7

The role of vocal tract gestural action units in understanding the evolution of phonology

Louis Goldstein, Dani Byrd, and Elliot Saltzman

7.1 Introduction: duality of patterning

Language can be viewed as a structuring of cognitive units that can be transmitted among individuals for the purpose of communicating information. Cognitive units stand in specific and systematic relationships with one another, and linguists are interested in the characterization of these units and the nature of these relationships. Both can be examined at various levels of granularity. It has long been observed that languages exhibit distinct patterning of units in syntax and in phonology. This distinction, a universal characteristic of language, is termed *duality of patterning* (Hockett, 1960). Syntax refers to the structuring of words in sequence via hierarchical organization, where words are meaningful units belonging to an infinitely expandable set. But words *also* are composed of structured cognitive units. Phonology structures a small, closed set of recombinable, non-meaningful units that compose words (or signs, in the case of signed languages). It is precisely the use of a set of non-meaningful arbitrary discrete units that allows word creation to be productive.[1]

In this chapter we outline a proposal that views the evolution of syntax and of phonology as arising from different sources and ultimately converging in a symbiotic relationship. Duality of patterning forms the intellectual basis for this proposal. Grasp and other manual gestures in early hominids are, as Arbib (Chapter 1, this volume) notes, well suited to provide a link from the iconic to the symbolic. Critically, the iconic aspects of manual gestures lend them a *meaningful* aspect that is critical to evolution of a system of symbolic units. However, we will argue that, given duality of patterning, *phonological* evolution crucially requires the emergence of effectively *non-meaningful* combinatorial units. We suggest that vocal tract action gestures are well suited to play a direct role in phonological evolution because, as argued by Studdert-Kennedy (2002a), they are

[1] Phonological units, in addition to being discrete and recombinable, must yield sufficient sensory perceptibility and distinctiveness to be useful for communication.

Action to Language via the Mirror Neuron System, ed. Michael A. Arbib. Published by Cambridge University Press.

inherently non-iconic[2] and non-meaningful yet particulate (discrete and combinable). The vocal organs form the innate basis for this particulate nature (Studdert-Kennedy, 1998; Studdert-Kennedy and Goldstein, 2003). In our proposal the lack of iconicity, rather than being a weakness of vocal gestures for language evolution (cf. Arbib, 2005; Chapter 1, this volume), is *advantageous* specifically for phonological evolution in that little or no semantic content would have been needed to be "bleached out" of these mouth actions to allow them to serve as the necessarily non-meaningful phonological units – they are ideally suited to phonological function.

A reconsideration of spoken language evolution that attends to syntactic and phonological evolution as potentially distinct may cast new light on the issues. In one proposed evolutionary scenario (Arbib, Chapter 1, this volume), the original functional retrieval of symbolic meaning from manual action required, at one point, a way of distinguishing similar limb, hand, or face movements. The combination of intrinsically distinct (but meaningless) vocal actions with members of a set of iconic manual gestures could have provided the necessary means of disambiguating otherwise quite similar and continuously variable manual protowords (see also Corballis (2003) and Studdert-Kennedy and Lane (1980)). Ultimately, vocal gestures became the primary method of distinguishing words, and manual gestures were reduced in importance to the role we see them play in contemporary spoken communication (McNeill, 1992). Thus, the syntactic organization of meaningful actions and the phonological organization of particulate non-meaningful actions can be speculated to have evolved symbiotically.[3]

The existence of duality of patterning as a requisite characteristic of human languages indicates that both components of the evolution of language – the syntactic and the phonological – are intrinsic to the structure and function of language. Significantly, however, syntax and phonology may have originally evolved through different pathways. This would be consistent with the fact that when their fundamental organizing principles are described as abstract formal systems, there is little overlap between the two.[4]

7.2 Phonology and language evolution

A hallmark property of human language, signed or spoken, is that a limited set of meaningless discrete units can recombine to form the large number of configurations that are the possible word forms of a language. How did such a system evolve? This hallmark property requires particulate units (Studdert-Kennedy, 1998; see Abler, 1989) that can

[2] MacNeilage and Davis (2005) argue that there are non-arbitrary mappings in some cases between vocal gestures and meaning, such as the use of nasalization for words meaning "mother." However, their proposed account for the non-arbitrariness is contextual not iconic per se.

[3] Clearly, a radical change over evolutionary time to spoken language as the primary communication mode would have required enormous restructuring in the neural system(s) involved to move from a dependence on the manual–optic–visual chain to the vocal–acoustic–auditory chain of communication. However, we are suggesting that the systems evolved in parallel complementary ways, and the existence of signed languages with syntactic structures comparable to spoken language indicates that a manual–optic–visual chain remains accessible for communication, given a suitably structured linguistic system, namely one having duality of patterning.

[4] Except where they directly interact with one another, in the prosodic form of utterances.

function as atoms or integral building blocks; further, it requires a "glue" that can bond these atomic units together into larger combinatorial structures or molecules. From an evolutionary perspective, we wish to consider the possibility that the basis for these (units and glue) was already largely present in the vocal tract and its control before the evolution of phonology, thus explaining how such a system might have readily taken on (or over) the job of distinguishing words from one another.

Traditional phonological theories have analyzed word forms in terms of segments (or phonemes) and symbolic bi- or univalent features differentiating these segments in linguistically relevant dimensions. Both segments and features are assumed to be abstract cognitive units, the complete set of which is defined by universal properties of phonological contrast and with specific segments/features defining lexical contrast within a particular phonological system. However, speech scientists have long noticed an apparent mismatch between the proposed traditional sequence of concatenated discrete units and the physical, observable characteristics of speech, which lacks overt boundaries (e.g., silences) between segments, syllables, words, and often phrases (e.g., Harris, 1951; Hockett, 1955; Liberman *et al*., 1959).

In response to this mismatch, the framework of articulatory phonology (Browman and Goldstein, 1992, 1995; Byrd, 1996a; Byrd and Saltzman, 2003) has been forwarded as an account of how spoken language is structured. This approach is termed *articulatory phonology* because it pursues the view that the units of contrast, i.e., the phonological units, are isomorphic with the units of language production, i.e., the phonetic units. This framework views the act of speaking as decomposable into atomic units of vocal tract action – which have been termed in the field of phonetics for the last two decades *articulatory gestures*. (It's important to note here that in presenting this framework, we will be using the term articulatory *gesture* to refer to goal-directed vocal tract actions, specifically the formation of auditorily important constrictions in the vocal tract; we will not generally be using the term to refer to anything manual, unlike many of the other papers in this volume.) Articulatory gestures are actions of distinct vocal *organs*, such as the lips, tongue tip, tongue dorsum, velum, glottis (Fig. 7.1). Articulatory gestures are simultaneously units of action and units of information (contrast and encoding). This approach, like some other current approaches to phonological evolution (see, e.g., MacNeilage and Davis, 2005), views phonological structure as arising from the structural and functional characteristics and constraints of body action in the environment.

Under this account, plausible ingredients for the particulate units and bonding ("glue") necessary as precursors to the evolution of phonology are hypothesized to be, respectively, *actions* of vocal tract organs existing outside the linguistic system (plausibly actions of oral dexterity such as involved in sucking and swallowing and/or actions producing affectual vocalizations) and the *dynamic* character of their interactions as they ultimately cohere into structured units.

As action units, gestures can be described further in a twofold manner. The first is in terms of the articulatory motions themselves. More specifically, a gesture can be defined as an equivalence class of goal-directed movements by a set of articulators in the vocal

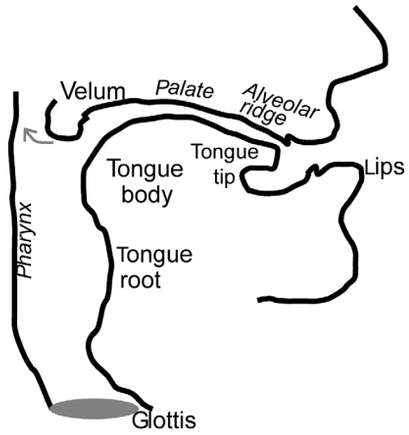

Figure 7.1 Constricting organs of the vocal tract (tongue tip, tongue body, tongue root, lips, velum, and glottis), and some of the potential places of constriction for these organs (palate, alveolar ridge, and pharynx).

tract (e.g., Saltzman and Munhall, 1989). For example, the bilabial gestures for /p/, /b/, /m/ are produced by a family of functionally equivalent movement patterns of the upper lip, lower lip, and jaw that are actively controlled to attain the speech-relevant goal of closing the lips. Here the upper lip, lower lip, and jaw comprise the *lips* organ system, or effector system, and the gap or aperture between the lips comprises the controlled variable, of the organ/effector system. The second manner in which gestures serve as action units is that they embody a particular type of dynamical system, a point-attractor system that acts similarly to a damped mass-spring system, that creates and releases constrictions of the end-effectors that are being controlled. Point attractors have properties useful for characterizing articulatory gestures. For example, regardless of their initial position or unexpected perturbations, articulatory gestures can reach their target (i.e., equilibrium position) successfully. When activated, a given gesture's point-attractor dynamics creates a pattern of articulatory motion whose details reflect the ongoing context, yet whose overall effect is to attain the constriction goal in a flexible and adaptable way.

Point-attractor models have been used to characterize many skilled human movements. Reviews of the use of point-attractor models for skilled movement can be found in Shadmehr (1995), Mussa-Ivaldi (1995), and Flash and Sejnowski (2001). In other words, we view the control of these units of action in speech to be no different from that involved in controlling skilled movements generally, e.g., reaching, grasping, kicking, pointing, etc. Further, just as a given pattern of articulatory gestures can be modulated by expressive or idiosyncratic influences, the pointing movements of an orchestral conductor can also be so modulated.

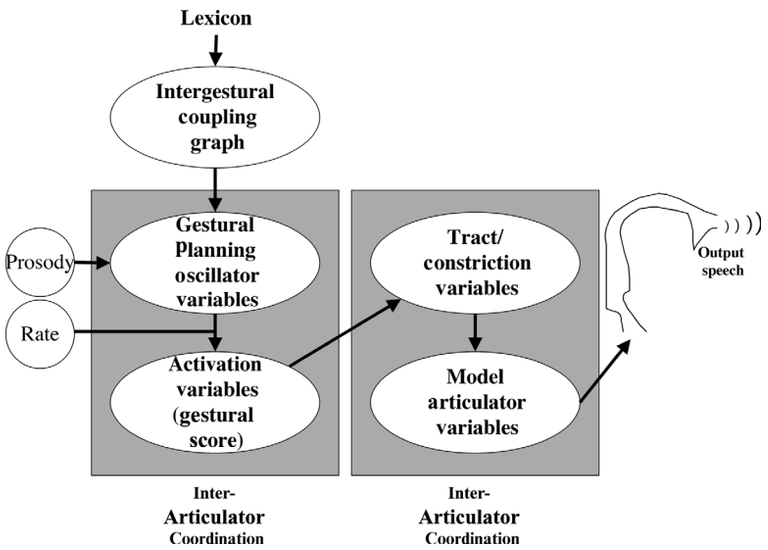

Figure 7.2 The organization of the task-dynamic model of speech production (Saltzman and Munhall, 1989; Browman and Goldstein, 1992; Nam and Saltzman, 2003).

7.2.1 Overview of a gestural, task-dynamic model of speech production

The organization of the gesture-based task-dynamic model of speech production that we have developed is shown in Fig. 7.2 (Saltzman, 1986, 1995; Saltzman and Kelso, 1987; Saltzman and Munhall, 1989; Browman and Goldstein, 1992; Nam and Saltzman, 2003).

The spatiotemporal patterns of articulatory motion emerge as behaviors implicit in a dynamical system with two functionally distinct but interacting levels (associated with the corresponding models shown as boxes in Fig. 7.2). The *interarticulator* coordination level is defined according to both *model articulator* (e.g., lips and jaw) variables and goal space or *tract-variables* (which are constriction based) (e.g., lip aperture (LA) and lip protrusion (LP)). The *intergestural* level is defined according to a set of *planning oscillator* variables and *activation* variables. The activation trajectories shaped by the intergestural level define a *gestural score* (an example of which is shown in Fig. 7.3a for the word "bad") that provides driving input to the interarticulator level.

The gestural score represents an utterance as a set of invariant gestures in the form of context-independent sets of dynamical parameters (e.g., target, stiffness, and damping coefficients) that characterize a gesture's point-attractor dynamics and are associated with corresponding subsets of model articulator, tract-variable, and activation variables. Each activation variable reflects the strength with which the associated gesture (e.g., lip closure) "attempts" to shape vocal tract movements at any given point in time. The tract-variable and model articulator variables associated with each gesture specify, respectively, the particular vocal-tract constriction (e.g., lips) and articulatory synergy

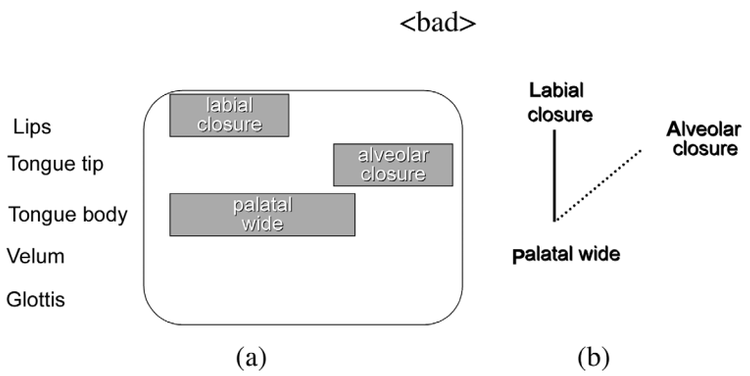

Figure 7.3 (a) A gestural score for the word "bad" showing the activation intervals for the three gestures composing the word and driving input to the interarticulator level. (b) The coupling graph for the word "bad" in which the lines indicate coupling relationships between pairs of gestures. Solid lines represent in-phase coupling, dashed lines represent anti-phase coupling.

(e.g., upper lip, lower lip, and jaw) whose behaviors are affected directly by the associated gesture's activation. The interarticulator level accounts for the observed spatiotemporal coordination among articulators in the currently active gesture set as a function of their dynamical parameter specifications. While each gesture is modeled with invariant point-attractor dynamics, the concurrent activation of multiple gestures will result in correspondingly context-dependent patterns of articulator motion. Thus, invariance in phonological specification lies not at the level of articulatory movements but in the speech tasks that those movements serve.

The intergestural level can be thought of as implementing a dynamics of *planning* – it determines the patterns of relative timing among the activation waves of gestures participating in an utterance as well as the shapes and durations of the individual gesture activation waves. Each gesture's activation wave acts to insert the gesture's parameter set into the interarticulator dynamical system defined by the set of tract-variable and model articulator coordinates (see Saltzman and Munhall (1989) for further details).

In the current model,[5] intergestural timing is determined by the *planning oscillators* associated with the set of gestures in a given utterance. The oscillators for the utterance are coupled in a pairwise, bidirectional manner specified in a *coupling graph* that is part of the lexical specification of a word. (There are also prosodic gestures (Byrd and Saltzman, 2003) that are not part of the lexical specification, but these will not be discussed further here.) For example, a coupling graph for the word "bad" is shown in Fig. 7.3b, where the

[5] In the original version of the model (e.g., Browman and Goldstein, 1990), the activation variables in gestural scores were determined by a set of rules that specified the relative phasing of the gestures and calculated activation trajectories based on those phases and the time constants associated with the individual gestures. The gestural score then unidirectionally drove articulatory motion at the interarticulator level. Thus, intergestural timing was not part of the dynamical system per se, and such a model was not capable of exhibiting dynamical coherence, such as can be seen, for example, in the temporal adjustment to external perturbation (Saltzman *et al*., 2000).

lines indicate coupling relationships between pairs of gestures (to be discussed further in Section 7.2.4). A set of equations of motion for the coupled oscillator system is implemented using the task-dynamical coupled oscillator model of Saltzman and Byrd (2000), as extended by Nam and Saltzman (2003). The steady-state output of the coupled oscillator model is a set of limit-cycle oscillations with stabilized relative phases. From this output, the activation trajectories of the gestural score are derived as a function of the steady-state pattern of interoscillator phasings and a speech rate parameter. The settling of the coupled oscillator system from initial relative phases to the final steady-state phases (over several cycles) can be conceived as a real-time planning process. Nam (in press) has found that the model's settling time in fact correlates well with speakers' reaction time to begin to produce an utterance across variations in phonological structure.[6]

This method of controlling intergestural timing can be related to a class of generic recurrent connectionist network architectures (e.g., Jordan, 1986, 1990, 1992; see also Lathroum, 1989; Bailly *et al*., 1991). As argued in Saltzman *et al*. (in press), an advantage of an architecture in which each gesture is associated with its own limit cycle oscillator (a "clock") and in which gestures are coordinated in time by coupling their clocks is that such networks will exhibit hallmark behaviors of coupled non-linear oscillators – entrainment, multiple stable modes, and phase transitions, all of which appear relevant to speech timing. As we attempt to show later, it is also possible to draw a connection between such behavioral phenomena and the qualitative properties of syllable structure patterning in languages. In addition, we will show that an explicit model based on these principles can also reproduce some subtle quantitative observations about intergestural timing and its variability as a function of syllable structure.

7.2.2 The basis of particulation in speech

Next, we must consider what makes articulatory gestures discrete (particulate) and what causes them to cohere structurally. Articulatory gestures control independent constricting devices or organs, such as the lips, tongue tip, tongue dorsum, tongue root, velum, and

[6] It is reasonable to ask why we would propose a model of speech production as apparently complex as this, with planning oscillators associated with each point-attractor vocal tract constriction gesture. Other models of speech production, at approximately this level, do not include such components. For example, in Guenther's (1995) model of speech production, timing is controlled by specifying speech as a chain-like *sequence* of phonemic targets in which articulatory movement onsets are triggered whenever an associated preceding movement either achieves near-zero velocity as it attains its target or passes through other kinematically defined critical points in its trajectory such as peak tangential velocity (e.g., Bullock *et al*., 1993; Guenther, 1994, 1995). Several kinds of observations made over the last 15 years appear to present problems for simpler models of this kind when they attempt to account for the temporal structure of speech – regularities in relative timing between units, stochastic variability in that timing, and systematic variability in timing due to rate, speaking style, and prosodic context. For example, one observation that poses a challenge to such a serial model is the possibility of temporal sliding of some (but not all) production units with respect to one another (Suprenant and Goldstein, 1998). Another challenging observation is the existence of systematic differences in segment-internal gestural timing as a function of syllable position (Krakow, 1993) or as a function of prosodic boundaries (Sproat and Fujimura, 1993; Byrd and Saltzman, 1998). While it is, of course, possible that the simple serial model could be supplemented in some way to produce these phenomena, they emerge naturally in a model in which gesture-sized units are directly timed to one another in a pairwise fashion (Browman and Goldstein, 2000; Byrd and Saltzman, 2003; Nam and Saltzman, 2003).

glottis. Articulatory gestures of distinct organs have the capacity to function as discretely different. Even neonates (with presumably no phonological system) show sensitivity to the partitioning of the orofacial system into distinct organs (Meltzoff and Moore, 1997). Young infants will protrude their tongues or lips selectively in response to seeing an experimenter move those organs, and will, when initiating such an imitation, cease to move organs that are not participating in the imitative movement (Meltzoff and Moore, 1997). Imitation of the organ's action is not always accurate and the infant will show improvement across attempts, but the organ choice is correct. While there is yet little direct evidence that this kind of somatotopic organization of the orofacial system is deployed in speech behavior, the basis for it is strongly supported by the mimicry data.

Particulate units are discrete and can therefore be combined without loss of their distinguishing characteristics (Abler, 1989). Studdert-Kennedy (1998) hypothesizes that the distinct organs of the vocal system are the basis for particulation in language. Indeed, it can be shown that contrasts between articulatory gestures of distinct organs are the primary contrasts used by the phonologies of human languages to differentiate word forms (Goldstein and Fowler, 2003), and that the ability to perceive as distinct the actions of distinct vocal organs may not decline in adulthood (Best and McRoberts, 2003), in contrast to the often noted loss of ability to discriminate contrasts not present in one's native language. Also, there is a recent report (Polka *et al*., 2001) of a consonant contrast that is not discriminated in an adult-like fashion at birth (unlike almost all others that have been tested) and that requires ambient language experience to develop full discriminability, and this is a within-organ contrast ("d" vs. "th").

Within one of these organs, articulatory gestures can be differentiated by the degree and location of a vocal tract constriction goal. For example, *tick*, *sick*, and *thick* all begin with a constriction made by the tongue tip organ but their constrictions can be differentiated in terms of constriction degree (*tick* versus *sick*) or in terms of constriction location (*sick* versus *thick*). Example gestural scores, schematically capturing the action representation of these words, are shown in Fig. 7.4. Each rectangle represents a gestural activation interval, which denotes the time interval during which a given gesture actively shapes movements of its designated organ/effector system (shown at the left of each row). Within the activation interval for each gesture is indicated the location within the vocal tract at which the organ (shown at left) is required to create a constriction if there is more than a single possibility (e.g., at the alveolar ridge or at the teeth) and the narrowness of that constriction (closure, narrow, or wide).

While these constriction parameters are, in principle, numerical continua, it is possible to model the emergence of discrete regions of these continua through self-organizing systems of agents that attune to one another (e.g., Browman and Goldstein, 2000; de Boer, 2000a, 2000b; Goldstein and Fowler, 2003; Oudeyer, 2003, 2005). These models divide a continuous constriction dimension into a large number of units that are evenly distributed across the continuum at the start of the simulation. The location of the units shifts during the simulation as a function of an agent's experience of its own productions and those of

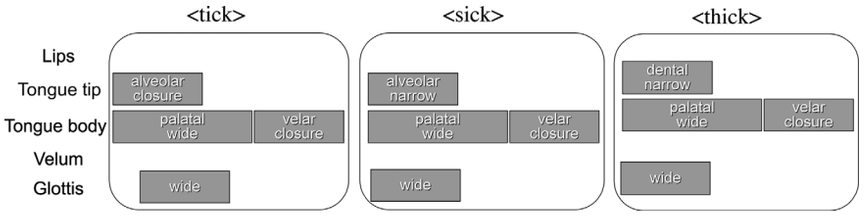

Figure 7.4 Gestural scores for "tick," "sick," and "thick."

its partner(s) via positive feedback – experience of a particular continuum value makes the agent more likely to produce that value again. Over time, the values of the units clump into a small number of modes. Critical to the clumping is a kind of categorical perception – an exact match is not required to inform the agent that some particular value has occurred in its experience. If exact matches are required (a behavior not found in real language perception), no clumping occurs.

When the agents interact through a constriction-acoustics map that harbors non-linearities (like those uncovered by Stevens, 1989), the nature of the clumping will be constrained by those non-linearities, and repeated simulations will produce modes in similar locations (Goldstein, 2003; Oudeyer, 2003). Languages would be expected to divide continua in these cases into similar regions, and this seems to be the case (e.g., constriction degree is associated with such non-linearities and is divided into comparable stop–fricative–glide regions in most languages). Other cases do not involve such non-linearities, and there may be different modal structures in different languages. For example, Hindi has a bimodal distribution of tongue tip constriction location values (dental vs. retroflex), where English has a single mode. It will be difficult, therefore, for English speakers to perceive the contrasting Hindi values as they are part of a single mode in English – they will be referred to the same mode. This can be the basis for Kuhl's *perceptual magnet effect* (Kuhl *et al*., 1992) and the *single category* case of Best's (1995) *perceptual assimilation model* (which attempts to account for which non-native contrasts will be difficult to discriminate and which easy). It may also be difficult to acquire enough experience to bifurcate a well-learned mode into two. While speakers may learn to produce the contrasting second-language values by splitting a native language mode on the basis of explicit instruction and orosensory feedback, speakers' perceptions (even of their own productions) may still be influenced by their native language's modal structure.

Evolutionarily, this view predicts that systematic differentiation of an organ's constriction goals evolved *later* than systematic use of distinct organs. Distinct organs and the sounds produced when they move during vocal expiration existed independently of and in advance of any phonology (though neural machinery for linking action and perception of distinct organs may have evolved at some point), while the process of differentiation appears to require some interaction among a set of individuals already engaged in coupling their vocal actions to one another (and therefore possibly already using a

primitive phonology). While this hypothesis cannot be tested directly, a parallel hypothesis at the level of ontogeny can be tested[7] – that children should acquire between-organ contrasts earlier than within-organ contrasts because organ differentiation requires that the infant must attune to her language environment. Studdert-Kennedy (2002b) and Goldstein (2003) find support for this hypothesis. Goldstein (2003) employed recordings of six children in an English language environment, ages $10:1 - 1:9$[8]) (Bernstein-Ratner, 1984) from the CHILDES database (MacWhinney, 2000). Word forms with known adult targets were played to judges who classified initial consonants as English consonants. The results indicate that for all six children, the oral constricting organ in the child's production (lips, tongue tip, tongue body) matched the adult target with greater than chance frequency, even when the segment as a whole was not perceived by judges as correct. That is, the errors shared the correct organ with the adult form, and differed in some other properties, usually in constriction degree or location. Some children also showed significant matching of glottis and velum organs with adult targets. However, no child showed matching of within-organ differentiation of constriction degree (i.e., stop, fricative, glide) with greater than chance frequency. These results support the organ hypothesis. The child is matching her own organs to those perceived and, in doing so, is using organs to differentiate lexical items.

In sum, organs provide a basis for discreteness or particularity of articulatory gestures in space. Within-organ goals can particulate through self-organization in a population of agents, though children join a community in which the "choices" have already been made.

7.2.3 The coherence of gestures

Next, we must turn to a consideration of how articulatory gestures cohere into words. Word forms are organized "molecules" composed of multiple articulatory gestures (the "atomic" units). Gestural molecules are systematically patterned and harbor substructures that are cohesive (i.e., resistant to perturbation) (see, e.g., Saltzman *et al*., 1995, 1998, 2000) and recurrent (appear repeatedly within a language's lexicon). The gestures composing a word can be organized in configurations that can differ somewhat from language to language (but as we argue below, these patterns are guided by a basic set of universal principles). Contrasting lexical items can be formed by replacing one atom with a different one, for example "bad" and "dad" in Fig. 7.5.

Note that the discreteness properties of gestures imply that these two words are phonologically and articulatorily equivalent except within the window of time in which

[7] See MacNeilage and Davis (2005) for an extended discussion of the validity of a recapitulationist position: "In both ontogeny and phylogeny, sound patterns are characterized by . . . [a stage in which they] are highly subject to basic constraints of biomechanical inertia and . . . [a stage of] partially overcoming these constraints in the course of developing lexical openness." They take the view that an advantage of vocal-origins theories of phylogeny is that they "can begin with the known outcome of language evolution – *current speech* – and use its *linguistic structure*, as seen in the course of ontogeny and in the structure of current languages, as a basis for inferring its phylogeny."

[8] $x:$ y means x years and y months of age.

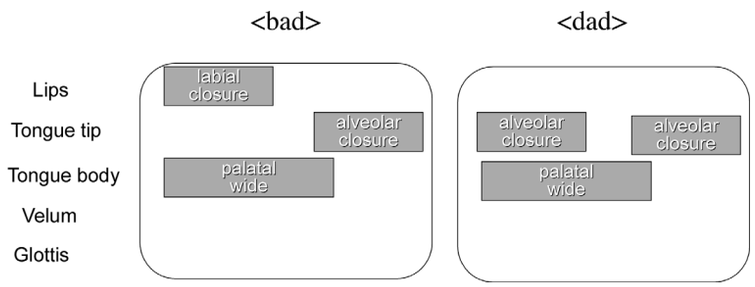

Figure 7.5 Gestural scores for "bad" and "dad," illustrating the informational significance of gestural selection.

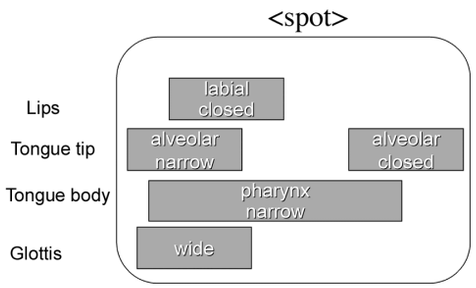

Figure 7.6 A gestural score for "spot" indicating the structure of the word initial consonants – two oral gestures and a single laryngeal abduction gesture.

the lips are being controlled in one and the tongue tip in the other. In this way, lexical differentiation is assured by gestural selection.

The organization of gestures into molecules appears to be an independent layer of structure that speakers of a language learn – it cannot be reduced to a concatenation of traditional segmental units. While traditional segments can be viewed as sets of gestures, it is not the case that the gestural molecule for a given word corresponds to the concatenation of the gesture sets for a sequence of segments. For example, in Fig. 7.6, we see the gestural score for "spot" in which a single gesture for voicelessness (a wide glottis) is coordinated with the oral gestures required to produce the two initial consonants – they do not each have a separate glottal gesture (as they do at the beginning of words like "saw" and "pa") and thus cannot be understood as the simple concatenation of an /s/ segment and a /p/ segment. Rather, the [sp] is an organized collection of two oral gestures and one glottal gesture centered between them. (This organization, incidentally, is the source of the lack of aspiration of voiceless stops seen in [sp], [st], and [sk] clusters, where segmental concatenation would otherwise lead one to expect aspiration.)

The pattern of intergestural relative timing for a given molecule is informationally significant and is represented by the pattern of activation intervals in a gestural score. For

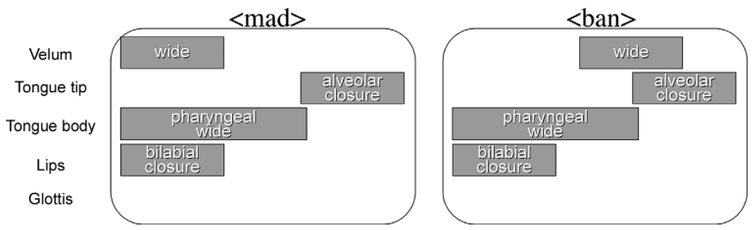

Figure 7.7 Gestural scores for the words "mad" and "ban," illustrating the informational significance of intergestural organization.

example, Fig. 7.7 shows that "mad" and "ban" are composed of the same gestures, but in a different organization.

What is the "glue" that allows articulatory gestures to be coordinated appropriately within a word form and that permits the establishment of lexically distinct coordination patterns across word forms? One possibility would be to hypothesize that gestures are organized into hierarchical segment and syllable structures that could serve as the scaffolding that holds the gestures in place through time. However, these relatively complex linguistic structures could only exist as part of an already developed phonology and could not be available pre-phonologically as part of an account of how articulatory gestures begin to be combined into larger structures.

Is there an alternative hypothesis in which the glue could exist in advance of a phonology? One such proposal is the syllable "frame" hypothesis, offered by MacNeilage and Davis (e.g., MacNeilage and Davis, 1990; MacNeilage, 1998) in which proto-syllables are defined by jaw oscillation alone. However, as will be discussed below, important aspects of syllable-internal structure are not addressed by such a frame-based account. In contrast to hierarchical structure or jaw oscillation hypotheses, our proposal is that relatively complex molecular organizations in adult phonologies are controlled dynamically via coupling relations among individual gestures (Saltzman and Munhall, 1989). This is shown schematically in the coupling graph for "bad" in Fig. 7.3b, where the lines indicate coupling relationships between pairs of gestures (solid and dashed are different modes of coupling as discussed below, in-phase and anti-phase, respectively).

Thus, learning to pronounce a particular language includes not only learning the atomic gestural components (the relevant tract variables/organs and the gestures' dynamic parameter specifications) of the language but also tuning the overall dynamical system to coordinate these atomic units into larger molecular structures. Coupled, non-linear (limit-cycle) oscillators are used to control (or plan) the timing of gestural activations within the molecules. The timing of the activation variables for a molecule's component gestures (i.e., the molecule's gestural score) is specified or planned according to the steady-state pattern of relative timing among the planning oscillators associated with the individual gestures. Gestural molecules are represented using coupling graphs in which nodes represent gestures and internode links represent the intergestural coupling

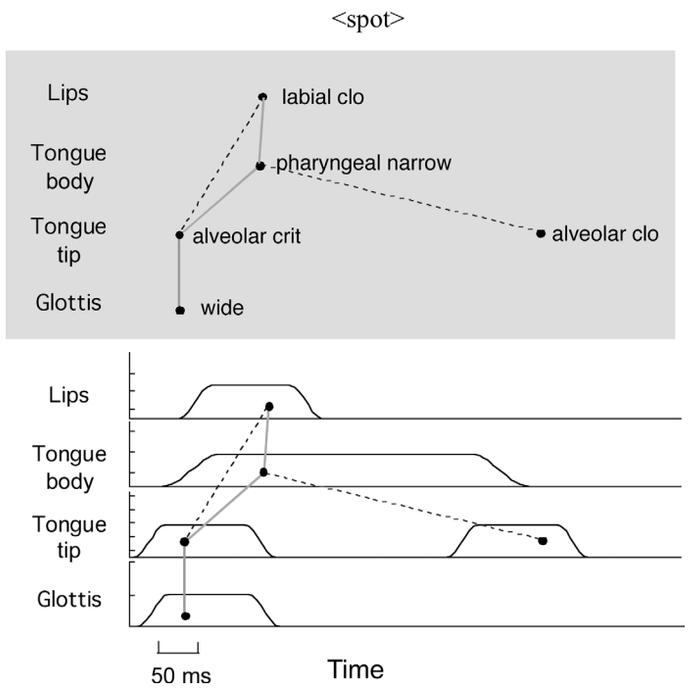

Figure 7.8 The coupling graph for "spot" (top) in which the tongue tip (fricative) gesture and the lip closure gesture are coupled (in-phase) to the tongue body (vowel) gesture, while they are also coupled to one another in the anti-phase mode. The pattern of gestural activations that results from the planning model is also shown (bottom). Lines indicate coupling relationships between pairs of gestures – solid and dashed are different in-phase and anti-phase coupling modes, respectively.

functions. In turn, these coupling graphs are used to parameterize a planning network for controlling gestural activations.

In the planning network, we adopt coupling functions originally developed by Saltzman and Byrd (2000) to control the relative phasings between pairs of gestures, and have generalized their model to molecular ensembles composed of multiple gestural oscillators (Nam and Saltzman, 2003), which can, in principle, exhibit competing phase specifications. For example, in the coupling graph for "spot" in Fig. 7.8 (top), both the tongue tip (fricative) gesture and the lip closure gesture are coupled (in-phase) to the tongue body (vowel) gesture, while they are also coupled to one another in the anti-phase mode. The basis for these couplings and evidence for them are discussed in the following sections. In Fig. 7.8 (bottom) the pattern of gestural activations that results from the planning model is added.

This generalized, competitive model has provided a promising account of intergestural phasing patterns within and between syllables, capturing both the mean relative phase values and the variability of these phase values observed in actual speech data (e.g., Byrd,

1996b). The emergence of cohesive intergestural relative phasing patterns in our model is a consequence of the glue-like properties of entrainment (frequency and phase locking) and multiple stable modes of coordination that characterize non-linear ensembles of coupled oscillators. We hypothesize that these properties exist pre-(or extra-)phonologically and could serve as the dynamical basis for combining articulatory gestures and holding gestural combinations together.[9]

7.2.4 Coupling and syllable structure

A fundamental property of phonological structure in human languages is the hierarchical organization of speech units into syllable structures. Internally, a syllable can be analyzed as composed of an onset (any consonants in a syllable that precede the vowel) and a rime (the vowel plus any consonants that follow it). The rime is then composed of a nucleus (usually simply the vowel) and a coda (any consonants following the nucleus). So, for example, in the word *sprats*, *spr* is the onset; *ats* is the rime; *a* is the vocalic nucleus, and *ts* is the coda. This organization has been shown to be relevant to many phonological processes and to guide the development of the children's phonological awareness of syllable constituency.

Coupling modes

We argue that this internal structure of the syllable can be modeled by a coupled dynamical system in which action units (gestures) are coordinated into larger (syllable-sized) molecules. In fact, we will show how syllable structure is implicit in coupling graphs such as the one shown in Fig. 7.8. The key idea is that there are intrinsically stable ways to coordinate, or phase, multiple actions in time, and in a model in which timing is controlled by coupling oscillators corresponding to the individual action units, the stable coordination possibilities can be related to stable modes of the coupled oscillators.

Ensembles of coupled non-linear oscillators are known to harbor multiple stable modes (Pikovsky *et al*., 2003). Christiaan Huygens, a seventeenth-century Dutch physicist, noticed that pendulum clocks on a common wall tended to synchronize with each other. They come to exhibit the same frequency (1 : 1 frequency-locking) and a constant relative

[9] As stated earlier, gestural molecules corresponding to words can be characterized as being composed of organized gestural substructures (e.g., generally corresponding to segments and syllables) that are both recurrent in a language's lexicon and internally cohesive. We hypothesize that the internal cohesion of these substructures may be attributed to their underlying dynamics – the dynamics of (sub)systems of coupled nonlinear oscillators. We further hypothesize that their recurrence in a lexicon can be attributed to their functional utility. When viewed from this perspective, these substructures – more accurately, their corresponding coupling graphs – appear to play the role of network motifs (e.g., Milo *et al*., 2002, 2004). Network motifs have been defined as recurring subpatterns of "interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks" (Milo *et al*., 2002, p.824), that perform specific functions in the networks and that have been identified in networks of systems as seemingly diverse as ecological food webs, biological neural networks, genetic transcription networks, the World Wide Web, and written texts' word-adjacency networks. Understanding the combinatorial properties of these molecules then becomes the challenging problem of understanding the dynamical properties of their corresponding, underlying subgraphs, i.e., of understanding their graph-dynamics (see Farmer (1990) and Saltzman and Munhall (1992) for a discussion of graph-dynamics and its relation to the dynamics of a system's state-variables and parameters).

phase (phase-locking). In human bimanual coordination, limbs that start out oscillating at slightly different frequencies will similarly entrain in frequency and phase (e.g., Turvey, 1990). Certain modes are spontaneously available for phase-locking in interlimb coordination; these are 0° (in-phase) and 180° (anti-phase). (These two modes can be found in many types of simple systems.) Further, as frequency increases, abrupt transitions are observed from the less stable of these two spontaneous modes – 180° – to the more stable, in-phase mode – 0° (Haken *et al*., 1985). Other phase-locks can be learned only with difficulty. In principle, arbitrary coupling relations can be learned, but we hypothesize that phonological systems make use of intrinsically stable modes where possible, and that the early evolution of phonological systems took advantage of these modes to begin to coordinate multiple speech actions.

Central to understanding syllable structure using a coupling model is the hypothesis that there are two basic types of gesture (in terms of their intrinsic properties, as discussed in the next section) – consonant and vowel – and that the internal structure of syllables results from different ways of coordinating gestures of these basic types. Consonant and vowel gestures can be hypothesized to be coordinated in either of the intrinsically stable modes: in-phase (the most stable) and anti-phase. We hypothesize that syllable-initial consonants and their following vowels are coordinated in phase with one another, and we can call this the *onset relation*.

Since the planning oscillators in our model determine the relative timing of the onsets of two coordinated gestures, in-phase coupling implies that the onsets of the two gestures should be synchronous. In the case of syllables that begin with a single consonant, we can see direct evidence for this in the speech kinematics. Figure 7.9 shows time functions of vocal tract variables, as measured using X-ray microbeam data, for the phrase "pea pots." Boxes delimit the times of presumed active control for the oral constriction gestures for /p/, /a/, and /t/, which are determined algorithmically from the velocities of the observed tract variables. As the figure shows, the onset of the lip gesture is approximately synchronous with the onset of the vowel gesture (within 25 ms).[10]

We hypothesize that this synchronous coordination emerges spontaneously in development because it is the most stable mode. Support for this is the early preference for consonant–vowel (CV) syllables (e.g., Stoel-Gammon, 1985). In contrast, we hypothesize that the coordination of a vowel with its following consonants – the coda relation – is an anti-phase coordination. As shown in Fig. 7.8, the onset of the /t/ in "pot" occurs at a point late in the control for the vowel gestures.

When *multiple* consonants occur in an onset, such as in the consonant cluster at the beginning of the word "spot," we assume that *each* of the consonants is coupled in-phase with the vowel (the syllable nucleus) – this is what makes them part of the onset. However, the consonant gestures must be at least partially sequential in order for the resulting form to be perceptually recoverable. Therefore, we have hypothesized

[10] In data from another subject with a receiver further back on the tongue, closer synchrony is observed. However, for this subject we do not have the parallel "spot" utterances for comparison below.

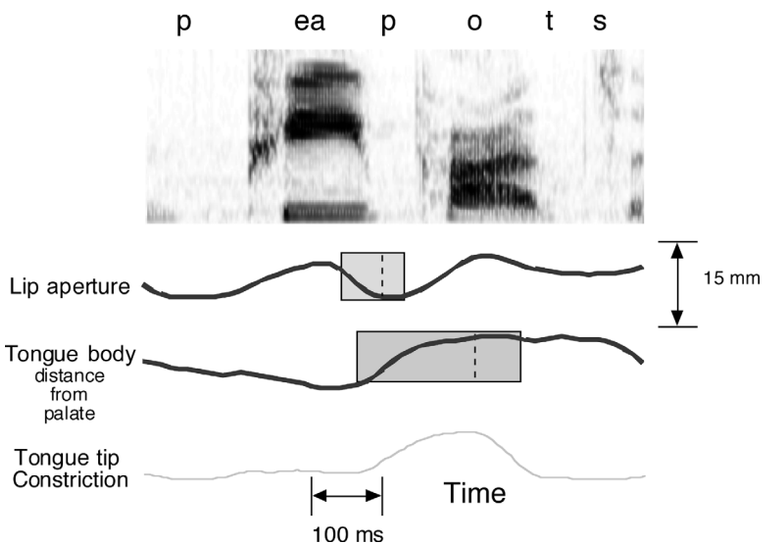

Figure 7.9 Time functions of vocal tract variables, as measured using X-ray microbeam data, for the phrase "pea pots" showing the in-phase (synchronous within 25 ms) coordination of the lip gesture for the /p/ in "pots" and the /a/ gesture for the vowel in "pots." Tract variables shown are *lip aperture* (distance between upper and lower lips), which is controlled for lip closure gestures (/p/ in this example) and *tongue tip constriction degree* (distance of the tongue tip from the palate), which is controlled in tongue tip gestures (/t/ and /s/ in this example). Also shown is the time function for the distance of the tongue body from the palate, which is small for /i/ and large for the vowel /a/, when the tongue is lowered and back into the pharynx. (The actual controlled tract variable for the vowel /a/ is the degree of constriction of the tongue root in pharynx, which cannot be directly measured using a technique that employs transducers on the front of the tongue only. So distance of the tongue body from the palate is used here as a rough index of tongue root constriction degree.) Boxes delimit the times of presumed active control for the oral constriction gestures for /p/ and /a/. These are determined algorithmically from the velocities of the observed tract variables. The left edge of the box represents gesture *onset*, the point in time at which the tract variable velocity towards constriction exceeds some threshold value. The right edge of the box represents the gesture *release*, the point in time at which velocity away from the constricted position exceeds some threshold. The line within the box represents the time at which the constriction *target* is effectively achieved, defined as the point in time at which the velocity towards constriction drops below threshold.

(Browman and Goldstein, 2000), that *multiple, competing* coupling relations can be specified in the network of oscillators in the coupling graph. For example, in the case of "spot," the oral constriction gestures of /s/ and /p/ are coupled in-phase to the vowel gesture and simultaneously anti-phase to one another, as shown in the coupling graph in Fig. 7.8. The coupled oscillator planning model (Nam and Saltzman, 2003) predicts that the onset of the vowel gesture should occur midway between the onset of the tongue tip gesture for /s/ and the lip gesture for /p/. As shown in Fig. 7.10 for the phrase "pea spots," kinematic data (for the same speaker as in Fig. 7.9) supports this prediction. Nam and

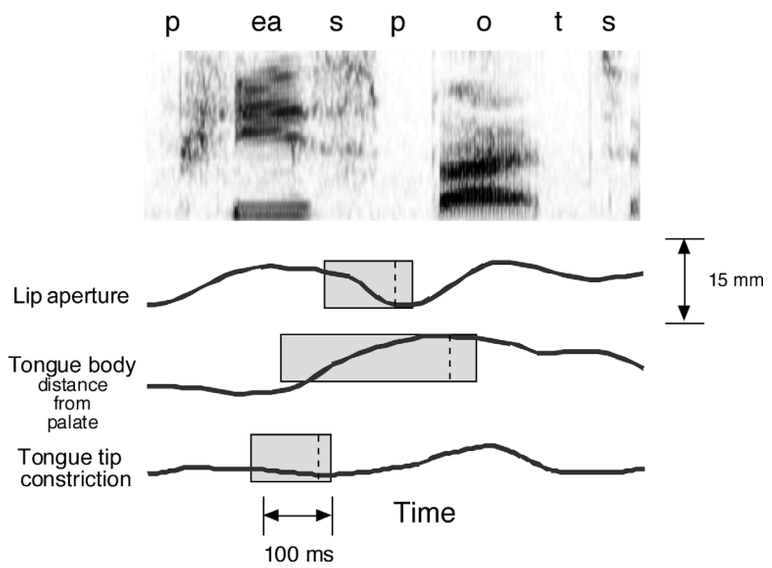

Figure 7.10 Kinematic data (for the same speaker as in Fig. 7.9) for the phrase "pea spots" showing that the onset of the vowel gesture occurs midway between the onset of the tongue tip gesture for /s/ and of the lip gesture for /p/. Boxes delimit the times of presumed active control for the oral constriction gestures for /s/, /p/, and /a/. (See Fig 7.9 caption for further details.)

Saltzman (2003) show that not only can the relative timing of onsets and codas be predicted by coupling networks with this kind of topology, but also their differential timing variability (Byrd, 1996b).

Coupling, combinatoriality, and valence

Traditional hierarchical syllable structure captures (but does not explain) systematic differences in combinatorial freedom of different syllable constituents. Syllable onsets and rimes combine relatively freely in most languages, as is seen in the English examples: *sight, blight, light, right, . . .; sip, blip, lip, rip*. Combinations of vowel and coda consonants are however somewhat more restricted than onset plus vowel combinations: for example, English allows a wider array of consonant sequences after a lax vowel: *ram, rap, ramp*, than after a tense vowel: *tame, tape, *taimp* (*indicates a sequence that is not permitted in the language). And finally, combinations of syllable consonants within syllable onsets (or within syllable codas) are also more restricted (and disallowed in many languages) than onset-vowel combinations: *sl sn pl pr* are allowed in English onsets, but **pn *sr * tl* are not.

We hypothesize that coupling 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

Table 7.1 *Characteristics of coordination and combination with respect to syllable positioning*

Initial C, V gestures (e.g. [*CV*])	Final V, C and within onsets and codas (e.g., [*VC*], [*CC*V*CC*])
In-phase coordination	Anti-phase (or other) coordination
Emerges spontaneously	Learning may be required
Free combination cross-linguistically	Restricted combination cross-linguistically

(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. The patterns are summarized in Table 7.1.

What is it about consonant and vowel gestures that allows them to be initiated synchronously (in-phase) while multiple consonant gestures (for example, the /s/ and /p/ in the "spots") are produced sequentially (anti-phase)? A minimal requirement for a gestural molecule to emerge as a stable, shared structure in a speech community is that the gestures produced are generally recoverable by other members of the community. If the tongue tip and lip gestures for /s/ and /p/ were to be produced synchronously, then the /s/ would be largely "hidden" by the /p/ – no fricative turbulence would be produced. The resulting structure would sound much the same as a lip gesture alone and almost identical to a synchronously produced /t/ and /p/ gesture. Similar considerations hold for other pairs of consonant gestures with constriction degrees narrow enough to produce closures or fricative noise. It is hard to imagine how structures with these properties could survive as part of the shared activity of community members – what would guarantee that everyone is producing the same set of gestures, rather than one of the several others that produce nearly identical sounds (and even nearly identical facial movements). However, vowel and consonant gestures can be produced in-phase and still both be recovered because of two key differences between them. (1) Vowel gestures are less constricted than (stop or fricative) consonant gestures, so the vowel can be produced during consonant production without interfering with the acoustic properties of the consonant that signal its presence (because the more narrow a constriction, the more it dominates the source and resonance properties of an acoustic tube). (2) Vowel gestures are formed more slowly and are active longer than consonant gestures, so they dominate the acoustics of the tube during a time with no overlapping or only weakly competing consonant gestures. Mattingly (1981) makes a very similar point, arguing that CV structures are efficient in that they allow parallel transmission of information, while remaining recoverable.

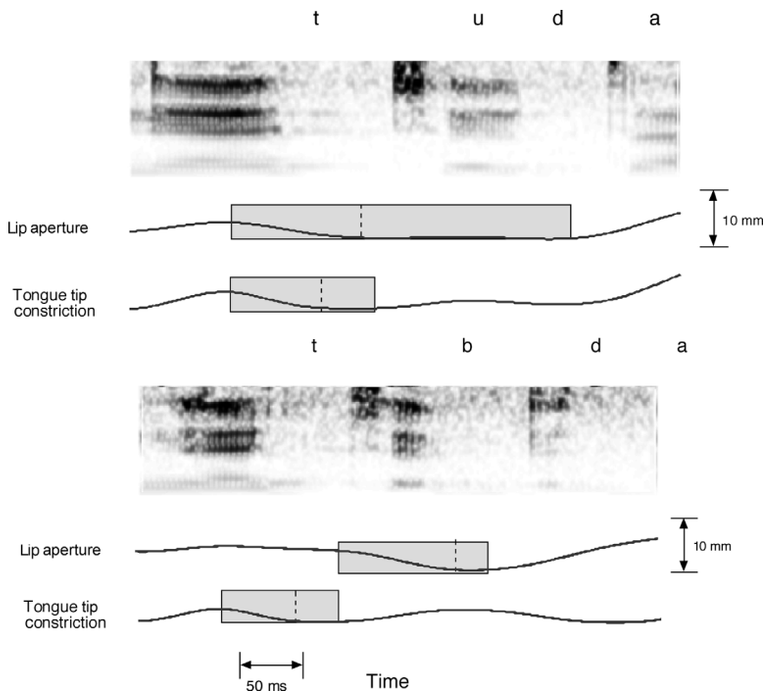

Figure 7.11 Example words /tuda/ "suffice" (top) and /tbda/ "begin" (bottom) from Tashlhiyt Berber, spoken in Morocco. The tongue tip and lip gestures are produced in-phase (synchronous at onset) in /tuda/, but not in /tbda/, where the production of the lip closure is initiated at the release of the [t]'s tongue tip gesture–anti-phase coordination. (See Fig. 7.9 caption for further details.)

To illustrate this point, consider an example from Tashlhiyt Berber (spoken in Morocco), a language that allows long strings of consonant gestures; words can even consist entirely of consonants. The Berber words /tuda/ "suffice" and /tbda/ "begin" are quite similar, differing primarily in the constriction degree of the lip gesture that follows the initial tongue tip closure and its duration – the lip gesture for /u/ is much less narrow than that for /b/ and is longer. As the kinematic data in Fig. 7.11 show,[11] the speaker produces the tongue tip and lip gestures in-phase (synchronous at onset) in /tuda/ (top part of figure), but not in /tbda/ (bottom part), where the production of the lip closure is initiated at the release of the [t]'s tongue tip gesture – anti-phase coordination.

Vowel gestures and consonant gestures are typically produced in-phase, as in this example and as in Fig. 7.8, but for multiple consonant gestures, this kind of coordination does not readily occur. Thus, the distinguishing properties of vowel and consonant gestures, together with the stability of in-phase coupling gives rise to their *valence* – they combine freely with each other in onset relations. Other reasons also support the

[11] These data were collected in collaboration with Catherine Browman, Lisa Selkirk, and Abdelkrim Jabbour.

development of such organizations: information rate (*parallel transmission*), as argued by Mattingly (1981) and biomechanical efficiency (MacNeilage and Davis, 1990, 1993).

While the grammars of most languages allow free combination of onsets and rimes (combining any onset with any rime will produce a grammatically acceptable possible word), statistical analysis of (adult, as well as child) lexicons reveals quantitative biases in favor of certain CV combinations (MacNeilage and Davis, 2000). For example, coronal (tongue tip) consonants combine more readily with front (palatal) vowels, velar consonants with back (velar and uvular) vowels, and labial consonants with central vowels. These biases can be understood, we suggest, as resulting from the interaction of the desired synchronous coupling of CV structures with the articulatory constraints of the particular pair of gestures. Because of the demands on shared articulators and anatomical limits on those articulators, not all CV pairs can be equally effectively produced in-phase. For example, producing a tongue tip gesture for /d/ while producing a back vowel may be difficult to achieve – the tongue body needs to be fronted to produce the tongue tip constriction (Stevens, 1999). So while control regimes for such a syllable might launch the production of /d/ and back vowel gestures synchronously, the articulator motion for the vowel in this case could be delayed. Structures in which the movements themselves do occur synchronously could emerge more easily. If so, this would account for the coronal–front vowel bias. The labial–central vowel bias can be conceptualized in a similar way. Raising the jaw for a labial stop while synchronously producing a front vowel could produce an unintended constriction of the front of the tongue against the palate, and synchronous production with a back vowel could produce an unintended velar constriction. Careful modulation of the amount of jaw raising in these contexts or reposturing of the tongue would be required to prevent this. Movements for labial consonants and central vowels, however, can be synchronized without these problems.

Evidence that these biases are indeed due to difficulties in synchronization can be found in the fact that the pattern of biases observed in adult languages in CV structures are absent in VC structures (MacNeilage *et al*., 2000). VC structures are hypothesized to be coupled in an anti-phase mode (above, p. 233), rather than an in-phase mode, so synchronization difficulties would not be predicted to be relevant. While these bias patterns are exhibited in VC structures in children's early words (Davis *et al*., 2002), this could be due to the overwhelming preference for reduplicative C sequences at this stage, so the CV biases are effectively transferred to the VC. This interpretation of the CV bias is compared to that of MacNeilage and Davis (below, pp. 237–239), along with a general comparison of views on the origins of the syllable and syllable structure.

Segment-level valences

While closure gestures of the lips, tongue tip, and tongue body cannot be produced synchronously with one another without compromising recoverability, there are other gestures (in addition to vowel gestures) that can successfully be produced synchronously with these constrictions. These are gestures of the velum (lowering) and glottis (abduction

or closure). The nature of velic and glottal gestures is such that their acoustic consequences are fairly independent of a concurrently produced lip, tongue tip, or tongue body constriction, so both gestures are readily recoverable even if they are synchronous at both onset and offset. And of course such combinations of gestures occur frequently among languages and are usually described as constituting single *segments*. The basis for the single-segment transcription may lie in the fact that these combinations are typically produced in-phase and are roughly equivalent in duration, so they occupy the same interval of time.

When a multi-gestural segment, such as a nasal stop (e.g., [m]), is produced in a syllable onset, all three necessary gestures can be (and are) produced in-phase – the oral closure, the velum lowering, and the vowel gesture – and this is clearly the most stable structure for that combination of gestures. Direct articulatory evidence for synchronous production of these events in onsets has been presented by Krakow (1993). If the nasal consonant is in coda, however, the oral constriction gesture for the consonant must be produced in an anti-phase relation to the vowel, as argued above, so any of the remaining possibilities of coordinating the three gestures is less stable than the onset pattern, and it is not clear which of these suboptimal patterns would be the most stable. If the velum lowering gesture is synchronized with the oral closure, then it necessarily is anti-phase with the vowel. Or if it is synchronized with the vowel, then it is anti-phase with respect to the consonantal oral gesture. This leads to the predictions that (a) the segment-internal coordination of gestures may be different in coda than in onset, and that (b) more than one stable pattern may be evidenced in coda cross-linguistically.

Evidence is available in support of both predictions. With respect to (a), a number of recent articulatory studies have demonstrated systematically different segment-internal gestural timing relations depending on whether a segment appears in the onset or coda of the syllable (see Krakow (1999) for a review). For example, velum and oral constriction gestures for nasals in English are coordinated sequentially in coda, rather than the synchronous coordination observed in onset. The multiple gestures of the English liquid consonant [l] (tongue tip raising and tongue rear backing) show a very similar difference in coordination pattern (Delattre, 1971; Sproat and Fujimura, 1993; Browman and Goldstein, 1995; Krakow, 1999) as a function of syllable position. These differences in coordination result in what have traditionally been described as examples of allophonic variation – nasalization of vowels before coda nasals and velarization of /l/ in coda. While in traditional formal analyses the allophonic variants for nasals and /l/s are unrelated to each other, they are in fact both superficial consequences of the deeper generalization that onset and coda relations involve distinct coupling modes (Browman and Goldstein, 1995).

Evidence for (b), the availability of multiple stable structures in coda, is found in the fact that languages may differ in which of the predicted stable patterns they employ. Direct articulatory evidence for anti-phase coordination between the velum and oral closure gestures has been found in English, and similar indirect evidence is available

from many languages in the nasalization of vowels in the context of a coda nasal (Schourup, 1973; Krakow, 1993). However, such nasalization of vowels in the context of coda nasals is not present in aerodynamic data from French (Cohn, 1993) and various Australian languages (Butcher, 1999). For the latter languages, the velum lowering and oral closure gestures are apparently coordinated synchronously in coda.

In the within-segment case, we can also see a correlation between combinatoriality and the stability of coordination mode, like that observed in the case of CV versus VC coordination. In onset, oral closures can combine relatively freely with glottal gestures or with velic gestures. Such combinatoriality is part of the basis for the traditional decomposition of segments into *features*. In coda, however, such combinatoriality may be lacking in some languages. Many languages, for example, allow only nasal consonants in coda, requiring that an oral closure and a velic lowering gesture occur together, or do not occur at all (for example Mandarin Chinese). Interestingly, the coda consonants in Australian languages that show relatively synchronous coordination of oral constrictions and velic lowering in coda (Butcher, 1999) are highly combinatorial. There are four to six different places of articulation (labial, dorsal, and as many as four distinct coronals), and these can all occur with or without nasalization in coda. Butcher attributes the synchronous coordination to the functional requirement for keeping the vowel in pre-nasal position completely non-nasalized, so that the formant transitions effectively distinguish the many places of articulation (Tabain *et al*., 2004). However, in a broader perspective, this observation can be seen as combinatoriality co-occurring with synchronous coordination.

Language differentiation

As is clear from several examples presented in the last two sections, patterns of coordination of speech gestures may differ from language to language. CV syllables have been argued to be the most stable coordination pattern, as the component actions are coordinated in-phase. Indeed such patterns occur in every language, and there are languages in which CV is the only syllable available (e.g., Hawai'ian). However, there are languages that also employ coda consonants (anti-phase coordination) and sequences of consonants, including languages like Georgian where sequences of three or four consonants can begin a syllable (Chitoran, 2000) and like Tashlhiyt Berber where syllables can be composed exclusively of consonants (Dell and Elmedloui, 1985, 2002). Indeed Tashlhiyt Berber can be analyzed as having no restrictions on segment sequencing (no phonotactic constraints). How do such hypothetically suboptimal structures (from the point of view of coordination stability) arise, and how are they sustained?

It is a reasonable speculation that CV syllables were the first to arise in the evolution of phonology (MacNeilage and Davis, 2005). Gestures such as lip smacks and tongue smacks, observed in non-human primates, could have spontaneously synchronized with phonated vowel-like constrictions to form early CV or CVCV words. Once a lexicon evolved, however, other forces could come into play. First, there is an inverse power law relation between word-length (measured in segments) and word frequency (Zipf, 1949).

Frequently used words can become shorter by loss of one or more of their gestures, for example, the final V in a CVCV structure. The result would be CVC structure. Loss of the first of the vowels would result in a CCV structure. Examples of vowel loss of these types do occur in historical sound change in languages (apocope and syncope, respectively); less stable coordination patterns could arise this way. While they are suboptimal with respect to stability, they satisfy a competing constraint on word form. In this way languages with different patterns of coordination and combination could arise. Second, the existence of a well-learned lexicon of molecules imparts its own stability to the shared forms. So even if a coda consonant is not the most intrinsically stable organization, repeated instances of this in the words of a language make it a sufficiently stable mode. In fact it is possible to view the occurrence of speech errors as the result of competition between the intrinsic modes and the learned patterns associated with particular lexical items.

The regularities underlying the coordination patterns (syllable structures) of a particular language form part of its phonological grammar, an account of speakers' knowledge of the possible word forms of the language and of the regular correspondences among words or morphemes in different contexts. One model of phonological grammar, *optimality theory*, calls on a set of rank-ordered constraints (Prince and Smolensky, 2004). Lower-ranked constraints may be violated to satisfy higher-ranked constraints. The hierarchy of constraint ordering can differ from language to language. Constraints include preferences for certain types of structures ("markedness" constraints) and for keeping the various phonological forms of a particular lexical item as similar as possible ("faithfulness" constraints). Cross-linguistic differences in constraint ranking result in different segmental inventories, different syllable structures, or different alternations in a word's form as a function of context. In the domain of syllables and syllabification, constraints include injunctions against coda consonants and against onset clusters. In recent work, some markedness constraints have been explicitly cast as constraints on gestural coordination (Gafos, 2002), and the preference for synchronous coordination has been argued to play a role in phonological grammar (Pouplier, 2003; Nam, in press) and could be the basis for the "Align" family of constraints that require edges of certain structures to be coincident. One phonological process that can be related to a preference for the most stable coordination pattern is *resyllabification*. A word-final consonant before a word beginning with a vowel can be resyllabified with the following vowel (so that *keep eels* is pronounced like *key peels)*, and increases in speaking rate can lead to a greater tendency for such resyllabifications (Stetson, 1951). This phenomenon can be analyzed as an abrupt transition to a more stable coordination mode (Tuller and Kelso, 1991; de Jong, 2001).

The emergence of the syllable: mandible-based frame or synchronous coupling?

Like the present view, the frame-content theory (e.g., MacNeilage, 1998; MacNeilage and Davis, 1999, 2000, 2005) attempts to explain the evolution of phonological structure and its emergence in infants on the basis of general anatomical, physiological, and

evolutionary principles, rather than on domain-specific innate endowments. In their theory, the syllable develops out of (biomechanical) mandibular oscillation, an organized behavior already present in non-human primates subserving the function of mastication. The infant's babbling and early words are hypothesized to result almost exclusively from jaw oscillation with no independent control of the tongue and lips at the timescale of consonants and vowels. This oscillation constitutes the syllable frame (one cycle of oscillation is a syllable), while control over individual consonants and vowels is the content, which is hypothesized to develop later. The theory predicts that the CV patterning in early syllables should not be random, but rather there should be systematic associations described above: coronals with front vowels, dorsals with back vowels, labials with central vowels. The basis of the prediction is that if the tongue happens to be in a relatively advanced static posture when the jaw begins to oscillate, the front part of the tongue will make contact with the palate when the jaw rises, thus producing (passively) a coronal constriction. When the jaw lowers, the fronted tongue will produce a shape appropriate for a front vowel. When there is no tongue advancement, the lips (rather than the tongue tip), will form a passive constriction, and the lowered jaw will produce a central vowel shape. A retracted tongue will produce a dorsal constriction and back vowel when the jaw oscillates. These predictions are borne out in data of CV combinations in babbling (Davis and MacNeilage, 1995) and early words (Davis *et al*., 2002); the predicted combinations consistently show ratios of observed to expected frequencies of greater than 1, while most other combinations show ratios less than 1.

As attractive and elegant as this theory is, it seems to us to suffer from some empirical problems connecting data and theory. While it is true that the preferred CV syllables *could* be produced by moving only the jaw,[12] there is no direct evidence that the infants are in fact producing them in that way. And there is some indirect evidence that they are not. First, there are many syllables that the infants produce that are not of one of the preferred types – the preferences are reliable but are relatively small in magnitude. Some independent control of tongue and/or lips is required to produce these other patterns. Second, there appears to be no developmental progression from exclusively preferred CVs early on to a more varied set later. Perhaps this is just a question of the appropriate analyses having not yet been performed. A preliminary test for age grading (comparing 6, 9, and 12 months) using the Haskins babbling database does not reveal such a developmental trend, but more data needs to be examined. Finally, MacNeilage *et al*. (2000) show that similar observed-to-expected ratios are observed in the lexicons of ten (*adult*) languages. But we know that adults do *not* produce CV syllables by moving only the jaw (even when they produce the preferred syllable types). They have independent control over vowel and consonant constrictions. While it is possible that mandibular oscillation is part of the explanation for this CV preference in adult languages (more on this below) and that the preferences are inherited from childhood, the empirical

[12] In MacNeilage and Davis (2005), the claim that certain syllables are produced exclusively by jaw movement is made in explaining why labials should be a preferred consonant – they can be produced with only jaw activity.

point here is that the existence of these CV preferences cannot constitute *evidence* for a jaw-only motor control strategy, since the preferences exist in adult languages but the jaw-only strategy does not.

Another weakness of the frame-content theory compared to the current view is that it provides no account of differences between onset and rime consonants – in their timing, their variability, and their combinatoriality. As we have argued above, these can be accounted for in a principled way in the oscillator coupling model. Relatedly, the coupling model predicts the lack of VC associations in adult languages (as discussed above), while this does not follow from the frame-content theory.

We outlined above (p. 234) an alternative account of the CV preferences in adult languages, based on the hypothesis that synchronous production of consonant and vowel constrictions can be successfully achieved more easily in the preferred CV patterns than in other combinations. This could account for the preferences in infants as well, if we assume, contra MacNeilage and Davis, that infants are synchronizing multiple actions. This account would not suffer the same problem as that of MacNeilage and Davis, in that no assumption of jaw-only control is made. Nonetheless, there is at least one potentially explanatory aspect of MacNeilage and Davis' jaw-based account that the current account lacks – an explanation for the overall duration of a syllable. In their account, syllable production rate should correspond to an oscillatory mode (natural frequency) of the jaw. Thus, it would be desirable to integrate the accounts in some way. One possibility is to hypothesize that infants are synchronously coupling constriction-directed activity of the tongue and lips along with jaw oscillation. Given the more massive jaw, its natural frequency would be expected to dominate in the coupling.

It is also interesting to consider how the jaw could play a role in determining the preferred CV patterns in this hybrid account. The preferred CV combinations not only afford in-phase productions of consonant and vowel, but they also have the characteristic that the compatibility of the vowel and consonant constrictions is such that when they are produced synchronously, jaw raising and lowering can assist in the production of both constrictions. In contrast, in a synchronous combination of a non-preferred pattern (for example coronal consonant and back vowel), competing demands on the jaw from the two synchronous constrictions would cancel each other out, and less jaw movement would be expected. In general, since the jaw is massive, system efficiency would be enhanced by being able to take advantage of jaw movement in production of consonant and vowel constrictions.

7.3 Phonology as gestural structure: implications for language evolution

The hypothesis that stored lexical forms are molecules built through coupling of dynamical vocal tract constriction gestures can provide an approach for understanding: how a combinatorial phonological system might have evolved, why phonologies tend to have the kinds of units that they have, and why phonologies tend to have the types of combinatorial structure that they have.

If we assume action recognition to be important in language evolution (and acquisition), an understanding of speech production as relying on dynamically defined and organized vocal tract actions has a direct bearing. We agree with Davis *et al*. (2002) that "mental representation cannot be fully understood without consideration of activities available to the body for building such representations . . . [including the] dynamic characteristics of the production mechanism." Phonological structures can emerge through self-organization as individuals interact in a community, though certain biological preconditions are necessary (Studdert-Kennedy and Goldstein, 2003). Emergent representations rely on a universal (i.e., shared) set of organs moving over time. This initial organ-based representational space constrains possible future phonologies. Later differentiation of this representational space allows for phonological grammar to emerge (Studdert-Kennedy and Goldstein, 2003).

A necessary precondition for the evolution of language is that there exist a compatibility and complementarity of perception and action (Turvey, 1990; Liberman, 1996). Such complementarity can be demonstrated in a much wider range of contexts than just language. Perception guides action in many obvious ways (e.g., obstacle avoidance in locomotion and in reaching), but action systems are also involved in perception (at least in humans), particularly when the perceived act is self-produced or is a type of action that could be self-produced, for example, an act produced by a conspecific. Galuntucci *et al*. (in press) have summarized evidence for such effects in a variety of domains.

7.3.1 Mirror neurons and the complementarity of perception and action

Given the functionally integrated nature of perception and action (see, e.g., Barsalou, 1999; Hommel *et al*., 2001, cited in Dale *et al*., 2003), it is not surprising to find common neural units active during both the perception and performance of the same or related acts (cf. Rizzolatti and Arbib, 1998). But there is no reason to think that complementarity is limited to those tasks for which mirror neurons have (so far) been uncovered or that the ones uncovered so far are in any way primary (see the discussion in Dale *et al*., 2003; Oztop *et al*., this volume). Also, while there has been a tendency to focus on the role of *visual* and particularly manual information as providing the sensory input to such mirror neurons, recent evidence has shown that auditory information associated with the performance of some acts may also elicit responses in mirror neurons active during the performance of those acts (Kohler *et al*., 2002; see also Romanski and Goldman-Rakic, 2002), that mouth movements and tactile stimulation of the mouth yield a response by specific inferior F5 neurons (Rizzolatti *et al*., 1981), and that there are mirror neurons relating ingestive behavioral and orofacial communicative acts (Ferrari *et al*., 2003). This is consistent with behavioral evidence of the functional equivalence in perception of the multiple lawful sensory consequences of an act (for example the "McGurk" effect in speech perception: McGurk and MacDonald, 1976; Massaro *et al*., 1996; Fowler and Dekle, 1991; Rosenblum and Saldaña, 1996).

Types of action and action recognition

Two kinds of action/action recognition scenarios might be relevant in considering the path followed in the evolution of language. First, an act *on* the environment, for example a grasp, might be executed in the presence of a perceiver. Such an act is generally visible, often limb-related, and may result in the movement of an object by an agent. The recognition of this act might be considered to proceed from the visual to the conceptual, and its semantic aspects are relatively transparently recoverable from the sensory information available to the perceiver. A different kind of action to consider might be an act *in*, rather than *on*, the environment – a body-internal act like vocalization would be an example. Such an act would likely (though not necessarily) involve body-internal, largely non-visible actions and result in a perceivable signal with at least the potential of arbitrary semantic content. Its recognition by a perceiver would, in the case of vocalization, proceed from the auditory to the conceptual, and if semantically arbitrary, would be *mediated* by a lexicon of sorts. The first type of action – action *on* the environment – may well be the evolutionary source of syntax, that is, the linguistic characterization of how meaningful words are patterned to convey information. The second type of action – action *in* the environment – is a possible evolutionary source of phonology, the linguistic characterization of how vocal tract actions are patterned to create words. It is this latter *phonological* type of action and action recognition that we focus on – in particular, how human articulatory gestural actions can encode words and the implications of these considerations for understanding the evolution of phonology. But we suggest that both these evolutionary processes (syntax and phonology) could have gone on in parallel and that vocal gestures and manual gestures could be produced concurrently. Thus, we do not view the process of language evolution as having necessarily undergone a massive shift from manual to vocal (a shift that MacNeilage and Davis (2005) also find implausible). Rather the modalities have been complementary.

Action tasks and perceptual objects

The compatibility and complementarity of perception and action suggests that it is profitable to decompose an animal's behavior into functionally defined *tasks* that can be given a formal description integrating both their motor and multi-modal sensory consequences. For example, in the case of speech, we hypothesize that gestural task-space, i.e., the space of articulatory constriction variables, is the domain of confluence for perception-oriented and action-oriented information during the perception and production of speech. Such a task-space would provide an informational medium in which gestures become the objects shared by speech production and perception, atomic or molecular (Liberman, 1996; Goldstein and Fowler, 2003; Galantucci *et al*., in press). In this regard, speech's task-space defines a domain for coupling acting and perceiving in a manner similar to that proposed for the locomotory navigation task-space of Fajen and Warren (2003) and to the common coding principle for perception and action of

Prinz (1997). It is possible that a mirror neuron system could be the neurophysiological instantiation of this cognitive coupling, though the identification of such a specific system in humans remains an outstanding challenge.

From this point of view, evolution of a new function (such as language) can be seen as a process in which previously existing tasks (or coherent parts thereof) are recruited and combined into a new pattern of organization (see Farmer, 1990; Saltzman and Munhall, 1992; and also footnote 9). This is the familiar self-organization approach to the ontogenesis of locomotion skills (Thelen, 1989) applied at a longer timescale: "development proceeds . . . as the opportunistic marshalling of the available components that best befit the task at hand. Development is function-driven to the extent that anatomical structure and neurological mechanisms exist only as components until they are expressed in a context. Once assembled in context, behavior is, in turn, molded and modulated by its functional consequences" (Thelen, 1989, p.947). In the case of language, we offer the possibility that elements of manual tasks (e.g., from grasping) and orofacial tasks (e.g., from food ingestion or emoting) are recombined in the evolution of language. Each of the tasks provides an important component – syntactic and phonological, respectively – of evolving language use.

Grasping tasks can be employed (in the absence of an actual object actually being present) to represent actions symbolically, thus providing a basis for reference and semantics. Pantomime can convey a great deal of information prior to the development of arbitrary conventions (Arbib, 2005; Chapter 1, this volume). Such information would have been crucial in developing the use of symbols, indexicality ("mine," "yours," "theirs"), and possibly concepts of events and complex cause-and-effect. This type of action was likely necessary for the evolution of syntax, i.e., the structured patterning of words to convey meaning. But pantomimic actions formed with the hands (and attached limbs) without more sophisticated elaboration (such as that found in signed *languages*, in which the actions are elaborated in signing space and are no longer pantomimic) do not provide a ready source of discreteness to differentiate similar actions with different symbolic meaning (e.g., similar-looking objects; actions that took place in the past or might occur in the future). The fingers are, of course, intrinsically discrete units, but it is hard to use them independently while still performing a grasping-related action with the hand as a whole. Further, there are several ways in which the grasping task does not provide an ideal scaffolding, at least alone, for the evolution of language (see, e.g., Corballis, 2002, 2003): a reliance on (proto)sign is problematic when there is no direct line of sight between individuals, a situation which must be assumed to have existed for our ancestors, in the trees, in the dark, or over moderate distance. In fact, it is in just these circumstances that communicating information about potential danger or food is likely to be most vital. Falk (2004; cited in MacNeilage and Davis, 2005) suggests that mothers' need to be able to undertake parental care at a distance fostered dyadic vocal communication. Manual communication is also problematic with occupied hands, such as during foraging, grooming, tool use, or child-care – all, one would think, frequent

activities of (somewhat) smart early primates. (See, however, Emmorey, this volume, for an alternative view.)

Tasks engaging the organs of the face and mouth might have been recruited to supplement manual tasks in the evolution of spoken language. The partitioning of the face and mouth into distinct organs and the association of their actions with reliable structuring of optic and acoustic media (lip-smack vs. tongue-smack; lip protrusion vs. no lip protrusion) afford the ability to distinguish ambiguous meanings by using discrete non-iconic actions. The relative independence of the organs when producing constrictions of the vocal tube, the existence of intrinsically stable modes of interorgan action coupling, and the physics of sound generation allows articulatory gestures to combine readily into larger combinations with distinctive structuring of the acoustic and visual media. The combinatorial possibilities would presumably increase with anatomical changes in hominid vocal tract, such as we see in ontogeny, with the tongue body and tongue tip becoming independently useful in constriction formation (Vihman, 1996). Further, organ-specific mirror neurons for orofacial actions in the F5 of macaque monkeys have been discovered with selective tuning to particular orofacial organs or their combination (Ferrari *et al*., 2003); different neurons appear to be sensitive to different orofacial tasks: grasping, ingesting of food, communicative acts.

Thus, we see a possible direct source of phonological evolution, i.e., the patterning of (non-meaningful) particulate action units to form (meaningful) words, *as distinct from* the development of symbolic thinking and/or syntax. This development could be expected to enhance syntactic evolution by (a) allowing a larger and more discrete set of word forms and grammatical markers and (b) freeing the specifically syntactic and semantic processes from the parallel responsibility of generating multiple word forms. Eventually, the rich phonological differentiation of words could have made accompanying manual iconic acts redundant (see also Corballis, 2003).

7.4 Summary

Whether or not our particular speculations deserve further serious consideration, we think it is important to draw attention to the more general consideration of the potential independence of phonological evolution in theorizing as to how language evolved. Arbib (2005; Chapter 1, this volume) addresses the question of language evolution by treating it as a progression from protosign and protospeech to languages with full-blown syntax and compositional semantics, but this view says little about the phonology of protosign and critically neglects to consider the emergence of duality of patterning as a hallmark characteristic of language.

In this chapter we have proposed that the evolution of syntax and of phonology arose from different sources and ultimately converged in a symbiotic relationship. We argue that phonological evolution crucially requires the emergence of particulate combinatorial units. We suggest that articulatory *gestures* are well-suited to play a direct role in

phonological evolution because, as argued by Studdert-Kennedy (2002a), they are typically non-iconic and non-meaningful, yet discrete. The lack of iconicity of vocal gestures, rather than being a weakness for language evolution, is *advantageous specifically for phonological evolution*. The combination of syntactically meaningful actions with phonologically particulate non-meaningful actions can be speculated to have symbiotically generated the evolution of language. The existence of duality of patterning as a hallmark characteristic of human languages indicates that both components of the evolution of language – the syntactic and the phonological – are robustly present and necessary. Significantly, however, syntax and phonology may have originally evolved along different pathways.

Acknowledgments

The authors gratefully acknowledge the support of the National Institutes of Health and thank Michael Arbib, Barbara Davis, and Karen Emmorey for their helpful comments. This work was supported by NIH grants DC-03172, DC-00403, and DC-03663.

References

Abler, W. L., 1989. On the particulate principle of self-diversifying systems. *J. Soc. Biol. Struct.* **12**: 1–13.

Arbib, M. A., 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behav. Brain Sci* **28**: 105–124.

Bailly, G., Laboissière, R., and Schwartz, J. L., 1991. Formant trajectories as audible gestures: an alternative for speech synthesis. *J. Phonet.* **19**: 9–23.

Barsalou, L., 1999. Perceptual symbol systems. *Behav. Brain Sci.* **22**: 577–660.

Bernstein-Ratner, N., 1984. Phonological rule usage in mother–child speech. *J. Phonet.* **12**: 245–254.

Best, C. T., 1995. A direct realist perspective on cross-language speech perception. In W. Strange and J. J. Jenkins (eds). *Cross-Language Speech Perception*. Timonium, MD: York Press, pp. 171–204.

Best, C. T., and McRoberts, G. W., 2003. Infant perception of nonnative contrasts that adults assimilate in different ways. *Lang. Speech* **46**: 183–216.

Browman, C. P., and Goldstein, L., 1990. Tiers in articulatory phonology, with some inplications for casual speech. In J. Kingston and M. E. Beckman (eds.) *Papers in Laboratory Phonology*, vol. 1, *Between the Grammar and Physics of Speech*. Cambridge, UK: Cambridge University Press, pp. 341–376.

1992. Articulatory phonology: an overview. *Phonetica* **49**: 155–180.

1995. Dynamics and articulatory phonology. In T. van Gelder (ed.) *Mind as Motion Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press, pp. 175–194.

2000. Competing constraints on intergestural coordination and self-organization of phonological structures. *Bull. Commun. Parlée* **5**: 25–34.

Bullock, D., Grossberg, S., and Mannes, C., 1993. A neural network model for cursive script production. *Biol. Cybernet.* **70**: 15–28.

Butcher, A. R., 1999. What speakers of Australian aboriginal languages do with their velums and why: the phonetics of the nasal/oral contrast. *Proceedings 16th International Congress of Phonetic Sciences*, Berkeley, CA, pp. 479–482.

Byrd, D., 1996a. A phase window framework for articulatory timing. *Phonology* **13**: 139–169.

1996b. Influences on articulatory timing in consonant sequences. *J. Phonet.* **24**: 209–244.

Byrd, D., and Saltzman, E., 1998. Intragestural dynamics of multiple phrasal boundaries. *J. Phonet.* **26**: 173–199.

2003. The elastic phrase: dynamics of boundary-adjacent lengthening. *J. Phoneti.* **31**: 149–180.

Chitoran, I., 2000. Some evidence for feature specification contraints on Georgian consonant sequencing. In O. Fujimura, B. Joseph and B. Palek (eds.) *Proceedings of LP 98*, pp. 185–204.

Cohn, A. C., 1993. The status of nasalized continuants. In M. Huffman and R. Krakow (eds.) *Nasal, Nasalization, and the Velum*. San Diego, CA: Academic Press, pp. 329–367.

Corballis, M. C., 2002. *From Hand to Mouth: The Origins of Language*. Princeton, NJ: Princeton University Press.

2003. From mouth to hand: gesture, speech, and the evolution of handedness. *Behavi. Brain Sci.* **26**: 199–260.

Dale, R., Richardson, D. C., and Owen, M. J., 2003. Pumping for gestural origins: the well may be rather dry. *Behav. Brain Sci.* **26**: 218–219.

Davis, B. L., and MacNeilage, P. F., 1995. The articulatory basis of babbling. *J. Speech Hear. Res.* **38**: 1199–1211.

Davis, B. L. and MacNeilage, P. F., and Matyear, C. L., 2002. Acquisition of serial complexity in speech production: A comparison of phonetic and phonological approaches to first word production. *Phonetica* **59**: 75–107.

de Boer, B., 2000a. Self-organization in vowel systems. *J. Phonet.* **28**: 441–465.

2000b. Emergence of vowel systems through self-organisation. *A.I. Commun.* **13**: 27–39.

de Jong, K., 2001. Rate induced re-syllabification revisited. *Lang. Speech* **44**: 229–259.

Delattre, P., 1971. Consonant gemination in four languages: an acoustic, perceptual, and radiographic study, Part I. *Int. Rev. Appl. Linguist.* **9**: 31–52.

Dell, F., and Elmedlaoui, M., 1985. Syllabic consonants and syllabification in Imdlawn Tashlhiyt Berber. *J. Afri. Lang. Linguist.* **7**: 105–130.

2002. *Syllables in Tashlhiyt Berber and in Moroccan Arabic*. Dordrecht, Nethelands: Kluwer.

Fajen, B. R., and Warren, W. H., 2003. Behavioral dynamics of steering, obstacle avoidance, and route selection. *J. Exp. Psychol. Hum. Percep. Perform.* **29**: 343–362.

Falk, D., 2004. Prelinguistic evolution in early hominids: whence motherese. *Behav. Brain Sci.* **27**: 491–503.

Farmer, J. D., 1990. A Rosetta Stone for connectionism. *Physica D* **42**: 153–187.

Ferrari, P. F., Gallese, V., Rizzolatti, G., and Fogassi, L., 2003. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* **17**: 1703–1714.

Flash, T., and Sejnowski, T., 2001. Computational approaches to motor control. *Curr. Opin. Neurobiol.* **11**: 655–662.

Fowler, C. A., and Dekle, D. J., 1991. Listening with eye and hand: cross-modal contributions to speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **17**: 816–828.

Fowler, C. A., Galantucci, B., and Saltzman, E., 2003. Motor theories of perception. In M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, 2 edn. Cambridge, MA: MIT Press, pp. 705–707.

Gafos, A., 2002. A grammar of gestural coordination. *Nat. Lang. Linguist. Theory* **20**: 269–337.

Galuntucci, B., Fowler, C. A., and Turvey, M., in press. The motor theory of speech perception reviewed. *Psychonom. Bull. Rev.*

Goldstein, L., 2003. Emergence of discrete gestures. *Proceedings 15th International Congress of Phonetic Sciences*, pp. 85–88.

Goldstein, L., and Fowler, C. A., 2003. Articulatory phonology: a phonology for public language use. In N. Schiller and A. Meyer (eds.) *Phonetics and Phonology in Language Comprehension and Production*. Berlin: Mouton de Gruyter. pp. 159–208.

Guenther, F. H., 1994. A neural network model of speech acquisition and motor equivalent speech production. *Biol. Cybernet.* **72**: 43–53.

1995. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychol. Rev.* **102**: 594–621.

Haken, H., Kelso, J. A. S., and Bunz, H., 1985. A theoretical model of phase transitions in human hand movements. *Biol. Cybernet.* **51**: 347–356.

Harris, Z. S., 1951. *Methods in Structural Linguistics*. Chicago, IL: University of Chicago Press.

Hockett, C., 1955. *A Manual of Phonology*. Bloomington, IN: Indiana University Press.

1960, The origin of speech. *Sci. American* **203**: 88–111.

Hommel, B., Musseler, J., Aschersleben, G., and Prinz, W., 2001. The theory of event coding (TEC): a framework for perception and action planning. *Behavi. Brain Sci.* **24**: 849–937.

Jordan, M. I., 1986. *Serial Order in Behavior: A Parallel Distributed Processing Approach*, Technical Repant No. 8604. San Diego, CA: University of California, Institute for Cognitive Science.

1990. Motor learning and the degrees of freedom problem. In M. Jeannerod (ed.) *Attention and Performance*. vol. 13 Hillsdale, NJ: Lawrence Erlbaum, pp. 796–836.

1992. Constrained supervised learning. *J. Math. Psychol.* **36**: 396–425.

Kohler, E., Keysers, C., Umiltà, M. A., *et al*., 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**: 846–848.

Krakow, R. A., 1993. Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate. In M. A. Huffman and R. A. Krakow (eds.) *Nasals, Nasalization, and the Velum*. New York: Academic Press, pp. 87–116.

1999. Physiological organization of syllables: a review. *J. Phonet.* **27**: 23–54.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B., 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science* **255**: 606–608.

Lathroum, A., 1989. Feature encoding by neural nets. *Phonology* **6**: 305–316.

Liberman, A. M., 1996. *Speech: A Special Code*. Cambridge, MA: MIT Press.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P. C., and Cooper, F. S., 1959. Minimal rules for synthesizing speech. *J. Acoust. Soc. America* **31**: 1490–1499.

MacNeilage, P. F., 1998. The frame/content theory of evolution of speech production. *Behav. Brain Sci.* **21**: 499–546.

MacNeilage, P. F., and Davis, B. L., 1990. Acquisition of speech production: achievement of segmental independence; In N. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modeling*. Dordrecht, Netherlands: Kluwer, pp. 55–68.

1993. A motor learning perspective on speech and babbling. In B. de Boysson-Bardies, S. Schoen, P. Jusczyk, P. MacNeilage, and J. Morton (eds.) *Changes in Speech and Face Processing in Infancy: A Glimpse at Developmental Mechanisms of Cognition*. Dordrecht, Netherlands: Kluwer pp. 341–352.

1999. Euolution of the form of spoken words. *Evol. Commun.* **3**: 3–20.

2000. Origin of the internal structure of word forms. *Science* **288**: 527–531.

2005. The frame/content theory of evolution of speech: a comparison with a gestural origins alternative. *Interaction Studies: Interaction Stud.* **6**: 173–199.

MacNeilage, P. F., Davis, B. L., Kinney, A., and Matyear, C. L., 2000. The comparison of serial organization patterns in infants and languages. Issue, *Infant Devel.* **71**: 153–163.

MacWhinney, B., 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum.

Massaro, D. W., Cohen, M. M., and Smeele, P. M., 1996. Perception of asynchronous and conflicting visual and auditory speech. *J. Acoust. Soc. America* **100**: 1777–1786.

Mattingly, I. G., 1981. Phonetic representation and speech synthesis by rule. In. T. Myers, J. Laver and J. Anderson (eds.) *The Cognitive Representation of Speech.* Amsterdam: North Holland, pp. 415–420.

McGurk, H., and MacDonald, J., 1976. Hearing lips and seeing voices. *Nature* **264**: 746–747.

McNeill, D., 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: University of Chicago Press.

Meltzoff, M., and Moore, K., 1997. Explaining facial imitation: a theoretical model. *Early Devel. Parent.* **6**: 179–192.

Milo, R., Shen-Orr, S., Itzkovitz, S., *et al*., 2002. Network motifs: Simple building blocks of complex networks. *Science* **298**: 824–827.

Milo, R., Itzkovitz, S., Kashtan, N., *et al*., 2004. Superfamilies of evolved and designed networks. *Science*, **303**: 1538–1542.

Mussa-Ivaldi, F. A., 1995. Geometrical principles in motor control. In M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press, pp. 434–438.

Nam, H., in press. A competitive, coupled oscillator model of moraic structure: split-gesture dynamics focusing on positional asymmetry. In J. Cole and J. Hualde (eds). *Papers in Laboratory Phonology*, vol. 9.

Nam, H., and Saltzman, E., 2003. A competitive, coupled oscillator of syllable structure. *Proceedings 12th International Congress of Phonetic Sciences*, Barcelona, pp. 2253–2256.

Oudeyer, P.-Y., 2003. L'auto-organisation de la parole. Ph.D. dissertation, University of Paris VI.

2005. The self-organization of speech sounds. *J. Theoret. Biol.* **233**: 435–449.

Pikovsky, A., Rosenblum, M., and Kurths, J., 2003. *Synchronization*. Cambridge, UK: Cambridge University Press.

Polka, L., Colantonio, C., and Sundara, M., 2001. A cross-language comparison of /d/-/D/ perception: evidence for a new developmental pattern. *J. Acoust. Soc. America* **109**: 2190–2201.

Prince, A., and Smolensky, P., 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford, UK: Blackwell.

Prinz, W., 1997. Perception and action planning. *Eur. J. Cogn. Psych.* **9**: 129–154.
Pouplier, M., 2003. The dynamics of error. *Proceedings 15th International Congress of the Phonetic Sciences*, pp. 2245–2248.
Rizzolatti, G., and Arbib, M. A., 1998. Language within our grasp. *Trends Neurosci.* **21**: 188–194.
Rizzolatti, G., Scandolara, C., Gentilucci, M., and Camarda, R., 1981. Response properties and behavioral modulation of "mouth" neurons of the postarcuate cortex (area 6) in macaque monkeys. *Brain Res.* **255**: 421–424.
Romanski, L. M., and Goldman-Rakic, P. S., 2002. An auditory domain in primate prefrontal cortex. *Nature Neurosci.* **5**: 15–16.
Rosenblum, L. D., and Saldaña, H. M., 1996. An audiovisual test of kinematic primitives for visual speech perception. *J. Exp. Psychol. Hum. Percept. Perform.* **22**: 318–331.
Saltzman, E. L., 1986. Task dynamic coordination of the speech articulators: a preliminary model – Generation and modulation of action patterns. In H. Heuer and C. Fromm (eds.) *Experimental Brain Research*, New York: Springer-Verlag, pp. 129–144.
1995. Dynamics and coordinate systems in skilled sensorimotor activity. In R. Port and T. van Gelder (eds.) *Mind as Motion*. Cambridge, MA: MIT Press, pp. 150–173.
Saltzman, E., and Byrd, D., 2000. Task-dynamics of gestural timing: phase windows and multifrequency rhythms. *Hum. Mov. Sci.* **19**: 499–526.
Saltzman, E. L., and Kelso, J. A. S., 1987. Skilled actions: a task dynamic approach. *Psychol. Rev.* **94**: 84–106.
Saltzman, E. L., and Munhall, K. G., 1989. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1**: 333–382.
1992. Skill acquisition and development: the roles of state-, parameter-, and graph-dynamics. *J. Motor Behav.* **24**: 49–57.
Saltzman, E., Löfqvist, A., Kinsella-Shaw, J., Kay, B., and Rubin, P., 1995. On the dynamics of temporal patterning in speech. In F. Bell-Berti and L. Raphael (eds.) *Producing Speech: Contemporary Issues for Katherine Safford Harris*. Woodbury, NY: American Institute of Physics, pp. 469–487.
Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., and Rubin, P., 1998. Dynamics of intergestural timing: a perturbation study of lip–larynx coordination. *Exp. Brain Res.* **123**: 412–424.
Saltzman, E., Löfqvist, A., and Mitra, S., 2000. "Glue" and "clocks": intergestural cohesion and global timing. In M. Broe and J. Pierrehumbert (eds.) *Papers in Laboratory Phonology*, vol. 5 Cambridge, UK: Cambridge University Press, pp. 88–101.
Saltzman, E., Nam, H., Goldstein, L., and Byrd, D., in press. The distinctions between state, parameter and graph dynamics in sensorimotor control and coordination. In A. Feldman (ed.) *Progress in Motor Control: Motor Control and Learning over the Life Span*. New York: Springer-Velag.
Schourup, A., 1973. A cross-language study of vowel nasalization. *Ohio State Univ. Working Papers Linguist.* **15**: 190–221.
Shadmehr, R., 1995. Equilibrium point hypothesis. In M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press, pp. 370–372.
Sproat, R., and Fujimura, O., 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *J. Phonet.* **21**: 291–311.
Stetson, R. H., 1951. *Motor Phonetics*. Boston, MA: College-Hill Press.
Stevens, K. N., 1989. On the quantal nature of speech. *J. Phonet.* **17**: 3–45.

1999. *Acoustic Phonetics*. Cambridge, MA: MIT Press.

Stoel-Gammon, C., 1985. Phonetic inventories, 15–24 months: a longitndinal study. *J. Speech Hearing Res.* **18**: 505–512.

Studdert-Kennedy, M., 1998. The particulate origins of language generativity. In J. Hurford, M. Studdert-Kennedy, and C. Knight, (eds.) *Approaches to the Evolution of Language*. Cambridge, UK: Cambridge University Press, pp. 202–221.

2002a. Mirror neurons, vocal imitation, and the evolution of particulate speech. In V. Gallese and M. Stamerov (eds.) *Mirror Neurons and the Evolution of Brain and Language*. Amsterdam: Benjamins, pp. 207–227.

2002b. Evolutionary implications of the particulate principle: imitation and the dissociation of phonetic form from semantic function. In C. Knight, M. Studdert-Kennedy and J. B. Hurford (eds.) *The Evolutionary Emergence of Language: Social Function and the Origin of Linguistic Form*. Cambridge, UK: Cambridge University Press, pp. 161–176.

Studdert-Kennedy, M., and Goldstein, L., 2003. Launching language: the gestural origin of discrete infinity. In M. H. Christiansen and S. Kirby (eds.) *Language Evolution: The States of the Art*. Oxford, UK: Oxford University Press, pp. 235–254.

Studdert-Kennedy, M., and Lane, H., 1980. Clues from the difference between signed and spoken languages. In U. Bellugi and M. Studdert-Kennedy (eds.) *Biological Constraints on Linguistic Form*. Berlin: Verlag Chemie, pp. 29–40.

Suprenant, A., and Goldstein, L., 1998. The perception of speech gestures. *J. Acoust. Soc. America* **104**: 518–529.

Tabain, M., Breen, J. G., and Butcher, A. R., 2004. VC vs. CV syllables: a comparison of Aboriginal languages with English. *J. Int. Phonet. Ass.* **34**: 175–200.

Thelen, E., 1989. The (re)discovery of motor development: learning new things from an old field. *Devel. Psychol.* **25**: 946–949.

Tuller, B., and Kelso, J. A. S., 1991. The production and perception of syllable structure. *J. Speech Hear. Res.* **34**: 501–508.

Turvey, M., 1990. Coordination. *Am. Psychologist* **45**: 938–953.

Vihman, M., 1996. *Phonological Development: The Origins of Language in the Child*. Cambridge, MA: Blackwell.

Zipf, G. K., 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.