Competing constraints on intergestural coordination and self-organization of phonological structures

CATHERINE P. BROWMAN

Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA

LOUIS GOLDSTEIN

Department of Linguistics, Yale University, New Haven, CT 06520, USA
and Haskins Laboratories, 270 Crown St., New Haven, CT 06511, USA

ABSTRACT

Within the Articulatory Phonology framework, the notion of gestural structure plays a central role. The nature of such structures has been the focus of two related lines of recent research that we present here. One line involves a proposal to enrich gestural structures to include an explicit representation of the relative cohesiveness of pairs of gestures within an utterance. As it turns out, this simple addition to the model has unexpected and interesting explanatory consequences, one of which involves the phonetic and phonological properties of syllable structure. The second line involves developing a research strategy to account for the observed properties that gestural structures exhibit in languages using principles of self-organization.

1. Introduction

The articulatory phonology framework (Browman & Goldstein, 1986, 1989, 1990, 1992) hypothesizes that a description of speech in terms of articulatory gestures can provide structures that capture both physical and phonological properties. In this approach, speech is decomposed into a set of gestures that control the actions of distinct articulator sets within the vocal tract. These gestures are simultaneously *units of action* and *units of information*. They are units of action in the sense that each gesture is a dynamic neuromotor system that guides the coordination of a set of multiple articulators and muscles in the formation of a characteristic vocal tract constriction. They are units of information in the sense that individual gestures may be used to distinguish, or contrast, utterances from one another, for example, by the presence vs. absence of a particular gesture, or by differences in the location and degree of a gestural constriction.

The computational instantiation of Articulatory Phonology produces, for any arbitrary utterance in English, a *gestural structure,* which consists of a set of gestures and a specification of how they are temporally coordinated with respect to one another. In previous work (referred to above), we have shown how such gestural structures can reveal generalizations that underlie of a variety of types of phonetic and phonological alternations.

An example gestural structure for the word "pen" is shown here Figure 1a. Rows of the display correspond to the distinct articulator sets that gestures can regulate. Gestures are represented by descriptors stand for numerical values, or ranges of values, of the dynamic control parameters that specify the goals of a given constriction. For example, the Tongue Tip gesture at the end of word is specified for a value of *constriction degree* that will produce complete closure, and a value of *constriction location* at the alveolar ridge.

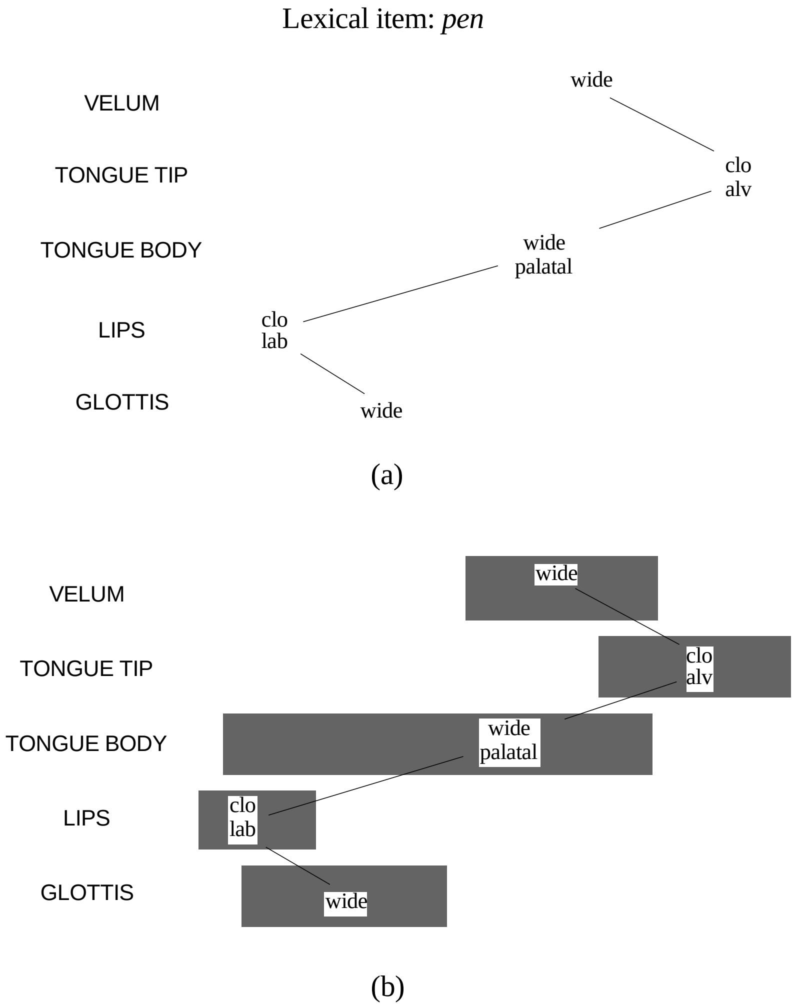


Figure 1: (a) Gestural structure for the lexical item *pen*. Lines connect gestures whose coordination is specified lexically in the computational model. (b) Gestural score. Horizontal extent of boxes represents temporal activation intervals. After Browman & Goldstein, 1995.

Lines shown in the figure connect the particular pairs of the word's gestures whose coordination is specified within the model of gestural structure (e.g., Browman & Goldstein, 1992). Coordination is accomplished by means of phasing: some phase of motion of one member of the pair is specified to occur simultaneously with some phase of the other member. For the purposes of this paper the specific hypothesis that coordination is specified in terms of phase, and the mechanics of this, are irrelevant. What is important is that coordination is specified in a local, pair-wise fashion. Given this fact, for an utterance with *n* gestures, specification of *n-1* pairs completely determines the temporal structure of an utterance. Thus, for "pen," there are five gestures, and 4 phasing specifications.

Given the n-1 phasing specifications and the intrinsic dynamic parameter specification of the individual gestures, temporal activation intervals for each gesture are calculated by the model. The result is what we have called a *gestural score* (Figure 1b). Note that gestural activation intervals are partially overlapping. While gestural **structures** are hypothesized to be fixed, lexical properties of a form, the quantitative values of the gestural parameters and the phase relations may be quantitatively scaled as a function of speaking conditions. The result of such scaling will be reflected in the gestural **score**. The gestural score is input to the task-dynamic model of Saltzman (1986), which calculates the response of a set of simulated articulators to the dynamic controls; the articulators move in a coordinated fashion to achieve the constriction goals, or tasks.

2. Competing phase relations and syllable structure

One problem with this model of gestural structure as originally proposed has been highlighted by Byrd (1996a,b). She showed that the phase relations in an utterance's gestural structure may differ from one another in the amount of constraint they appear to impose on the relative timing, or overlap, of the gestures involved. For example, she found that while oral constriction gestures for consonants in a syllable onset exhibit little variability in overlap, and thus highly constrained phasing, consonant gestures in a coda and across syllable boundaries show significantly more variability in overlap. She proposed a "phase window" model (Byrd, 1996b) to account for this, and other observations, extending Keating's (1990) work in the spatial domain to the temporal domain.

An alternative method of handling such observations that we have been pursuing is to associate every phase relation within a lexical unit with a bonding (or coupling) strength that represents the degree of cohesion of those gestures. Computationally, then, the sources of scalar variation in gestural overlap (e.g., due to speaking rate, style, prosody) would influence a given pair of gestures in inverse proportion to their bonding strengths. Thus, variation in overlap among coda consonant gestures could result, for example, from local changes in speaking rate. Such changes in rate would exert much less effect on onset consonants, due to their higher bonding strength.

The proposed bonding analysis could also account for the environments in which variation in overlap is extreme enough to be perceptible. For example, Browman & Goldstein (1990) showed that gestures could "slide" in casual speech styles so as to produce the perception of consonant deletions and assimilations. In fact, in all these cases, the gestures that slide with respect to one another are not part of the same lexical unit. If we hypothesize that post-lexical phasing between gestures at the end of one lexical unit and beginning the next has weak (or possibly non-existent) bonding strength, then the site of these assimilations and deletions is accounted for.

Adding bonding strength to the model also has an interesting formal consequence. If each phase relation is associated with a bonding strength, then it is no longer the case that we are limited to specifying n-1 phase relations in a gestural structure. Incompatible, or competing, phase relations, can be specified, and the actual temporal pattern that surfaces in the gestural score can be computed as the one that maximizes satisfaction of the competing constraints, as weighted by their bonding strengths. And, rather than being a formal curiosity, it turns out that competing phase relations can do useful work. Here we will see how it can account for some subtle, but reliable, properties of syllable structure in English (and perhaps other languages).

In our earlier work, the behavior of consonant constriction gestures in onset position has been described as a possible exception to local, gesture-to-gesture phasing specification. A number of studies (Browman & Goldstein, 1988; Honorof & Browman, 1995; Byrd, 1995) have found that gestures in the onset reliably exhibit what we have dubbed the “c-center” effect, in which the oral constriction gestures constituting the onset appear to be phased as a single unit with respect to the vowel gesture. Evidence for this can be best understood by looking at the sample data Figure 2. The graphs show articulatory and acoustic data for a speaker producing the utterances: “cuff SAYED,” “cuff PAID,” and “cuff SPAYED.” (The subjects were instructed to accent the capitalized words). The articulatory data are time functions of Tongue Tip Constriction Degree and Lip Aperture (vertical distance between upper and lower lips) estimated from X-ray microbeam data. Shaded areas represent the intervals of time during which the tongue tip or lips are at their presumed target values for the initial consonant gestures. The curves are aligned at the time of maximal tongue tip constriction for final consonant gesture in these words, which is represented by the solid black lines. We assume that the coordination of the vowel gesture with this final consonant gesture does not change as a function of initial consonant, and therefore, aligning the curves with respect to the final consonant gesture is equivalent to aligning them with respect to some point in their vowel gestures, which cannot be directly measured here.

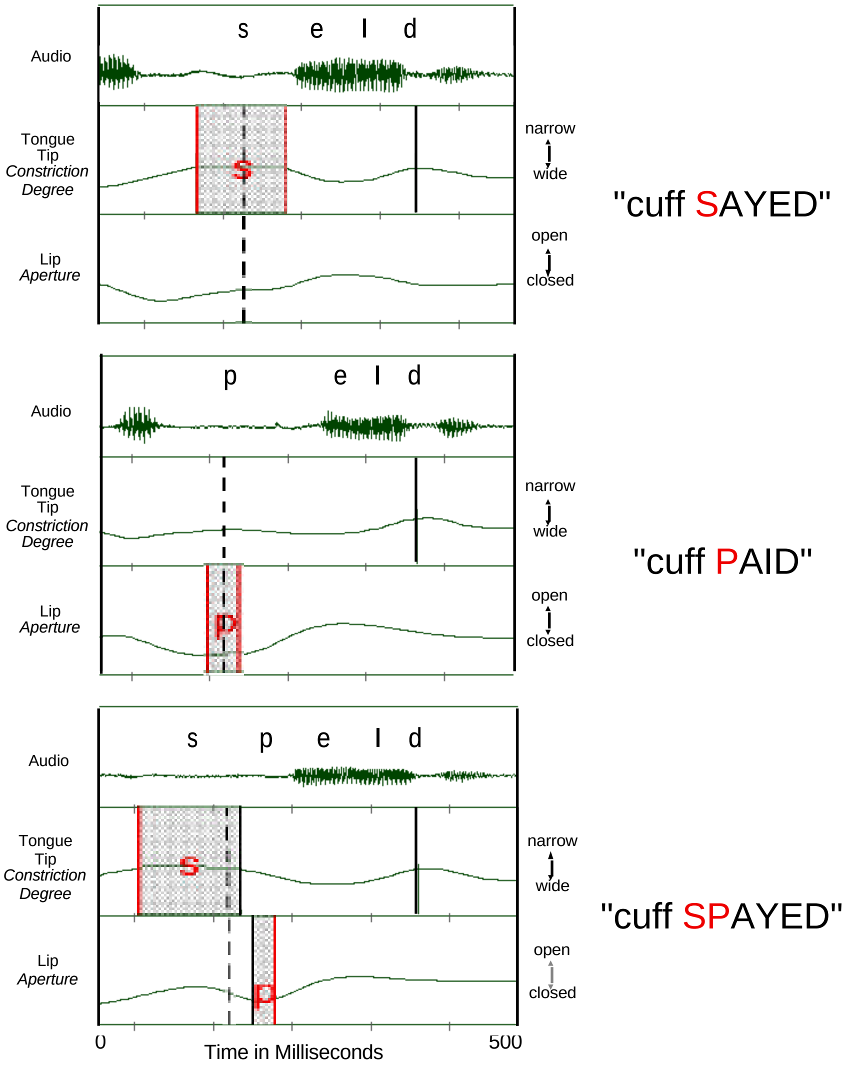


Figure 2: Audio waveforms and time functions for Tongue Tip Constriction Degree and Lip Aperuture (vertical distance between upper and lower lips) for utterances "cuff SAYED", "cuff PAID", and "cuff SPAYED." Time functions estimated from x-ray microbeam data. Shaded boxes represent intervals of time during which tongue tip and lips are at their presumed target positions. Dotted lines are "c-centers" of initial consonants (see text). Curves are aligned at solid black lines, which represent the point in time at which the tongue tip reaches its maximum degree of constriction for the final /d/,

When we look at the data for “sayed” and “paid,” we see that the centers of the target intervals for these words (indicted by the dotted vertical lines) coincide fairly well with each other, and by the above assumption, are coordinated with the same phase of the vowel gesture. But now look at “spayed.” Neither the tongue tip gesture, nor the lip gesture is aligned to the rest of the word as it is in “sayed” and “paid.” The tongue tip gesture occurs earlier than it does in “sayed,” and the lip gesture later than it does in “paid.” However, c-center of the two gestures (indicated by the dotted line) is aligned in the same way as it is in the words with a single onset consonant. The c-center is calculated as the mean of the centers of the target intervals of the individual gestures. Thus, as gestures are added to the onset, the mean of the target intervals of the entire ensemble of gestures retains a stable temporal relation to the vowel. This is the c-center effect. This generalization cannot be captured by a local gesture-to-gesture phase relation. Rather, it appears to be a global property of the oral constriction gestures that constitute the onset. Onsets composed of three consonants (e.g., “splayed”) appear to function in exactly the same way.

Using competing phase relations and bonding, however, we can give an analysis of the c-center effect that retains gesture-to-gesture phasing for onsets, and provides a more principled account. Let us hypothesize that each consonant gesture in a complex onset bears exactly the same phase relation to the vowel (let’s call it the *C-V* relation) that it bears in a syllable in which it is the lone onset consonant. If this were the only specified phase relation, consonant gestures in an onset would all be synchronous (Figure 3a). The gestural score for “spayed” would not be stable with respect to recoverability (cf. Mattingly, 1981)—listeners could not reliably hear both tongue tip and lip gestures.

Let us hypothesize an additional phasing specification (the *C-C* relation) that phases the consonant gestures to each other so as to allow them to be recoverable. This relation must have greater bonding strength than the C-V relation, in order to overcome their tendency to synchronize. But in order to minimize violation of both phase relations, the consonant gestures should be equally displaced in time from the point specified by the *C-V* relation. For example, Figure 3b shows a gestural score for “spayed” in which the lip gesture retains the C-V relation and the tongue tip gesture is advanced by the C-C relation. This gestural score shows greater (mean-squared) violation of the C-V constraint than the gestural score in Figure 3c in which the two gestures are equally displaced from the point specified by the C-V relation. The configuration is exactly what we observe in the data, which we have described as the c-center effect.

There are a number of advantages to this competing phase relations account of the c-center, compared to the analysis in which the consonant gestures in an onset are phased as a unit to the vowel gesture. First, no non-local phase specification is required and all phase relations coordinate one gesture to another gesture. Second, each of the competing phase relations can be viewed as a constraint related to an underlying principle. *C-V* phasing ensures parallel transmission of vowels and consonants (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967), and *C-C* phasing ensures recoverability (Mattingly, 1981; Silverman, 1995) . Third, this may help explain the greater stability in overlap that has observed for onset consonants than for coda consonants. There is no reliable evidence that coda consonants show the c-center effect (evidence is negative in some studies or variable in others). Thus, final consonants may *not* be attracted into simultaneity by a *V-C* relation (parallel to the *C-V* relation). If this is correct, then there would be no need for a strong *C-C* bonding to prevent coda consonants from synchronizing. This asymmetry between onsets and codas could also constitute the basis for the general weightlessness of onsets (cf. Davis, 1988): because of the C-V relation, adding a consonant to an onset increments the syllable duration by an amount less than the duration of the added consonant. If there is no V-C relation for codas, adding a consonant there could increment syllable duration by the full duration of the added consonant.

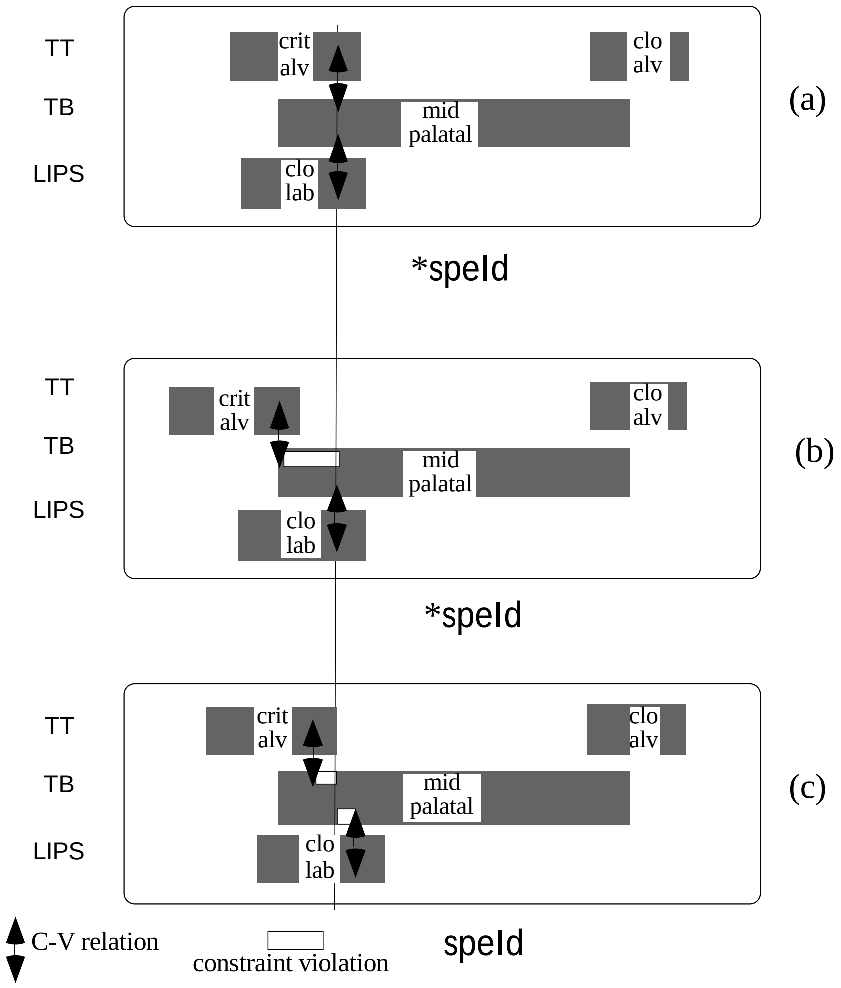


Figure 3: Hypothetical and actual gestural scores for "spayed". Arrows show coordination imposed by C-V relation. Vertical line marks hypothesized point in vowel gestures to which the C-V relation coordinates the consonants. White bars represent violations of C-V relation imposed by C-C constraint. (a)-(b):Hypothetical (non-occurring) gestural scores (c) Actual gestural score (see text).

3. Emergence of gestural structures

From one point of view, the addition of bonding strength to gestural structures comes with a theoretical cost: it adds several more potential degrees of freedom to the description of phonological objects. Since it appears that the gestural structures actually employed by languages involve a small subset of the *a priori* possible values of phasing and bonding, it becomes increasingly important to address the issue of how gestural structures come to have this particular restricted set of properties. Put another way, we would like to understand how bonding among gestural primitives or “atoms” creates a restricted set of larger stable structures whose “molecular” properties underlie traditional units such as segments, syllable constituents and syllables. We think that an approach based on principles of *self-organization* could be used to begin to answer such questions. In the recent past, principles of self-organization (spontaneous emergence of order) have been identified that cut across domains such as physics, biology, economics, sociology, anthropology (cf. Kauffman, 1995). While there here has been some application of related ideas to aspects of phonological structure (Lindblom, MacNeilage &

Studdert-Kennedy, 1983; Lindblom, 1983; Maddieson, 1995), there has been no attempt to understand structural relations among gestural primitives in these terms.

Together with Stuart Kauffman of the Santa Fe Institute, we have been developing a research strategy for investigating the self-organization of gestural structures. The goal is to discover the attractors in the space of possible gestural configurations, that is, the kinds of gestural configurations to which a communication system based on gestures would tend, given certain assumptions.

The strategy works like this. Gestures are treated as “atoms” and combined into larger and larger structures, or “molecules,” with random phasing. Under some set of conditions in which combination occurs, structures with stable properties of phasing and bonding like those seen in languages may emerge from this random combination. Thus, we can use this technique to identify the boundary conditions that are necessary for the emergence of gestural structures.

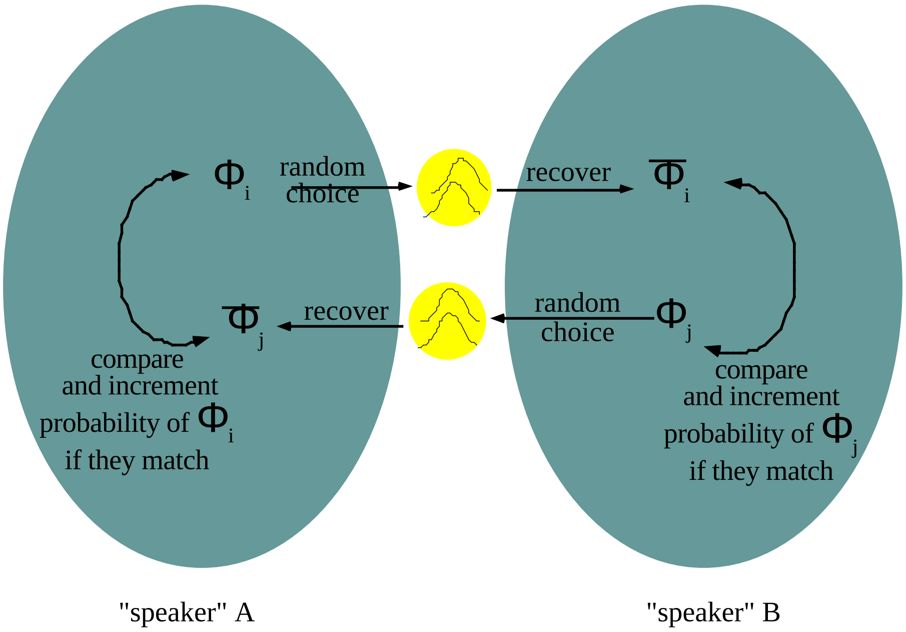
We have begun to explore the plausibility of this enterprise by undertaking a very simple computational experiment that illustrates the role of boundary conditions. In this experiment, we have two “speakers”, each of which produces the phasing of two consonant gestures completely at random, subject to two conditions:

Accommodation condition

The two “speakers” want to act like each other.

Recoverability condition

There are limits on ability of a “speaker” to recover the gestures from sound.



Φ_i
random choice
recover
Φ_i
Φ_j
recover
random choice
Φ_j
compare and increment probability of Φ_i if they match
compare and increment probability of Φ_j if they match
"speaker" A
"speaker" B

Figure 4: Schematic representation of computational experiment on self-organization of gestural structures.

The experiment investigates the phasing of two gestures, A and B. A continuum of phasing, from gesture A leading to gesture B leading, is divided into equal intervals, and at the beginning of the experiment, each speaker produces each of the resulting values with equal probability. That is, there is no bonding between the gestures at this point. The experiment proceeds as follows: Each speaker chooses a phase at random, and then compares its own production to the one recovered from the other speaker. If they match (within some specified criterion), then that speaker increments the probability of producing that phase again. This is illustrated schematically in Figure 4. Two versions of the experiment have been run that test different assumptions about the **recovery** condition. In this “toy’ experiment, recovery is simulated by taking the produced value and adding random gaussian noise. In one version, recovery is assumed to be equally accurate throughout the entire continuum (gaussian noise employed to derive recovered value has the same standard deviation throughout the continuum). In the second version, recovery is assumed to be less accurate (more noisy) when the two gestures are approximately synchronous, as would be the case when the two gestures are closures, and therefore capable of "hiding" one another. This is simulated by using gaussian noise with a higher standard deviation for produced values around zero.

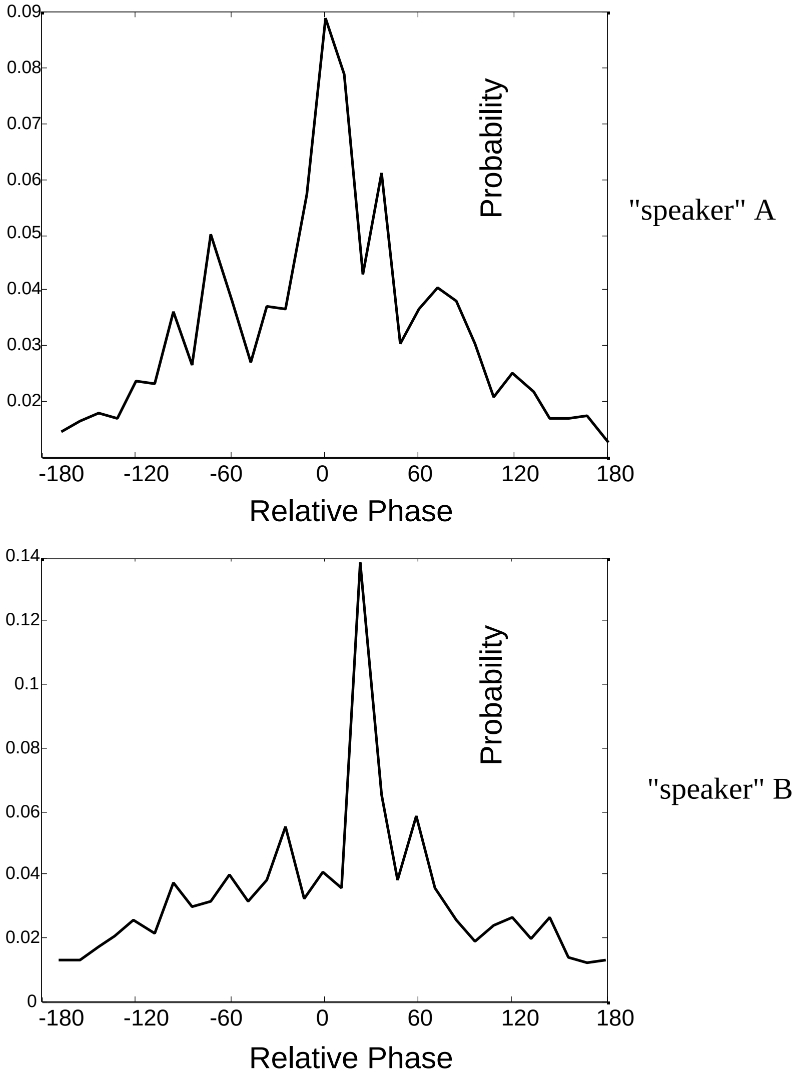


Figure 5: Sample result of computational experiment illustrated in Figure 4. Graphs show probability of "speakers" emitting a particular phase relation. In this experiment, recoverability is set to be equally accurate across the entire phasing continuum.

For the equal recoverability condition, both "speakers" converge on a fairly narrow range of phase values (after roughly 12,000 iterations, where the probability is incremented by .01 for a successful match). A typical result is shown in this Figure 5. Stable bonding between the gestures has emerged. Note that the modal phase value is not identical for the two speakers, but it is, of course, within the error criterion chosen. The emergent modal value tends to occur near the middle of the continuum, because the error criterion is relatively broad (as much as 90 degrees in some simulations) and non-directional. So for a given error criterion, ε, points near the middle of the continuum can match values in the range Φ–ε to Φ+ε, a range of 2ε, while a continuum endpoint can only match values in a range of ε adjacent to that endpoint (Φ-ε or Φ+ε).

A typical result for the condition in which recoverability is poorer in the center of the continuum is shown in Figure 6. Here, two phasing modes are found, one on either side of the region of increased recovery error. In this case, not only has bonding emerged, but a contrast in serial order has emerged: A leading B and B leading A. Thus, recoverability may play an important role in constraining the kinds of contrastive gestural patterns that languages employ. Certain phase relations are discouraged, and a contrast in serial order emerges spontaneously.

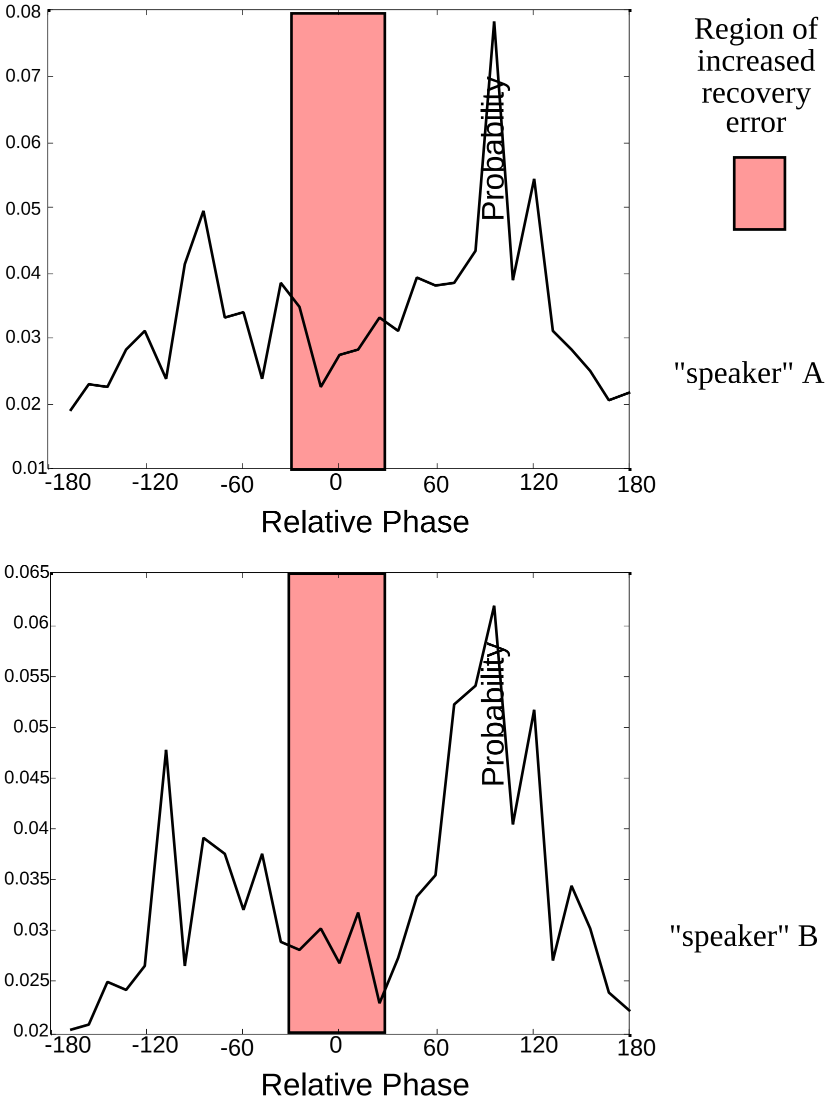


Figure 6: Sample result of computational experiment illustrated in Figure 4. Graphs show probability of "speakers" emitting a particular phase relation. In this experiment, recoverability is set so it is more error-prone in the shaded region of the continuum, as might be the case if gestures were both stop closures.

This experiment is a "toy" designed to explore the plausibility of the basic approach, and to demonstrate the roles of the accommodation and recoverability conditions in the emergence of patterns of phasing and bonding. We are, nonetheless, encouraged by the results and plan to do systematic simulations in which the "speakers" generate utterances using our gestural (production) model and recover gestures and their phasing from the resulting acoustics, using an automatic algorithm.

Acknowledgements

This work was supported by NIH grant HD-01994 and NSF grant SBR-9514730 to Haskins Laboratories. We thank Stuart Kauffman for his inspiration and suggestions.

References

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

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

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 151-206.

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

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*(3-4), 155-180.

Browman, C.P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In T. van Gelder & B. Port (Eds.), *Mind as motion* (pp. 175-193). Cambridge, MA: MIT Press.

Byrd, D. (1995). C-centers revisited. *Phonetica, 52*, 285-306.

Byrd, D. (1996a). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*, 209-244.

Byrd, D. (1996b). A phase window framework for articulatory timing. *Phonology, 13*, 139-169.

Davis, S. (1988). Syllable onsets as a factor in stress rules. Phonology, 5, 1-19.

Honorof, D. N., & Browman, C. P. (1995). The center or edge: How are consonant clusters organized with respect to the vowel? In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences.*, (Vol. 3, pp. 552-555). Stockholm: KTH and Stockholm University

Kauffman, S. (1995). *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*. Oxford: Oxford University Press.

Keating, P.A. (1990). The window model of coarticulation: Articulatory evidence. In J. Kingston & M.E. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 451-470). Cambridge: Cambridge University Press.

Liberman, A.M., Cooper, F.S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431_461.

Lindblom, B. E. (1983). Phonetic universals in vowel systems. In J. Ohala (Ed.), *Experimental phonology*, . New York: Academic Press.

Lindblom, B. E., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of language universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations for language universals*, (pp. 181-203). The Hague: Mouton.

Maddieson, I. (1995). Gestural Economy. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences.*, (Vol. 4, pp. 574-577). Stockholm: KTH and Stockholm University.

Mattingly, I. G. (1981). Phonetic representation and speech synthesis by rule. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech*, (pp. 415-420). Amsterdam: North Holland.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Experimental brain research*, (pp. 129-144). New York: Springer-Verlag.

Silverman, D. (1995). *Phasing and recoverability.* Unpublished PhD, UCLA.