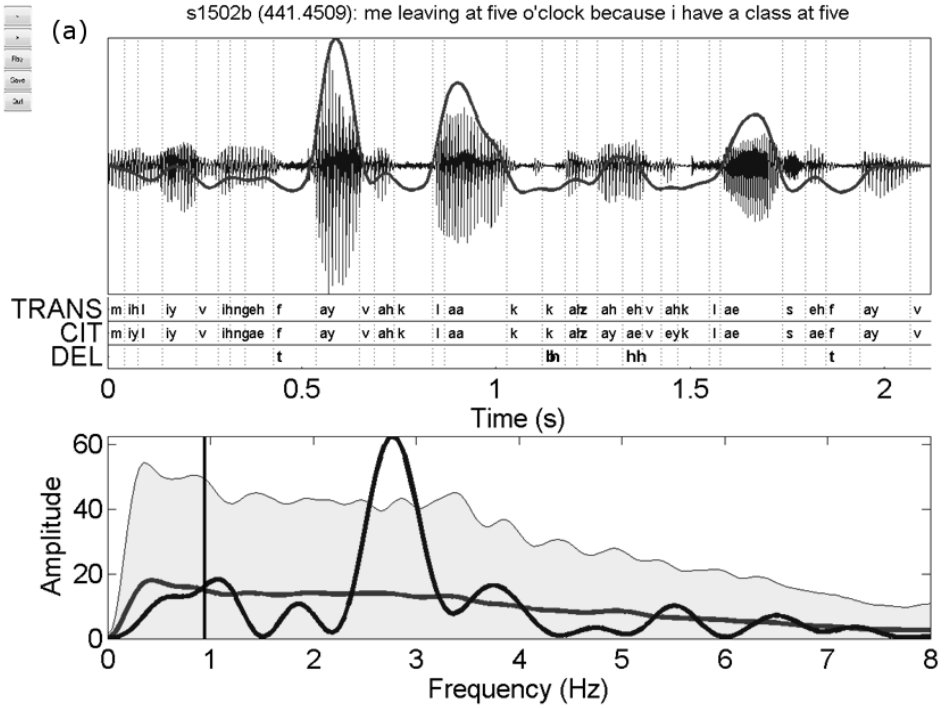
Figure 34. Example chunk of speech. (Top) waveform and amplitude envelope, along with citation, transcription, and deletions. (Bottom) Power spectrum (peaked line) against average (flatter line) and 2.5 standard deviation region (filled) for the entire corpus.

The vertical line in Figure 34 is at twice the frequency corresponding to the duration of the chunk (for a 2.5 s chunk, this is 0.8 Hz). Spectral peaks below this line are not analyzed because they do not represent true periodicities in the signal, i.e. amplitude variations that occurred at least twice. These false peaks arise due to the zero-padding, which is necessary to achieve detailed frequency resolution. Note that the range of chunk durations analyzed should accord with the rhythms one intends to study: analysis of longer chunk durations reveals lower-frequency phrasal rhythms while blurring higher-frequency syllabic rhythms. The 2-3 s range of chunk durations used here were chosen because this range is suitable for analysis of syllabic rhythms that occur on the timescale of several metrical feet. Note that chunks in the deletion datasets are centered on the deletion, so that the amplitude envelope before and after the deletion location contribute approximately equally to the spectrum.

Not all deletions identified in the corpus occur with the same frequency. A fair number of pseudo-deletions are attributable to instances in which the citation form of a word, which is taken from a standard pronunciation dictionary, includes a segment that never occurs or only very seldom occurs in that word in the dialects of the corpus speakers. Alternatively, some of these deletions may be categorical deletions subject to phonological rules. Here we are interested in phonetic deletions whose probability of occurrence is closer to chance. To avoid confounding categorical deletions with chance ones, only "active" deletions that occur between 25% and 75% of the time in their respective words are considered in the datasets below.

5.1.4 Artificial epenthesis

Rhythmic analysis can also be conducted by manipulating a stretch of speech in which a deletion has occurred. The manipulation, dubbed *artificial epenthesis*, involves the insertion of a segment approximately where the deletion occurred. This allows on to investigate relations between speech rhythm and vowel deletion with maximal control over the surrounding context. Chunks with a vowel deletion can be compared to identical chunks in which a schwa-like vowel has been artificially epenthesized. This method is unique to this thesis, and its advantage is that it allows for very precise and otherwise impossible degree of control over the context in which a deletion does/does not occur. Schwa-like vowels were constructed by synthesizing vowels with evenly-spaced formants at 500, 1500, and 2500 Hz. The amplitudes of schwas were modulated such that the maximum amplitude of the vowel was half of the maximum amplitude in a given chunk, reflecting the normally reduced intensity of these vowels in conversational speech. The schwa signals were Tukey-windowed (r = 0.25) and have a duration of 100 ms, which is fairly representative of the average duration of fully-articulated schwa. Figure 35 (a) shows the waveform with superimposed amplitude envelope for the utterance "...me leaving at five o'clock because I have class at five," in which the speaker omitted the first syllable of the word "because", and shows the rhythm spectrum below. Figure 35 (b) shows the same utterance in which a schwa has been artificially epenthesized.



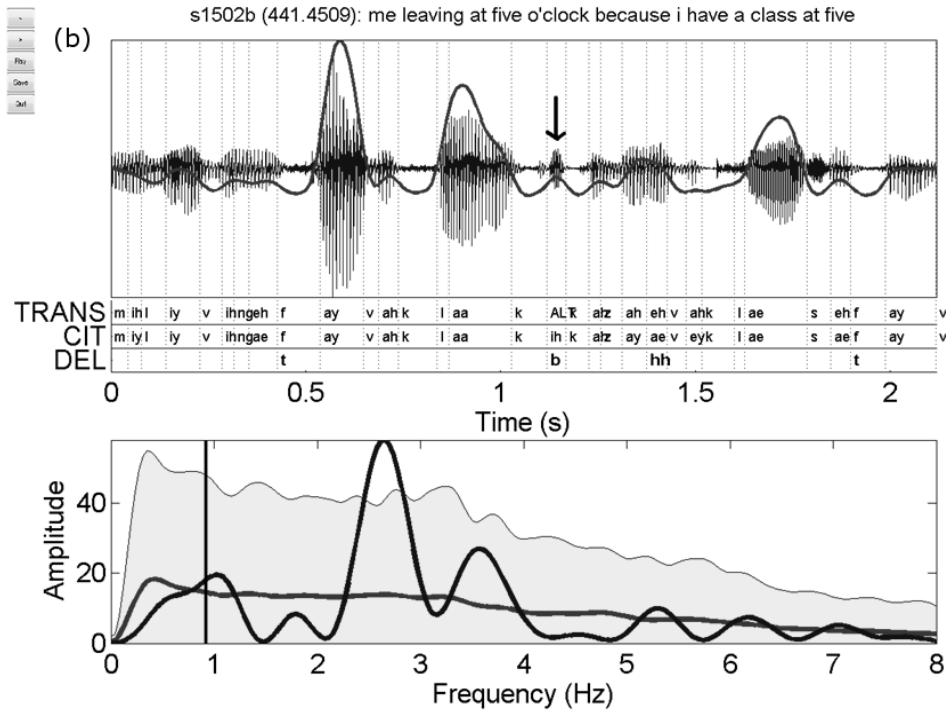


Figure 35. Example of artificial epenthesis. (a) original waveform with deletion of first syllable in "because," along with rhythm spectrum. (b) waveform with artificial epenthesis of a schwa in "because".

The rationale behind the artificial epenthesis method is to control for the speech context in the neighborhood of a deletion as much as possible. This approach effectively reverses the deletion of a vowel, allowing the rhythm spectra of no-deletion and deletion chunks to be compared in a paired manner. A potential drawback to this approach is that information derived from consonants surrounding the vowel (whether deleted or preserved) is not accurately represented in the waveform. Further, there may be changes in nearby undeleted vowels due to a schwa-elision; artificial epenthesis does not reverse these changes. However, the advantage to the approach—enabling much more precise control over surrounding context—may outweigh these drawbacks.

5.2 Results

The analysis of rhythmic patterns presented here employs a two-dimensional "rhythm space" in which the frequency and amplitude values of the *n* highest peaks from each spectrum represent the rhythmic characteristics of a dataset (here $n = 1$). These values are used to compute a 2-dimensional histogram, which is then transformed into a density distribution using a Gaussian kernel density estimator (this procedure is analogous to smoothing the histogram). Figure 36 shows two such density distributions, one derived from a dataset of chunks centered upon a consonant deletion (left) and the other for chunks centered upon a vowel deletion (right).

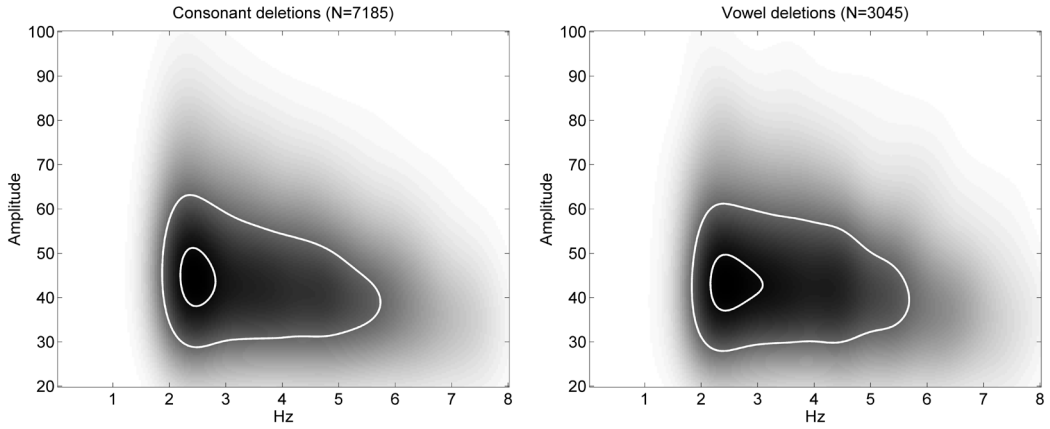


Figure 36. Rhythm space density distributions. (Left) consonant deletions, (right) vowel deletions.

The density distributions in Figure 36 indicate that the amplitude values of the highest peaks taken from each spectrum in the deletion datasets cluster around 40-50 normalized amplitude units, and the frequency values range from 2-6 Hz. White and black lines show 50% and 90% density contours. The distributions are more useful analytic tools when compared to other density distributions. They are particularly revealing when compared to the density distribution for a set of spectra taken from chunks without a deletion, as in Figure 37.

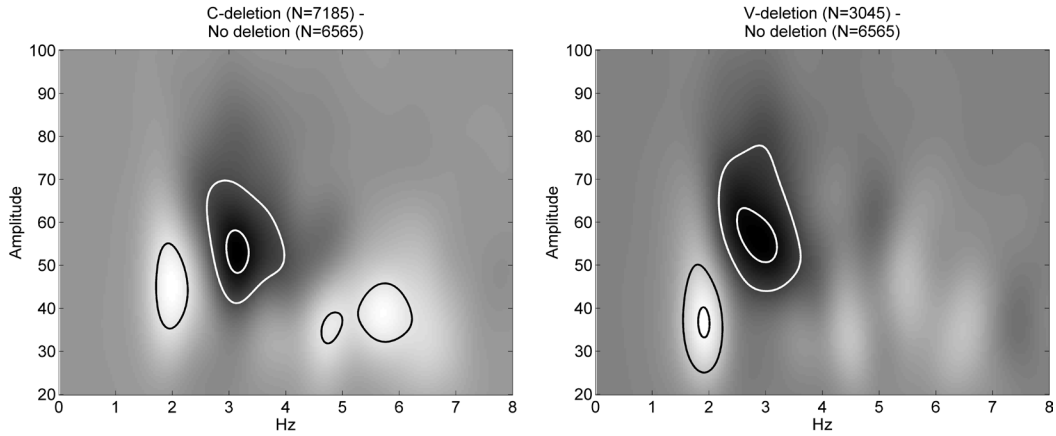


Figure 37. Rhythm space density differences between the no-deletion subset and the consonant (left) and vowel (right) deletion subsets. Light areas indicate a relative predominance of peaks from chunks with deletion, dark areas a predominance of no deletion. Lines trace 50% and 95% density difference contours.

Figure 37 shows that C-deletion is associated with low frequency (2 Hz) and high frequency (5-6 Hz) rhythms. V-deletion is strongly associated only with low-frequency rhythms. The light areas indicate a statistical predominance of spectral peaks taken from chunks with a deletion, and dark areas indicate a paucity of deletion (the absence of deletion, i.e. *preservation*). These analyses show that segmental preservation is associated with high-amplitude rhythms in the 2.5-3.5 Hz range.

It is convenient to use linguistically meaningful labels to refer to frequency ranges, even if there is a certain degree of arbitrariness in the definition of those ranges. For example, one can say that both C- and V-deletion tend to occur relatively more frequently in chunks with a dominant rhythm on the slow foot (1-2.2 Hz) timescale. C-deletion also tends to occur more frequently on the fast syllable (4-6 Hz) timescale. Preservation tends to occur more commonly on a fast Ft (2.2-3.5 Hz) and slow syllable (3-4 Hz) timescales.

The existence of a region of the rhythm space where spectral peaks from no-deletion chunks are relatively more prevalent suggests that there may exist a sort of "stability zone" in which segmental deletion becomes less likely, perhaps due to increased stability in timing of articulatory gestures. We will speculate on explanations for this and other patterns later on.

The analysis of Figure 37 does not take into account one important consideration mentioned earlier: namely, that we must distinguish speech rate from speech rhythm. In order to accomplish this, deletion and no deletion data subsets were constructed in which only chunks where speech rate (syllables per second) is within one standard deviation of the mean speech rate (approximately 5 σ/s, calculated over all datasets). This ensures that all speech considered will be within a normal range of speech rates, and makes the amplitude dimension of the rhythm space more meaningfully represent periodicity.

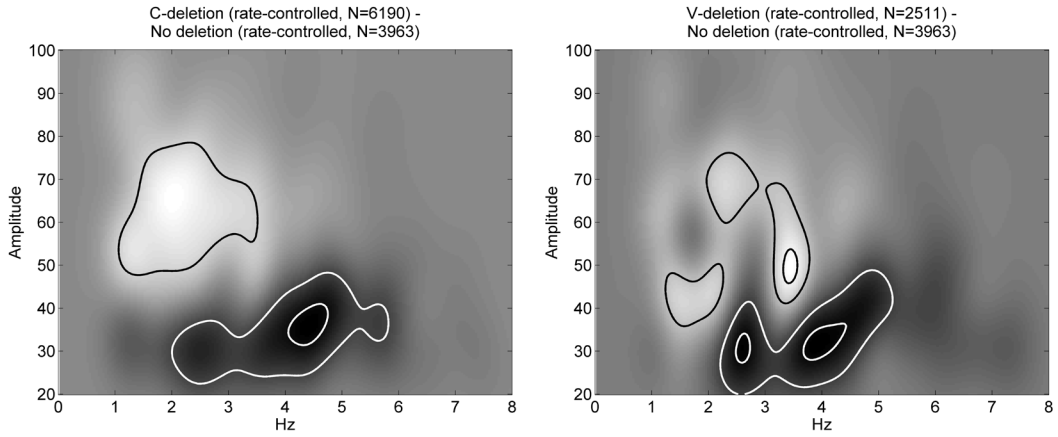


Figure 38. Rhythm space density differences between rate-controlled (± 1 s.d.) no-deletion subset and the consonant (left) and vowel (right) deletion subsets. Light areas indicate a relative predominance of peaks from chunks with deletion, dark areas a predominance of preservation (no deletion). Lines trace 50% and 95% density difference contours.

The rate-controlled density differences in Figure 38. show a markedly different pattern than that of Figure 37. Here we see that consonant deletions are associated with high-amplitude peaks in the slow-foot and fast-foot ranges (1-3 Hz). Vowel deletions are associated with higher-amplitude peaks (i.e. more periodic rhythm) in the slow-syllable range (3-4 Hz) and to a lesser extent in the fast-foot timescale (2-3 Hz). Preservation of both types of segments is more likely to occur with lower amplitude peaks (i.e. less periodic rhythms). This means that the effect of periodicity on consonant and vowel deletion is frequency-dependent, and this dependence differs between consonants and vowels.

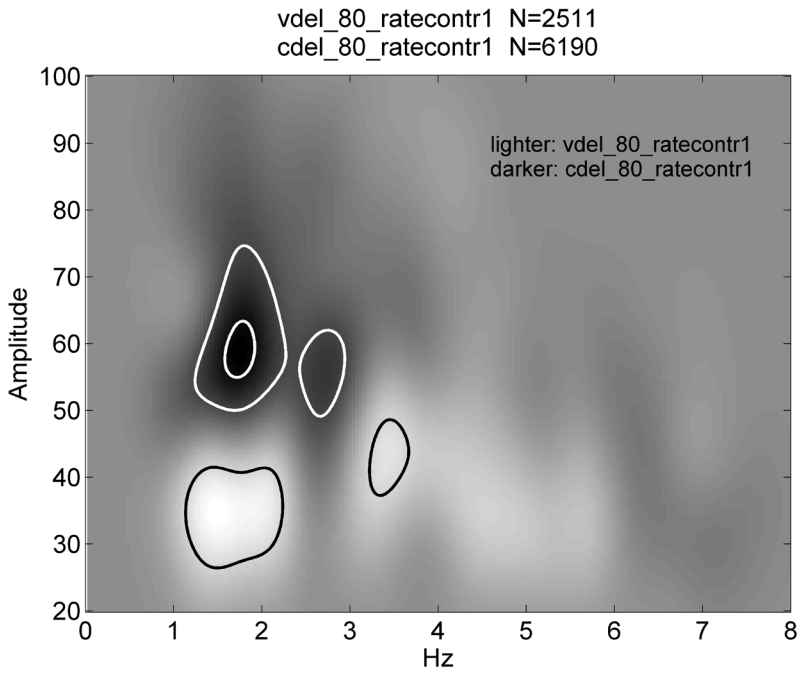


Figure 39. Rhythm space density difference between rate controlled vowel deletion and consonant deletion datasets.

The rhythm space density differences between rate controlled vowel deletion and consonant deletion datasets in Figure 39 further illustrate that C-deletion is much more strongly associated with high-amplitude foot timescale rhythms, and V-deletion is somewhat more strongly associated with slow-syllable rhythms and low-amplitude foot rhythms.

Now consider the density differences in Figure 40, which contrast a dataset with the first vowel of the word "because" deleted with a no deletion dataset (left) and a dataset in which an schwa has been artificially epenthesized (right). There are interesting similarities and differences between the two comparisons. The comparison between the deletion dataset and the artificial epenthesis dataset suggests that deletions of this vowel are associated with fast-syllable rhythms, while the artificial epenthesis of the vowel appears to bring out a slow-foot rhythm. The same correlations can be seen the comparisons of natural speech data, but there are intriguing additional patterns: vowel deletion is also associated with fast-foot timing and vowel preservation with slow-syllable timing.

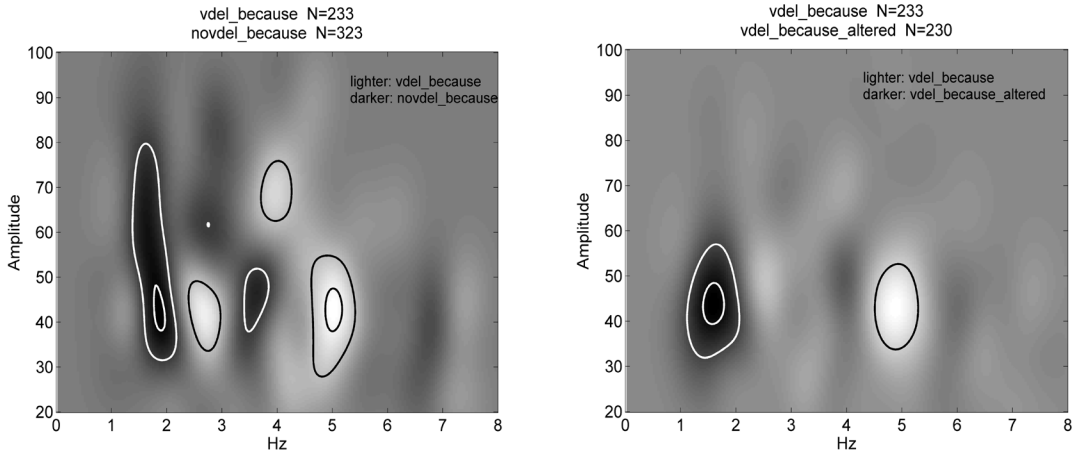


Figure 40. Rhythm space density differences between a vowel deletion dataset (the unstressed vowel in "because") and natural vowel preservation (left) and preservation through artificial epenthesis (right).

Why the rhythm space density difference involving the artificially epenthesized vowels lacks the additional density concentrations is hard to determine. One possibility is that the schwas that are preserved in the natural speech exhibit amplitude and durational variation that lends either to slow-foot or fast-syllable rhythms, whereas the artificial schwas contribute only to the former.

5.3 Discussion

The main findings from this corpus study are reiterated below:

- In data not controlled for speech rate, there exists a frequency range in the rhythm spectrum—corresponding to fast Ft timing—where deletion is less likely.
- In rate-controlled data, more rhythmic speech contains more deletion.
- Consonants and vowels differ with regard to the second finding. C-deletion is more strongly associated with high-amplitude rhythms in the Ft-timescale, while V-deletion is more strongly associated with high-amplitude rhythms in the slow-syllable timescale.

The first finding is somewhat difficult to interpret due to the lack of control for speech rate. Highly rhythmic speech in the fast foot and slow syllable timescales was strongly associated with preservation of segmental articulations. Without controlling for rate, we cannot know if the relative absence of deletion at those frequencies is due merely to high propensities for deletion in very slow, perhaps disfluent speech, as well as in very fast, hypoarticulated speech. In other words, rhythmicity (periodicity) and speech-rate may be confounded and no strong conclusions can be drawn. Indeed, if one takes the view that the patterns are due specifically to speech rate, then preservation may occur in the midrange of frequencies precisely because very slow speech is disfluent and very fast speech is hypoarticulatory. One way to address this in future work would be to do separate analyses for slow and fast speech.

The second finding is more theoretically interesting. It raises an important question: why is rhythmicity associated with deletion? One possible account hypothesizes an interaction

between rhythm and gestural phasing. This account can be best understood in the framework of task dynamics using concepts of dynamical systems theory. It posits that rhythmic and gestural timing are both regulated by systems of synchronized coupled oscillators, and further assumes that rhythmic systems interact with gestural systems through coupling. On occasion this has the effect of producing *rhythmically-driven gestural overlap*. This explanation is inspired, on the one hand, by work on rhythmic timing which has successfully incorporated the concept of dynamic coupling to explain observed patterns between syllables, feet, and phrases, and on the other hand, by work on gestural dynamics which has likewise captured observed temporal patterns.

To understand correlations between deletion and rhythm in this context, we need a mechanism that increases the likelihood of gestural overlap that is substantial enough to result in the perception of deletion. With a dynamical model of rhythmic and gestural systems, it becomes possible to explore interactional effects involving both domains. There are several logical possibilities in such a model. First, there may exist no interaction between rhythmic and gestural systems—in which case we would expect no correlations between rhythmicity and likelihood of deletion, assuming no other mechanism of rhythmic-gestural interaction is present. A second possibility is that rhythmic systems drive gestural systems, in which case rhythm should affect gesture, but not vice versa. The third possibility is the reverse: gestures drive rhythms. A fourth possibility is that rhythmic and gestural systems mutually interact. The interactional forces could be symmetric or nearly symmetric, or dominated by systems in one or the other domain.

The observed correlation between rhythmicity and deletion in the corpus data can be understood in the following way. Assume that rhythmic systems drive gestural ones—which means that there are forces which compel gestural systems to synchronize in-phase with syllables and feet. Lexical specifications between gestures work against these forces, by compelling adjacent gestures to synchronize in an anti-phase relation. Stochastic noise, transient perturbations, and contextual influences can subvert the intergestural phase specifications, especially if driving forces exerted by rhythmic systems on gestural systems become stronger. So, if one assumes that rhythmo-gestural interactional forces increase when speech is more rhythmic (i.e. when inter-rhythmic coupling forces are stronger), then gestures would be more likely to overlap. In other words, more rhythmic speech exhibits more gestural overlap, resulting in more deletion—precisely what was observed in the rate-controlled dataset. Chapter 7 explores this idea in further detail, and relates it to the covariability of rhythmic and intergestural timing observed in the experiment reported in chapter 4.

An alternative account of the deletion-rhythm correlation hypothesizes that that transcribers were perceptually biased to perceive deletion in more rhythmic speech. It is not inconceivable that given some indirect and non-causal rhythm-deletion association, listeners expect more deletion in more rhythmic speech. In that case, given acoustically similar sequences of segments, the sequence in a more rhythmic context is more likely to be transcribed without a segment.

While this *perceptual bias* hypothesis cannot be rejected outright, there are several reasons to be suspicious of its validity. For one, the corpus was initially automatically segmented, and then human transcribers were trained to use visual spectrographic information to correct the segmentation. This information and the automatic segmentation should have mitigated against any expectations based purely upon auditory information. Moreover, the effect of any such bias is unlikely to be large enough in magnitude to account for the full extent of the patterns.

A third account views speech rhythmicity as an emergent quality that arises from segmental deletion. There are two versions of this *emergent rhythmicity* hypothesis. In the teleological version, speakers employ deletion in order to make their speech more rhythmic, perhaps for stylistic reasons. In the non-teleological version, segmental deletion occurs randomly, and the effect of it happens to be more rhythmic speech. In both cases, some mechanism is needed to understand *why* the effect of deletion would be more rhythmic speech.

This mechanism may be comprehensible again in a dynamical framework. However, in this account, gestures drive rhythmic systems (or exert a stronger, asymmetric influence on them). Because the gestures are often anti-phase synchronized, their interactional forces upon rhythmic systems will have competing effects. These competing forces will induce greater variability in the timing of syllables and feet. The reasoning behind emergent rhythmicity is as follows. Segmental deletions arise from complete or extensive overlap of gestures, which occurs because the gestures have become less strongly anti-phase synchronized (for whatever reason), and hence more in-phase synchronized. In that case, the gestures would exert more coherent coupling forces on rhythmic systems, which means that the relative timing of those systems will be less variable—hence more rhythmic.

Both the rhythmically-driven gestural overlap account and the emergent rhythmicity account utilize concepts from dynamical systems theory, but differ with regard to whether gesture or rhythm is considered responsible for the observed correlations between deletion and rhythmic speech. It is possible, but perhaps more difficult to articulate, that both effects exist simultaneously.

The third finding of this study was that consonant and vowel deletion were differently associated with rhythmic frequencies. Caution should be exercised in interpreting this pattern, because of inhomogeneity in the dataset. For example, some vowel deletions correspond to the loss of a syllable, while others are offset by the syllabification of a following liquid or nasal. Likewise, some consonant deletions result in vowel hiatus, others occur phrase-finally, and various effects on syllable structure can result from such deletion. The diversity of these deletion patterns calls into question any account which distinguishes only between consonants and vowels.

5.4 Summary and future directions

Using spectral analysis to characterize speech rhythm, we have seen that there are strong tendencies for deletion to occur more often in more rhythmic speech. This is an important finding because it indicates that rhythm and articulation interact. It is incumbent upon speech researchers to understand the nature and causes of this interaction.

There are some noteworthy limitations on interpreting the patterns in the current study, which derive from the inhomogeneity of the deletion datasets. Future corpus analyses may remedy this problem in several ways. One way is to study rhythmic differences between deletion/non-deletion on a word-by-word basis, which introduces a much higher degree of control over type of deletion. Preliminary single-word analyses have unexpectedly revealed remarkably different rhythm-space distributions of deletion in different words; however, sample sizes were small—for this approach to be fruitful a large corpus is required. Another approach is to simulate deletion by removing a vowel from the signal, or to epenthesize a vowel where a deletion occurred. However, preliminary results using this method indicated that artificial

epenthesis only partly captures the rhythmic consequences of natural segmental preservation. It would be fruitful to further distinguish between types of vowel and consonant deletions, according to factors such as position within the syllable/foot/phrase, the featural content of the segment (e.g. voiced or voiceless, fricative vs. stop,), and segmental context (intervocalic vs. preconsonantal, etc.).

In addition, spectral analysis techniques for measuring rhythm may be further refined, which would improve our ability to conduct studies of the sort reported here. That said, this work has demonstrated the utility of rhythm spectrum analysis and illuminated a set of issues and directions of research that will hopefully lead to a better understanding of speech rhythm and articulation.

Chapter 6: Spatial patterning and episodic memory in speech planning

This chapter reports on a study that investigates some of the *spatial* aspects of speech planning. A primed vowel shadowing paradigm was used for this purpose. Spatial patterning is relevant to how we understand the preparation of articulatory targets. These targets can be understood as locations in a coordinate space, defined either by perceptual or articulatory dimensions. Articulatory targets are commonly thought of as static, invariant aspects of the speech production system. However, the experiment reported here demonstrates that contemporaneously planned vowel targets interact in such a way that the targets themselves are subject to change. These results can be understood in the context of an exemplar model of production using the concept of intergestural inhibition.

In what way are such patterns "spatial"? Consider that in preceding chapters we saw evidence that rhythmic systems interact with articulatory ones. These interactions are most naturally understood with temporal reasoning, and the evidence for them is temporal as well: rhythmic-gestural covariability is a matter of correlated variability in phasing and durational patterns, and segmental deletion can be thought of as the reduction of segmental duration (or perceived segmental duration) to zero. Indeed, a dynamical model of these phenomena, presented in Chapter 7, incorporates no reference to spatial dimensions.

Any model that lacks a spatial organization cannot be cognitively plausible, for the simple reason that the neural ensembles involved in the planning of speech gestures and other speech systems exist in cortical space. Because of somatotopy in primary motor cortex and somatosensory cortex, there are presumably spatial regularities in premotor cortex. One consequence of this cortical spatial patterning is that languages use different articulatory gestures. However, the meaning of "spatial" in this context is far from the low-dimensional sort of space we are used to experiencing with our bodies. There is no straightforward way to intuitively understand the organization of the *planning space*. It is very high-dimensional on the scale of neural connections, although a fair degree of reduction in dimensionality can be used to help conceptualize how these speech planning systems are embedded in the cortical space, and how they interact with each other.

Another way that "space" has been conceptualized in speech planning involves the notion of similarity. A similarity space is composed of sub-spaces, each of which has its own similarity metric; hence a global similarity metric can be defined by integrating component similarities. This is very useful for modeling, because it gets directly at our intuitive associations between proximity in space and similarity. In other words, to be close is to be similar. Thinking in terms of similarity space also avoids the thorny dilemma that the spatial organization of neural ensembles encoding speech gestures is not well understood. Both the similarity-based conception of space and the cortical space conception are useful and important for understanding how contemporaneously planned articulations interact.

6.1 Background

One reason that spatial considerations are important is that the formation of memories involves spatial organization, and memory formation has interesting effects on speech behavior. An important issue in modeling speech perception and production is how episodic memory

contributes to the representation of speech percepts and targets. Episodic (or exemplar-based) memories store various details associated with a specific event. In the case of a speech event, an episodic memory (or *exemplar*) includes information about the talker, location, time, associated emotions, as well as detailed auditory traces of the words in the utterance. One key difference between episodic and other forms of memory is that episodic memories refer to specific events; in contrast, non-episodic memories are believed to arise from integration of multiple experiences—hence in an adult listener, exposure to a speech event creates in episodic memory new examples of the linguistic categories involved in the utterance, but only very slightly alters the non-episodic representations of those categories.

Here we are concerned foremost with the question of how episodic memories influence speech production. For these purposes, we will sidestep the issue of whether speech targets should be conceptualized primarily as perceptual or motor memories. Undoubtedly, both perceptual and motor domains play a role in the cognitive representation of a speech target, and any satisfactory theory of speech requires some form of a mapping between these two domains (c.f. Liberman & Mattingly 1985). With this agnostic stance, the issue of perceptual vs. motor representation of targets becomes more about *when* and *how* perceptual representations are mapped to motor representations.

Regarding *when*, models of speech production can differ with respect to whether the mapping occurs online or offline in the planning and production of speech movements. This difference corresponds to distinct predictions about whether very recent perceptions can influence productions. Fowler et. al (2003) demonstrated that this mapping can occur online: in comparing response times in simple and choice shadowing tasks, they found relatively small (~ 25 ms) differences between the simple and choice conditions. This finding suggests that recent perceptions are mapped to gestures very quickly, because latencies in choice conditions—in which perceptual stimuli must be mapped to articulatory plans—do not greatly exceed those in simple conditions, where articulatory plans are prepared prior to the stimuli.

Regarding *how* the mapping occurs, the question we will address here is whether episodic memories (exemplars) or non-episodic perceptual representations are mapped to motor representations. In other words, do representations of speech targets incorporate subphonemic details of the sort that are retained in episodic memory, or are the targets derived exclusively from abstract, non-episodic memories? One can imagine that the perceptual-motor mapping involves the transformation of a relatively invariant and abstracted representation of a percept to similarly invariant gestural coordinates, in which case very recent perceptions should not influence productions. Alternatively, if episodic perceptual representations are mapped to gestures, then very recent perceptions may exert observable influences on productions.

In the domain of perception, there exists ample evidence to support the view that perceptual representations of speech include episodic memories; a variety of perceptual phenomena are usefully understood as arising from episodic memory of speech events. For example, Hintzman et. al. (1972) found that voice details of isolated spoken words persist in memory. Goldinger, Pisoni, & Logan (1991) found improved recall for 10-speaker wordlists compared to 1-speaker wordlists presented at slow rates, which suggests that interspeaker voice variation is stored in long-term memory along with lexical information. Palmeri, Goldinger, & Pisoni (1993) found that phonetic details are retained for several minutes, and Goldinger (1996) found similar effects lasting up to a week.

Models of speech perception which utilize episodic memory for the representation of speech can account nicely for such observations. Goldinger (1998) showed he could account for

his findings using the Hintzman (1986) MINERVA 2 model of episodic memory. The most relevant models for our purposes will be the speech perception model described by Johnson (1997, 2006), and the production model described by Pierrehumbert (2001, 2002), in which an exemplar is a set of associations between auditory properties (the output of the peripheral auditory system, including phonetic details) and category labels, which describe the speaker, setting, and various linguistic categories such as words or phonemes. Exactly what sort of information can constitute a category label and exactly how category labels arise is still not well understood, although presumably multisensory integration and associative mechanisms are involved in their formation. In these approaches, a new item is classified by comparing it to previously stored exemplars: each exemplar becomes activated according to its similarity to the new item, and the summed activation of the exemplars in a given category constitutes evidence the new item belongs to that category. Each exemplar has a base level of activation, which decays over time. Consequently more recent exemplars contribute more than older exemplars to the categorization of new percepts.

Exemplar approaches have several nice aspects not shared by non-episodic approaches. For one, an exemplar model obviates the need for talker normalization algorithms in perception. Because exemplar representations include category labels referencing age, gender, and other social variables, the process of perception is biased to perceive speech inputs that reflect the presence of those categories in a given context. Johnson (2006) shows that cross-linguistic variation in the effect of gender on vowel formants cannot be accounted for by hard-wired normalization algorithms, while an exemplar-based model can handle such variation. Exemplar storage can account for word-frequency effects in lexical recognition (Goldinger 1998) and phonetic categorization (Ganong 1980); assuming that the base activation level of an exemplar decays over time (memories decay), more frequent words will have higher summed activation levels because they are associated with more exemplars—hence all other things being equal, a new item is more likely to be classified as a member of a more frequent category.

In the domain of production, there are a variety of effects that suggest an influence of episodic memory. For one, subphonemic influences on productions were found in the shadowing task conducted by Fowler et. al. (2003). In their experiment, there were two conditions, a simple task in which listeners shadowed the first V of a model VCV sequence and then produced a predetermined CV, and a choice task in which listeners shadowed the entire model VCV sequence. In addition to finding only a relatively small response time difference between the simple and choice tasks, the researchers found adjustments in VOT toward the model VOT—hence a subphonemic detail of the percept was found to influence production on a very short timescale. This suggests that an episodic representation of the model VCV was quickly mapped to a motor representation that incorporated subphonemic information. Goldinger (1998) also presented evidence for subphonemic influences on production in shadowing wordlists produced by a variety of talkers. Recently, Babel (2009) showed that exposure to auditory stimuli shifts productions in the direction of those stimuli, but the degree of this shift is dependent on phonetic and social factors.

Word-specific phonetic patterns argue for including episodic memory in production models too. For example, Yaeger-Dror and Kemp (1992) and Yaeger-Dror (1996) showed that in Montreal French certain semantically or culturally-defined groups of words display idiosyncratic vowel qualities, having failed to undergo an otherwise regular shift. As Pierrehumbert (2001) points out, for any production model to account for this sort of pattern, production targets must

reference the specific lexical items they are associated with in a given utterance—the sort of association that can be attributed to episodic memory.

Pierrehumbert (2001, 2002) presented an exemplar model of production that can account for such findings. In this model, perception occurs basically as described in Johnson (1997): new items are categorized according to their similarity to existing items in an exemplar space, which includes dimensions for each phonetic value that the perceptual system stores in memory (exactly what phonetic values are involved in this process remains an open question). A production target is determined by randomly selecting one exemplar from a subgroup of exemplars, and then taking an activation-weighted average of nearby exemplars. The activation strength is a gradient value which influences the probability an exemplar will be chosen for a production goal. Because the activations of exemplars decay over time, more recent exemplars are more likely to be selected. Due to the weighted averaging, more recent and frequent exemplars also contribute more to the determination of production targets.

6.2 Primed vowel-shadowing

To address whether recent perceptions affect speech production, this study employed a primed shadowing task with two vowel phonemes, /a/ and /i/. On half of all trials, subjects first heard a "cue" vowel (prime), and then after a short delay, heard a target vowel; on these trials, subjects repeated the target vowel as quickly as possible. However, on a quarter of the trials, the target stimulus was a pure tone beep ("no-target" trials), in which case subjects repeated the cue vowel. These trials encouraged subjects to plan to produce the cue vowel to a greater extent than the non-cue vowel, because the cue vowel was twice as likely as the non-cue to be the required response. In the remaining control trials, subjects heard a beep for the cue stimulus, and then shadowed the target vowel. Note that the control condition is comparable to a two-choice vowel-shadowing task.

The stages of the different types of trials used in the experiment are schematized in Table 1. The vowels [a] and [i] were used for both target and cue stimuli; F1- and F2-centralized versions of these vowels, [a]* and [i]*, were used for the cue stimuli as well. The differences between the normal vowels and their centralized counterparts were subphonemic, on the order of 50-70 Hz for both formants. On each trial, subjects heard a short stretch of white noise preceding the cue. Interstimulus delays of 100 and 800 ms were balanced across conditions.

Table 1. Trial types and percentages in the primed vowel-shadowing task

		1	2	3	4	5
Trial Type (%)	Cue Type	Noise	Cue	Delay	Target	Response
concordant (25%)	normal	"shh"	[V_1]	100/800 ms	[V_1]	V_1
	centralized	"	[V_1]*	"	"	"
discordant (25%)	normal	"	[V_1]	"	[V_2]	V_2
	centralized	"	[V_1]*	"	"	"
no-target (25%)	normal	"	[V_1]	"	BEEP	V_1
	centralized	"	[V_1]*	"	"	"
no-cue (25%)		"	BEEP	"	[V_1]	V_1

Trials in which the cue and target stimuli belong to the same phoneme will be called *concordant* trials, and those in which cue and target belong to different phonemes will be called *discordant* trials. Inclusion of no-target trials in the design had the effect that whenever the cue was a vowel, the probability of that same vowel being the required response was 2/3, and the probability of the other (non-cue) vowel being the required response was 1/3. This bias, along with a motivation to respond quickly, encouraged subjects to plan to say the cue vowel to a greater extent than the non-cue vowel; subjects also may have adopted a biased articulatory configuration prior to hearing the target—this possibility will be discussed in section 4. If productions are influenced by phonemic or subphonemic details of the cue stimuli, then we should conclude that very recent perceptual experiences can exert effects on the formation of speech targets.

The exemplar model of production described in Pierrehumbert (2001, 2002) allows for recently perceived subphonemic details to influence the determination of production targets. This is because more recent exemplars are more highly active and thus more influential in the activation-weighted averaging of phonetic values through which production targets are determined. The comparison of primary interest in this experiment is between response vowel formants on normal-cue and centralized-cue concordant trials, and is made explicit by the following hypothesis:

> Hyp. *Rapid subphonemic perceptual-motor mapping:* on concordant trials, responses made after centrally-shifted cue stimuli will be more central than those made after normal cues.

This hypothesis entails, for example, that [a] responses after centralized [a]* cues will tend to be more central in F1-F2 space than those made after normal [a] cues. This would indicate that the sub-phonemic differences in the cue stimuli were perceived and influenced the planning of vowel targets. We will in addition consider whether cross-phonemic priming effects occur, i.e. whether vowel responses will differ between concordant and discordant trials.

6.3 Method

6.3.1 Subjects

Subjects were 18 to 40 year-old native speakers of American English with no history of speech or hearing problems. 12 subjects participated, 6 male, 6 female. All subjects took part in 2 or 3 one-hour sessions, over the course of which they performed a total of 25 to 40 blocks of 32 trials, which amounts to around 800-1200 trials for each subject.

6.3.2 Stimuli

Vowel stimuli were constructed with the following procedure: a speaker of Midwestern American English who makes no distinction between a low back vowel /ɑ/ and a mid back vowel /ɔ/ produced sets of approximately 100 tokens each of the vowels /ɑ/ and /i/. The vowel /ɑ/ is

normally more front than the IPA [ɑ] in most American English dialects and so will here be represented as /a/. The tokens closest to the mean F1 and F2 of each set were selected as base tokens (vowel formants were estimated using an LPC algorithm implemented in Matlab, c.f. section 2.4 for details). Using PSOLA resynthesis, the pitch contours of both vowels were changed to 105 Hz with a slightly falling contour using the formula: F0 = 105 – 20*t*, where *t* is the time in seconds from the onset of the vowel. The first 250 ms (over which the pitch fell from 105 Hz to 100 Hz) of the signals were windowed using a Tukey window with r = 0.25 to reduce the salience of onset transients, and all stimuli were normalized to have the same signal energy. Hence all vowels were 250 ms in duration.

Centralized versions of the stimuli were constructed using a method of formant shifting described in Purcell & Munhall (2006). The base tokens produced with the procedure described above were bandpass filtered in a narrow range of frequencies above or below the formant being shifted, and likewise the signals were bandstop filtered in a range of frequencies containing the center of the formant. The two filtered signals were resynthesized to produce a signal in which the manipulated formant was shifted in the direction of the bandpass filter. This method was chosen over LPC resynthesis because it produces more natural-sounding formant-shifted stimuli. De-emphasis was accomplished with a 3rd-order elliptical filter with 2 dB of passband ripple and 50 dB of stopband attenuation, and emphasis was accomplished with a 3rd-order elliptical filter with 0.5 dB of passband ripple and 15 dB of stopband attenuation. The filtered signals were resynthesized at a 1:1 ratio. Passbands and stopbands for F1 and F2 were, respectively (relative to the unshifted formants, in Hz), $F1_{pass}$([a]) = [-150 0], $F1_{stop}$([a]) = [0 150], $F2_{pass}$([a]) = [50, 250], $F2_{stop}$([a]) = [-250, 50], $F1_{pass}$([i]) = [25, 350], $F1_{stop}$([i]) = [-75, 25], $F2_{pass}$([i]) = [-350, -50], $F2_{stop}$([i]) = [-50, 200].

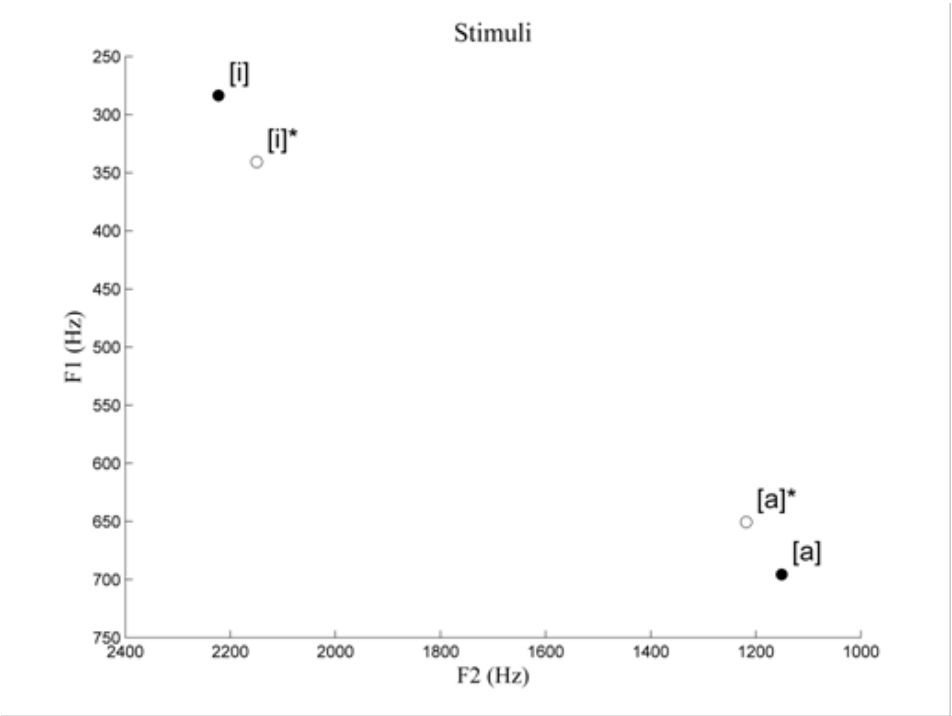


Figure 41. Vowel space locations of normal and centralized stimuli (formants were averaged over the middle third of each vowel).

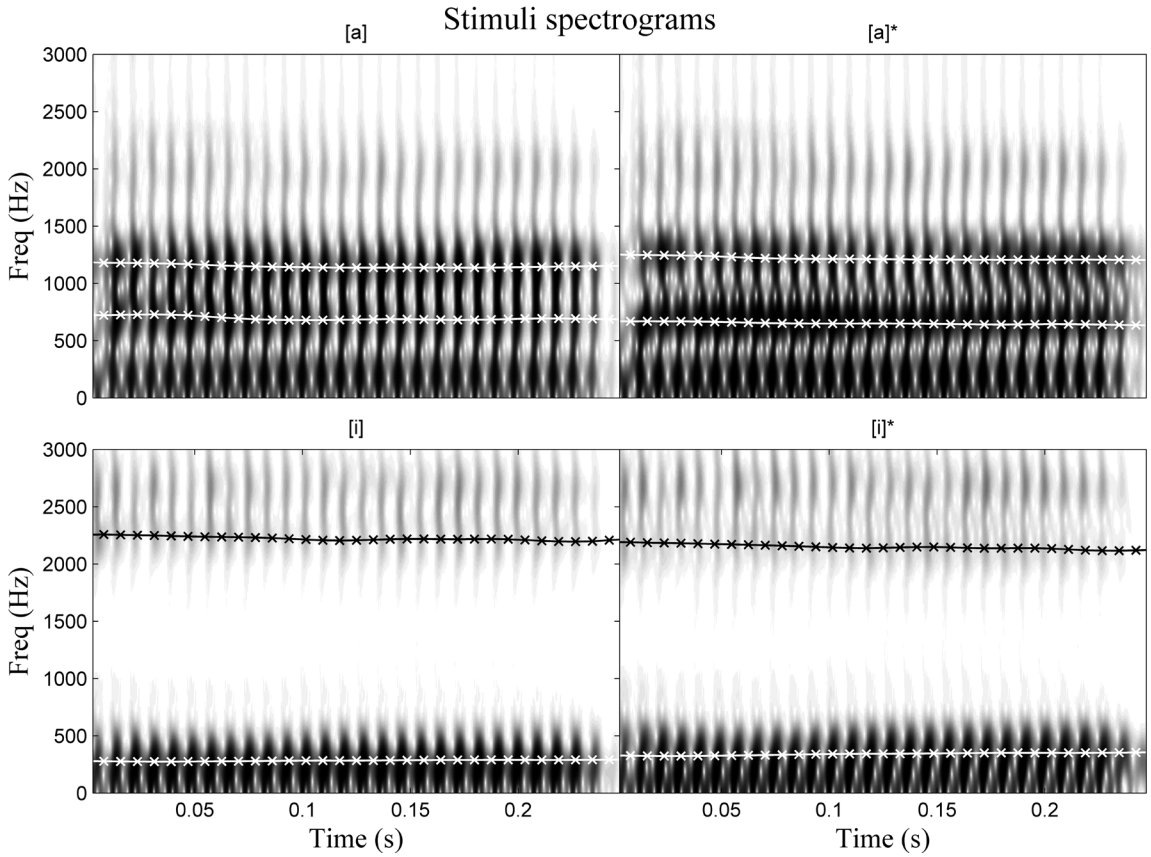


Figure 42. Spectrograms of normal and centralized stimuli.

The locations of the stimuli formant values (averaged over the middle third of each vowel) are shown in Figure 41 in F1-F2 space. Figure 42 shows spectrograms of the four stimuli. Looking carefully at the spectrograms, one can see that F1 and F2 are slightly closer for [i]* than [i], and slightly further apart for [a]* than [a]. The formants of [a] fall slightly; F2-[i] decreases across the vowel, while F1-[i] rises slightly. The mean LPC-estimated F1 and F2 of the four vowel stimuli were: [a] (696, 1151), [a]* (651, 1218), [i] (284, 2223), and [i]* (341, 2150). The differences between the normal and centralized stimuli were thus (-45, 67) for [a] and (57, -73) for [i]. These differences are within the normal range of formant variation that one would expect of these vowels in casual speech, and were unlikely to have been consciously perceived by unsuspecting, untrained ears. The impression one gets in hearing these stimuli is that the difference between the centralized [i]* and normal [i] is slightly more difficult to hear than the difference between centralized [a]* and normal [a]. There was a remote possibility that the centralized [a]* could have been perceived as a low and central /ʌ/, but no subjects reported hearing a third vowel or responded so as to indicate such a perception, and moreover the task instructions encouraged them to think of their responses as either one of the two vowel phonemes /a/ and /i/.

6.3.3 Procedure

Each trial began with an interval of white noise of random duration from 1000-4000 ms, followed by a 100 ms interval of silence. The white noise was intended to disrupt any residual effects from the preceding trial. The duration of the noise was randomized in order to avoid the establishment of a rhythm from the onset of the noise to the cue stimulus. After the 100 ms interval of silence, subjects heard the cue stimulus (for 250 ms), which was either a beep or one of the four vowels described in the preceding section.

Following the cue was a delay of either 100 ms or 800 ms. These provide approximately 500 ms and 1200 ms intervals between the end of the cue stimulus and the typical response onset. These durations were chosen to provide intervals of time differing in the relative extent to which planning of the cue stimulus might influence subsequent response planning and execution. After the interstimulus delay, subjects heard the target stimulus, which was either a beep or one of the two unshifted vowels, [a] and [i]. Note that if the cue stimulus was a beep, the target stimulus was restricted to a vowel, so that beep-beep trials never occurred. Following the target stimulus was a 2000 ms interval in which the subject responded.

Each trial can be characterized the three control parameters: cue stimulus, interstimulus delay, and target stimulus. Each block of trials consisted of 32 trials, 16 of which represented all permutations of the four cue vowels, two delay conditions, and two target vowels (these are the concordant and discordant trials). 8 trials consisted of a beep cue followed by the two target vowels in both delay conditions, all repeated twice in each block (no-cue/control trials). The remaining 8 trials consisted of the four cue vowels with a beep target in both delay conditions (no-target trials). Note that hearing any given cue vowel made that vowel twice as likely as the non-cue vowel to be the required response. This imbalance was expected to encourage greater planning of the cue vowel on all trials. The order of trials was randomized within each block to discourage subjects from guessing at the next response.

After each block of trials (except for the first two of each session), subjects received feedback regarding the speed of their responses in the block. This feedback came in the form of rating numbers which indicated how quickly they responded relative to their past response times in the session. The rating numbers were computed by using the means of the response times in the last completed block as arguments to the inverse cumulative normal distribution function, the parameters of which were estimated from the means of the response times in all prior blocks in the session. Thus the ratings ranged from 0 to 100, with values near 50 indicating that the average response time for in the last block was close to the average mean response time in all preceding blocks. This system had the advantage that as subjects responded more quickly, it became more difficult to achieve higher ratings, and thus they had to maintain a high level of attention over the course of a session if they wanted to get high ratings. In order to facilitate concentration on the task, subjects were given a five minute break halfway through each session.

6.3.4 Data Analysis

The primary dependent variables in this experiment were F1 and F2. Secondary dependent variables were response time (RT) and response duration (Dur). Response time was defined as the time from the onset of the target stimulus to the onset of vocal fold vibration. The first block that each subject completed was not included in the analysis. Overall, less than 5% of

the entire dataset was excluded for the reasons mentioned above. Most of the exclusions were of trials on which subjects responded late (more than three standard deviations greater than their mean RT), representing around 2% for most subjects. Trials were also excluded when subjects responded early, failed to respond, if a response duration was too short (less then 120 ms), the wrong response, mixed (i.e. a transition from [a] to [i] or vice versa), or had some other abnormality, such as being obfuscated by a non-speech vocalization. The mixed and incorrect responses occurred mostly on discordant trials, less than 0.5% of the time for most subjects. When one discounts the late responses, less than 2% of all trials were excluded. Keeping the late responses makes no qualitative differences in the formant results presented in the following section. Because quick response times are indicative of attention to the task, and because lack of attention is possibly a confounding factor, it was judged best to exclude these RT outliers.

Formants were estimated using an LPC algorithm implemented in Matlab. Responses were recorded at 44100 Hz and downsampled to 11025 Hz. 12 and 10 LPC coefficients were used for the male and female subjects, respectively; LPC coefficients were computed for 40 ms windows at steps of 5 ms. The first 10 ms of each response was skipped to avoid transient perturbations due to the onset of phonation. A pre-emphasized signal was used for measuring F1 and F2 of both responses, except for F1 of [i], where the pre-emphasis was found to occasionally interfere with the detection of a peak corresponding to F1.

Some further steps were taken to ensure robust formant measurement. When a formant for a given frame did not fall within a reasonable range, its value was interpolated from nearby formants; if reasonable formants were not found in ten consecutive frames or twelve frames total in the vowel, the LPC algorithm was considered to have failed and the trial was discarded. Finally, the formants were smoothed and averaged across frames. Most of the LPC algorithm failures (< 0.2% of trials) involved low amplitude tokens in which the F1 and F2 spectral peaks of an [a] were indistinct or the F2 and F3 of an [i] were indistinct. Additionally, tokens with F1 or F2 values more than four standard deviations away from the mean for each subject were also discarded.

For the comparisons of mean formant values between conditions, means were averaged over the first, middle, and last third of each token. In the presentation of results below, the analysis of the mean formants from the middle third of each vowel is shown, unless otherwise stated. Analyses of variance (unbalanced, repeated measures) were conducted using type III sum of squares. Subjects were treated as fixed factors. Lack of balance in the data resulted primarily from differing numbers of observations between subjects. A very small imbalance within conditions was due to the excluded trials. Where contours are presented below, these contours were normalized for duration within subjects such that each response had the same number of LPC analysis frames as the mean number of frames.

6.4 Results

6.4.1 Subphonemic priming effects

Analyses of variance revealed significant mean F1 and F2 differences between response vowels produced after normal and centralized cues on concordant trials. The effect of centralization was significant for F2-[a] {$F(1,1428) = 12.08$, $p < 0.001$}, F1-[i] {$F(1,1446) = 62.54$, $p < 0.001$}, and F2-[i] {$F(1,1446) = 6.04$, $p < 0.014$}. Figure 43 shows 95% confidence

ellipses for the bivariate means on normal-cue, centralized-cue, and control (no-cue) trials, for each subject. These ellipses represent regions in which one can be 95% confident the true population mean vector is located. The tilt of the ellipses relative to the coordinate axes reflects the correlation between F1 and F2, and the lengths of the major and minor axes of the ellipses correspond to the variability of the samples in the directions of those axes. The figure shows that most subjects exhibit a tendency to produce relatively more centralized responses after a centralized cue stimulus, particularly for the F2 of [a] and for both F1 and F2 of [i].

Several patterns can be seen in the relations between the normal, centralized-cue, and control trial means. Note that because there were twice as many control (no-cue) trials, their confidence regions are smaller. Where normal vs. centralized-cue differences were present, the most common pattern is for the control trial means to be relatively similar to the normal trials (e.g. [a]-*m2*; [i]-*f1, f2, f3, f6, m1, m2, m3, m5*). For some subjects, however, the control trial means are noticeably distinct from the normal and centralized-cue trials (e.g. [a]-*f1, m1, m3, m5*; [i]-*f4, f5, m4, m6*), although most of these remain more similar to the normal trials than the centralized-cue trials. However, for other subjects the control trial means are located between the normal and centralized-cue trial means (e.g. [a]-*f2, f5, m1, m6*; [i]: *f1, f5, m1, m2*).

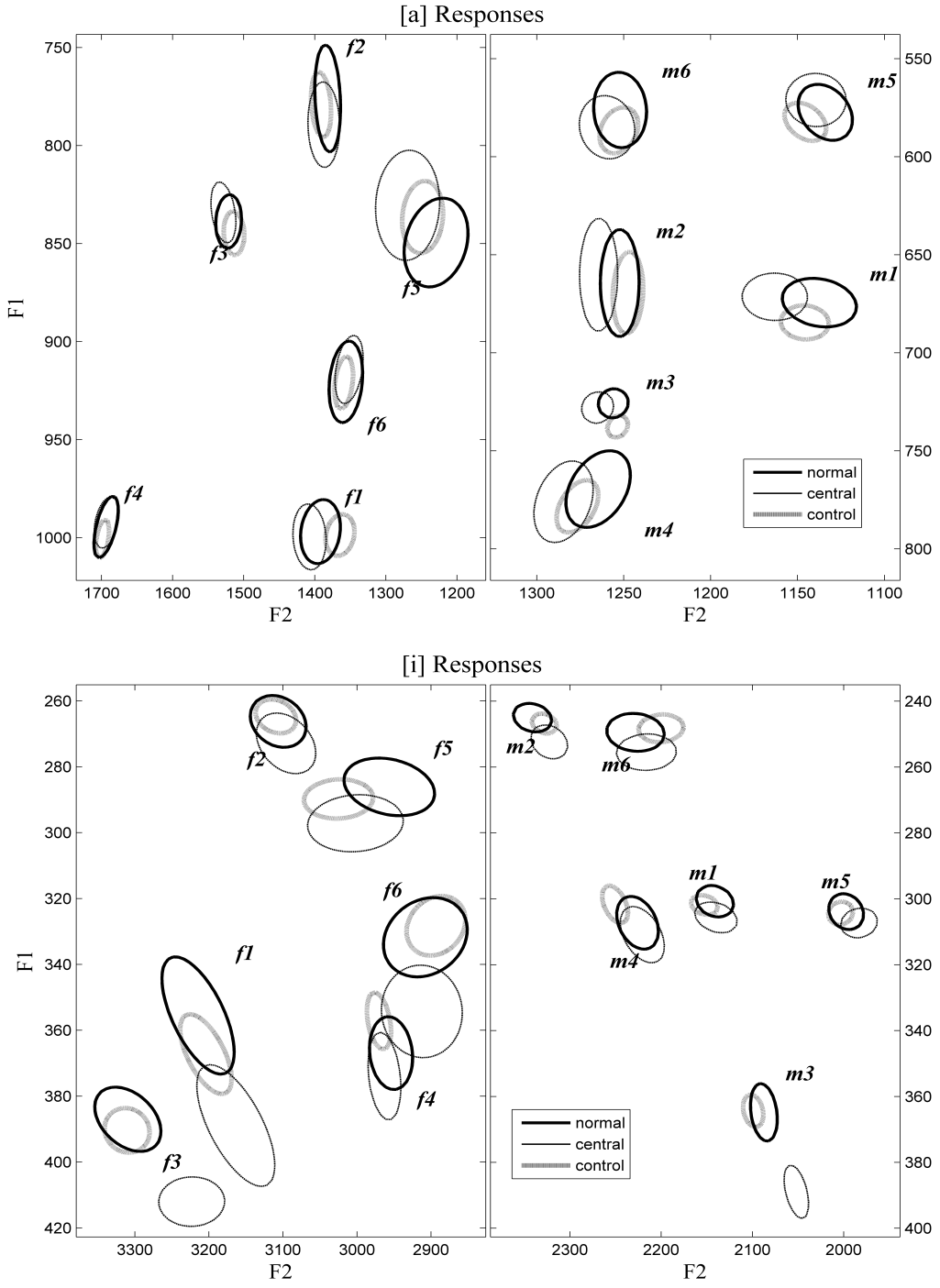


Figure 43. Confidence regions for [a] responses (top) and [i] responses (bottom) after normal-cue, centralized-cue, and control trials. Ellipses show 95% confidence regions for bivariate means of normal trials (dark bold), centralized-cue trials (thin), and control trials (light thick).

Within-subject comparisons of normal and centralized-cue trial formants were conducted using 1-tailed t-tests (cf. the appendix for a table of within-subject statistics). 5 subjects exhibited

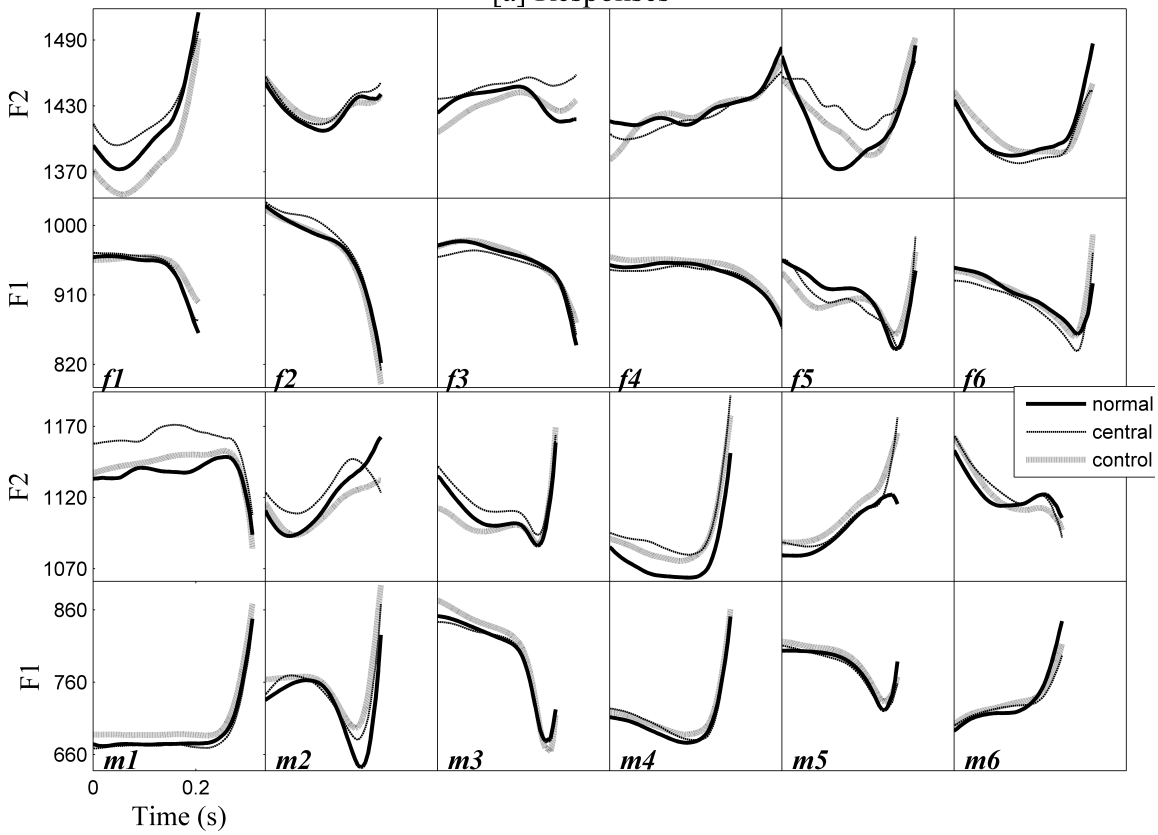
individually-significant differences for F2-[a], and all but two contributed to a trend for centralization. 9 subjects exhibited marginal or significant differences for F1-[i], and a clear trend can be seen across subjects to produce responses with higher F1 after shifted cue stimuli. Likewise, for F2-[i], 5 exhibited significant differences, and 9 followed the same trend.

Figure 44 shows F1 and F2 contours for [a] and [i] responses made on normal, centralized-cue, and control trials. In general the patterns of differences between normal and centralized-cue trials accord with the analyses of variance, but inspection of the trajectories allows for more detailed temporal patterns to emerge. First examining [a] responses (where lower F1 and higher F2 are more central), we see that the majority of subjects exhibited negligible differences in F1 throughout the contours, yet several subjects had relatively lower F1 in the first half of the contour (*f3, f5, f6*), suggesting centralization, while 2 subjects showed relatively higher F1 during some portions (*f2, m2*).

For [a]-F2, all male subjects tended to produce higher contours for part or all of the vowel, and several female subjects did so as well. One common pattern was for the centralized-cue trial formants to begin and remain central relative to the normal trials (*m1,m3, m4*). A different but no-less-prevalent pattern was for the trajectories to begin apart, then eventually converge (*f1, f4, m5, m6*;). Another pattern is for the formants to begin near one another, but then the centralized-cue responses diverge centrally (*f3*), and in the case of (*f5*), subsequently reconverge.

For [i] responses, most of the male and female subjects clearly demonstrated centralization in F1, particularly during the first half of the response. As with [a], some subjects produced a relatively constant difference in formant value between the normal and centralized-cue conditions, while for others the formants converge eventually. The majority of subjects produced responses with more central F2, although a couple produced responses with more peripheral F2 (*f4, f5*).

[a] Responses


F2
1490
1430
1370
F1
1000
910
820
f1
f2
f3
f4
f5
f6
normal
central
control
F2
1170
1120
1070
F1
860
760
660
m1
m2
m3
m4
m5
m6
0
0.2
Time (s)

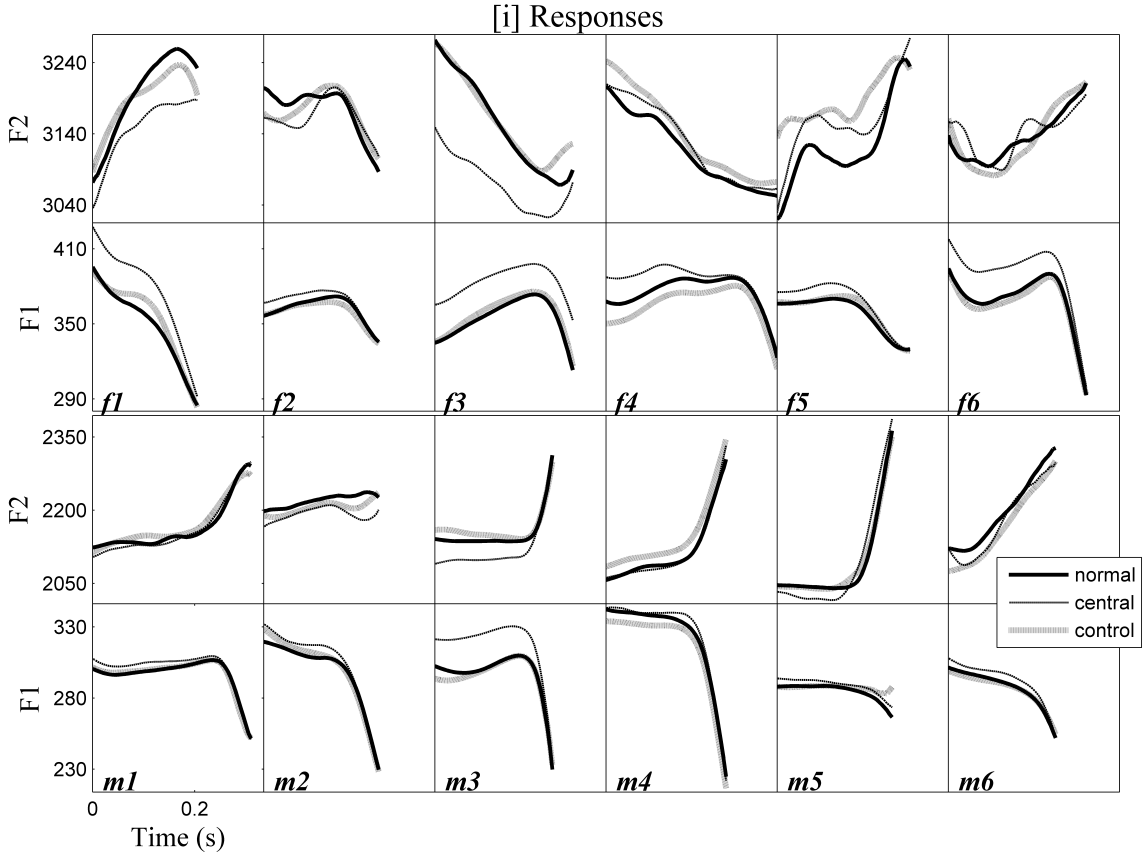


Figure 44. Average formant trajectories for [a] responses (top) and [i] responses (bottom) after normal (bold), centralized-cue (thin), and control (light thick) trials. Bold lines show trajectories for normal-cue trials; thin dashed lines shows. All panels in the same row have an identical frequency scale, but different absolute formant ranges.

Analysis of no-target trials, which also offer a comparison between trials with normal and centralized cues, revealed the same pattern. The effect of centralization was significant for F2-[a] {$F(1,1420) = 5.56$, $p < 0.019$}, F1-[i] {$F(1,1409) = 49.34$, $p < 0.001$}, and F2-[i] {$F(1,1409) = 10.14$, $p < 0.001$}. Note that centralization effects on formants in the first and last thirds of responses were significant for the same vowel-formant pairs. Centralization had no significant effects on response time. For [i] responses only, cue-centralization had a significant effect on response duration {$F(1,1446) = 5.57$, $p < 0.018$}, with responses tending to be longer after normal cues. Interstimulus delay did not exhibit any main effects on vowel formants on concordant trials, nor any interaction effects with centralization.

6.4.2 Discussion

The results strongly support the hypothesis that subphonemic details of the cue stimuli would be perceived and integrated into response vowel planning, indicating that episodic memories influence target planning. This was true in particular for F2-[a], F1-[i], and F2-[i]. Although not every subject exhibited significant differences individually, the trends were highly

significant across subjects. Because the same pattern held for comparisons between no-target trials, the effect does not depend upon the recency of the unconscious perception of a difference between a centralized cue and normal target stimulus. Hence very recent perceptions (i.e. over the past 350-1500 ms) can have an impact on production.

The substantial intersubject variation observed in the relations between control and concordant trial responses suggests that subjects may have adopted differing response strategies. It is reasonable to assume that subjects have the ability to pre-plan both vowel responses, prior to and during each trial, and further that subjects control the *extent* of preplanning in order to minimize their response times. It follows that upon hearing the cue vowel, the subject prepares that response vowel to a greater extent than the non-cue vowel. In contrast, on control trials subjects prepare both vowels to similar extents, because both responses are equally probable. Differences in response preparation may account for fine-grained intersubject variation, but the details of such an account are beyond the scope of the present experimental data.

6.4.3 Cross-phonemic priming effects

Analysis of variance revealed significant mean F1 and F2 differences between response vowels produced on concordant and discordant trials. Only trials with normal cues are considered here. The effect of concordance was significant for F2-[a] {$F(1,1416) = 9.37$, $p < 0.002$} and F1-[i] {$F(1,1430) = 4.02$, $p < 0.045$}. Unexpectedly, these concordance effects were *dissimilatory*: formants were more peripheral when the cue and target vowels were different phonemes, i.e. dissimilation was observed on discordant trials relative to concordant trials. These concordance effects were even stronger for the mean formants of the first third of each response, where a marginal effect on F2-[i] was observed {$F(1,1432) = 2.58$, $p < 0.108$} in addition to concordance effects on F2-[a] {$F(1,1418) = 12.07$, $p < 0.001$} and F1-[i] {$F(1,1432) = 11.35$, $p < 0.001$}.

Figure 45 shows confidence ellipses for within-subject F1,F2 bivariate means on concordant, discordant, and control trials. Comparing the concordant trials (bold ellipses) to discordant trials (thin ellipses), one can see that a number of subjects exhibited noticeable dissimilatory differences ([a]: *f1, f4, f6, m1, m2, m3, m4, m6*; [i]: *f3, f4, f5, f6, m1, m3, m4*). Here "dissimilation" should be read not in a phonological or historical sense, but in a more literal, phonetic sense, entailing less similarity to the cue vowel: [a] responses tended to be acoustically less like [i] after an [i] cue, and vice versa, [i] responses were less like [a] after an [a] cue.

Where the concordant and discordant trial responses tended to differ, several patterns can be seen in the relations between the control trials and the concordant and discordant trials. One such pattern is for the control trial formants to be relatively more similar to the concordant trials (e.g. [a]: *f4, f6, m1, m2, m4*; [i]: *f2, f3, f6, m1*). A comparably common pattern is for the control trials mean vectors to be located either between the concordant and discordant means (e.g. [a]: *f1, m2;* [i]: *f1, f5, m3*), or to be more similar to the discordant trial means (e.g. [a]: *f5, m3, m6*; [i]: *m4, m6*).

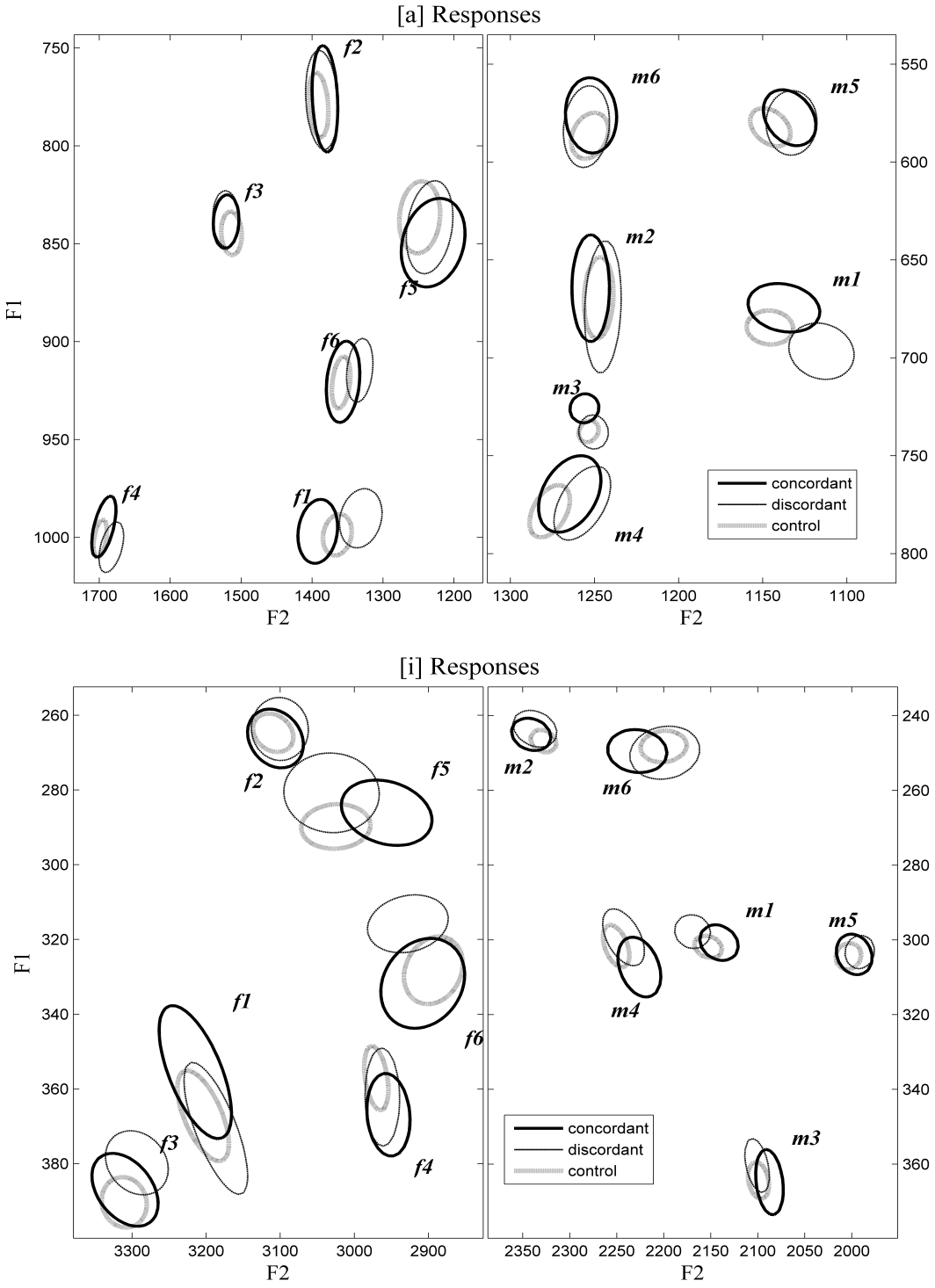


Figure 45. Confidence regions for [a] response (top) and [i] response (bottom) mean formant vectors from concordant (bold), discordant (thin), and control (thick light) trials. Formants averaged over the first third of each response.

Within-subject comparisons of mean F1 and F2 from the first third of responses from concordant and discordant trials showed that most subjects exhibited dissimilatory patterns (cf.

the appendix for a table of statistics). For [a]-F1, where there was no main effect of concordance, two subjects showed individually significant differences that were assimilatory in nature; one showed a significant dissimilatory difference. For [a]-F2, 5 subjects had a significant dissimilatory tendency to lower F2 in [a] responses made after an [i], and 10 subjects contributed to the overall trend. For [i]-F1, 4 subjects showed significant dissimilatory patterns, while 9 contributed to the significant trend across subjects. For [i]-F2, 2 subjects showed significant dissimilation.

Figure 46 shows average F1 and F2 contours from concordant, discordant, and control trials. Although most subjects exhibit some subtle idiosyncrasies, several distinct types of patterns can be seen. First examining the F1 of [a] responses (where higher F1 and lower F2 indicate peripherality/dissimilation), we see that the majority of subjects had negligible differences between discordant and concordant trials, while two produced on discordant trials somewhat more peripheral responses (*m1, m2*) and two others more central responses (*f1, f5*), particularly in the first portion of the vowel. For [a]-F2, the majority pattern was a dissimilatory lowering after discordant cue stimuli (*f1, f4, f5, f6, m1, m2, m3, m4*); in some cases this lowering persisted throughout the vowel, in others it occurred primarily in the first portion. In examining [i] responses, lower F1 and higher F2 indicate peripherality/dissimilation. For [i]-F1, we see that a majority of subjects produced more peripheral responses on the discordant trials (*f3, f4, f5, f6, m1, m3, m4*), particularly in the first part of the responses. Peripheralization was observed for these same subjects in [i]-F2 (though less clearly for *f3*). There are two instances in which the discordant trials were associated with noticeable assimilatory changes in F2 (*f1, m6*).

One noteworthy aspect of the comparison of formant contours is that there exists some variation with regard to the initial differences between concordant and discordant trial formants as well as the subsequent divergence or convergence of those formants. For example, one pattern is *separation*, where the responses tend to initially differ and remain relatively constantly separated throughout the vowel ([a]-F2: *f1, f4, f6*; [i]-F2: *f5, m4, m6*). A more common pattern is *separation-convergence*, where an initial (usually dissimilatory) separation is followed by eventual convergence ([a]-F1: *f1, f5*; [a]-F2: *m1, m3, m4*; [i]-F1: *f1, f3, f6, m3, m4*; [i]-F2: *m1, m3, m4*). A third pattern is *proximity-divergence(-convergence)*, in which formants are initially similar but then eventually diverge ([a]-F1: *m1, m2*; [a]-F2: *m2,m*6; [i]-F2: *f2, m4, m5*), and in some instances converge subsequently ([i]-F1: *f1, f5*).

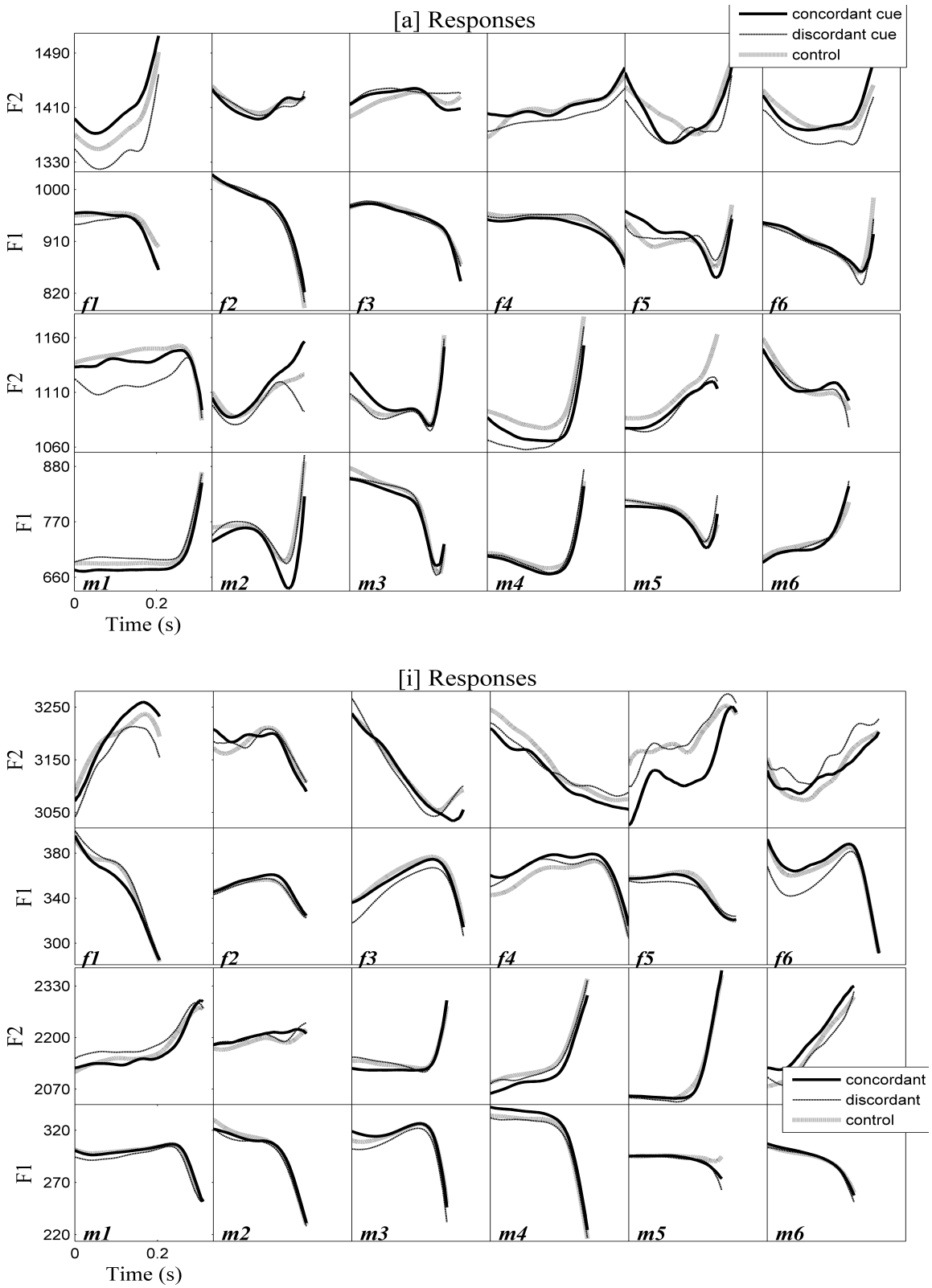


Figure 46. Average formant trajectories for [a] responses (top) and [i] responses (bottom) after concordant, discordant, and control (normal cue, no-target) trials. Bold lines show trajectories for concordant trials; thin dashed lines shows

trajectories for discordant trials; light, thick lines show trajectories for control trials. All panels have the same scale, but different absolute formant ranges.

Concordance affected response time and duration differently for [a] and [i]. It had an effect on the RT of [a] responses {$F(1,1392) = 6.548$, $p < 0.01$}, but not [i] responses. [a] responses on concordant trials tended to be made more quickly than on discordant trials. Note that the subjects with the two longest mean response times (*f1*, *f5*) were responsible for the only two cases of significant assimilatory differences, which occurred in [a]-F1. Conversely, concordance had a reliable effect on [i] duration {$F(1,1406) = 6.48$, $p < 0.01$}, but not on [a] duration. Examination of within-subject comparisons indicates that the effects on [i]-duration were reliably observed in only two subjects (*m1, f4*), who happened to have the two longest mean response durations of all subjects, in the range of 325-400 ms (for comparison, the other subjects exhibited mean response durations in the range of 230-325 ms). Interstimulus delay did not exhibit any significant main or interaction effects on vowel formants between concordant or discordant trials, nor on response times.

6.4.4 Discussion

Unexpectedly, for a number of subjects dissimilatory trends were observed in response formants on discordant trials compared to concordant trials. While this finding begs for an explanation, it would be prudent to conduct further investigation before developing a model. In section 4, a brief sketch of an explanation will be presented, but this sketch does not account for the substantial intersubject variation observed in the cross-phonemic priming. This variation was manifested in several ways: in whether dissimilatory or assimilatory patterns were produced, in the specific vowels and formants that were affected, and in the time-course of the effects, which are visible in formant contours.

There was more variation for some vowel-formant combinations than others. For [a]-F1, significant assimilatory and dissimilatory patterns were seen, albeit in only three subjects. A telling observation is that the two subjects who produced assimilatory patterns (*f1, f5*) were also the two subjects with the longest RTs, and *f5* also had an anomalously high RT variance—these correlations suggest that the assimilatory patterns produced by f1 and f5 may have been due to a lack of attention to the task. For [i]-F2, only two subjects produced significant dissimilatory patterns, and the concordance-by-subject interaction was significant. The subject with the largest (but not significant) assimilatory difference in [i]-F2 happened to show an abnormal propensity to respond early, i.e. before the cue. The relations between the assimilatory patterns and RT patterns suggest that either the assimilatory patterns are not general, or that a certain degree of attention is required for the dissimilation to emerge in the task.

The somewhat motley patterns in [a]-F1 and [i]-F2 were matched by robust trends in [a]-F2 and [i]-F1, where all but a few subjects evidenced dissimilation. It thus appears that whatever mechanism is responsible for these dissimilatory patterns, the mechanism can have differing effects on F1 and F2. This could indicate either that F1 and F2 are governed by independent planning mechanisms, or that these variables are not the parameters most directly manipulated by the production system. Note that some subjects exhibited more consistency in dissimilating: for example *m3* and *m1* produced significant dissimilatory patterns in 3 and 4 of the vowel-formant combinations, respectively, while others showed in only 1 or 2 cases.

As in the centralized-cue comparisons, the control trials did not consistently pattern with one or the other condition, although the most common pattern was for the controls to be similar to the concordant trials. Control trial variation may arise from differences in task strategies adopted by subjects, particularly with respect to balancing maintenance of the cue in working memory with preparation for a 2-choice shadowing task. Yet another source of intersubject diversity was variation in the time-course of the difference between concordant and discordant mean formant trajectories. Careful observation of these trajectories in Figure 46 reveals three general patterns of relations: separation, separation-convergence, and proximity-divergence(-convergence).

6.5 General discussion

The subphonemic priming effects reported in section 3.1 argue for a speech production model that incorporates episodic memories into production targets. In an exemplar model, the recency of the cue vowel endows its associated exemplar with an observable influence upon the subsequently-planned production target. Non-episodic production models, in which targets are abstractions integrated over many instances of a category, do not allow for substantial changes in targets to occur rapidly or be influenced by subphonemic details of recent percepts. Section 4.1 presents two alternative possibilities for when the subphonemic perceptual-motor mapping occurs in the course of a trial, and shows how an exemplar-based model can account for either possibility. Section 4.2 will sketch an explanation for cross-phonemic priming effects, which, although perhaps a more remarkable finding, cannot be unambiguously interpreted as the result of planning processes and episodic memory. In chapter 5, an alternative dynamical planning field model will be presented.

6.5.1 Subphonemic priming effects in an exemplar model

There are two possible explanations for how subphonemic priming effects arise. The key difference between them is *when* in the course of each trial the mechanism(s) responsible for the subphonemic priming effects induce articulatory shifts. Figure 47 presents a simple box-schema of how planning and articulatory processes may occur over the time-course of a concordant trial with a centralized cue. Recall that prior to the cue stimulus, the vowels [a] and [i] are equally likely to be the required response. After the cue stimulus, the cue vowel is twice as likely to be the required response. After sufficient exposure to the task, subjects incorporate these probabilities into their pre-trial expectations. Their expectations then change depending upon the cue stimulus, and this affects their response planning. Specifically, after a cue vowel, the activations of the cue and non-cue vowel targets are weighted proportionally to their associated expectations.

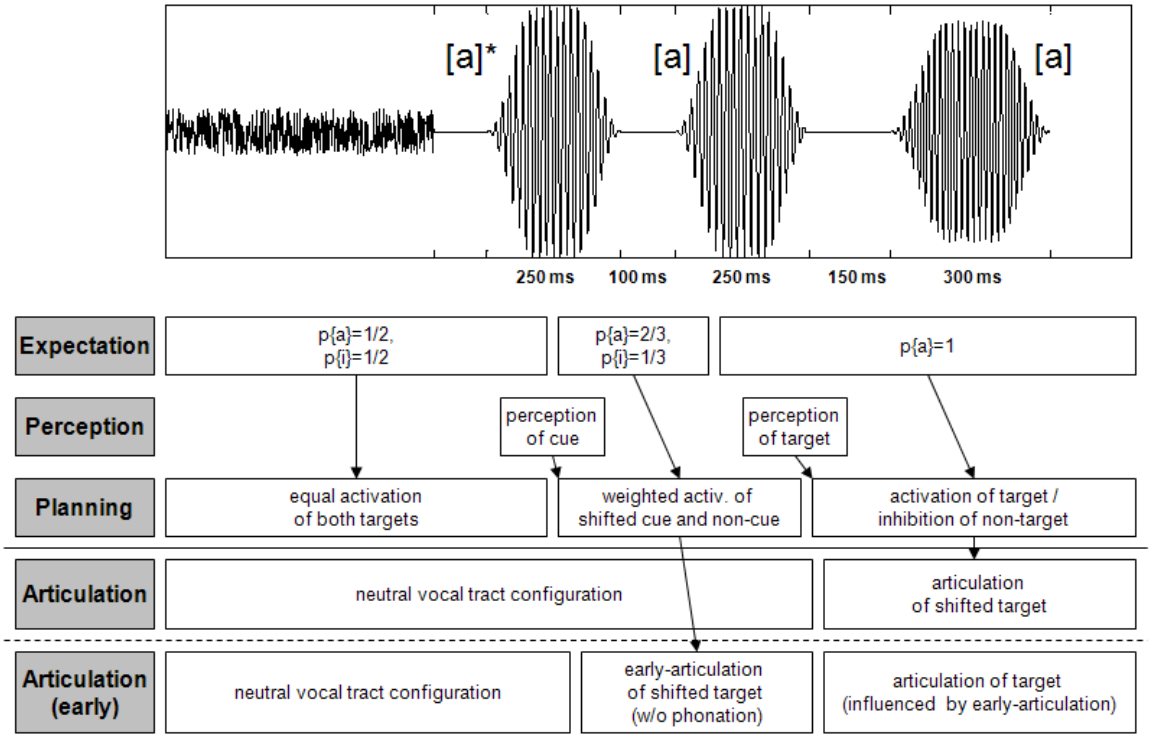


Figure 47. Schematic representation of the time-course of a shifted-cue concordant trial, and hypothesized relations between cognitive processes. (Top) noise, cue [a]*, target [a], and response vowel [a]. Articulatory events in a planning-interaction account are shown above the dashed line; the alternative early-articulation account is shown below the dashed line.

On the one hand, the priming effects may arise from changes exclusively in the planning of speech targets, which are then manifested as articulatory changes immediately prior to (and/or during) the production of the response vowel. The "articulation" level in Figure 47 represents this interpretation. The activation (or planning) of the response is influenced by the prior weighted-activation of a shifted cue vowel. This account holds that before perception of the target stimulus (or perhaps, before the initiation of articulation), the shape of the vocal tract does not vary—in other words, only the speech targets themselves have been altered due to the priming. Experimental effects are then attributed solely to interactions between planned articulations, particularly the influence of pre-target planning upon post-target planning. This *planning-interaction* account places the cause of differences between experimental conditions entirely in the realm of speech planning/target selection.

On the other hand, priming effects may arise from changes in vocal tract configuration that take place soon after the cue stimulus has been perceived. Because knowledge of the cue stimulus creates a probabilistic response bias, subjects may pre-shape their vocal tracts, either for the sake of minimizing their response times, or for some other reason (the “early articulation” level in Figure 47 represents this idea). The pre-configuration of the vocal tract could reflect subphonemic details of the cue stimulus, and would then influence the subsequent production. This *early-articulation* account locates the direct cause of the priming effect earlier in the trial.

It is crucial to understand that regardless of which account is correct, subphonemic priming results from the influence of episodic memory upon production. The only difference is precisely when this influence occurs. Because articulatory data were not collected in this experiment, one cannot prove that either planning-interaction or early-articulation is the sole correct explanation for observed patterns (in future work articulatory data would be highly informative, because they would resolve whether articulatory posturing occurs during the interstimulus period). Furthermore, these two explanations are not mutually exclusive—both may play a role in our understanding of the priming effects.

Figure 48 presents schematic illustrations of the exemplar models of perception and production described in Pierrehumbert (2001; 2002), which are built upon the exemplar model of perception described by Johnson (1997). The left side of the figure depicts how a new sensory stimulus (unfilled circle) is categorized according to its similarity to (phonetic distance from) nearby exemplars in a phonetic space. Any number of phonetic dimensions can be used to define a space with a distance metric, but for current purposes we consider only F1 and F2. In categorizing a new stimulus, phonological category labels (boxes) are activated, and the strength of this activation is the sum of all the activations of the exemplars of the category. Furthermore, each exemplar has its own base-level of activation (represented by the shades of the circles), which decays over time. The effect of this decay is that recent exemplars contribute disproportionately to the categorization of new stimuli. For example, although exemplar_1 is closer to the new percept than exemplar_2, exemplar_1 corresponds to an older memory and in this case fails to contribute to categorization of the stimulus. The most highly-activated category then becomes associated with the new stimulus, which is stored as a new exemplar in long-term memory.

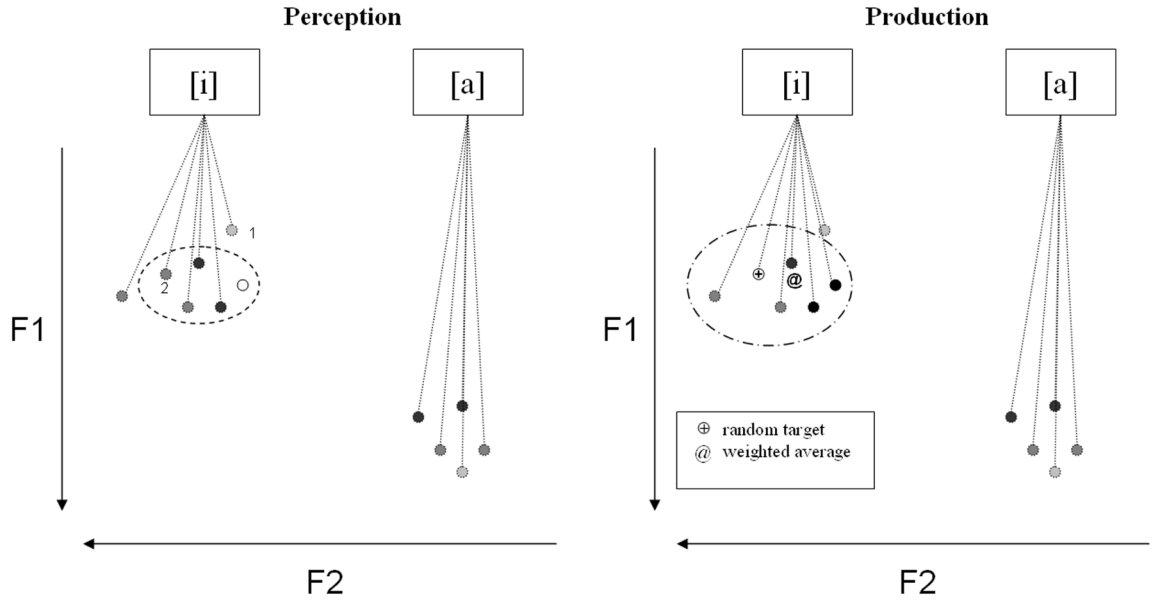


Figure 48. Simplified schematization of exemplar models of production and perception as described by Johnson (1997) and Pierrehumbert (2001; 2002). (Left) perception model; unfilled circle: new stimulus; (2) is further from the new stimulus than (1), but plays a role in perception due to memory-weighting. (Right) production model; + randomly selected target; @ weighted average after entrenchment.

The exemplar-based approach to production is based upon the perception model, but operates in reverse (although c.f. Pierrehumbert (2002) for an elaboration of lexical and phonological category influences). The first step in obtaining a production target is for the desired phonological category to be chosen. This "choosing" corresponds to an intention to plan a production target. Next an exemplar associated with the category is randomly selected (Figure 48, right: +); the probability of an exemplar being chosen is a function of its recency, which reflects the temporal decay of exemplar memories. Then in a process of activation-weighted averaging, the phonetic characteristics of a nearby group of exemplars are averaged. In the Pierrehumbert (2001) model, this group consists of the *n*-nearest exemplars to the randomly selected exemplar (where distance is activation-weighted). The memory-weighting of the distance function makes more recent exemplars more likely than equally-distant older exemplars to be integrated into the production target. Additionally, the averaging of phonetic characteristics within the neighborhood of the selected exemplar is memory-weighted, so that more recent exemplars contribute more than older ones to the determination of target values (this is analogous to the role recent exemplars play in the perceptual model).

Now consider how such an exemplar model relates to the box-schema of the time course of perception, planning, and production processes depicted in Figure 47. For simplicity, we can assume that target planning occurs in the same acoustic exemplar space as perception, which is consistent with the models outlined in Pierrehumbert (2001, 2002). (A different possibility is that

the planning occurs in motor or gestural coordinates, but this would not substantially change the conceptual structure of the account).

The exemplar production model operates in the primed-shadowing task as follows. First, prior to the cue, both categories are activated to equal extents and both targets are planned, i.e. exemplars are selected. After the cue has been perceived and stored as an exemplar, two things occur: 1) the activations of the category labels are reweighted according to their likelihood of being the required response, and 2) the targets are replanned—the fresh exemplar potentially plays a role in determining the phonetic parameters of the new targets. Then, after the target stimulus is perceived, the corresponding vowel becomes fully activated, and a new production target is planned and executed. Articulation and target planning can potentially operate in parallel, meaning that target planning may affect response articulation in mid-production (this provides one way to account for the variation in formant trajectories). In other words, even after articulation and phonation have begun, target planning effects may still emerge and disappear.

The planning interaction account holds that the differences between normal and centralized-cue trials are due to the relatively high activation of the cue exemplar during the planning of the response. Because of its recency, this exemplar contributes disproportionately to target planning. When the cue is an [a]*, the activation-weighted average of [a] exemplars will be biased toward more central formant values, resulting in a more central response.

The early-articulation account differs from the planning-interaction account with respect to when and how articulation occurs. This approach holds that following perception of a centralized [a]* cue, the vocal tract is rapidly pre-configured so as to favor the articulation of the [a]*. Presumably a slight difference between the configurations made after [a] and [a]* cues arises in early-articulation due to incorporation of subphonemic details of the cue exemplar. More precisely, the suballophonic details play a role in selection of the exemplar target. Although normal and centralized cues differed by about 50 and 70 Hz in F1 and F2, activation-weighted averaging of exemplars should tend to decrease the difference, and indeed this is what was observed.

To address whether the early-articulation or planning-interaction account should be preferred, consider whether either one makes any predictions that the other does not. The early-articulation account relies on the assertion that there are differences in vocal tract geometry between concordant normal and centralized-cue trials (prior to the target stimulus), and that these are maintained up through at least part of the response. Even if we assume a fair degree of nonlinearity in the mapping from acoustics to vocal tract configuration, this account predicts that the formant differences between centralized and normal-cue trials will either decrease or stay constant over the production, depending upon what assumptions are made about the dynamics of target planning. In no case should formant differences increase or exhibit more complex temporal patterns. In other words, the early-articulation account predicts that formants will differ most between the two conditions at the beginnings of response contours, or at least will not become substantially more different later in the response.

For the majority of subjects, this prediction of the early-articulation account is confirmed, but there are a number of cases where it is violated, too many not to call into question the role of early-articulation as the only explanation for the observed priming effects. In these cases ([a]-F1: *f5*, [a]-F2: *f3, f5*; [i]-F1: *m2*, [i]-F2: *f1, m6*), the normal and centralized-cue trial formants begin in close proximity, but then diverge such that the centralized-cue responses become more central. In some of these cases, they reconverge. The existence of these patterns suggests that—at least

for some subject-vowel-formant combinations—early articulation is not the sole cause of the subphonemic priming effect.

In contrast, the planning-interaction account may accommodate such differences in the following way: individual phonetic values of exemplars can be conceptualized as vectors, specifying values in an acoustic dimension over time (both Johnson (1997) and Pierrehumbert (2001) allow for this). Thus exemplar targets are actually trajectories. It is conceivable in this approach that there exist differences in how production targets are determined, with respect to their temporal dimensions. In other words, subjects may attend to subphonemic details of the cue differently in different phases of the cue. Yet this speculation leaves unanswered questions about why the internal dynamics of the productions are so complex and speaker-dependent. In the absence of data to resolve these issues, the question is better left to future investigation.

It is worth noting that variation in which specific formants are affected may arise due to a number of factors, such as individual differences in perceptual ability/attention, memory formation and decay-time, linguistic or functional aspects of the vowels involved, differences in vocal tract geometry, etc. The formant-specificity of the centralization effects may be affected by nonlinearities in the perception of acoustic differences. Frequency-discrimination is more approximately linear across lower frequencies (<1000 Hz) corresponding to [i]-F1 and [a]-F1, and decays at higher frequencies, so that a 70 Hz difference is more psychoacoustically salient around [a]-F2 than the same difference is around [i]-F2. This may explain why the effect on F2 was observed for more subjects in [a] responses. However, a 50 Hz difference is only slightly more salient for [i]-F1 than the same difference is for [a]-F1, so this cannot explain the general absence of a robust effect on [a]-F1. It is likely that for linguistic-functional reasons, certain formants in speech can bear higher functional loads than others and this influences perceived differences; one could speculate that [a]-F2 is linguistically more salient than [a]-F1 and that this is responsible for the presence/absence of effects in the F2/F1 of [a]—such an explanation would however be pure speculation at this point. Relatedly, higher-level linguistic modulation of perception could lead to a factoring out of variation in acoustic/articulatory dimensions which are correlated with the most salient dimension. In other words, subjects may attribute any centralization perceived in one dimension to centralization perceived in the highly-correlated, more salient one. It is also entirely possible that the acoustic measures of F1 and F2 are only indirect associates of some other set of acoustic or motor variables that are the primary targets of the production system.

Regardless of what the ultimate source of such variations may be, subphonemic priming clearly demonstrates that rapid incorporation of subphonemic detail into target planning is possible. Moreover, such phonetic details must be understood as aspects of episodic memory, and so we should conclude that episodic memories play a significant role in speech planning and production.

6.5.2 Cross-phonemic priming effects in an exemplar model

A brief sketch is presented here of one mechanism through which dissimilatory patterns may arise in cross-phonemic priming. However, there are several issues (mentioned below) with this interpretation, and these issues cannot be resolved by the experimental data. For that reason, this sketch is only meant to provide a starting point for future investigation.

It has been observed in studies of ocular and manual motor behavior that eye movement (saccade) and reaching trajectories tend to deviate slightly away from distractor targets to which movements have been previously planned (Sheliga et. al. 1994; Walker & Doyle 2001; Van der Stigchel & Theeuwes, 2005). These findings have been modeled by some researchers as arising from the neural inhibition of motor plans associated with a previously-planned saccade or reach. (Houghton & Tipper 1996; Tipper, Howard, and Houghton 1999). The basic idea is that in order to saccade to or reach for a target, movement plans to other possible targets must be selectively inhibited. Motor plans to distinct targets are coded by overlapping populations of neurons, and inhibition of an alternative plan has a small but observable effect upon the executed target movement. Further, the more highly activated the alternative plan(s), the stronger the selective inhibition needs to be.

This model applies fairly straightforwardly to speech target planning in the primed vowel-shadowing task. On a discordant trial, the cue vowel will be more highly activated than on the concordant trials, and hence production of the target vowel will require greater *intergestural inhibition* of the cue. This greater inhibition will shift the distribution of neurons coding for the vowel away from the cue, resulting in dissimilation relative to concordant trials. In the context of an exemplar model, this can be conceptualized as diminished activation of exemplars most similar to the cue, resulting in a production target that is dissimilated from the discordant cue vowel. Such an effect would not occur between vowels that differ only subphonemically.

An alternative interpretation of the dissimilation pattern is that articulator movement amplitude is diminished on the concordant trials due to early articulation of the cue. If the vocal tract is biased toward production of the cue prior to the target stimulus, movement amplitude and velocity may be diminished, the effect being hypoarticulation on the concordant trials. Likewise on discordant trials, if the vocal tract is pre-shaped with a cue-vowel bias, movement amplitude and velocity may increase, resulting in hyperarticulation. One or both of these effects could explain the dissimilatory patterns, and do so in a way that would involve no difference in target planning between the concordant and discordant trials.

The most reliable way to distinguish the inhibitory and early articulation accounts would be to collect articulatory data during the task, but since this was not done in the present experiment, no definitive claims should be made. The issue is worth resolving in future work, because intergestural inhibition, if it exists, may inform our understanding of a variety of linguistic patterns, including boundary-effects on gestural amplitude and duration, neighborhood density effects, and diachronic forces such as contrast preservation.

6.6 Summary and future directions

The results observed in this primed vowel-shadowing experiment showed clear effects of subphonemic priming on vowel formants. These effects argue for a role for episodic memory in the formation of speech targets, and are well-understood in the context of exemplar models of perception and production (Johnson 1997; Pierrehumbert 2001, 2002). Subphonemic details of a cue stimulus are stored in memory as an exemplar, i.e. a set of associations between phonetic values and various linguistic and non-linguistic categories. The recency of the cue vowel exemplar endows it with a relatively large influence over the subsequently planned production target, resulting in relatively centralized production after a centralized cue vowel. Furthermore, this priming effect can happen rapidly, on timescales as short as 300 ms.

Observed cross-phonemic priming effects were primarily dissimilatory, with responses made on discordant trials being more peripheral in F1-F2 space than those made on concordant trials. While some form of intergestural inhibition may be useful in understanding these and other speech patterns, an account based upon early articulation cannot be ruled out. Only further experimental work that measures articulation directly can address this issue.

Primed shadowing is a method for investigating cognitive planning mechanisms involved in speech, and in the present case has demonstrated that effects of recent percepts can exert substantial influences on subphonemic details of articulation. These findings indicate that speech production involves episodic memory, and confirm predictions of exemplar production models. This experimental paradigm has also provided some intriguing cross-phonemic priming results that merit further exploration.

Appendix 6.A

Table 6.A.1 shows within-subject comparisons of normal and centralized-cue trial formants conducted using 1-sided t-tests. Table 6.A.2 shows within-subject comparisons of the mean F1 and F2 of responses from concordant and discordant trials, where formant measurements were taken from the first third of each response. Confer sections 3.1 and 3.3 for discussion of these patterns.

Table 6.A.1
Within-subject F1 and F2 comparisons between normal and centralized-cue trials

	F1								F2						
	Normal		Central.		Normal-Centralized				Normal		Central.		Normal-Centralized		
Subj.	Hz	(σ,N)	Hz	(σ,N)	Δ	p <		Subj.	Hz	(σ,N)	Hz	(σ,N)	Δ	p <	
/a/								/a/							
f5	849	(62,49)	830	(77,49)	-19	0.09	+	f6	1356	(64,47)	1352	(55,50)	-5	0.65	
m5	577	(42,54)	571	(39,52)	-6	0.21		f4	1694	(47,48)	1693	(49,50)	-1	0.53	
f6	921	(56,47)	914	(48,50)	-6	0.28		m5	1134	(45,54)	1139	(49,52)	5	0.28	
f3	839	(37,48)	834	(43,50)	-5	0.29		f2	1382	(64,78)	1387	(78,77)	6	0.32	
m2	664	(96,77)	660	(100,76)	-4	0.40		f3	1521	(50,48)	1529	(49,50)	7	0.23	
m1	674	(43,76)	671	(43,77)	-3	0.33		m6	1252	(41,48)	1260	(44,49)	8	0.20	
f4	994	(43,48)	992	(36,50)	-2	0.39		m3	1256	(31,80)	1265	(33,80)	9	0.04	*
m3	726	(27,80)	728	(29,80)	2	0.69		m2	1252	(39,77)	1264	(38,76)	12	0.03	*
f1	957	(54,68)	960	(56,69)	3	0.61		f1	1392	(92,68)	1407	(77,69)	15	0.15	
m4	770	(55,51)	776	(57,49)	6	0.72		m4	1265	(52,51)	1285	(47,49)	20	0.02	*
m6	576	(52,48)	585	(44,49)	9	0.81		m1	1137	(74,76)	1163	(66,77)	26	0.01	*
f2	776	(96,78)	789	(76,77)	13	0.83		f5	1229	(124,49)	1269	(124,49)	40	0.06	+
/i/								/i/							
m5	304	(16,53)	307	(13,51)	3	0.12		f3	2810	(123,49)	2723	(122,49)	-86	0.01	*
m4	307	(23,52)	311	(24,52)	4	0.22		f1	3215	(165,72)	3164	(177,71)	-51	0.04	*
m1	301	(17,81)	306	(17,82)	5	0.04	*	m3	2088	(52,78)	2052	(48,79)	-36	0.01	*
m6	249	(16,48)	255	(15,48)	6	0.03	*	m2	2341	(73,76)	2323	(70,76)	-18	0.06	+
m2	245	(15,76)	252	(18,76)	7	0.01	*	m5	1997	(55,53)	1983	(57,51)	-14	0.10	+
f2	286	(28,79)	293	(33,80)	7	0.09	+	m6	2228	(86,48)	2216	(90,48)	-12	0.25	
f4	367	(30,49)	374	(37,50)	7	0.15		f2	3106	(134,79)	3096	(143,80)	-11	0.32	
f5	286	(23,46)	297	(24,49)	11	0.01	*	m4	2126	(65,52)	2121	(69,52)	-5	0.34	
f6	332	(33,50)	354	(39,50)	23	0.01	*	m1	2141	(74,81)	2140	(85,82)	-1	0.46	
m3	365	(31,78)	389	(29,79)	24	0.01	*	f6	3008	(157,50)	3012	(152,50)	5	0.56	
f3	387	(27,49)	412	(21,49)	25	0.01	*	f4	2954	(80,49)	2963	(62,50)	9	0.73	
f1	356	(60,72)	389	(62,71)	33	0.01	*	f5	2956	(162,46)	3002	(177,49)	46	0.90	

* : > 95% confidence in a significant difference between population means, + : > 90% confidence. 1-sided t-tests.

Table 6.A.2
Within-subject formant comparisons between concordant and discordant trials

	F1								F2						
	Concordant		Discordant		Discordant-Concordant				Concordant		Discordant		Discordant-Concordant		
Subj.	Hz	(σ,N)	Hz	(σ,N)	Δ	p <		Subj.	Hz	(σ,N)	Hz	(σ,N)	Δ	p <	
[a]								[a]							
f5	876	(64,49)	850	(72,47)	-26	0.07	+	f1	1375	(101,69)	1332	(86,69)	-43	0.01	*
f1	960	(51,69)	943	(49,69)	-18	0.04	*	f5	1284	(119,49)	1258	(106,47)	-26	0.26	
f6	952	(58,48)	945	(59,47)	-7	0.58		m1	1133	(73,76)	1113	(62,79)	-20	0.07	+
m6	567	(50,47)	563	(55,44)	-5	0.67		f4	1690	(57,49)	1672	(66,49)	-17	0.17	
f2	803	(83,78)	806	(87,79)	2	0.86		m4	1277	(50,51)	1261	(47,50)	-16	0.10	+
m3	749	(29,82)	753	(36,79)	4	0.42		m3	1275	(38,82)	1261	(38,79)	-15	0.02	*
f3	851	(34,50)	856	(41,48)	5	0.51		f6	1376	(71,48)	1363	(65,47)	-14	0.34	
m4	793	(62,51)	798	(58,50)	5	0.69		m2	1239	(42,77)	1233	(35,73)	-6	0.35	
f4	994	(38,49)	1001	(32,49)	7	0.30		m5	1121	(39,53)	1119	(39,49)	-1	0.87	
m5	588	(42,53)	596	(50,49)	8	0.39		m6	1272	(47,47)	1273	(36,44)	1	0.88	
m2	668	(97,77)	683	(107,73)	15	0.38		f3	1512	(54,50)	1515	(54,48)	3	0.81	
m1	674	(47,76)	695	(50,79)	21	0.01	*	f2	1403	(67,78)	1410	(81,79)	7	0.56	
[i]								[i]							
f6	331	(34,50)	308	(22,49)	-23	0.00	*	m6	2163	(102,48)	2136	(101,48)	-27	0.19	
f3	364	(33,50)	352	(28,49)	-13	0.04	*	f1	3118	(151,72)	3097	(147,69)	-21	0.40	
m3	361	(32,77)	348	(31,80)	-13	0.01	*	f2	3100	(189,79)	3089	(179,76)	-11	0.70	
m4	312	(20,52)	303	(19,51)	-9	0.02	*	m5	1999	(60,54)	1996	(49,53)	-3	0.79	
m1	298	(18,81)	292	(18,78)	-5	0.07	+	m2	2322	(54,76)	2320	(77,73)	-2	0.82	
f4	353	(29,49)	350	(33,48)	-3	0.66		f4	3025	(63,49)	3034	(53,48)	9	0.45	
f5	286	(30,46)	282	(29,48)	-3	0.57		f3	2904	(132,50)	2916	(122,49)	12	0.64	
m2	253	(20,76)	251	(19,73)	-2	0.45		m4	2103	(74,52)	2125	(62,51)	22	0.11	
f2	277	(24,79)	275	(21,76)	-2	0.60		m1	2129	(77,81)	2157	(76,78)	28	0.02	*
m6	257	(15,48)	256	(22,48)	-1	0.77		m3	2088	(56,77)	2117	(64,80)	29	0.01	*
m5	305	(19,54)	305	(16,53)	0	0.95		f6	2994	(179,50)	3023	(165,49)	29	0.41	
f1	378	(60,72)	389	(59,69)	10	0.30		f5	2931	(185,46)	2991	(189,48)	60	0.13	

* : 95% confidence in a significant difference between population means, + : 90% confidence. 2-sided t-tests.

Chapter 7: Modeling hierarchical spatiotemporal patterns in speech rhythm and articulation

This chapter explores how a multifrequency system of coupled oscillators—waves—can account for temporal phenomena reported in Chapters 4 and 5, and describes how these wave systems can be embedded within interacting dynamical fields, which can be used to explain the spatial patterns in Chapter 6. The patterns of articulatory-prosodic interactions reported in Chapters 4 and 5 are predominantly temporal in that they involve the relative timing of events, whereas the patterns reported in Chapter 6 are predominantly spatial in that they involve the location of vowel targets in an hypothesized articulatory/acoustic space. This chapter suggests that other linguistic phenomena can be well understood in the context of this wave and field theory, and proposes that general cognitive mechanisms of coupling and stability underlie the seemingly modular and hierarchical organization of speech.

7.1 Multifrequency coupled oscillators model of rhythmic-gestural interaction

The hierarchical node and connection representation of the relations between prosodic systems and articulatory gestural systems does not generate predictions with sufficient temporal accuracy to address phenomena such as rhythmic-gestural covariability and correlations between rhythmicity and segmental deletion. In contrast, a dynamical model in which prosodic and gestural systems interact in the course of planning and production can account quite nicely for these effects. Chapter 3 presented most of the conceptual background for this model. Here more technical details are discussed, and simulation results are provided. We are first concerned specifically with the covariability observed between phrase-foot timing and [s]-[p] intergestural timing in the phrase *take on a spa* (Chapter 4), but the theoretical framework can be generalized to other covariability patterns.

An important assumption must be made that all relevant prosodic units—phrase, foot, syllable—and all relevant gestural units C_1[s], C_2[p], and V[a]—correspond to oscillatory planning systems with phase variables (θ_{Phr}, θ_{Ft}, θ_{σ}, θ_{C1}, θ_{C2}, θ_{V},) and radial frequencies ($2\pi\omega_{Phr}$, $2\pi\omega_{Ft}$, etc.). This is more than an assumption, really. It is a reconceptualization of what these "units" *ARE*, i.e. an existential heuristic that allows us to think of units not as objects or containers, but as patterns of change, trajectories in an abstract space. The systems are presumed to correspond to the transient states of the averaged dynamics of distributed, overlapping, contextually variable neural ensembles associated with speech planning (cf. Chapter 3).

There are many possibilities for how to mathematically describe a group of coupled oscillatory systems. What form should the oscillator equations take? What should the coupling functions be? How does noise enter into the picture? There are no unambiguous answers to these questions, but empirical data can inform these choices, and a guiding principle is to strike a balance between maximization of explanatory power and minimization of conceptual complexity. The model presented below draws heavily from Keith & Rand (1984), Saltzman & Byrd (2000), Nam & Saltzman (2003), Haken et al. (1996), and many others. However, for reasons described below, a first-pass instantiation of the model posits no amplitude dependence of coupling strength. The idea of a multi-timescale dynamical model integrating rhythmic and gestural speech systems is not novel, but to my knowledge this is the first time such a model has been used to explore relations between rhythmic and gestural variability. Much of the model is

inspired by work done to understand rhythmic interlimb coordination, which has obvious relevance to speech.

However, one of the key differences between the observables in the rhythmic-gestural covariability experiment and those in studies of rhythmic interlimb coordination is the lack of a measurable amplitude of oscillation. With swinging legs or wagging fingers, the motions of the phase variables correspond to physical motions that are both cyclic and have easily-measured positions in space. This is not so with the rhythmic measures derived from p-centers, which are spaceless points in time. By the Cummins & Port (1998) and Port (2003) hypotheses the locations of p-centers are attracted to pulses which are phase-locked to systems of coupled oscillations. Yet there is no known way to observe the amplitudes of these oscillations. With rhythmic interlimb coordination, measurable characteristics of the limbs (such as amplitude) can be reasonably attributed to the oscillatory systems hypothesized to govern those movements. In contrast, it would be hasty to endow speech-rhythmic oscillators with varying amplitudes in the absence of a directly related observable.

The same problem applies to the gestural systems. Physical correlates of gestures do not exhibit oscillatory amplitude variation, but rather, behave like systems with point-attractors. They have been modeled as overdamped mass-spring systems in the task dynamic approach (Saltzman & Munhall, 1989). Subsequently, Browman & Goldstein (2000) and Saltzman & Byrd (2000) hypothesized that gestural planning oscillators determine the relative timing of gestural onsets, and Nam & Saltzman (2003) modeled them as such. However, the gestural planning oscillators do not govern the amplitudes of the gestural systems, nor can their amplitudes be observed. For these reasons, the first pass model presented below does not attempt to account for the influence of oscillator amplitude on coupling strength (cf. Peper, Boer, de Poel, & Beek, 2008); instead coupling strengths are based entirely on phase relations and all systems have unit radial amplitude.

Not only do both rhythmic and gestural oscillatory systems in speech lack observable amplitudes, but their phases are only observable indirectly and at limited points. Foot-systems, for example, can be stipulated to have phase 0 at stressed syllable p-centers, but their phases are not easily inferred at other points in time. Gestural planning system phases can be stipulated to correspond to special points of corresponding gestures, such as position and velocity extrema, but these are not well-substantiated assumptions. Under the weakened assumption that the relative timing of gestural onsets is determined by gestural planning oscillator relative phases (Nam & Saltzman, 2003), then as long as gestural stiffness parameters (i.e. velocities, Saltzman & Munhall, 1989) are relatively invariant in the speech cycling task, differences in intergestural relative phase translate into temporal differences in timing of positional and velocity extrema associated with the gestures. In other words, we assume that variability in the gestural planning oscillators can be observed indirectly through measurement of kinematic landmarks. This view can be thought of as associating the gestural planning systems with "premotor" cognitive activity, which is further removed from movement control, and associating the mass-spring gestural systems with "motor" dynamics, which is more directly involved in motor control.

7.1.1 Model equations and parameters

Each oscillator in the model can be described in a simple fashion by differential equations governing its phase and amplitude, shown in Eqs. (7.1). These equations treat the instantaneous

phase velocity of each oscillator as the inherent frequency of the oscillator, ω, plus the sum of phase changes due to coupling forces exerted on the oscillator (*F*). The coupling forces are weighted by corresponding coupling strengths (*A*). A noise term (η) is present as well, adding Gaussian noise to the change in phase. The polar amplitudes (radii) of the oscillators are constant, and so in the absence of noise and coupling, the equations in (7.1) correspond to polar forms of harmonic oscillators. These systems can be visualized very straightforwardly as points moving around unit circles, which will speed up and slow down due to interactions with other systems. More biologically plausible models usually incorporate some form of nonlinear damping, such as Van der Pol and/or Rayleigh damping. For current purposes, these nonlinear terms are unnecessary, because the entirety of the conceptual workload can be carried by the notions of phase coupling and synchronization.

Eqs. 7.1—7.3:

$$(1) \qquad \dot{\theta}_i = \omega_i + \sum_j A_{ij} F(\phi_{ij}) + \eta_i$$

$$\dot{r}_i = 1$$

$$(2) \qquad \phi_{ij} = \omega_i \theta_j - \omega_j \theta_i \mod 2\pi$$

$$(3) \qquad F(\phi_{ij}) = -\frac{dV_{ij}}{d\phi_{ij}}$$

$$V(\phi_{ij}) = -\cos(\phi_{ij} - \Phi_{ij})$$

The coupling function *F* takes as its argument a generalized relative phase difference (φ) between a pair of oscillators. The generalized relative phase difference is defined in Eq. (7.2), and is modulo 2π like the phase variable. Oscillator frequencies (ω) are restricted to integer multiples of the lowest frequency, which belongs to the phrase oscillator. The phrase oscillator can be defined to have an inherent frequency of 2π with the introduction of an appropriately normalized time variable. It is important to note that fixing the phrase oscillator frequency and treating all other frequencies as harmonics of the phrase frequency is has important theoretical consequences—namely, that the phrase timescale is the most basic and the other timescales are derived from the phrase timescale. It could alternatively be the case that a shorter timescale should be considered the most basic, or, most likely, that *no* timescale is basic, and that harmonicity between frequencies is an approximated idealization emerging from the multitude of phase coupling interactions involved in the system. While this latter possibility is the most attractive from the perspective of self-organization, it is also the most challenging to parameterize and implement computationally—hence for convenience, the phrasal timescale has been taken to be basic here.

The choice of frequency relations between rhythmic systems is fairly straightforward when modeling a phrase repetition task. In accordance with Cummins & Port (1998) and Port (2003), the ratio of frequencies ω_{Phr}:ω_{Ft} is 1:2 in $\Phi_{0.5}$, and either 1:2 or 1:3 in the higher order Φ. The ratio of foot to syllable frequencies is similarly parameterizable as 1:*n*, where *n* is the number of syllables in the relevant interstress interval. The vowel gestural planning systems can

be assumed to adhere to a 1:1 frequency relation with the syllabic systems. There are several arguments for this: 1) the centrality of the vowel to the syllable, 2) the greater dynamic flexibility of the vowel duration, 3) the relatively greater perceptual load of the vowel, 4) the relatively stronger association of suprasegmental articulations (tone, stress, nasality, etc.) with the vowel. Diphthongal and triphthongal vowels—multi-moraic structures—can be modeled with anti-phase relations. The consonantal gestural systems can also be assumed to adhere to a 1:1 relation to the syllable, although justifications for this assumption are wanting.

Following Haken (1983) and Saltzman & Byrd (2000), the coupling force between a pair of oscillators is the negative of the derivative of the potential function (V) governing their generalized relative phase, shown in Eqs. 7.3. The potential force is zero when the relative phase between two oscillators is equal to the target relative phase, Φ_{ij}, which defines a stable equilibrium. The absolute value of the derivative of the potential corresponds to the speed with which relative phase changes, and is greatest when the relative phase is at $\Phi_{ij} \pm \pi/2$, halfway between the stable and unstable equilibria ($\Phi_{ij} \pm \pi$). The target relative phase Φ_{ij} could take any value, but it is theoretically much more parsimonious to limit Φ_{ij} to either in-phase or anti-phase coupling (0 or $\pm\pi$).

The coupling forces can either speed up or slow down the phase velocity of an oscillator. For any pair of oscillators, osc_1 and osc_2, the strengths of the coupling forces exerted upon osc_1 by osc_2 and vice versa are represented separately in the coupling strength parameter matrix (A_{ij}). This allows for either oscillator to drive the other one in a unidirectional, asymmetric manner, or for mutual interaction, which can be symmetric or asymmetric. Note that many people reserve the term "coupling" for only the bidirectional case. General hypothetical parameter matrices relevant for rhythmic-gestural interaction in productions of the phrase *take on a spa* are shown in Table 1.

Table 7.1 Coupling strength and target relative phase parameter matrices relevant to the syllable *spa* in the phrase *take on a spa*.

		Coupling strengths* A						Generalized relative phase targets Φ					
		Phr	Ft	σ	$C_{[s]}$	$C_{[p]}$	$V_{[a]}$	Phr	Ft	σ	$C_{[s]}$	$C_{[p]}$	$V_{[a]}$
$\omega_{Phr} = 2\pi$	Phr		a						0				
$2\omega_{Phr}$ / $3\omega_{Phr}$	Ft	a		a	b	b	b	0		0	0	0	0
$2\omega_{Ft}$ / $3\omega_{Ft}$	$\sigma_{/spa/}$		a		c	c	c		0		0	0	0
ω_{σ}	$C_{[s]}$					d	e					-π	0
ω_{σ}	$C_{[p]}$				d		e				π		0
ω_{σ}	$V_{[a]}$				f	f					0	0	

*Coupling strengths represent the influence of row *i* upon column *j*.

In the coupling structure in Table 7.1, the syllable [spa] belongs to its own foot, and interactions between this syllable and others are not considered. Moras have been omitted, as

well as many other gestures involved. C[s] represents the tongue blade protrusion gesture associated with [s], C[p] the bilabial gesture, and V[a] represents a tongue body gesture associated with [a]. A target relative phase of 0 indicates in-phase synchronization, and ±π anti-phase synchronization. Although there are two Ft systems involved in production of the phrase, they are synchronized in-phase and collapsed into a single system in the present treatment. Ft and σ may be coupled 2:1 or 3:1 to Phr and Ft, respectively, depending upon the rhythmic structure of the utterance, as well as the words involved. The only anti-phase specifications in the model adhere between the [s] and [p] gestures, and are responsible for a c-center effect relative to the vowel (cf. Chapter 3).

The parameters instantiated in the coupling strength matrix correspond to a subset of all possible hypotheses for such a model. Parameters appearing on both sides of the diagonal indicate either symmetric or asymmetric coupling. For example, the parameters *a* and *d* correspond to symmetric coupling, meaning that the phrase and foot systems, and also the consonantal gestures, mutually influence each other to the same extent. The parameters *e* and *f* allow for the possibility of asymmetric coupling between consonants and vowels, so that one may influence the other to a greater extent. Where parameters appear only on one side of the main diagonal, this indicates a unidirectional coupling (often called *driving*), which means that the system in the corresponding row influences the system in the corresponding column, but not vice versa. In the coupling structure above, *b* and *c* represent foot and syllable driving influences upon gestural systems.

7.1.2 Model simulations

Covariability between rhythmic and gestural systems can be modeled parsimoniously with variation of coupling strength between rhythmic systems (*a*). Decrease of rhythmic coupling strength reflects changes in rhythmic difficulty or increased interference between 1:2 and 1:3 frequency-locking associated with the higher-order Φ. More difficult rhythms are associated with greater variance in rhythmic timing, and the model can reproduce this effect with a decrease in inter-rhythmic coupling.

One way to conceptualize how rhythmic variability arises is from a coupling strength-frequency order relation, whereby the amplitude of the relative phase potential is higher for systems with lower-order frequency-locking (i.e. 1:2 > 1:3 > 1:4, etc.). This is essentially the strategy used by Haken et al. (1996): higher amplitude potentials in 1:2 coupling exert stronger forces to correct noise-induced perturbations in relative phase. As for why such a relation occurs, one reason may be a form of "interference" between potentials corresponding to different frequency locks. In any given rhythmic condition, potential forces corresponding to both 1:2 and 1:3 frequency-locking are operative, although one mode dominates. Because 1:2 is of lower-order, it exerts greater interference on 1:3 than the reverse.

Phase velocity noise plays a crucial role in how decreased inter-rhythmic coupling is translated to increased intergestural variability. Due to its Gaussian nature and the central limit theorem, the noise normally has only a small effect on oscillator phases (which translates to small effects on relative phase), but on occasion the noise has a larger effect. With relatively strong coupling forces between rhythmic systems, large noise-induced fluctuations in oscillator phases and relative phases are corrected more quickly, making rhythmic relative phases less variable. This in turn produces a more coherent, stronger overall force on gestural systems

through the rhythmic-gestural coupling. When there is more noise, this stabilizing effect is of larger magnitude.

An alternative way in which rhythmic-gestural covariability can arise is through a relation between the order of frequency-locking and noise. If higher-order frequency-locking is subject to more noise (perhaps due to interference from lower-order potentials), then whenever the 1:3 ratio dominates the coupling between phrase and feet, variability would be higher for both rhythmic and gestural systems.

Numerical simulations verified that the multifrequency system of coupled oscillators described above can replicate the experimentally observed rhythmic-gestural covariability. Figure 49 presents an example simulation of the model. The top panel shows oscillator positions (i.e. cos θ_i) and selected relative phases over time. For both rhythmic and gestural systems, the oscillators begin out of phase but eventually synchronize. This synchronization can be observed in the stabilization of generalized relative phases (φ). Initial oscillator phases in this example were θ = [1.02, 5.75, 0.30, 5.40, 0.09, 2.20] radians, corresponding in degrees to [58°, 329°, 17°, 309°, 5°, 126°]. Regardless of the initial phases of the oscillators, the same relative phase configuration is eventually reached; initial phases only have an effect on how long relative phases take to stabilize. Following Saltzman & Nam (2003), Φ_{C1C2} was set lower than $-\pi$ (ideal anti-phasing), in this case $\Phi_{C1C2} = -\pi/2$. Further details regarding simulation parameters can be found in Appendix A. For the Phr, Ft, and σ systems, the pattern is such that the Ft repeats twice within the Phr period, the σ repeats four times, and both of these oscillators hit a peak (i.e. $\theta = 0 = 2\pi$) approximately when the Phr oscillator does. For the C_1, C_2, and V systems, C_1 phase precedes V by the same amount that C_2 follows V—this exemplifies a c-center effect.

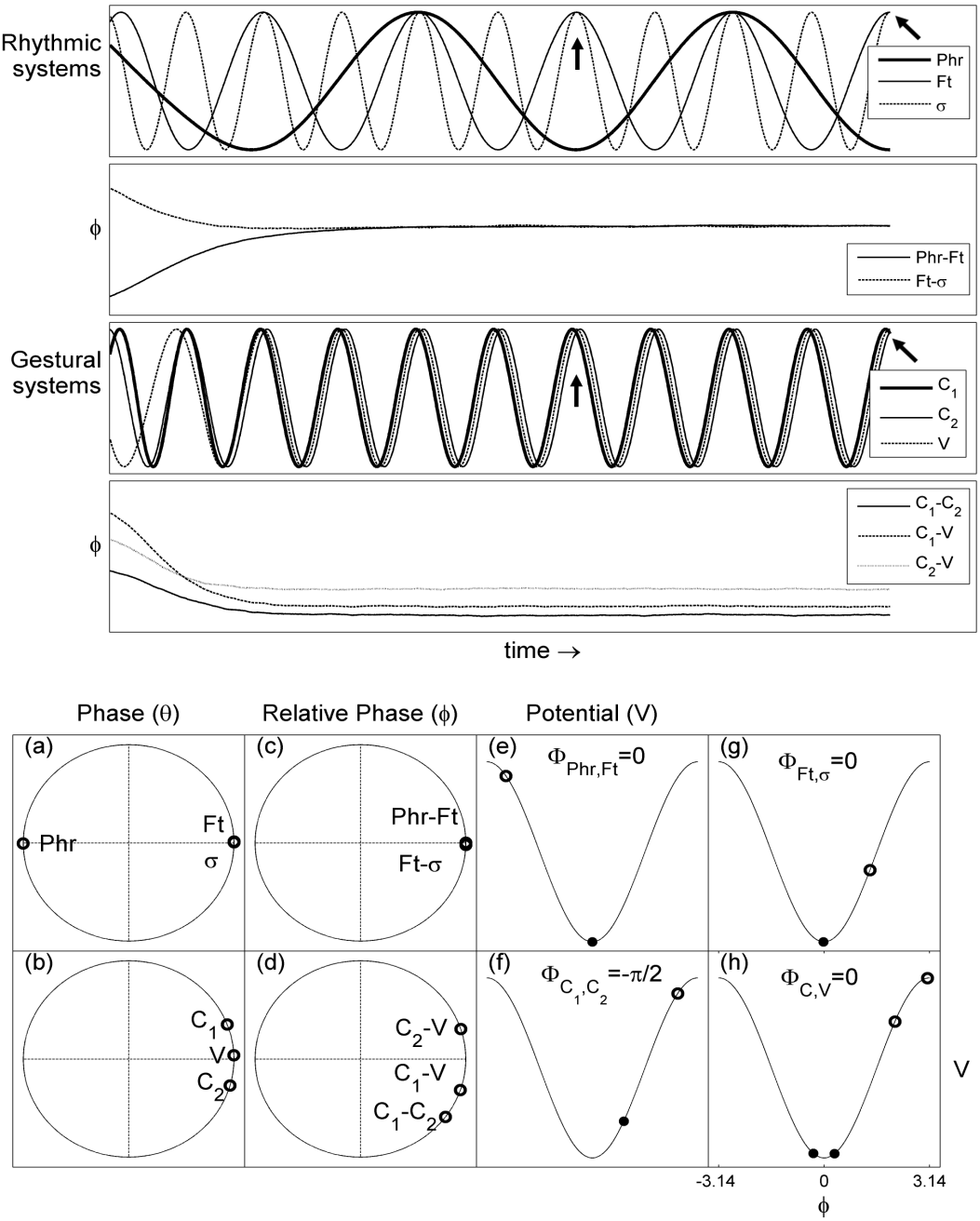


Figure 49. Example simulation. Oscillator motions and relative phases are shown over time; arrows indicate a point where relative phases are sampled. Simulation-final phases (a-b) and relative phases (c-d) are shown on the unit circle. Initial and final relative phases are also shown in their corresponding potentials (e-h), where open circles show starting relative phases and filled dots final relative phases.

The simulation in this example ends when Phr reaches a minimum ($\theta = \pi$), and thus the Ft, σ, and V oscillators are at a peak. Figure 49 shows the final oscillator phases on the unit circle (a,b), where phase angles increase counterclockwise. Relevant simulation-final generalized

relative phases are shown in (c,d). The initial and final relative phases are also shown in potentials (e-h). The potentials show that rhythmic relative phases reach their stable fixed points at the minima of their corresponding potential functions. The potentials in (f) and (h) show that φ_{C1C2} and φ_{CV} do not attain their lowest possible energy states; this occurs because the gestural forces influencing each C system oppose one another: each C system experiences a force toward in-phase synchronization with V, but simultaneously experiences a repulsive force away from the other C. The final compromise corresponds to a balance between these forces. As expected, the φ_{CV} are equally displaced from their target ($\Phi_{CV} = 0$) in opposite directions.

Simulated variabilities are presented below from four conditions: with 1:2 or 1:3 phrase-foot frequency ratios and with relatively high or low noise levels. Standard deviations (in radians) of φ_{C1C2} were calculated from 1500 samples of φ_{C1C2} in each condition (for further details see Appendix 8.A). Figure 50 shows that as inter-rhythmic coupling (*a*) is increased, intergestural relative phase φ_{C1C2} becomes more variable. In addition, if noise levels are taken to be frequency-dependent, then 1:3 inter-rhythmic coupling incurs more intergestural variability than 1:2 coupling. Otherwise, if noise-levels are independent of frequency no difference is observed (not shown)—although inter-rhythmic coupling strength has the same effect.

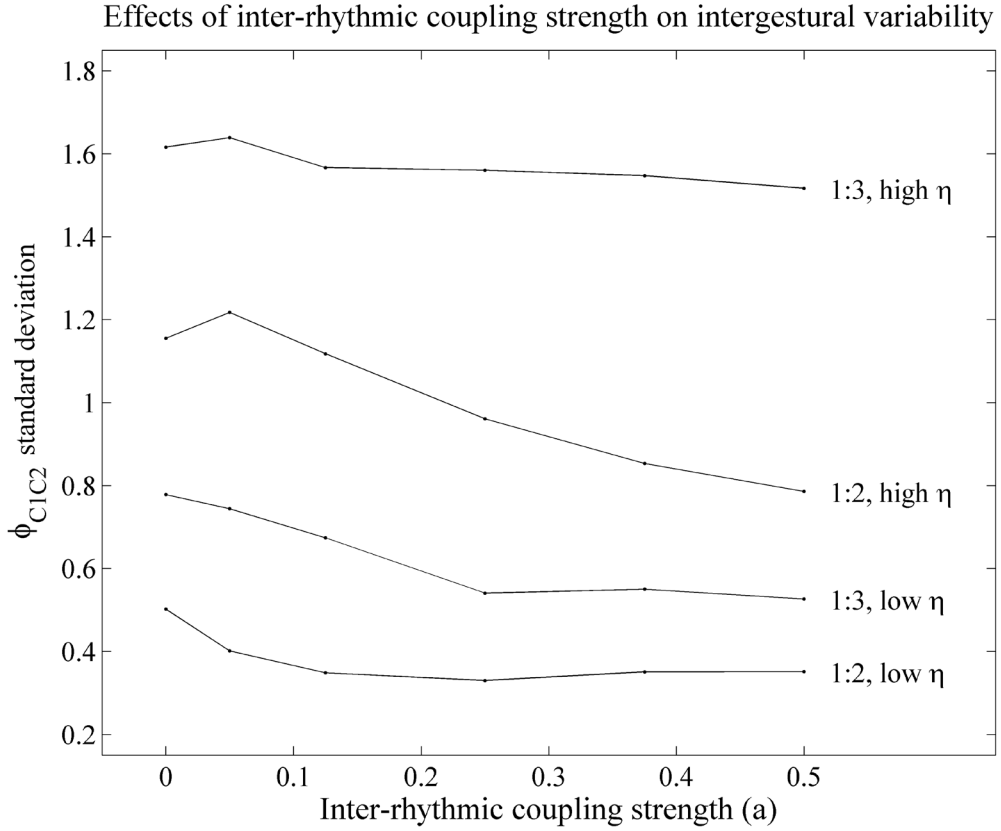


Figure 50. Effects of inter-rhythmic coupling strength on intergestural variability: st. dev. φ_{C1C2}. Standard deviations of φ_{C1C2} are plotted for a range of *a* (inter-rhythmic coupling strengths) for 1:2 and 1:3 phrase-foot coupling in relatively high and low noise-levels (η).

To translate the standard deviations in Figure 50 into quantities which can be more directly compared to the experimental data, some assumptions have to be made. A reasonable possibility is that one period of the syllable oscillator corresponds to the average duration of the syllable ($\Delta\sigma$), and so simulated durations can be derived by assuming $\Delta\sigma = 2\pi$. Then if $\Delta\sigma \approx 250$ ms, the standard deviations in Figure 50 range from about 16 ms to 66 ms, which agrees fairly well with the empirical range (cf. Appendix 8.A, Table 8.A.1). The changes in standard deviations observed within a given model as coupling strength is varied are smaller, on the order of 8 to 16 ms. These ranges are in line with the observations made for many subjects, although there are some who exhibited larger differences in intergestural standard deviations between Φ conditions; these may require additional mechanisms not captured by the model, or may be found in other regions of the model parameter space.

These simulations demonstrate that either one of two possible mechanisms may be responsible for rhythmic-gestural covariability. Weaker coupling between rhythmic systems leads to less coherent and weaker forces acting to stabilize the relative timing of gestural systems, resulting in increased intergestural variability; this effect is larger with greater noise levels. In addition, relatively greater noise-levels associated with the higher-order 1:3 frequency ratio result in greater intergestural variability. Importantly, rhythmic-gestural coupling strengths were held constant in all the simulations ($b = c = 0.5$), so effects are attributable entirely to changes in inter-rhythmic coupling (a). If rhythmic-gestural coupling is similarly increased in proportion to inter-rhythmic coupling, then the covariability effects become even larger.

In the present instantiation of the model, target relative phases of gestural onsets are derived from sampling gestural planning system relative phases, as in Nam & Saltzman (2003). However, an alternative is to couple linearly damped gestural oscillators to the limit cycle gestural planning oscillators, as has been proposed for interlimb coordination by (Beek, Peper, & Daffertshofer, 2001).

There are numerous ways in which the model presented above can be restructured or differently parameterized, some of which could be challenging—but not impossible—to distinguish experimentally. Regardless of which is correct, the importance of the model lies in its dynamical approach, particularly in its use of the concepts of stochastic influence, coupling, generalized relative phase, and multi-timescale/multifrequency interaction. These concepts are general and will likely outlive major changes in specific instantiations of the model. They allow for a coherent understanding of an otherwise perplexing experimental observation: the correlation of variability in intergestural timing with variability in rhythmic timing.

7.2 Modeling relations between speech rhythmicity and segmental deletion

With some additional assumptions, the model of rhythmic-gestural interaction described above can account for relations observed between speech rhythm and segmental deletion in the Buckeye corpus (Chapter 5). The deletions of interest here are phonetic deletions, rather than categorical phonological deletions.

Consider the following: if segmental deletion occurs because of increased gestural overlap, and increased intergestural variability occasions instances of increased gestural overlap, then the wrong prediction is made: more rhythmic speech should decrease intergestural variability and thus be correlated with decreased likelihood of segmental deletion. But this is the opposite of what was found: increased rhythmicity increases the likelihood of segmental

deletion. If the empirical findings are trustworthy, then either the model is wrong, or one of these assumptions is incorrect.

First let us assume that some segmental deletion does indeed occur because of increased gestural overlap. Next, consider the assumption that increased overlap arises from increased intergestural variability. This is undoubtedly sometimes true, but of course not necessarily true, as increased intergestural variability will equally often result in decreased overlap. Moreover, we need not rule out the possibility that other mechanisms may increase overlap too.

Another way by which increased intergestural overlap may occur is when intergestural coupling becomes weak relative to other relative coupling strengths. For example, as two consonants become more strongly coupled to a vowel or syllabic system, their mean stabilized relative phase becomes smaller. As long as this relative phase remains above some threshold which allows both gestures to be executed and perceptually recovered, no deletion occurs. However, below a threshold, this may change. In other words, *strong* rhythmic-gestural coupling can induce a high degree of overlap.

This predicts there are two ways such deletions can occur. One is through articulatory gestural overlap, the other involves a more premotor form of deletion that could arise from the inhibition of one of the gestural planning systems of a pair whose relative phase has fallen below a threshold. Figure 51 illustrates the gestural overlap scenario:

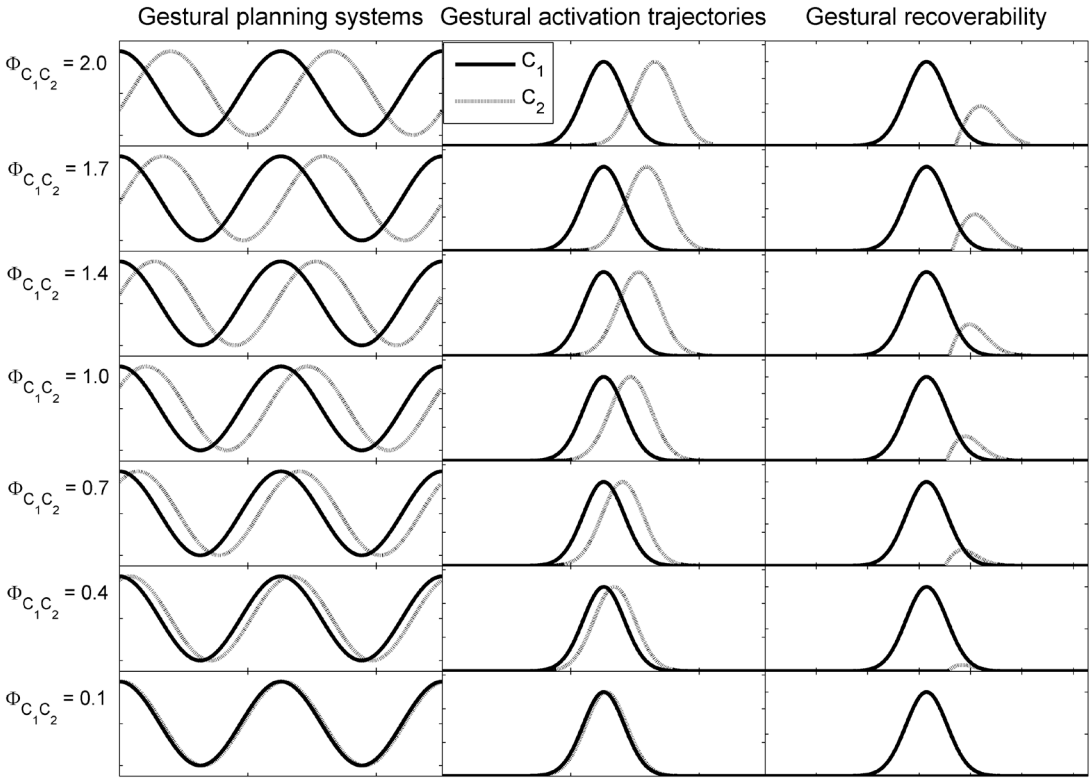


Figure 51. Changes in gestural overlap and recoverability as the relative phase of gestural planning systems decreases.

As the relative phase of the gestural planning systems decreases, their activation trajectories overlap to a greater extent. If the second gesture can be "hidden" by the first, for example when a tongue blade gesture is hidden by a bilabial closure, then its recoverability (perceptibility) decreases with relative phase. If the recoverability of the C_2 gesture falls below some threshold, it will not be perceived and a deletion has occurred. This scenario can be generalized to more complex situations, such as when a vowel of short duration is hidden by adjacent consonantal gestures. Such gestural hiding has indeed been observed in some kinematic studies (Browman & Goldstein, 1986), and is a standard account of how phonetic deletions arise.

Why, then, would gestural overlap be associated with highly rhythmic speech? As noted previously, if stronger rhythmic coupling decreases intergestural variability, then less gestural overlap should occur in more rhythmic speech. However, stronger rhythmic coupling to gestural systems also has the effect of decreasing intergestural relative phase. This occurs because the gestural system, being coupled to the rhythmic systems, experience relatively strong forces toward in-phase synchronization with the rhythmic systems. These forces are strong compared to the anti-phase coupling forces between the gestural systems. This leads to a high degree of gestural planning system overlap, which in turns leads to overlap of gestural activation trajectories and potentially diminished recoverability. Hence, both very rhythmic speech and very arrhythmic speech may lead to deletion, but for different reasons.

The second possibility, inhibition between planning systems, has not been proposed in previous literature, to my knowledge. It requires a fairly substantial addition to the model described above, but as will be discussed below, this addition is warranted for other reasons and greatly improves the explanatory scope of the model. The addition, in short, is intergestural inhibition (cf. Chapter 6, and later in this chapter). Intergestural inhibition is probably the key mechanism needed to account for how anti-phase coupling arises in the context of a neural network model. Importantly, this inhibition not only functions to keep certain gestural systems from overlapping, but alters the amplitudes of those functions. Recall that the multifrequency system of coupled oscillators presented above treats all planning system amplitudes as constant and non-interacting. What is being proposed here is that the model needs to be made more complicated by allowing for inhibitory and excitatory interactions between planning systems. This has not been fully implemented in the models reported here, but others have described related approaches (e.g. Schöner, 2002) for rhythmic interlimb coordination.

Figure 52 schematizes the hypothesized effect of intergestural inhibition in a CC cluster by assuming that the inhibitory force exerted by the C_1 planning system on the C_2 planning system is proportional to the activation of the C_1 planning system.

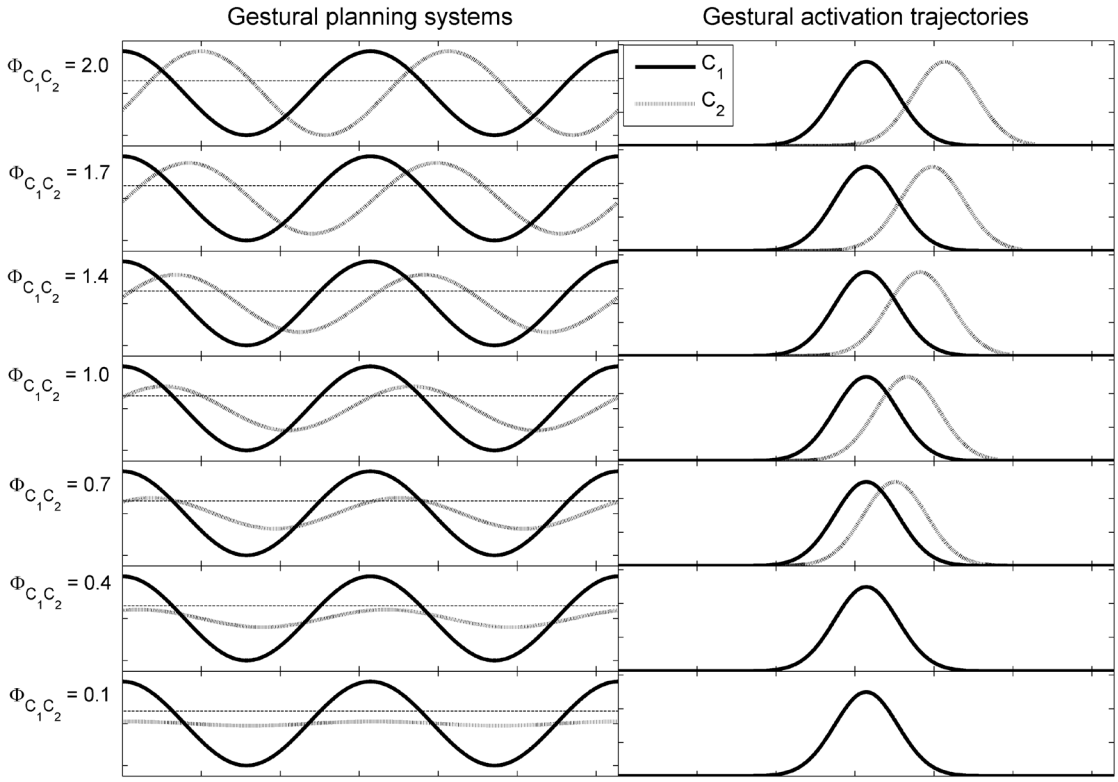


Figure 52. Deletion of a gesture caused by failure of a gestural planning system to reach threshold.

As the relative phase between the gestures decreases, the inhibitory effect of the first gesture upon the second increases. This decreases the amplitude of the gestural planning system associated with the second gesture. Normally the gestural planning amplitude is suprathreshold and hence triggers activation of the gestural system, but if the decrease in gestural planning system amplitude is large enough, the activation is subthreshold and no gestural activation is triggered.

This intergestural inhibition account can also explain the correlation observed between speech rhythmicity and segmental deletion. Highly rhythmic speech leads to gestural planning system overlap, which in turn leads increased intergestural inhibition. The inhibition potentially makes gestural planning system activity fall below an action threshold, in which case no gesture is produced.

If both accounts of deletion described above are correct, then there should be at least two types of phonetic deletion. Gestural overlap is a form of purely temporal deletion, in which the spatial properties of gestural planning are unaffected. In contrast, intergestural inhibition is spatiotemporal, because it involves a reduction in the spatial extent of gestural planning. How so? Perhaps the planning system amplitudes (e.g. the wave amplitudes in Figure 52) correspond to an integration of spatial patterns. The next section further develops this idea, which also turns out to be very useful in understanding the subphonemic priming and dissimilatory cross-phonemic priming reported in Chapter 6.

7.3 Dynamical field model of subphonemic and cross-phonemic priming

So far the model presented in this chapter has only dealt with the timing of speech rhythm and articulation, not the specification of targets. As shown in Chapter 6, gestural target specifications also have a dynamic component. Target specifications depend on recent perception and the interactions between contemporaneously active articulatory plans. This section explains how dynamical fields can accounts for these results. The integrated model—which incorporates the dynamical coordination of waves and the spatial organization of fields—accounts for all aspects of the experimental results in this thesis. Before presenting the dynamical field approach to modeling gestural planning, we examine some relevant results from other domains of motor control. A dynamical field approach has been used to understand these results, as well.

7.3.1 Selective inhibition in oculomotor and reaching movements

Sheliga et. al. (1994) report several experiments on eye movement (saccade) trajectories which provide a nice basis for understanding the dissimilatory patterns in cross-phonemic priming. Figure 53 details several stages in a trial of one such experiment. At the beginning of each trial, subjects first fixated on a central location on a screen, and then were directed to attend to a horizontally-oriented location in the peripheral visual field without relocating their fixation (a). Then one of two things would occur: either a visual imperative stimulus would appear in the cued horizontal location, or an auditory imperative stimulus would be heard. Subjects were instructed to make an up or down saccade to vertically oriented targets based upon which of the stimuli was perceived (b). Sheliga and colleagues found that the vertical saccade trajectories deviated *away* from the horizontal locations to which attention had been directed (c).

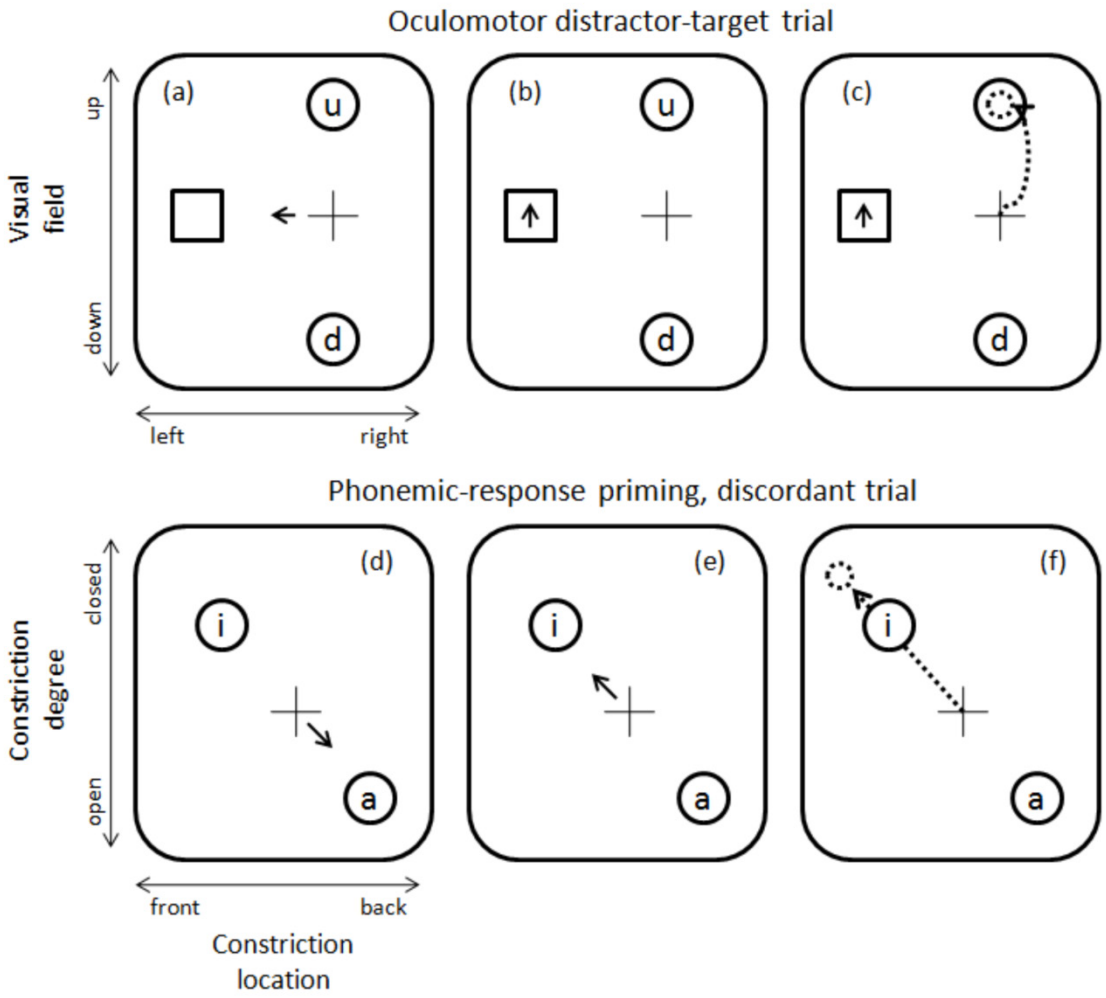


Figure 53. Schematization of distractor-target paradigm and phonological phonemic-response priming, discordant trials.

There are several main differences between oculomotor experiments of this sort (c.f. also Walker & Doyle, 2001; Van der Stigchel & Theeuwes, 2005) and the priming experiment in Chapter 6. For one, the fixation-to-distractor axis in the Sheliga et. al. (1994) experiment was oriented perpendicular to the two possible response trajectories. In contrast, in the study presented here, discordant trial distractors and targets shared approximately the same axis of articulatory motion (d, e), (assuming that the vocal tract was preshaped in an intermediate, schwa-like position). Accordingly, saccade deviations are measured in a plane perpendicular to the fixation-target axis, and hence deviate relative to more direct trajectories observed in control trials without distractors. In contrast, dissimilation and coarticulation are indirectly measurements of acoustic target overshoot or undershoot, which is a deviation of articulatory movement amplitude rather than a deviation of trajectory perpendicular to the main axis of motion. Despite these dissimilarities, there are suggestive similarities between the results observed in these experiments.

The premotor theory of attention (Rizzolatti, 1983) holds that the mechanisms of visual attention involve some of the same neural populations as those of saccade and reach planning;

this theory has been used to understand saccade and reaching trajectory deviations away from distractors. It follows from the theory that attention to the location of a distractor stimulus entails the planning of a saccade and reach to that location. Deviation away from a distractor (attended location) is held to result from the "selective inhibition" of motor plans associated with a saccade or reach to the distractor (Houghton & Tipper, 1994; Tipper and Houghton, 1996; Tipper, Howard, and Houghton, 1999). In this approach, movement targets are determined from the integrated activity of overlapping populations of neurons. The basic idea behind selective inhibition is that in order to saccade to or reach for a target, movement plans to competing targets must be selectively inhibited. Furthermore, more salient distractors evoke stronger selective inhibition. Strong inhibition of the population encoding the planning of a distractor response can thus shift the target response further away from the distractor, because their populations overlap to some extent:

> "Because each neuron's activity is broadly tuned, each cell will contribute to a variety of reaches. Thus, when two objects are present that both evoke reaches, the cell activities coding their directions can overlap, that is, some cells will be activated by both reaches. Inhibitory selection of one reach over the other may shift the population distribution in such a way that it affects the final reach to a target" (Tipper, Howard, & Houghton, 1999: 226).

If some version of the premotor theory of attention applies to speech movements, then there exists a partial analogy between oculomotor and reaching trajectory deviations away from distractors and the dissimilatory patterns in the phonemic-response priming task. The basic correspondence is this: just as saccades and reaches towards the distractors are planned and inhibited, the cue vowel response is planned and, in discordant trials, inhibited. Assuming there exists some overlap between the neural populations encoding the cue and non-cue response targets, then inhibition of the cue response would shift the articulatory target further away from the cue, causing dissimilation.

The subphonemic priming effects of the cue stimuli also have analogues in oculomotor and reaching studies. Van der Stigchel & Theeuwes (2006) cite a number of oculomotor studies in which deviations *toward* distractors occur when distractor and target are located close enough together (e.g. 20° to 30° of the visual field). Saccade endpoints in these cases are usually in-between the target and distractor stimulus. Ghez et. al. (1997) have reported similar findings for manual reaching. To model such findings, Tipper, Howard, & Houghton (1999) hold that when the distractor is relatively weak or located close to the target, no selective inhibition occurs and the response will be a compromise between the target and distractor. Erlhagen and Schöner (2002) present a dynamical field model capable of producing this result, in which multiple responses are represented by distributions of activity in a movement-planning field; when the distributions are close enough, they are both integrated into the response.

7.3.2 Dynamical field model of gestural planning, with intergestural inhibition

The model postulates two abstract vowel-planning spaces, a perception space and a planning space. For illustrative purposes, the spaces will be modeled here as two-dimensional, but presumably the results can be extended to higher-dimensional spaces. The perceptual space

can be defined in acoustic coordinates that correspond to F1 and F2, and the motor-planning space can be defined in either vocal tract coordinates that represent constriction degree and location or articulatory coordinates of tongue height and frontness/backness. We can for purposes of simplicity pretend that the coordinates are linear and the spaces are uniform, with the understanding that a more realistic model would introduce nonlinearities.

In addition, a mapping between the perceptual and motor-planning spaces provides a way for the two spaces to interact. The mapping need not be specified in detail for our purposes here, but one must be assumed to exist. The interaction of these spaces corresponds to the function of premotor-temporal/parietal mirror systems in which perception of an intentional gesture (e.g. a vowel) evokes premotor simulation of the same gesture (Rizzolatti & Arbib, 1998). This interaction is crucial for explaining the effects of subphonemic (i.e. subcategorical) priming. The idea that perception of a speech gesture relies on some of the same cognitive systems as the production of that gesture is the hallmark of any motor theory of speech perception (Liberman & Mattingly, 1985).

The workhorses of the model are two separate scalar fields, each defined over every point in their respective perceptual and motor spaces. The values of the scalar fields are activations, which are energy-like quantities that can be thought of as similar to neural potentials. Activation is interpreted differently depending upon which space is being considered: in the perceptual space, it represents the extent to which a given F1/F2 combination is perceived or attended to, and in the planning space, it represents the extent to which a given vocal tract or articulatory target is being planned. Smooth activation functions can be defined over both spaces at every instant in time. These functions can be used to visualize how phenomena such as subphonemic priming, coarticulation, and dissimilation may arise.

As with almost all models of speech production, we must invoke an "intention" to initiate a given categorical action. Intention should be thought of not as the conscious, willed act of an agent, but rather, as the spontaneous emergence of collective dynamics in prefrontal neural populations, arising from a complex interplay of brain-internal and external systems. In other words, intention is not "intentional" in the conventional sense of the word. The important role of intention in the field model is to excite and inhibit regions of vowel planning space that correspond to phonemic targets. The integration of excitation and inhibition constitutes the activation field, which describes how activation is distributed in the vowel-planning space from an arbitrary initial time to an arbitrary end time. Intentional excitation and inhibition can be associated with lexical and/or phonological representations—the representations should be thought of as specifications of distributions of excitation and inhibition.

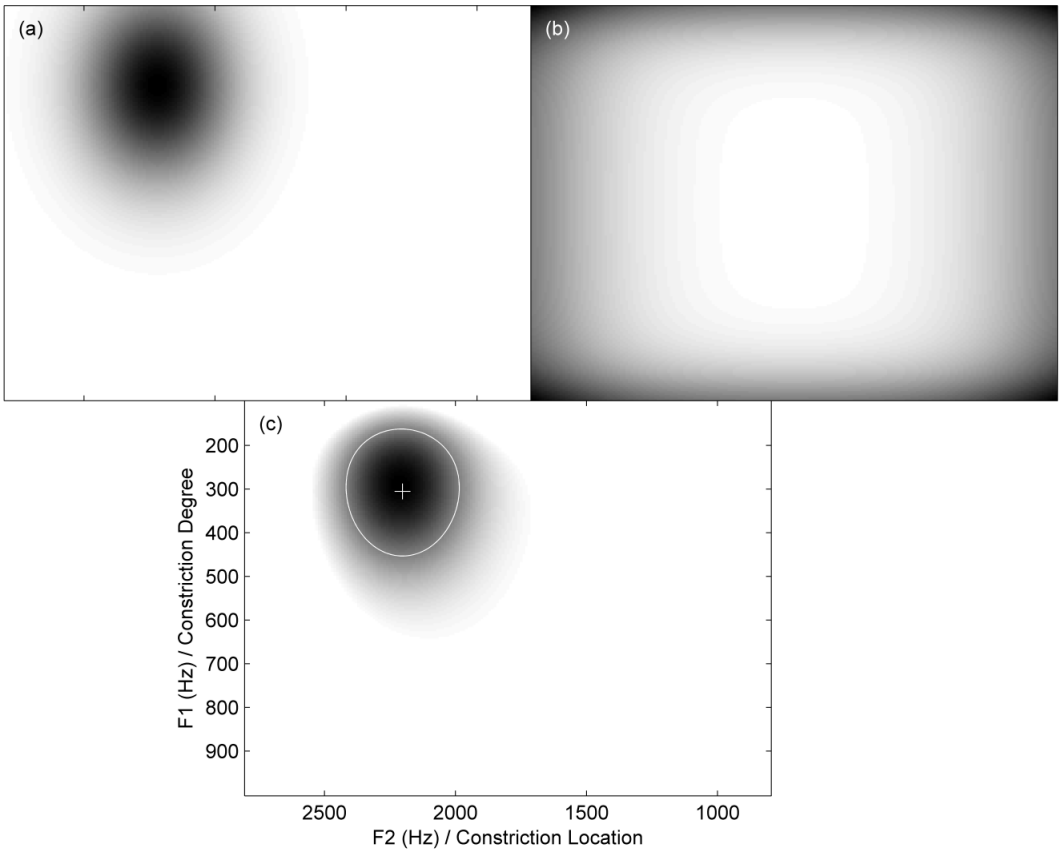


Figure 54. Activation field and components for production of /i/. Lexical excitation (a), boundary-constraint inhibition (b), and integrated planning activity with 50% threshold and vowel target (c) are shown; see text for details.

Figure 54 illustrates two components of an activation field, and their integration. The axes in panels (a), (b), and (c) are identical. Note that they are labeled with both acoustic and articulatory coordinates, reflecting agnosticism whether the activation field belongs to tract variable, articulatory, and/or perceptual spaces; for convenience the simulations presented here use acoustic formant values. Figure 54 (a) depicts the lexical excitation component of the vowel planning field, i.e. activation due to the intention to plan a vowel (which normally belongs to some lexical item). In this case the vowel is /i/. The lexical excitation components of activation fields are modeled here with bivariate Gaussian probability density functions that are modulated by a time-dependent driving function. There is no a priori reason for choosing a bivariate Gaussian rather than any number of other functions that could accomplish the same purpose. The main advantages of using bivariate Gaussians are the relatively few parameters needed to characterize them, the ease of extending them to higher-dimensional spaces, and the general familiarity with and widespread use of them.

Eq. (7.4) shows the probability density for a bivariate Gaussian function with zero correlation between the two variables, which yields elliptical contours of equal density. Two parameter vectors are necessary to describe the bivariate Gaussian, a mean vector (μ) that specifies the peak of the lexical excitation in planning space, and a standard deviation vector (σ) that specifies the spread of the excitation. Both parameter vectors have one element for each formant.

$$\mathrm{E}_{LEX}(F1, F2, t) = D(t)e^{-\left[\frac{(F1-\mu_{F1})^2}{{\sigma_{F1}}^2}+\frac{(F2-\mu_{F2})^2}{{\sigma_{F2}}^2}\right]} \quad (7.4)$$

The lexical excitation function E_{Lex} is the product of a bivariate Gaussian and a time-dependent driving function, *D(t)*, that describes when, and the extent to which, the lexical system excites vowel planning space. Choosing an exact form of the driving function *D(t)* is not necessary here—for current purposes it is sufficient for this function to exhibit some nonlinear growth and subsequent decay, representing the switching on and off of lexical excitation.

Speaker-specific constraints on the boundaries of the vowel planning space can be represented as an inhibition that diminishes from arbitrarily determined boundaries. It is beyond the scope of this paper to address the source of these constraints, particularly the question of whether they are purely cognitive or arise from muscular and physiological constraints on motor control (c.f. Liljencrants & Lindblom (1972) for similar constraints). Figure 54 (b) shows how such inhibition is distributed in planning space when it is the sum of sigmoidal functions of distance from the boundaries (Eq. 7.5). The parameters for this function are the boundaries (B), i.e. minimum and maximum formant values of vowel targets, and repulsive factors (R) associated with each of those boundaries that describe how far from the boundary (in Hz) the sigmoidally decaying inhibition reaches half of its maximum. Note that Eq. (7.5) attributes no temporal component to the inhibition, although a more powerful version of the model should do so.

$$\mathrm{I}_{BOUND}(F1, F2) = \sum_{i=1}^{2}\left[\frac{1}{1+e^{\frac{|F1-B_{F1i}|}{R_{F1i}}}}\right] + \sum_{i=1}^{2}\left[\frac{1}{1+e^{\frac{|F2-B_{F2i}|}{R_{F2i}}}}\right] \quad (7.5)$$

$$\dot{A}_{plan}(F1, F2, t) = E_{LEX}(F1, F2, t) - A_{plan}(F1, F2, t) - I_{BOUND}(F1, F2) \quad (7.6)$$

Eq. (7.6) presents a general form of the activation field equation. The temporal dynamics of the field are determined primarily by the growth and decay rates parameterized in the driving function that modulates lexical excitation. The field equation treats each point in the field as independent from every other point, but a more complex model might incorporate local interactions. For ease of implementation, the minimum value of any point in the field is zero.

Some additional mechanisms are required to translate from the vowel planning activation field to an articulatory target. An arbitrary threshold determines a subset of the field that contributes to the target. Figure 54 (c) shows a 50% activation contour enclosing the region of the planning space where the field is above the 50% activation threshold. There is no principled reason for a 50% threshold value per se, but varying the threshold generally produces only qualitative differences. All above-threshold points contribute to the determination of a production target, which is defined by the means of the formant values of the points inside the above-threshold region. The simulated [F1,F2] target vector in Figure 54 (c) is [305, 2202] Hz. To model variability in observed values, a random error component could contribute relatively small, normally distributed perturbations to the vowel targets; however such errors could also arise from lower levels of the motor control system, and thus are not included in Eq. (7.4).

The vowel targets (in articulatory coordinates) can serve as input to a production model, such as the task-dynamics model of speech gestures (Saltzman & Kelso, 1983; Saltzman, 1986; Saltzman & Munhall, 1989). In the task-dynamics approach, tongue dorsum constriction location and tongue dorsum constriction degree are tract variables whose motions in an abstract task space are determined by tract variable targets, which are dynamically turned on and off by gestural activation. This activation is the introduction of a driving force in a second-order mass-spring system that changes the equilibrium positions of the tract variables (masses). Tract variables can then be transformed into articulator motions (if an arbitrary weighting of articulators is assumed). In order to integrate the field model formant targets into this system, they could be transformed into task variables of dorsal constriction degree and location (although the details of such a transform would be quite complex).

The application of the field model to the results of the phonemic-response priming task employs one more crucial mechanism: selective inhibition. As described in the previous section, inhibition has been used to account for trajectory deviations away from planned responses in oculomotor and reaching experiments. In applying this concept to vowel planning, we utilize the mechanism of *intergestural inhibition*: selective inhibition of the planning field that arises from the production of a gesture and is focused on a contemporaneously planned gesture. In the present scenario, intergestural inhibition applies to coplanned vowel gestures. The source of intergestural inhibition should be viewed as external to the planning and perception fields, i.e. not as lateral inhibition within the fields, but rather, as a form of lexical and phonological inhibition—in other words, the inhibition is part of the long-term memory specific to a gesture.

Recall that the cross-phonemic primed vowel shadowing results in Chapter 6 showed that productions of /a/ and /i/ were dissimilated when the other vowel had been planned, i.e. on discordant trials. Figure 55 illustrates how intergestural inhibition can produce dissimilation between a planned cue vowel and a discordant response. The left side of the figure corresponds to a concordant trial with an /i/ cue and target; the right side to a discordant trial with an /ɑ/ cue and /i/ target. After the cue stimulus, the subject has prepared the possible responses to extents that reflect their probabilities of being the required response; to represent this, the lexical excitation components of the activation fields shown in panels (a) and (b) are driven by $D(t) = 2/3$ and $D(t) = 1/3$ for the cue and noncue responses, respectively. Panels (c) and (d) show the intergestural inhibition associated with production of the target, /i/. The precise time course of the inhibition, i.e. exactly when it takes effect relative to the production of the inhibiting gesture, is perhaps better left to future empirical investigation and development of the model. Following Houghton & Tipper (1999), the strength of the inhibition is greater for a more actively planned (or salient) nonresponse vowel—/ɑ/ in this case—and thus the intergestural inhibition is more influential on the discordant trial than the concordant one. The inhibition (Eq. 7.7) constitutes an additional term in the activation field equation (Eq. 7.8). It takes the form of a bivariate Gaussian that is modulated by the parameter α_{Inh}, which corresponds to the strength of the inhibition.

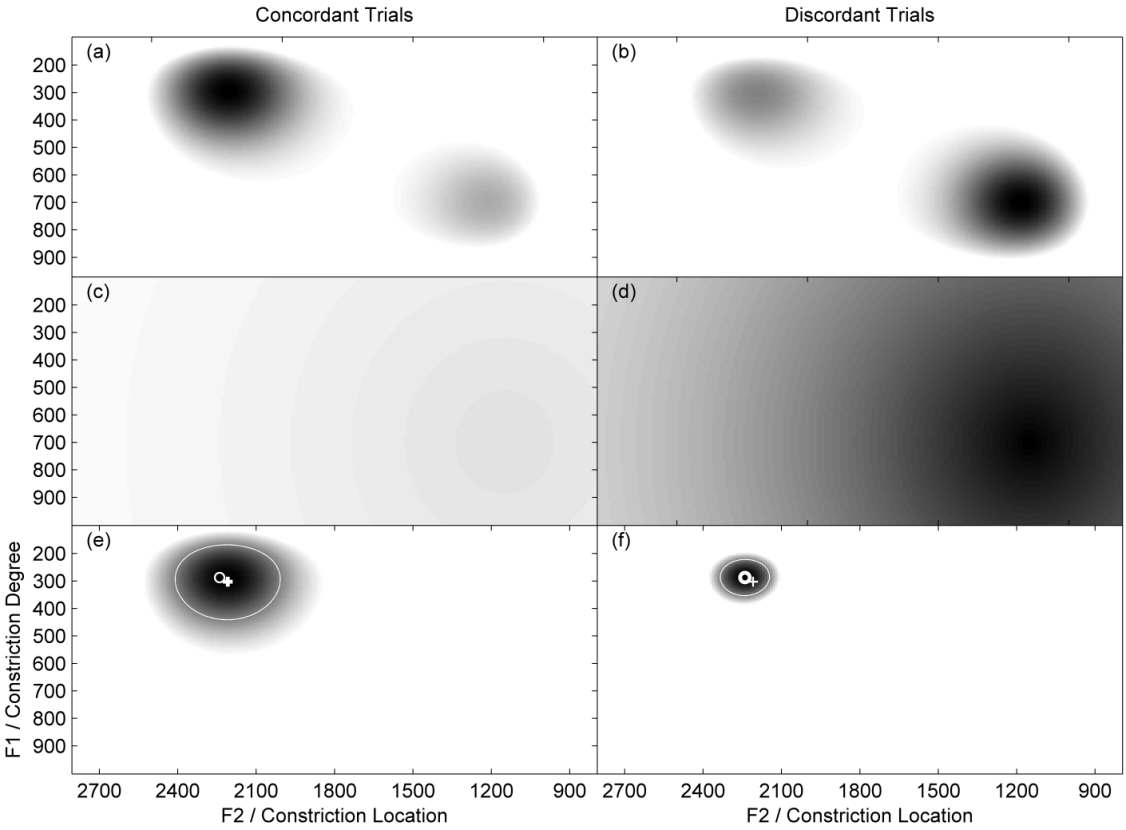


Figure 55. Activation fields at different stages of concordant and discordant trials. (a,b) post-cue activation on concordant and discordant trials. (c,d) intergestural inhibition. (e,f) response activation fields. White lines show 50% activation contours used to determine concordant targets (+) and discordant targets (o).

$$\mathrm{E}_{Intergestural}(F1, F2, t) = \alpha_{Inh} D(t) e^{-\left[\frac{(F1-\mu_{F1;NR})^2}{\sigma_{F1}{}^2} + \frac{(F2-\mu_{F2;NR})^2}{\sigma_{F2}{}^2}\right]} \quad (7.7)$$

$$\dot{A}_{plan}(t) = E_{LEX}(t) - A_{plan}(t) - I_{BOUND} - I_{Intervowel}(t) \quad (7.8)$$

On the discordant trial, after the /i/ target is known, the lexical excitation of /i/ rapidly increases (i.e. the driving force $D_{/i/}(t)$ increases to 1), and the lexical excitation of /ɑ/ gradually decays (i.e. $D_{/ɑ/}(t)$ changes slowly from the post-cue level of 2/3 to 0). Crucially, the rate of this decay is not by itself fast enough to prevent residual nonresponse /ɑ/ planning from being incorporated into the production target. Because of this, the intergestural inhibition affects the activation field more quickly than the decay of the nonresponse vowel excitation. The intergestural inhibition potentially eliminates any effect of competing vowel planning activity on the target.

Panels (e) and (f) in Figure 55 show the result of subtracting the intergestural inhibition from the post-target activation field (not shown), as well as 50% activation contours and the

resulting vowel targets. The discordant trial target ('o') is more peripheral than the concordant trial target ('+') because the intergestural inhibition is stronger and has a steeper gradient across the region of planning space corresponding to the response vowel, /i/. This results in a region of above-threshold activity that is relatively smaller and farther away from locus of inhibition. The activation field [F1,F2] vowel targets for the simulations shown in Figure 55 were [302, 2209] Hz for the concordant trial, and [287, 2240] Hz for the discordant trial, which constitutes a quasi-dissimilatory pattern of ΔF1 = 15 Hz and ΔF2 = -31 Hz.

Weaker intergestural inhibition, as shown in Figure 56, produces a quasi-coarticulatory difference between concordant and discordant trials. The weaker inhibition allows for some of the nonresponse vowel activity to remain above threshold (because of its relatively slow decay), especially in the discordant trial. The intergestural inhibition is altered by changing the parameter α_{Inh} in Eq. (7.7). Whereas for the simulation in Figure 55, concordant trial α_{Inh} = 0.5 and discordant trial α_{Inh} = 4, in the simulation shown in Figure 56, these parameters were reduced by a factor of 100, giving concordant trial α_{Inh} = 0.005 and discordant trial α_{Inh} = 0.04. This relatively weak intergestural inhibition is not sufficient to eliminate all of the residual planning activity of the cue vowel in the discordant trial. The targets for the simulations in Figure 56 were [305, 2202] Hz and [339, 2115] Hz for the concordant and discordant trials, resulting in a quasi-coarticulatory differences of ΔF1 = -34 Hz and ΔF2 = 88 Hz.

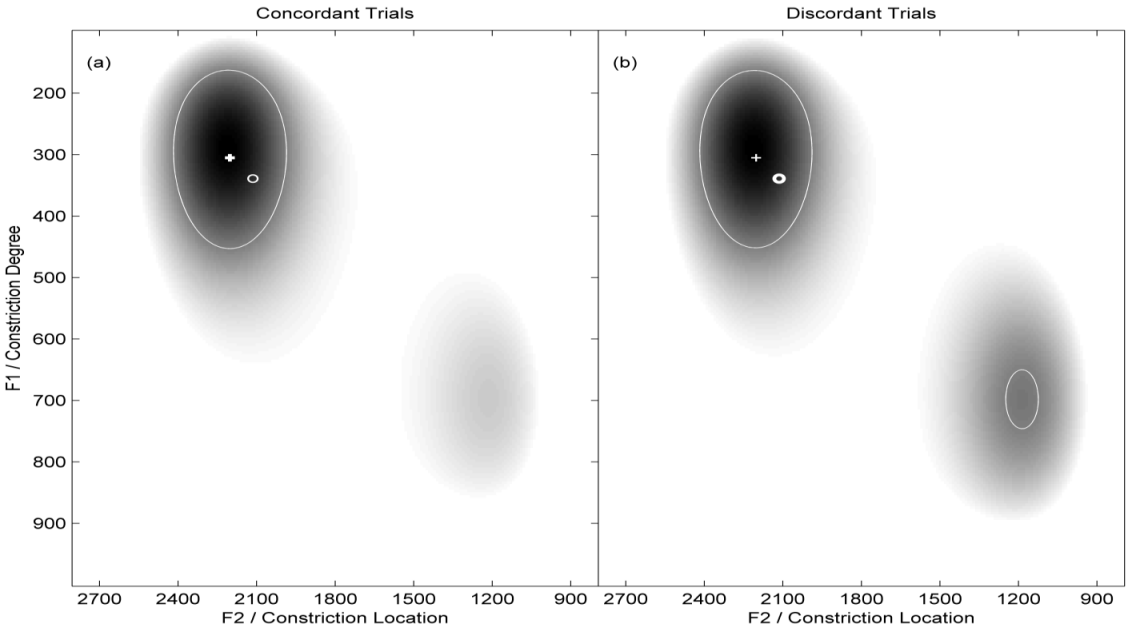


Figure 56. Response activation fields in concordant and discordant trials with relatively low levels of intergestural inhibition. White lines show 50% activation contours used to determine concordant targets (+) and discordant targets (o).

A useful aspect of this model is that variation in a single parameter, α_{Inh}, which represents the strength of intergestural inhibition, can produce three qualitatively distinct articulatory behaviors. A relatively large value of this parameter mimics dissimilation via intergestural inhibition of the region of the activation field that codes for the response vowel. A relatively low value of the parameter reproduces coarticulation by failing to sufficiently diminish activity in the

response region of the field. An intermediate value of the parameter results in a balance that leads to insubstantial degrees of coarticulation or dissimilation. In order to model the cases in which one formant exhibits dissimilation or coarticulation and the other exhibits no such trend, the σ parameters of the excitation and inhibition functions in Eqs. (7.4) and (7.7) can be manipulated to alter the range over which lexical excitation and intergestural inhibition have substantial impact on the field; these manipulations can affect the F1 and F2 dimensions independently, and could be used to obviate the α_{Inh} parameter altogether, although this option is not explored here.

To model the subphonemic priming effects of shifted cues, the mapping between the perceptual and planning spaces must be utilized. Perception of the cue stimulus activates a region of the perceptual field in a manner that is similar to the lexical excitation of the planning field. The perceptual activity is then mapped to planning activity. It is important that a sufficient level of phonetic detail be preserved in the mapping from the perceptual field to the planning field. The planning activity is then integrated with the lexical excitation from the response vowel. No intergestural inhibition is applied between the gestures because they belong to the same response category. Hence the model predicts that the response target will tend to fall in-between the cue and target stimuli. This is analogous to the results of the oculomotor and reaching experiments when the distractor and target are nearby in space. Indeed, one might associate with each gesture a region of target space within which no two target locations exhibit intergestural inhibition.

7.4 Extensions and further issues

The logical next step in understanding the speech phenomena reported in this dissertation is to explicitly (with equations) combine the dynamical field approach with a multifrequency system of coupled oscillators. The most direct way to do this is to posit the presence of a dynamic field for each articulatory and perceptual parameter, and then to treat the field activation functions corresponding to articulatory gestures as inherently oscillatory. In this approach, the rhythmic/prosodic systems would not require distinct fields: they would modulate activation across a number of the fields. In other words, think of the waves as descriptions of the activation in a region of a field. The activation waves correspond to planning an articulatory gesture, and to rhythmic and prosodic modulation of that planning. The dimensions of the field correspond to acoustic dimensions such as F1 and F2, or to vocal tract dimensions such as constriction degree and location.

An explicit formalization of this idea has not been attempted, because there are too many open questions and alternative possibilities that must be considered before the model should be implemented. Below we consider several of these.

7.4.1 Determining the set of fields and their (in)dependence.

One important issue is exactly what fields should be instantiated in the model. For now, let us put aside the complication that each field probably has distinct articulatory and perceptual realizations, and probably multiple realizations in each of those domains. Note that in the model above, a 2-dimensional field was used, but because the dimensions are independent, this model can be decomposed into two separate fields on two separate spaces. The question then is how

many of these 1-dimensional fields are needed? Intuitively, the fields should correspond to the articulatory and perceptual dimensions which are relevant to speech, perhaps precisely those that are "controlled" and that serve as "cues". It is likely that a suitable set of fields can be determined empirically, although procedures for doing so are needed. Perhaps subphonemic and cross-phonemic priming can inform these choices.

A second, even more challenging issue, is the interdependence of the fields. In the above treatment, the F1 and F2 fields were considered independent, avoiding many complications. However, it is highly unlikely that the assumption of independence is correct. Instead, the fields are probably dependent in very complex ways. For example, the planning activity associated with F1 probably interacts with the planning activity associated with F2, for several reasons. First, constraints on how articulators are moved and positioned in the vocal tract give rise to correlations between these dimensions. Second, proximity and overlapping of both motor and premotor cortex neural populations encoding articulatory gestures ensures that gestural planning in one dimension will interact with many others.

7.4.2 Transience of gestural planning activity

The multifrequency system of coupled oscillators described in section 7.1 is unrealistic because the oscillator equations do not produce amplitude variation. It was suggested above that a mechanism of intergestural inhibition impacts oscillator amplitudes, but there is an even more basic reason why amplitude change should be inherent to the system: gestural planning is transient. Without a doubt, planning must begin and end. The timecourse of this planning is presumably determined by higher-level semantic and syntactic planning, as well as intentional forces. Any model of gestural planning requires some assumptions about planning in semantic and syntactic domains.

One simple assumption is that gestural planning is "turned on" and "turned off" according to dynamics of semantic and syntactic planning. The equations below allow one to visualize the consequences of such semantic activation and deactivation on a single gestural planning system that is not coupled to any others (Eq. 7.9). This system differs noticeably from Eq. (7.1) in that the radial velocity is not constant, but instead governed by a potential function F, which represents the sum of all external semantic and syntactic influences on the amplitude of the gestural planning system. We will see presently that this potential function F is what accomplishes activation and deactivation. Note that Eq. (7.10) expresses the relation between polar coordinates and the Cartesian position-velocity coordinates used in the figures below.

$$d\theta/dt = \omega \qquad (7.9)$$
$$dr/dt = F(r)$$

$$x = r \cos \theta \qquad (7.10)$$
$$v = dx/dt = r \sin \theta$$

$$V(r) = \tfrac{1}{2} k_1 r^2 + \tfrac{1}{4} k_2 r^4 \qquad (7.11)$$

$$F(r) = -dV/dr = -k_1 r - k_2 r^3 \qquad (7.12)$$

Eq. (7.11) defines the potential function representing semantic and syntactic activation forces, along with the corresponding force function F in Eq. (7.12). The potential is merely a combination of polynomials, with the parameters k being crucial to the amplitude dynamics of the system.

Figure 57 shows an deactivated gestural planning system, where r_0 (initial amplitude) is approximately 0, $k_1 > 0$, and $k_2 > 0$. Whenever $k_1 > 0$, the potential function represents a system with a point attractor. If there were no noise in the system, the amplitude would eventually approach zero. However, the addition of a small amount of noise keeps this from happening. One can imagine the noise as providing small kicks to the amplitude in the potential function. The result is a low amplitude oscillation.

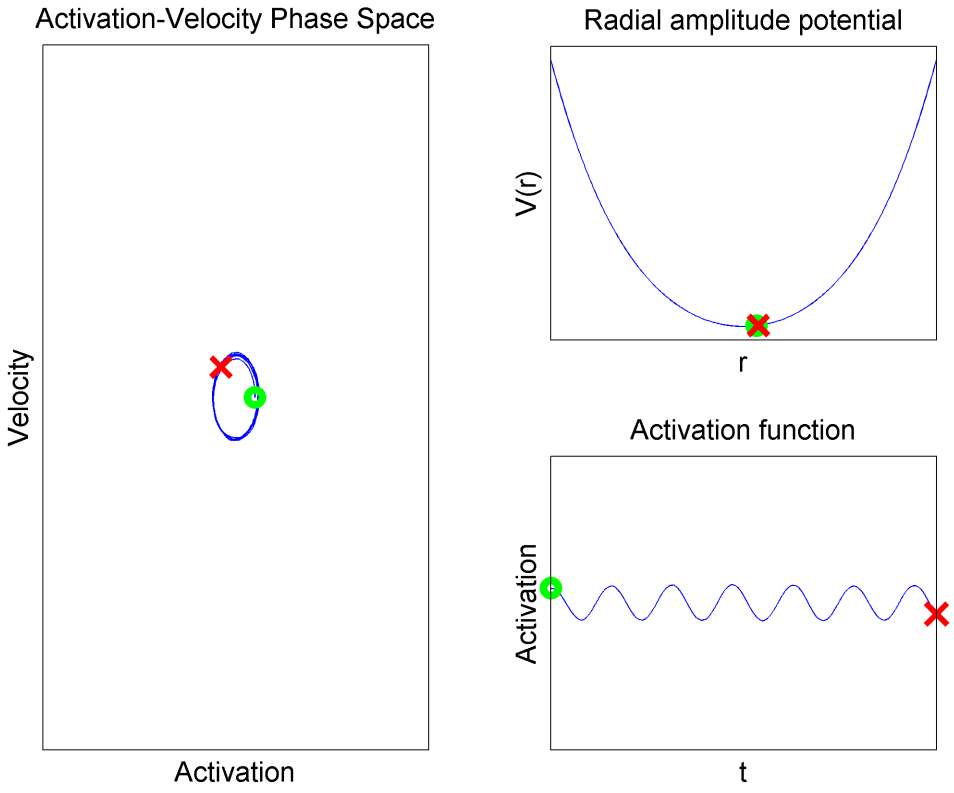


Figure 57. Deactivated gestural planning system. (o) initial state; (x) final state.

Figure 58 shows what happens when the system is activated. The activation or "turning on" is accomplished by changing the sign of parameter $k_1 < 0$, which is known as a supercritical Andronov-Hopf bifurcation. This gives the system a limit cycle attractor, and one can see that in activation-velocity phase space, the system spirals out toward the limit cycle. The ratio of k_1/k_2 determines how quickly this will occur and what the amplitude of the limit cycle will be.

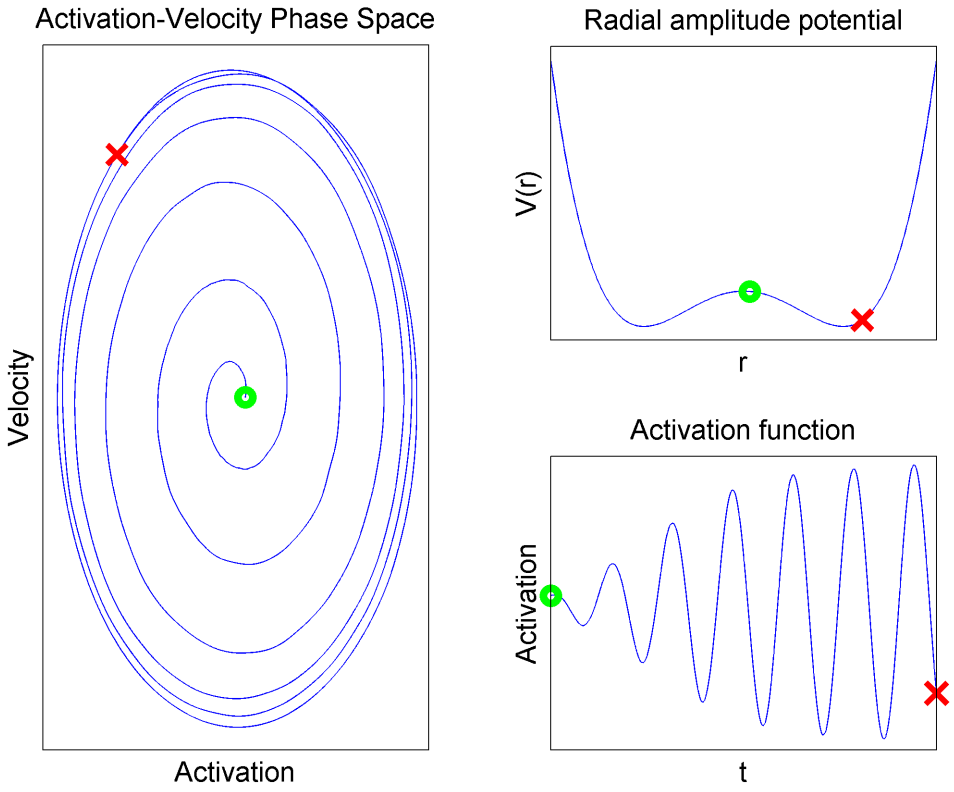


Figure 58. Activation of a gestural planning system. (o) initial state; (x) final state.

Exactly how quickly the amplitude grows and how long the limit cycle attractor persists will depend upon the semantic and syntactic forces. Eventually the gestural planning will be deactivated, i.e. $k_1 < 0$, and the state of the system will spiral back toward the point attractor (Figure 59). As with activation, the time course of deactivation will depend upon the forces exerted by semantic and syntactic systems.

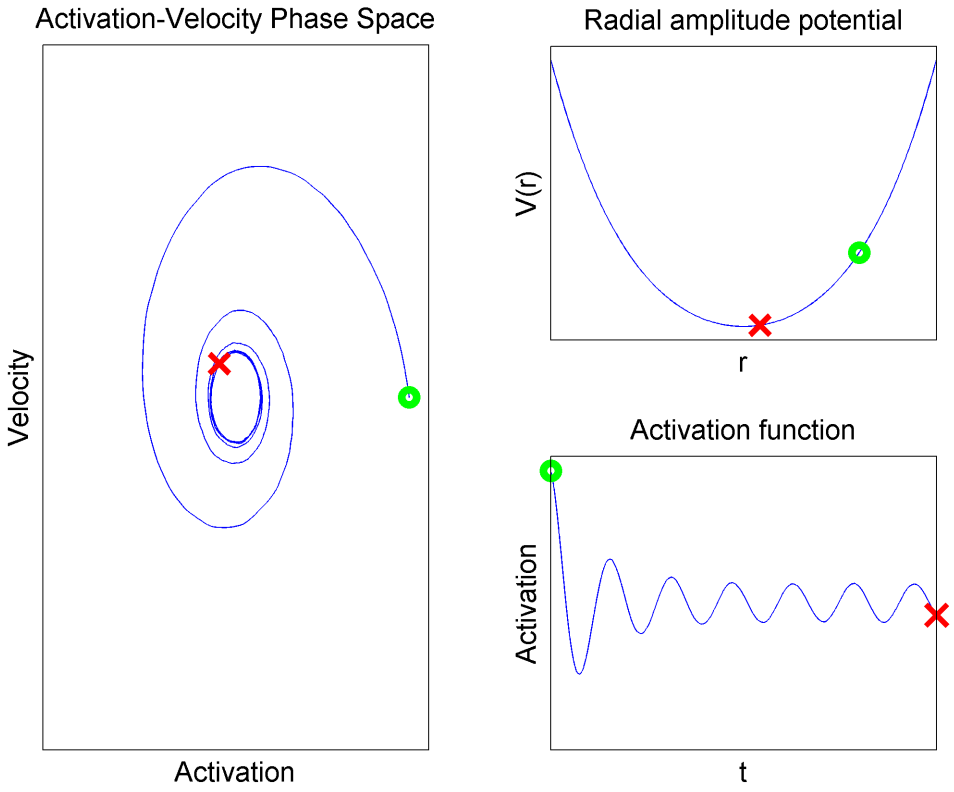


Figure 59. Deactivation of previously activated system. (o) initial state; (x) final state.

The picture that emerges when considering activation of planning systems is that the potential functions governing planning system amplitudes must also be dynamic. As a result, the planning systems exhibit transient jumps from a deactivated phase to an activated phase and back again to a deactivated phase.

Why go to the trouble of modeling these phase transitions, why complicate the model? The multifrequency system of coupled oscillators was fairly successful with constant amplitudes, which represent the behavior of the system in the limit of infinite time, i.e. system dynamics after amplitudes reached potential minima. It is one thing to argue for the complication by observing that gestural planning must begin and end, but what evidence is there that this complication increases the predictive or explanatory power of the model?

Currently there is no unambiguous evidence of transient oscillatory dynamics. However, interesting predictions can be made, some of which may already be supported. First, consider that in spontaneous speech, there may be little time between the activation of a set of gestural planning systems and their execution (execution is discussed in more detail below). This of course depends upon how quickly the speaker is talking and the extent to which their speech has been pre-planned or pre-formulated. In laboratory speech paradigms, these factors can be partly controlled; however, giving speakers an auditory stimulus to shadow or visual stimulus to read are not very natural situations. Regardless, if the time between the activation of gestural planning and the execution of those gestures is short enough, then limiting behavior may not be reached. I will refer to this phase of the dynamics as an *onset amplitude transience period*.

Now, if initial oscillator phases are effectively random (i.e. unpredictable, perhaps because the timecourse of activation of semantic systems is more chaotic), then there is necessarily some onset phase transience. If we further assume that phase coupling potentials are modulated by oscillator amplitudes, then the effects of onset phase transience would be heightened. To illustrate, imagine that two inactive gestural systems coupled in-phase are activated by semantic systems. The systems are coupled such that the strength of relative phase coupling depends upon the amplitudes of the systems. In other words, as the amplitudes of the planning systems grow, so do coupling forces. Now, the initial relative phase of the two systems is random, but will eventually approach 0. Because the initial amplitudes are small, phase-coupling forces are relatively small, and thus the onset phase transience period is relatively longer than in the case where amplitude and phase coupling strength are independent.

Given the above reasoning, here is the prediction: if execution occurs before the end of the onset phase transience period (before stabilization), then intergestural, inter-rhythmic, and rhythmic-gestural phasing should all be more variable over a number of samples. Anecdotally, this prediction seems right on: less planned, quicker speech is more errorful and variable. But an actual test of this idea in an experimental context should be considered before any conclusions are drawn about the validity of this idea.

An amplitude-phase coupling interaction may be useful in explaining observed differences between prepared speech and unprepared speech. Sternberg et al. (1988) report that in various studies of prepared speech, response latencies increase linearly with the number of stress groups in the pre-planned utterance. Using the same paradigm but with less response preparation, Wheeldon & Lahiri (1997) found that response latencies increased linearly with the number of units in the first stress group (or prosodic word) of the utterance. This suggests that given enough planning time, longer timescale prosodic systems associated with feet and phrases stabilize and would be expected to exert greater effects on utterance initiation. In contrast, with limited planning time, utterance initiation may be less effected by these lower frequency systems. If utterance initiation is furthermore held to involve the inhibition of competing responses, as Sternberg and colleagues have argued, then the asymmetry between prepared and unprepared speech can be seen to follow from differences in the amplitude of low-frequency prosodic systems when selection of the first unit of the utterance occurs.

7.4.3 Harmonicity of gestural planning activity

Another consideration is that rhythmic and gestural systems (and perhaps semantic and syntactic ones too) are not restricted to oscillate at a single inherent frequency (approximately), but rather, always exhibit harmonic and subharmonic oscillations of various amplitudes. This may be one way to reconcile the pervasiveness of delta (1.5-4 Hz), theta (4-8 Hz), alpha (8-13 Hz) and beta band (14-30 Hz) oscillations with a model whose oscillatory frequencies are associated more closely with behavioral periods (1-10 Hz). Furthermore, it allows for simultaneous in-phase and anti-phase synchronization, which provides another mechanism (in addition to generalized relative phase coupling) to explain how the constituents of a unit, i.e. the component systems, are bound together, yet mutually inhibitory. For example, although two vowel gestures in a diphthong are anti-phase synchronized, they could potentially be in-phase synchronized at harmonic and subharmonic frequencies.

The consequences of this idea have not yet been fully worked out, but it could be relevant to understanding a phenomenon recently reported by Goldstein et al. (2007) known as "gestural intrusion". Goldstein et al. conducted an experiment to produce such phenomena with gestural and segmental systems. The basic paradigm of the experiment is to have subjects repeat (along with a metronome) two words that differ with respect to one segment, e.g. *cop top cop top*… (cf. Figure 60). This sort of tongue-twister induces errors in the onset consonant gestures. In this example the onset consonant of the first word involves a tongue dorsum gesture and the second word a tongue tip gesture. The errors one might expect thus involve misplaced dorsal or tongue tip gestures. Goldstein et al. indeed found evidence for such insertions and deletions.

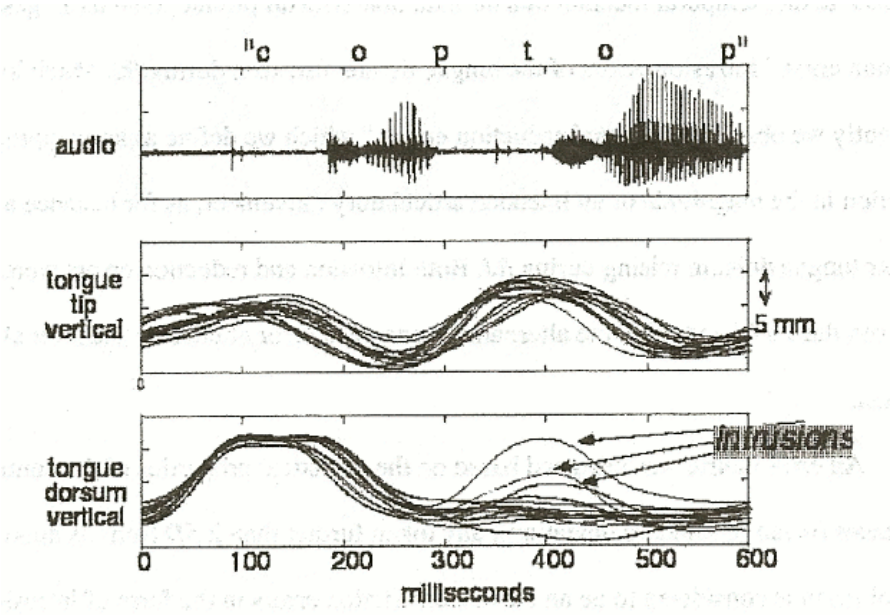


Figure 60. *Overlay of repetitions of cop top from Goldstein et. al (2007)*. (top) audio. (middle) tongue tip height. (bottom) tongue dorsum height.

Because they collected kinematic data, Goldstein et al. were able to distinguish between degrees of insertion and deletion. They found such errors to be gradient. In their analysis (and terminology), insertions are categorized as *full intrusions* or *partial intrusions*, and deletions are either *full reductions* or *partial reductions*. They found that partial errors are more common than full errors, which may suggest that such abnormal articulations are much more common than previously suspected in normal speech. By far the most frequent type of errors were intrusions. Also, errors increased with token repetition number within a trial and with repetition rate.

In a second experiment, coda consonants associated with multiple gestures were used, e.g. *bang bad bang bad...* In this case, there are potentially two gestures which may intrude upon the coda of the second word: the dorsal gesture and velar gesture of /ng/. They found that either or both gestures can potentially intrude. When both gestures intrude, the error can be considered segmental. Segments are generally considered to be gestural "molecules" in task-dynamic approaches.

The Goldstein et. al. interpretation of their experimental results employs the notion of a transition from 1:2 frequency-locking to 1:1 frequency-locking. Higher-order modes of frequency locking are known to be less stable than lower-order ones (cf. Chapter 3). Haken et. al.

(1996) describes the dynamics of such transitions. For *cop top*, for example, they observe that in an intrusion-free production, lip aperture cycles are in a 2:1 frequency relation with both tongue dorsum and tongue tip cycles. Intrusions of tongue dorsum or tongue tip gestures can thus be seen as transitions to more stable 1:1 frequency-locking between gestural planning oscillators (cf. Figure 61).

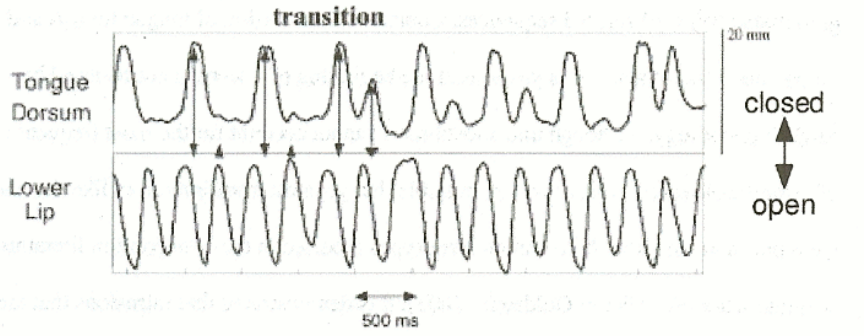


Figure 61. Transition in kinematic data from 2:1 to 1:1 frequency locking, from Goldstein et. al (2007)

A slightly different interpretation of full and partial intrusion phenomena involving harmonic oscillations is also possible. Rather than assuming that the tongue dorsum and lower lip planning systems normally are restricted to oscillations at a single frequency, it could be that both systems exhibit oscillations at both frequencies. In that case, the normal situation would require that the planning system frequencies not corresponding to the behavioral frequencies would be of relatively low amplitude. However, if for some reason the second harmonic of the tongue dorsum planning system were to exhibit increased amplitude, then intrusions such as observed in Figure 61 could arise. The cause of this increased amplitude might even be decreased inhibition between gestures.

7.4.4 Execution timing

The most problematic aspect of the modeling ideas in this chapter is the glaring omission of a mechanism determining how and when, precisely, gestures are physically executed. This can be thought of as a mechanism for translating the premotor planning system activity into motor system activity directing gestural performance. If the gestural planning systems are oscillating, even transiently, then what prevents gestures from being executed on each planning system cycle? It was assumed heretofore that the relative phasing hypothesized to be present between planning systems translates fairly faithfully to the timing of motor execution, but no well-developed model for this has been presented.

Indeed, our understanding of this crucial aspect of the model—the semantic and syntactic forces exerted upon gestural planning—is wanting as well. For current purposes, these systems might as well be tiny imps which decide when activation and deactivation occur, and how quickly limiting behavior is reached. But the lack of motivated assumptions about how these systems behave does not constitute counterevidence against the model. Additional inhibitory mechanisms and thresholding may play important roles in this regard. If anything, the model brings to the fore a new way of conceptualizing how semantics and syntax interact with speech rhythm and articulation. What is more, the very same mechanisms that have been argued to govern rhythmic and gestural planning and production may describe how morphosyntactic and semantic systems evolve in the timecourse of planning an utterance.

7.5 General mechanisms for speech planning

In the preceding sections, we have discussed four aspects of the modeling that warrant further consideration: field structure, transience, harmonicity, and execution timing. Despite these unresolved issues, the application to speech of the core ideas behind the model—multifrequency oscillations, planning fields, coupling, stochastic forces, and stability—are a solid first step toward reframing object- and connection- views of speech structure in dynamical terms that address planning from a cognitive perspective. There is furthermore a tantalizing and profound possibility that cognitive behaviors even further removed from movement can be usefully understood with similar interactions. Like prosodic structure in speech, syntactic structure has traditionally been represented using hierarchical branching schemas. Without committing oneself too much to any particular syntactic model, it is fair to say that one commonality among world languages appears to be the division of sentences into "constituents" that correspond to a subject noun phrase, one or two object noun phrases, and a verb phrase. Many syntactic theories hold that subject and object noun phrases are subconstituents of the verb phrase. The dynamical approach instead posits that each of these core constituent phrases: subject, verb, and object(s), correspond to oscillatory systems on the same timescale. Their subconstituent word classes (e.g. determiners, nouns, and verbs) correspond to oscillatory systems on a faster timescale. Figure 62 (a) schematizes one way in which the utterance "this sentence contains an object" may be organized dynamically.

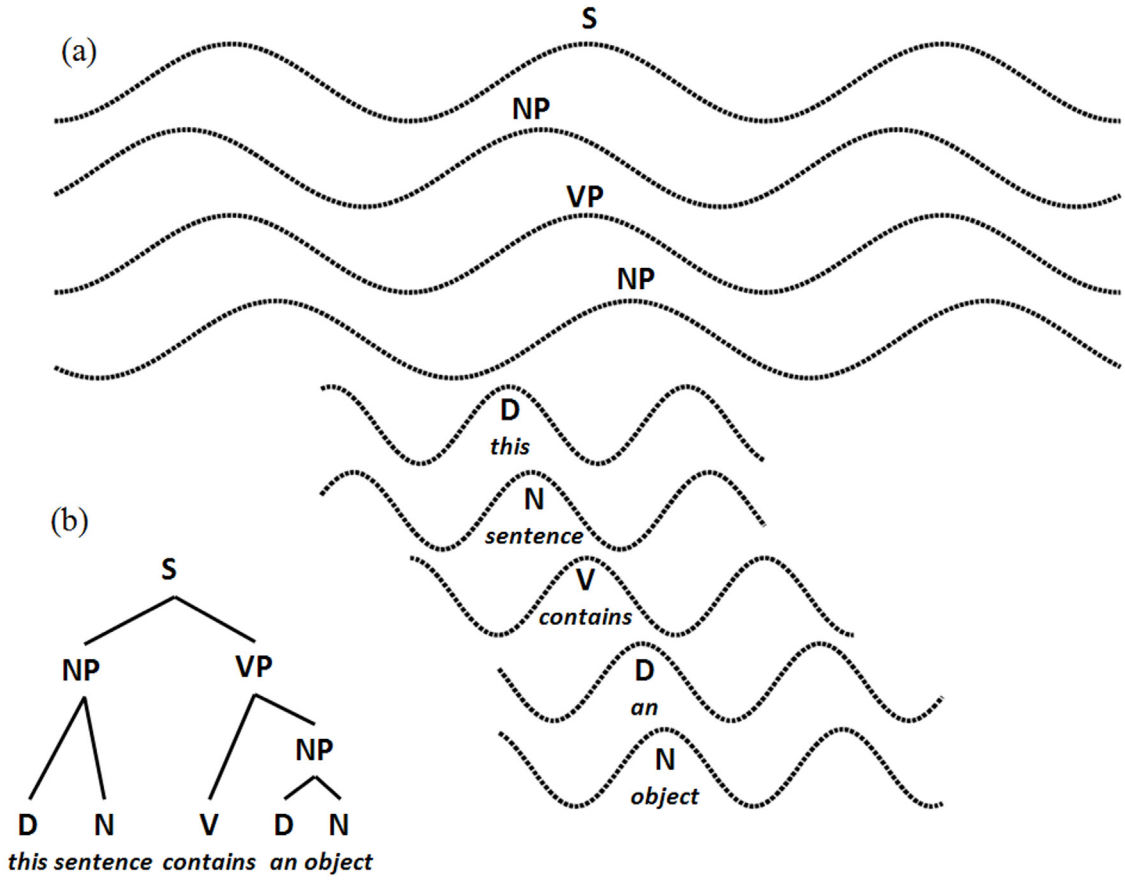


Figure 62. Dynamical organization of syntactic constituents of a sentence. (a) Oscillations of systems corresponding to sentence (S), noun phrases (NP) and verb phrase (VP), and word classes (Determiner, Noun, Verb). (b) Generic syntactic tree representation of the sentence.

One of the ideas presented in Figure 62 (a) is directly analogous to the c-center effect in CCV and CCCV systems. All three phrasal systems—subject, object, and verb phrase—are coupled synchronously to the sentence (or clause); in addition, anti-phase specifications adhere between the phrase systems. These conditions are general enough to dynamically generate the variety of phrasal patterns observed cross-linguistically. Patterns such as S-V-O, S-O-V, V-O-S, etc., can be generated by distinct coupling strength and target phase asymmetries which correspond to distinct regions of parameter space. Such asymmetries may be used to explain typological patterns. For example, in many languages there is evidence that objects are more closely related to verbs, and while there are certainly semantic reasons for this, these reasons may be realized dynamically through stronger verb-object coupling.

The subconstituents of phrases (word classes such as determiners, nouns, and verbs) can be organized in a fashion similar to the phrases themselves. These elements would be in-phase coupled to the phrasal systems that dominate them and anti-phased relative to systems within the same constituent. In contrast to rhythmic and gestural dynamics, it is not obvious that the effects of dynamical interactions between syntactic systems would be observed in 10-100 ms variation in the relative timing of phrases, words, and/or morphemes. Instead these effects would be

manifested in other ways, perhaps as constraints on structure derived from stability, as paradigmatic relations between constructions derived from changes in relative phasing, and as influences on the use of anaphoric devices from patterns of coherence or interference between contemporaneously oscillating systems. The general appeal of such an approach is that the same organizing principle can be understood to structure patterns in a variety of domains. That principle is two-fold: all systems are coupled in-phase to higher-level systems, and systems on the same timescale, if coupled, are anti-phased. Pattern regularity follows from coupling strength and relative phase asymmetries (which arise from broken symmetries), and pattern variability follows from stochastic forces.

7.5 Conclusion

This dissertation has attempted to show how hierarchical structure in speech can be reconceptualized as natural consequence of a system in which all subsystems exhibit wave-like dynamics and interact with each other through coupling forces. Undoubtedly the theoretical and modeling work presented in this chapter has raised more questions that it has answered. Despite this, the basic idea of the theory holds much promise for explaining how a very wide variety of linguistic patterns can arise from a small set of relatively simple cognitive mechanisms.

Appendix 7.A

Numerical simulations of the model used to produce Figure 49 were conducted using a 4^{th}-order Runge-Kutta algorithm in Matlab. Each simulation ran for 100 seconds of model time, with 2^5 iterations per second. Initial phases of each oscillator were randomly taken from a uniform distribution on the interval [0, 2π]. Oscillator frequencies ω_i for the Phr, Ft, σ, C_1, C_2, V were $2\pi \times$ [1, 2, 4, 4, 4, 4] and $2\pi \times$ [1, 3, 6, 6, 6, 6]. Phase and relative phase variables were wrapped on the unit circle; note that for clarity of conceptualization in Figure 49 phases were depicted on the interval [-π, π]. All relative phase targets were 0, except for Φ_{C1C2}, which following Nam & Saltzman (2003) was set at 3π/2, or equivalently, -π/2. Relative phases φ_{C1C2} were sampled from each foot-peak in each simulation after 10 s of model time had elapsed, which allows relative phases ample time to stabilize. The noise term η_i corresponds to the standard deviation of a Gaussian distribution that modulates oscillator frequency. Hence the frequency terms in Eq. (1) can be rewritten $\omega_i (1 + a_\eta \eta_i)$. For the high-noise condition, $\eta = 2$, and for low-noise $\eta = 1$. These are both fairly high noise levels relative to the oscillator frequencies, but on average only produce low-amplitude relative phase perturbations. Coupling effects of consonant oscillators on the vowel oscillator were $e = 0.1$ and of the vowel on the consonants were $f = 0.2$. Coupling between consonants (d) was $d = 0.25$ (see Table 1 for definition of variables). The change of oscillator phase due to each coupling term was taken simply as the $2\pi \times$ the negative sine of the difference between relative phase and target relative phase, modulated by the coupling strength. It is common in more sophisticated models to use the inverse Jacobian to partition a change of relative phase into changes of phase in component systems—although not strictly necessary, the Jacobian strategy is more theoretically attractive because it distinguishes explicitly between change of oscillator phase and change of relative phase, which is more precisely what is described by the negative of the derivative of the potential.

Appendix 7.B

The parameters for the simulation in Figure 55 were: $\mu_{V1ExcF1,F2} = [700, 1150]$, $\mu_{V2ExcF1,F2} = [280, 2220]$, $\sigma_{ExcF1,F2} = [150, 200]$, $R_{F1lower} = R_{F1upper} = 50$, $R_{F2upper} = R_{F2lower} = 200$, $B_{F1lims} = [100, 1000]$, $B_{F2lims} = [800, 2800]$, $\sigma_{InhF1} = |\mu_{V1ExcF1} - \mu_{V2ExcF1}|$, $\sigma_{InhF2} = |\mu_{V1ExcF2} - \mu_{V2ExcF2}|$, $\alpha_{Inh,conc} = 0.5$, $\alpha_{Inh,disc} = 4$. The temporal dimensions of the equations were not included in these simulations: the figures are static representations of the activation fields and components with $D_{/V:response/}(t) = 1$, and $D_{/V:nonresponse/}(t) = 1/3$ on the concordant trial and 2/3 on the discordant trial, corresponding to the hypothesized slow decay of cue planning activity. A 50% activation threshold was used for determining targets. Parameters for the simulation in Figure 56 were identical to those in Figure 55, with the exception that $\alpha_{Inh,conc}$ and $\alpha_{Inh,disc}$ were reduced by a factor of 100, corresponding to relatively weak inhibition.

Acknowledgements

If Keith Johnson had not come to the Berkeley Linguistics Department in my third year and taught an eye-opening seminar on speech production in the Spring of 2006, I would never have been introduced to the conceptual frameworks that inspired this work, and I would never have thought to devise the experiments herein. Keith provided crucial guidance on a range of topics, without which the quality of this work would have greatly suffered. Even more importantly, he did not discourage me from asking questions to which the answer is unknown. In my estimation, there are challenging questions that underlie this dissertation, addressing them required delicate and speculative experiments, and modeling the results of these experiments was an overly ambitious undertaking. Berkeley is a special place because I had the freedom to take these risks, but I would not have had the courage to follow through on them if Keith had not told me I was on to something.

Much credit is due to Sharon Inkelas, whose advice has substantially improved both the design of this thesis research and the presentation of results. Her graduate level morphology class was one of the best courses I have ever taken. That class was my first introduction to the range of complexity of linguistic systems. The aim of understanding that extraordinary complexity is the reason why I have pursued the domain-independent perspective of dynamical systems theory in my theoretical approach to modeling hierarchical patterns in prosodic and articulatory systems.

My research has benefitted greatly from the bottomlessly deep knowledge of experimental design available to me in discussions with Rich Ivry, whose perspective has shaped how I approach linguistic questions experimentally and theoretically.

Many thanks go to Peter Watson and Ben Munson, who generously enabled me to conduct a portion of this dissertation research at the Department of Speech-Language-Hearing Sciences at the University of Minnesota. Without their contributions of advice and lab facilities, I would not have been able to conduct the electromagnetic articulometry study that is a core part of this thesis.

Others have contributed in no small way to this dissertation or to my graduate education. Ronald Sprouse deserves much gratitude for technical help in the Berkeley Phonology Lab, and even more so for developing useful robust LPC and dynamic formant tracking algorithms. Without Ron being excited about the gritty details of signal processing, I would not have learned so much. Hosung Nam helped me in implementing a computational model of intergestural timing that appears in the dissertation. Ning Vei donated many hours of her time for elicitation of Kuki Thaadow that appears in a M.A. qualifying paper. Belén Flores deserves much thanks for helping me figure out how to navigate my way through the administrative bureaucracy of graduate school, and Natalie Babler and Paula Flores were helpful on numerous occasions.

There are a number of professors in the Berkeley Linguistics Department from whose instruction I have benefitted in the course of my graduate education: Larry Hyman, Sharon Inkelas, Keith Johnson, George Lakoff, Lynn Nichols, and John Ohala. I have also learned a lot from faculty for whom I have been a teaching assistant: Andrew Garrett, Larry Hyman (twice), Keith Johnson, Lynn Nichols, and Eve Sweetser. Thanks also to Alice Gaby and Ian Maddieson.

Two of my colleagues in the Linguistics Department, Molly Babel and Grant McGuire, have been of great assistance in numerous ways during the last several years of my graduate career. A number of other colleagues have also been helpful and supportive in various ways at various times: Lisa Bennett, Christian DiCanio, Yuni Kim, Russell Lee-Goldman, Joe Makin,

Mischa Park-Doob, Anne Pycha, Ryan Shosted, Maziar Toosarvandani, and Yao Yao. A handful of undergraduate students have helped me as well: Tyler Frawley, Marcele Januta, Emily Moeng, Kevin Sitek, and Anthony Wong.

Some of the research presented in this dissertation was made possible by a doctoral dissertation improvement grant from the National Science Foundation. I would like to thank the NSF for this support.

When approaching an important crossroads in my academic career, I was steered by Louis Goldstein and Dani Byrd from a less certain course onto a much more promising one. For this I am extremely grateful.

Finally, I would like to thank my mom, my dad, and my grandparents, Velma and Phil Veneziano, and Joyce and Robert Tilsen, for all of the diverse opportunities I had when growing up, without which I would not have found this path. Cade Tilsen has contributed in profound ways to my thinking on language. Most of all, for reasons that are too many to mention, I thank Kim Tilsen. Thank you, Kim.

References

Abbs, J. & Gracco, V. (1984). Control of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. *Journal of Neurophysiology*, 51:4, 705-723.

Abercrombie, D. (1967). *Elements of General Phonetics*. Chicago: Aldine.

Alfonso, P. & Van Lieshout, P. (1997). Spatial and temporal variability in obstruent gestural specification by stutterers and controls: comparisons across sessions. In W. Hulstijn, H. Peters, & P. Van Lieshout (Eds.), *Speech Production: Motor Control, Brain Research, and Fluency Disorders*, 151-160. Amsterdam, The Netherlands: Elsevier Publishers.

Allen, G. (1972). The location of rhythmic stress beats in English: an experimental study, parts I and II. *Language and Speech*, 15:72-100, 179-195.

Allen, G. (1975). Speech rhythm: Its relation to performance universals and articulatory timing. Journal of Phonetics, 3, 75–86.

Babel, M. (2009). Phonetic and Social Selectivity in Speech Accommodation. Doctoral dissertation. University of California, Berkeley.

Barbosa, P. (2002). Explaining Cross-Linguistic rhythmic variability via a coupled-oscillator model of rhythm production. Proceedings of Speech Prosody 2002.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty, & D. R. Ladd (Eds.), *Papers in laboratory phonology II: gesture, segment, prosody,* 68-86. Cambridge: Cambridge University Press.

Beek, P., Peper, C., & Daffertshofer, A. (2002). Modeling rhythmic interlimb coordination: beyond the Haken-Kelso-Bunz model. *Brain and Cognition*, 48, 149-165.

Beek, P., Peper, C., & Stegeman, D. (1995). Dynamical models of movement coordination. *Human Movement Science*, 14, 573-608.

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. Artificial Intelligence, 72, 173-215.

Bell-Berti, F. & Harris, K. S. (1976). Some aspects of coarticulation, *Haskins Laboratories Status Report on Speech Research,* SR45/46, 197-204.

Bernstein, N. (1967). *The co-ordination and regulation of movements*. Pergamon Press: London.

Blevins, J. (1995). The syllable in phonological theory. In *The Handbook of Phonological Theory*, John Goldsmith, ed. Blackwell: Cambridge, Mass, 206-244.

Bolinger, D. (1965). Pitch accent and sentence rhythm, *Forms of English: Accent, morpheme, order*. Cambridge, MA: Harvard University Press.

Browman, C. & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45, 140-155.

Browman, C. & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Bulletin de la Communication Parlee*, 5, 25-34.

Browman, C. P. & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. In M. Beckman & J . Kingston (Eds.), *Papers in laboratory phonology I* : *Between the grammar and the physics of speech*, 341-376. Cambridge: Cambridge University Press.

Butcher, A. & Weiher, E. (1976). An electropalatographic investigation of coarticulation in VCV sequences. *Journal of Phonetics ,* 4: 59-74.

Byrd, D. & Saltzman, E. (1998). Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics*, 26:173-199.
Byrd, D. & Saltzman, E. (2003). The elastic phase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31: 149–180.
Carson, R., Goodman, D., Kelso, J., & Elliot, S. (1995). Phase transitions and critical fluctuations in rhythmic coordination of ipsilateral hand and foot. *Journal of Motor Behavior*, 27:33, 211-224.
Castellano, C., Fortunato, S., & Loreto, V. (2007). Statistical physics of social dynamics. arXiv:0710.3256.
Chatfield, C. (1975). *The analysis of time series*. London: Chapman and Hall.
Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, 32: 141–176.
Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, 29(2): 155–190.
Cummins, F. & Port, R (1996). Rhythmic constraints on English stress timing. In *Proceedings of the fourth International Conference on Spoken Language Processing,* (H. T. Bunell & W. Idsardi (Eds.), 2036-2039. Wilmington, Delaware: Alfred duPont Institute.
Cummins, F. & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26, 145-171.
Dauer, R. (1983). Stress timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, 51-62.
de Jong, K. (1994). The Correlation of P-center adjustments with articulatory and acoustic events. *Perception & Psychophysics*, 56:4, 447-460.
de Poel, H., Peper, C., & Beek, P. (2007). Handedness-related asymmetry in coupling strength in bimanual coordination: furthering theory and evidence. *Acta Psychology*, 124, 209-237.
Doyle, M. & Walker, R. (2001). Curved saccade trajectories: Voluntary and reflexive saccades curve away from irrelevant distractors. *Experimental Brain Research*, 139: 333–344.
Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, 89(1): 369–382.
Erlhagen, W. & Schöner, G. (2002). Dynamic Field Theory of Movement Preparation. *Psychological Review*, 109(3): 545-572.
Fink, P., Foo, P, Jirsa, V., & Kelso, J. (2000). Local and global stabilization of coordination by sensory information. *Experimental Brain Research*, 134, 9-20.
Fletcher, J. (2004) . An EMA/EPG study of vowel-to-vowel articulation across velars in Southern British English. *Clinical Linguistics & Phonetics*, 18(6): 577-592.
Fowler, C. A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica,* 38: 35-50.
Gabrielsson, A. (2003). Music performance research at the millennium. *Psychology of Music*, 31:3, 221-272.
Gallese, V. & Lakoff, G. (2005). The Brain's Concepts: The Role of the Sensory-Motor System in Conceptual Knowledge. *Cognitive Neuropsychology*, 21.
Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6: 110-125.
Gay, T. (1974). A cinefluorographic study of vowel production. *Journal of Phonetics,* 2: 255-266.
Gay, T. (1977). Articulatory movements in VCV sequences. *Journal of the Acoustical Society of America,* 62: 183-193.

Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research,* 115: 217-233.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22 (5): 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105 (2): 251-279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(1): 152-162.

Goldstein, L. & Fowler, C. (2003). Articulatory Phonology: a phonology for public language use. In Schilller, N.O. & Meyer, A.S. (Eds.), *Phonetics and Phonology in Language Comprehension and Productions*, 159-207. Mouton de Gruyter.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In Arbib, M. A. (Ed.), *Action to Language via the Mirror Neuron System*. Cambridge Univ. Press.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). *Dynamic Action Units Slip in Speech Production Errors*. Cognition, 103, 386-412.

Gordon, P.C. & Meyer, D.E. (1984). Perceptual-Motor Processing of Phonetic Features in Speech. *Journal of Experimental Psychology: Human Perception and Performance*, 10(2): 153-178.

Haken, H. (1975). Cooperative phenomena in systems far from thermal equilibrium and in nonphysical systems. *Review of Modern Physics*, 47, 67-121.

Haken, H. (1983). Synergetics, an Introduction: Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry, and Biology, 3rd ed. New York: Springer-Verlag.

Haken, H. (1993). Advanced Synergetics: Instability Hierarchies of Self-Organizing Systems and Devices. New York: Springer-Verlag.

Haken, H., Kelso, J., & Bunz, H. (1985). A theoretical model of phase transitions in human hand movement. *Biological Cybernetics*, 51, 347-356.

Haken, H., Peper, C., Beek, P., & Daffertshofer, A. (1996). A model for phase transitions inhuman hand movements during multifrequency tapping. *Physica D*, 90, 179-196.

Halle, M. & Vergnaud, J. (1978). *Metrical structures in phonology*. Unpublished manuscript, MIT.

Hayes, B. (1995). Metrical stress theory: principles and case studies. University of Chicago Press.

Hintzman, D. L. (1986). "Schema Abstraction" in a multiple-trace memory model. *Psychological Review*, 93: 411-428.

Hintzman, D. L., Block, R., & Inskeep, N. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, 12: 741-749.

Hoole, P. (1996). Issues in the acquisition, processing, reduction and parameterization of articulographic data. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation*, München, 34, 158-173

Houghton, G. & Tipper, S. (1996). Inhibitory Mechanisms of Neural and Cognitive Control: Applications to Selective Attention and Sequential Action, *Brain and Cognition*, 30: 20-43.

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope. *Perceptual Psychophysics*, 43:1, 90-93.
Huffman, M. K. (1986). Patterns of coarticulation in English, *UCLA Working Papers in Phonetics,* 63: 26-47.
Hyman, L. (2006). Word-prosodic typology. *Phonology*, 23:2, 225-258.
Jenkins, G. M. & Watts, D. G. 1968. *Spectral analysis and its applications*. San Francisco: Holden-Day.
Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (Eds.) *Talker Variability in Speech Processing.* San Diego: Academic Press. pp. 145-165.
Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics,* 34: 485-499.
Keith, W. & Rand, R. (1984). 1:1 and 2:1 phase entrainment in a system of two coupled limit cycle oscillators. *Journal of Mathematical Biology*, 20, 122-152.
Kelso, J. (1981). On the oscillatory basis of movement. *Bulletin of the Psychonomic Society*, 18, 63.
Kelso, J. (1984). Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology; Regulatory, Integrative, and Comparative*, 15, R1000-R1004.
Kelso, J. (1995). Dynamic Patterns: the Self-Organization of Brain and Behavior. MIT: Cambridge, MA.
Kelso, J., Saltzman, E., & Tuller, B. (1986). The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14, 29-59.
Kelso, J., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, 10:6, 812-832.
Kopell, N. (1988). Toward a theory of modeling central pattern generators. In Cohen, A., Rossignol, S., & Grillner, S. (Eds.), *Neural Control of Rhythmic Movement in Vertebrates*, 369-413. John Wiley & Sons: New York.
Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5 (3), 253-263.
Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition* 21: 1-33.
Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry,* 8:3, 249-336.
Liberman, M. (1975). *The Intonational System of English*, Ph.D. dissertation, MIT. Garland Publishing Co., N.Y.
Liljencrants, J. & Lindblom, B. (1972). Numerical simulations of vowel quality systems: the role of perceptual contrast. *Language ,* 48: 839-862.
Magen, H. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics,* 25: 187-205.
Manuel, S. (1990). The role of contrast in limiting vowel-to-vowel coarticulation in different languages. *Journal of the Acoustical Society of America,* 88 , 1286 – 1298
Manuel, S. (1999). Cross-language studies: relating language-particular coarticulation patterns to other language-particular facts. Ch. in Hardcastle, W. & Hewlett, N. (Eds.)

Coarticulation: Theory, Data and Techniques, 179-198. Cambridge: Cambridge University Press.

Manuel, S. Y. & Krakow, R. A. (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. *Haskins Laboratories Status Report on Speech Research,* SR77/78: 69-78.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual Centers (P-centers). *Psychological Review*, Vol. 83, No. 5, 405-408.

Nam, H. (2007). Articulatory modeling of consonant release gesture. *International Congress on Phonetic Sciences XVI*, 625-628.

Nam, H., & Saltzman, E. (2003). A competitive, coupled oscillator model of syllable structure. In *15th International Congress of Phonetic Sciences*, 3, 2253-2256.

O'Dell, M. & Nieminen, T. (1999). Coupled Oscillator Model of Speech Rhythm. In *Proceedings of the XIV International Congress of Phonetic Sciences,* San Francisco, 2, 1075-1078.

O'Dell, M. (1995). *Intrinsic timing in a quantity language*. Licentiate Dissertation, Department of Linguistics, University of Jyväskylä, Finland.

Ohala, J. (1975). The temporal regulation of speech. In: G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and the perception of speech*. New York: Academic Press. 431 - 453.

Ohala, J. J. (1993). The phonetics of sound change. Ch. In Jones, C. (ed.) *Historical Linguistics: Problems and Perspectives*. London: Longman.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America,* 39: 151-168.

Palmer, C. (1997). Music performance. *Annual Review of Psychology*, 48, 115-38.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19 (22): 309-328.

Parush, A., Ostry, D. J. & Munhall, K. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustical Society of America,* 74: 1115-1123.

Patel, A., Löfqvist, A., & Naito, W. (1999). The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem. *Proceedings of the 14th International Congress of Phonetic Sciences*, 1, 405-408.

Peper, C. & Beek, P. (1998). Distinguishing between effects of frequency and amplitude on interlimb coupling in tapping a 2:3 polyrhythm. *Experimental Brain Research*, 118, 78-92.

Peper, C., Boer, B., de Poel, H., & Beek, P. (2008). Interlimb coupling strength scales with movement amplitude. *Neuroscience letters*, 437, 10-14.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.) *Frequency effects and the emergence of lexical structure*. John Benjamins, Amsterdam. 137-157.

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology VII*, Mouton de Gruyter, Berlin. 101-139.

Pike, K. L. (1945). *The intonation of American English*, Ann Arbor, MI: University of Michigan Press.

Pikovsky, A., Rosenblum, M., & Kurths, J. (2001). *Synchronization: A universal concept in nonlinear sciences.* Cambridge Press: Cambridge.

Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication,* 45 (1), 89-95.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics*, 17, 175–192.

Port, R. (1986). Translating linguistic symbols into time. *Research into Phonetics and Computational Linguistics, Report No. 5*, Department of Linguistics and Computational Linguistics, Indiana University, Bloomington, Indiana.

Port, R. (2003). Meter and Speech. *Journal of Phonetics*, 31, 599-611.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora-timing in Japanese. *Journal of the Acoustical Society of America*, 81 (5), 1574±1585.

Port, R., Cummins, F., & Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. *Technical Report No. 105*, *Indiana University Cognitive Science Program*, Bloomington, Indiana.

Purcell, D. & Munhall, K. (2006). Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*, 119(4): 2288-2297.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition,* 73: 265-292.

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *Journal of the Acoustical Society of America*, 76(6): 1624-1635.

Recasens, D., Pallares, M.D., & Fontdevila, J. (1997). A model of lingual coarticulation based on articulatory constraints. *Journal of the Acoustical Society of America*, 102(1): 544-561.

Repp, B. (1980). A range-frequency effect on perception of silence in speech. *Haskins Laboratories Status Report on Speech Research*, SR-61, 151-165.

Rizzolatti, G. & Arbib, M. (1998). Language within our Grasp. *Trends in Neuroscience,* 21(5).

Rizzolatti, G. (1983). Mechanisms of selective attention in mammals. In: Ewert, J.P., Capranica, R. R., & Ingle, D.J. (Eds.), *Advances in vertebrate neuroethology*, 261-297. London: Plenum Press.

Roach, P. (1982). On the distinction between `stress-timed' and `syllable-timed' languages. In D. Crystal, *Linguistic controversies*, London: Edward Arnold.

Saltzman, E. & Byrd, D. (1999). Dynamical simulations of a phase window model of relative timing. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey, (Eds.), *Proceedings of the XIVth International Congress of Phonetic Sciences*, 2275-2278. New York: American Institute of Physics.

Saltzman, E. & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, 19, 499-526.

Saltzman, E. & Kelso, J. (1983). Skilled actions: a task dynamic approach. *Haskins Laboratories Status Report on Speech Research*, SR-76, 3-50.

Saltzman, E. & Kelso, J. (1987). Skilled actions: a task-dynamic approach. *Psychological Review*, 94, 84-106.

Saltzman, E. & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, 333-382.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. *Experimental Brain Research Series*, 15: 129-144. Springer-Verlag: Berlin.

Schmidt, R., Carello, C., & Turvey, M. (1990). Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology, Human Perception and Performance*, 16:2, 227-47.
Schöner, G. Haken, H., & Kelso, J. (1986). A stochastic theory of phase transitions in human hand movement. *Biological Cybernetics*, 53:44, 247-257.
Scott, S. (1993). P-centers in Speech: An Acoustic Analysis. *Unpublished doctoral dissertation*, University College London.
Scott, S. (1998). The point of P-centres. *Psychological Research,* 61, 4-11.
Sheliga, B.M., Riggio, L., & Rizzolatti, G. (1994). Orienting of attention and eye movements. *Experimental Brain Research*, 98: 507-522.
Srivastava, M. S. (2002). *Methods of Multivariate Statistics*, New York: Wiley.
Sternad, D., Turvey, M., & Saltzman, E. (1999). Dynamics of 1:2 coordination: sources of symmetry breaking. *Journal of Motor Behavior*, 31:3, 224-235.
Tilsen, S. & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*, 124:2, EL34-39.
Tilsen, S. (2008). Relations between speech rhythm and segmental deletion. *Paper presented at the 44th annual meeting of the Chicago Linguistic Society*.
Tipper, S., Howard, L. & Houghton, G. (2000). Behavioral consequences of selection from neural population codes. In Monsell, S. & Driver, J. (Eds.), *Control of Cognitive Processes*, 225-245. MIT Press.
Treffner, P. & Peter, M. (2002). Intentional and attentional dynamics of speech-hand coordination. *Human Movement Science*, 21, 641-697.
Treffner, P. & Turvey, M. (1995). Handedness and the asymmetric dynamics of bimanual rhythmic coordination. *Journal of Experimental Psychology, Human Perception and Performance*, 21, 318-333.
Van der Stigchel, S. & Theeuwes, J. (2005). The influence of attending to multiple locations on eye movements. *Vision Research.*
Van der Stigchel, S., Meeter, M., & Theeuwes, J. (2006). Eye movement trajectories and what they tell us. *Neuroscience Biobehavioral Review*.
Van Lieshout, P. (2004). Dynamical systems theory and its application in speech. In B. Maassen, R. Kent, H. Peters, P. van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech*, 51–82. Oxford: Oxford University Press.
Vorberg, D. & Wing, A. (1996). Modeling variability and dependence in timing. Ch. 4 in *Handbook of Perception and Action, Volume 2: Motor Skills*, Heuer, H. & Keele, S. (eds.), Harcourt Brace & Company.
Welsh, T. & Elliot, D. (2005). The effects of response priming on the planning and execution of goal-directed movements in the presence of a distracting stimulus. *Acta Psychologica,* 119: 123-142.
Westbury, J., Lindstrom, M., & McClean, M. (2002). Tongues and lips without jaws: a comparison of methods for decoupling speech movements. *Journal of Speech, Language, and Hearing Research*, 45:4, 651-662.
Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, 18: 3-35.
Winfree, Arthur T. (1980). *The Geometry of Biological Time*. Springer-Verlag: New York.
Wing, A. & Kristofferson, A. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, 14:1, 5-12.

Yaegor-Dror, M. & Kemp, W. (1992). Lexical classes in Montreal French. *Language and Speech* 35: 251-293.

Yaegor-Dror, M. (1996). Phonetic evidence for the evolution of lexical classes: the case of a Montreal French vowel shift. In G. Guy, C. Feagin, J. Baugh, and D. Schiffrin (Eds.) *Towards a Social Science of Language*, 263-287. Philadelphia: Benjamins.