Self-organization of Syllable Structure: a Coupled Oscillator Model

Hosung Nam, Louis Goldstein & Elliot Saltzman

1. Syllable structure

It has been generally claimed that every language has syllables with onsets (CV structure), while languages may or may not allow coda consonants (VC structure). While recent work on Arrernte (Breen & Pensalfini, 1999) has cast doubt on the absolute universality of onsets, it is clear that there is a significant cross-linguistic preference for CV structure (Clements & Keyser, 1983; Clements, 1990). In addition, evidence from phonological development shows that CV structure is typically acquired before VC (e.g., Vihman & Ferguson, 1987; Fikkert, 1994; Demuth & Fee, 1995; Gnanadesikan, 1996; Salidis & Johnson, 1997; Levelt et al., 2000). This preference for CV structure in distribution and acquisition has been understood as due to universal grammar (UG: Chomsky, 1965) where CV is the unmarked core syllable structure. Yet, the UG hypothesis does not answer the question “Why is CV the most unmarked structure?” This study aims to provide a rationale, grounded in dynamical systems theory, for why CV is favored across languages by simulating the self-organization of syllable structure in phonological development using a model in which syllable structures are defined by the coupling graph in a system of gestural planning oscillators that control patterns of intergestural relative timing. The simulation shows that, due to the hypothesized stronger coupling inherent in CV compared to graphs, CVs emerge earlier than VC graph. The model of syllable structure based on coupled planning oscillators (see section 2 below) has been developed to account for a variety of empirical observations about speech production (Nam & Saltzman, 2003; Goldstein et al., 2006; Nam et al, submitted a, b), independently of any consideration of acquisition facts.

In addition to the explanatory weakness of the UG hypothesis, two additional empirical observations about phonological development are not easily accommodated by the UG hypothesis. One is that the delay in the emergence of coda consonants varies across languages as a function of the
frequency of coda consonants in adults’ word production (Roark & Demuth 2000). Thus, both intrinsic and extrinsic factors interact in the development of syllable structure. The second observation is that, unlike the acquisition patterns of single consonants, it has been shown in several languages that consonant clusters appear earlier in word- (or syllable-) final position than word-initial position (Mexican-Spanish: Macken, 1977; Telugu: Chervela, 1981; German and Spanish: Lleó & Prinz, 1996; Dutch: Levelt, Schiller, & Levelt 2000; English: Templin, 1957, Paul & Jennings, 1992, Dodd, 1995, Watson & Scukanee, 1997, McLeod et al. 2001, Kirk and Demuth 2003).

In the self-organization model presented in this paper, both of these facts are readily accounted for. The first follows from hypothesizing a self-organizing process that includes both intrinsic dynamic constraints on planning intergestural timing and attunement of speaker/listener agents to the behavior of other agents. The second follows from the independently motivated hypothesis (Browman & Goldstein, 2000; Nam & Saltzman, 2003) that the production of consonant clusters can involve competition between the coupling of (all of) the consonants to the vowel, and the coupling of the consonants to one another. As we will see in the rest of the paper, the fact that CV coupling is stronger than VC coupling makes it easier to learn to produce single onset Cs than coda Cs, but that competition provided by the stronger CV coupling in onset makes it more difficult to learn the CC coordination.

2. A coupled oscillator model of syllable structure

Within the framework of articulatory phonology (e.g., Browman & Goldstein, 1992; 1995), word forms are analyzed as molecules built up by combining discrete speech gestures, which function simultaneously as units of speech production (regulating constriction actions) and units of (phonological) information. In these molecules, the relative timing of gestures is also informationally or phonologically significant. For example, the words ‘mad’ and ‘ban,’ are composed of the identical set velum, tongue tip, tongue body, and lip gestures. The only difference between these two words is in velum gesture’s timing with respect to the other gestures. Thus, there must be some temporal glue in speech production that keeps the gestures appropriately coordinated. As with the gestures themselves, this glue appears to have both a regulatory and an informational function.
2.1. A coupled oscillator model of intergestural timing

We have been developing a model of speech production planning in which dynamic coupling plays the role of temporal glue (Saltzman & Byrd, 2000; Nam & Saltzman, 2003; Goldstein et al, 2006; Nam, in press; Nam et al., submitted a,b). The central idea is that each speech gesture is associated with a nonlinear planning oscillator, or clock, and activation of that gesture is triggered at a particular phase (typically 0°) of its oscillator. A pair of gestures can be coordinated in time by coupling their corresponding oscillators to one another so that the oscillators settle into a stable pattern of relative phase phasing during (internal) planning. Once this pattern stabilizes, the activation of each gesture is triggered by its respective oscillator, and stable relative timing of the two gestures is achieved.

There are two sources of evidence for the hypothesis that the relative timing of gestures is controlled by coupling their planning oscillators. The first comes from experiments on phase-resetting. When subjects repeat a particular word, the gestures composing that word exhibit stable relative phasing patterns. When the ongoing repetition is mechanically perturbed, the characteristic pattern of phasing is quickly re-established (reset) in a way that is consistent with the behavior of coupled oscillators (Saltzman et al., 1998). When a word is produced only once instead of repeatedly, qualitatively similar phase-resetting is also observed.

The second source of evidence comes from experiments on the kinematics of speech errors (Goldstein et al., in press). The subjects repeat phrases like ‘cop top,” in which the tongue dorsum gesture for /k/ and the tongue tip gesture for /t/ alternate. Over time, subjects’ productions tend to shift to a new pattern, in which tongue tip and tongue dorsum gestures are produced concurrently at the beginning of both words (causing the perception of speech errors, Pouplier & Goldstein, 2005). These errors have been interpreted as resulting from a shift to a more stable mode of frequency-locking among the gestural timing oscillators that compose these words. Specifically, in normal productions, there is a 1:2 relation between the frequency of the tongue tip (or dorsum) oscillators, and the oscillators for vowel or final C, whereas the new (errorful) pattern exhibits a more stable 1:1 frequency relation. Such shifts to more stable modes of frequency-locking have been observed in several types of human bimanual coordination (Turvey, 1990; Haken et al., 1996). Again, very similar kinds of changes are observed in tasks that do not involve any overt repetition (Pouplier, in press), when the shifts must occur in planning process.
To illustrate the planning model, consider Fig. 1. The word “bad” is composed of three gestures: a Lip closure, a wide palatal constriction of the Tongue Body, and a Tongue Tip closure. A typical arrangement of these gestures in time, as can be observed from kinematic data, is shown in the gestural score in (a). The horizontal extent of each box represents that gesture’s activation interval, the time during which its dynamics controls the appropriate constricting system (lips, tongue body, tongue tip). These activation intervals are the output of the coupled oscillator model of planning. The input to the planning model is a coupling graph, shown in (b). The graph specifies how the oscillators controlling the gestures’ timing are coupled to one another. The graph shows that the oscillator for the palatal Tongue Body (/a/) gesture is coupled to both the Lip (labial closure) and Tongue Tip (alveolar closure) oscillators. The solid line connecting the TB gesture with the Lip gesture indicates that the coupling target of those gestures is specified as an in-phase relation between the oscillators, while the dotted line connecting the TB gesture to the Tongue Tip gesture indicates that an anti-phase coupling target is specified.

<bad>

Figure 1. (a) Gestural score for ‘bad.’ Time is on the horizontal axis. (b) Coupling graph for ‘bad.’ Solid line indicates in-phase coupling target, dotted line indicates anti-phase coupling target. After Goldstein et al, 2006.

At the onset of the planning simulation for an utterance, each (internal) oscillator is set to an arbitrary initial phase, and the oscillators are set into motion. Over time, the coupling among the oscillators causes them to settle into a stable pattern of relative phases. The model that accomplishes this, the task dynamics model of relative phase first developed by Saltzman & Byrd (2000) for single pairs of gestures, has been extended to a network of
multiple couplings (Nam & Saltzman, 2003; Nam et al., submitted a, b). The coupling between each pair of gestures is controlled by a (cosine-shaped) potential function defined over their relative phase, with a minimum at the target (intended) relative phase value. When the relative phase of two oscillators differs from its target, forces derived from the potential functions are applied to the individual oscillators that have the effect of bringing their relative phase closer to the target value. The steady-state output of this planning process is a set of oscillations with stabilized relative phases. From this output, the gestural score (for example, the one for ‘bad’ in (a)) is produced in which gestural onset and offset times are specified as a function of the steady-state pattern of inter-oscillator phasing. So, because the TB and Lip oscillators (Fig. 1b) settle into a steady-state in-phase pattern, their activation intervals (Fig. 1a) are triggered synchronously. The TB and TT oscillators settle into an anti-phase pattern, and their activations show the TT gesture is triggered substantially later than the TB gesture.

Gestural scores can be input to the constriction dynamics model (Saltzman & Munhall, 1989), which generates the resulting time functions of constriction (tract) variables and articulator trajectories. Articulator trajectories can then be used to calculate acoustic output in our vocal tract model.

2.2. Intrinsic modes of coupling

One theoretical advantage of modeling timing using oscillator coupling graphs is that systems of coupled nonlinear oscillators can display multiple stable modes. These modes have been shown to play a role in the coordination of oscillatory movements of multiple human limbs (fingers, arms, legs; see Turvey, 1990, for a review). Such experiments show that when asked to oscillate limbs in a regular coordinated way, subjects can do so readily, without any training or learning, as long as the task is to coordinate them in-phase (0° relative phase) or anti-phase (180° relative phase) patterns. Other phase relations can be acquired as part of learned skills, for example complex drumming, but only with significant training.

While these two modes of coupling are (intrinsically) available without training, they are not equally stable. This has been demonstrated in experiments in which the frequency (rate) of subjects’ oscillation is manipulated. When subjects oscillate two limbs in an anti-phase pattern and the frequency is increased, the relative phasing undergoes a spontaneous transition to the in-phase pattern. However, if subjects begin oscillating in the
in-phase pattern, increases to oscillation frequency have no effect on the relative phase. From such results, it has been concluded that in-phase is the more stable mode.

These experimental results were the basis for Haken, Kelso and Bunz’s (1985) model of coordinated, rhythmic behavior, which can be understood as a self-organized process governed by low-dimensional nonlinear dynamics. They developed a simple potential function (aka the HKB potential function) that can account quantitatively for the results of these experiments.

\[ V(\psi) = -a \cos(\psi) - b \cos(2\psi) \quad (\psi = \phi_2 - \phi_1) \]

Figure 2. HKB potential function. \( a \) is varied from 1 (left) to 2 (center) to 4 (right).

The function, shown in Fig. 2, is the sum of two cosine functions of relative phase, one of which has half the period of the other. \( a \) and \( b \) are weighting coefficients of the two cosine functions respectively and their ratio, \( b/a \), determines the shapes of the potential landscapes. The left-most example in Fig. 2 shows the shape of the function for \( b/a = 1 \). There are two potential minima at 0 and 180 degrees and, depending on initial conditions, relative phase can stabilize at either of the two minima, making them attractors. However, the valley associated with the in-phase minimum is both deeper and broader and, thus, has a larger basin in the technical sense that there is a larger range of initial values \( \psi \) that will eventually settle into that minimum.

The experimental results suggest that frequency of oscillation (rate) is a control parameter for the system: as it is scaled upward in a continuous fashion, the behavior of the system will undergo a qualitative change at some critical point. If the value of \( b/a \) is specified as an inverse function of oscillation frequency, then the HKB model in Fig. 2 predicts an abrupt phase transition from the anti-phase to in-phase pattern as rate increases. This can be seen by comparing the function for the different values of \( b/a \) shown. As the \( b/a \) decreases from 1.0, the basin of the anti-phase mode becomes shallower and eventually the attractor disappears when \( b/a = 1/4 \).
At that point, the anti-phase pattern become unstable, and relative phase will be attracted to and captured by the minimum at 0°.

2.3. Syllable structure and coupling modes

For a system like speech, which is acquired early in a child’s life and without explicit training, it would make sense for the early coordination of speech gestures to take advantage of these intrinsically available modes. It has been proposed (e.g., Goldstein et al., 2006) that phonological systems make use of the in-phase and anti-phase modes, which form the basis of syllable structure in phonology. If we treat phonology as a fundamentally combinatorial system, consider the problem of coordinating two gestures—an consonant gesture with a vowel gesture—given the predisposition to exploit the presence of the intrinsically distinct in-phase and anti-phase modes. Goldstein et al. (2006) have proposed the coupling hypothesis of syllable structure: in-phase coupling of C and V planning oscillators underlies what we observe as CV structures, and more generally underlies the relation between onset and nucleus gestures; the anti-phase mode of coupling planning oscillators underlies CC structures and the relation between the nucleus and coda gestures.

Evidence for the in-phase mode in CV structures can be found in the fact that the constriction actions for the C and V gestures in CVs are initiated synchronously. For example, in Fig. 1, the activation of the Lip gesture and the Tongue Body gestures begin at the same time. This synchrony follows from the hypothesis that the oscillators associated with the C and V gestures are coupled so as to settle into an in-phase mode, together with the model assumption that gestural activation is triggered at phase 0 of a gesture’s oscillator. The idea that consonant and vowel gestures are triggered synchronously goes back to the pioneering work of Kozhevnikov & Chistovich (1965). Kinematic data for V₁pV₂ and V₁bV₂ utterances presented by Löfqvist & Gracco (1999) show that the onset of lip movement for /p/ or /b/ and the onset of tongue body movement for V₂ occur within 50 ms of one another, across all for 4 subjects and all six different V₁V₂ patterns, with only 2 outlier values. In the case of a coda /p/, the relation to the vowel is obviously not one of synchrony. The evidence that the oscillators exhibit the anti-phase relation is necessarily indirect. The anti-phase relation implies that the final /p/ will be triggered at 180° of the vowel gesture. The point in time that corresponds to 180° will, of course, depend on the frequency of the vowel oscillator. In Nam et al. (submitted, b), we show...
that simple hypotheses about the frequencies of vowel and consonant oscillators, combined with the hypothesis of in-phase CV and anti-phase VC coupling can account for a rich set of quantitative phonetic timing data.

The *coupling hypothesis* can also be used to explain a variety of qualitative properties of CV and VC structures. These include the following:

- **Universality.** CV syllables occur in all human languages, while VC ones do not. Since the in-phase mode is more stable, stronger (the potential well in Fig. 2 is deeper for in-phase), and has a larger basin of attraction than the anti-phase mode, it follows that the in-phase (CV) mode should always be available for coordinating Cs and Vs in a language, while the anti-phase (VC) mode may not be.

- **Combinatoriality.** Even in languages which allow VC structures, V and codas often exhibit restrictions on combination. Onsets and rimes, in contrast, can usually combine completely freely in languages. Indeed, their relatively free combinatoriality is the major source of phonological generativity, and the basis for the traditional decomposition of the syllable into onset and rime. Goldstein et al. (2006) propose that there is a relation between the stability/strength of coupling and combinatoriality. The idea is that it is possible to jointly perform *any* two actions as long as they are coordinated in-phase because this coordination is intrinsically the most stable. Even though anti-phase coordinations are more stable than other non-in-phase modes, speakers may need to learn *not* to use the more stable in-phase coordination for those forms that have coda Cs. Some combinations may fail to be learned, leading to combinatorial restrictions.

- **Re-syllabification.** Single, intervocalic coda Cs may be re-syllabified into onset position in running speech, particularly as speech rate increases (Stetson, 1951; Tuller & Kelso, 1991; de Jong, 2001a, b). This follows automatically from the HKB model (Fig. 2), when CV is defined as in-phase and VC is defined as anti-phase.

A further key hypothesis in the coupled oscillator planning model is that incompatible coupling specifications can compete with one another and that, during planning, the system of oscillators settles to a set of steady-state relative phases that is the result of the competition. The use of competitive coupling was originally proposed for CC and CV coupling for onset clusters (Browman & Goldstein, 2000). *All C* gestures in an onset were hypothesized to be coupled in-phase with the vowel (this is what defines an
onset consonant). For some combinations of C gestures, such as oral con-
strictions with velum or glottis gestures, synchronizing multiple C gestures
results in a recoverable structure, and the result is what is usually analyzed
as a multiple-gesture segment (nasal, lateral, voiceless stop). In other cases
(for example clusters such as /sp/), synchronous coordination does not pro-
duce a recoverable structure, and the two gestures must be at least partially
sequential (Goldstein et al, 2006; Nam, in press). Therefore, the oral conso-
nant gestures must also be coupled anti-phase to each other. Browman and
Goldstein (2000) proposed that this competitive structure could account for
a previously observed generalization about the relative timing of consonant
and vowel gestures in forms with onset clusters (the so-called ‘c-center’
effect, Browman & Goldstein, 1988; Byrd, 1995). As Cs are added to an
onset, the timing of all Cs relative to the vowel is shifted: the C closest to
the vowel shifts rightward to overlap the vowel more; while the first C
slides leftward away from the vowel. The temporal center of the sequence,
the c-center, maintains a relatively invariant timing to the vowel. This ef-
fect has since been modeled in coupled oscillator simulations (Nam &
Saltzman, 2003; Nam, et al, submitted-b). In-phase coupling of two onset
Cs with the V and anti-phase coupling of the two Cs with each other results
in an output in which the phasing of C₁ and C₂ to the V is -60° and 60°,
respectively, and the C₁C₂ phasing is 120°, a pattern consistent with available
data.

However, available evidence suggests that this kind of competitive
structure may or may not be found in coda Cs, depending on the language
(Nam, in press), or possibly on the speaker. In English, coda clusters do
not exhibit the c-center effect consistently (Honorof & Browman, 1995),
though it may be found for some speakers (Byrd, 1995). Browman & Gol-
dstein (2000) hypothesized a non-competitive structure for codas in English:
the first coda C is coupled anti-phase with the vowel and the second coda C
is coupled anti-phase with the first. More recent work (Nam & Saltzman,
2003; Nam, et al, submitted-b) has shown that this hypothesized difference
between onset and coda clusters for English accounts for the lack of coda c-
center effect and also for the fact that gestures in onset clusters exhibit less
variability in relative timing than do gestures in coda clusters. When noise
is added to the coupled oscillator simulation, the competitively-coupled
onset oscillators exhibit less trial-to-trial variability than do the noncom-
petitively-coupled coda oscillators.

Cross-language differences in the presence of competitive vs. non-
competitive coupling structure in coda have been proposed (Nam, in press)
in order to account for the differing moraic status of coda Cs across languages. English, and other languages in which coda Cs are moraic, are modeled with a non-competitive structure in coda: adding Cs to a coda is predicted not to decrease the acoustic duration of the vowel, and so the added C increases the duration of the entire syllable substantially, which in turn adds weight to the syllable for metrical purposes. Languages in which coda Cs are not moraic (e.g. Malayalam), are modeled with a competitive structure in coda, which causes acoustic vowel shortening as Cs are added to the coda, and therefore a lack of weight associated with the added C. Nam (in press) showed how these hypothesized coupling differences can account for acoustic data from these language types (Broselow et al., 1997).

It is striking that onset Cs are have been described as non-moraic in almost every language of the world (one exception is Ratak (a dialect of Marshallese; Bender, 1999). Thus, there appears to be a strong asymmetry between onsets and codas. Onsets always have a competitive structure, while this may be lacking in codas. However, regardless of topological differences in the coupling structures of onsets and codas, onset CVs are characterized by in-phase couplings while coda VC are characterized by anti-phase couplings. This asymmetry between onsets and codas led to the hypothesis (Nam, in press) that, due to the greater intrinsic strength of in-phase coupling, all prevocalic Cs are pulled into the in-phase relation with the V, whereas coda Cs can (and do in some languages) escape the pull of anti-phase coupling with the V.

In this paper, we show that the difference in coupling modes between onsets and codas can also account for the difference in acquisition time between CV and VC structures. To do this, we performed simulation experiments investigating the self-organization of syllable structure, in which the only relevant pre-linguistic structure attributed to the child is that (s)he comes equipped with (a) the HKB potential function for the pairwise coordination of multiple actions and (b) the ability to attune its behavior to the behaviors of others in its environment.

3. Self-organization of syllable structure

We investigated the self-organization of syllable structure in a series of simulations with a computational agent model. Agent models have been employed to investigate several aspects of phonological and phonetic structure, including: partitioning of physical continua into discrete phonetic
categories (Oudeyer, 2002, 2005, 2006; Goldstein, 2003), the structure of vowel systems (deBoer, 2001; Oudeyer, 2006), consonant-vowel differentiation (Oudeyer, 2005), and sequentiality in consonant sequences (Browman & Goldstein, 2000). In these simulations the agents interact using a very simplified set of local behaviors and constraints, and through these interactions, the agents’ internal states evolve, as do the more global properties of the system. Depending on the choice of constraints, the system may evolve to have quite different properties. Thus, for example, the importance of some constraint \( k \) in the evolution of some property of interest \( P \) can be evaluated by contrasting the results of simulations with and without that constraint. The models are not meant to be faithful simulations of the detailed process of (phylogenetic or ontogenetic) evolution of some property, but rather a way of testing the natural attractors of a simple system that includes the constraint of interest.

The models employed here involve a child agent with no syllable structure, and an adult agent with a developed syllable structure. The child comes to the simulation with two distinct classes of actions (C and V), and it attempts to coordinate them in time. The existence of distinct C and V actions early in the child’s development (e.g. during babbling) has been denied in the frame-content model of speech production development (MacNeilage and Davis, 1998); that model’s treatment of syllable structure and its emergence contrasts, therefore, with the one proposed here. This conflict will be addressed further in the Discussion. At this point, however, we note that even if the frame-content view is correct in excluding a C-V distinction during the babbling stage, this distinction could still have evolved by the time the child is producing words with onset and coda consonants, which is the age we are simulating here.

3.1. Emergence of CV vs. VC structures

Learning in this model is accomplished by self-organization under conditions imposed by both intrinsic constraints on coordination and attunement to the coordination patterns implicit in, and presumably recoverable from, the acoustic environment structured by the ambient language. A Hebbian learning model was employed, in a manner similar to that used by Oudeyer (2006) to model the emergence of discrete phonetic units. The simulation includes a child agent and an adult agent. Both have a representation of the relative phase between a pair of gestures that evolves over
time. That representation is used to select a relative phase value when the agent intends to coordinate the pair. The adult representation includes modes corresponding to CV (in-phase) and VC (anti-phase), where the relative strength of these modes can differ from language to language, corresponding to the relative frequency of CV and VC structures in that language. At the outset of learning, the child’s representation does not include any CV or VC modes, so the child displays no preference in producing some phases over others; the phases produced are random. As the result of the learning process, modes develop that correspond to the modes found in the adult speakers’ representations and to their relative strength. What we predict is that even though the child will develop the same modes as the adult partner, the rate at which the CV mode develops is faster than the VC mode, regardless of the ultimate relative strength of the modes.

The learning simulation proceeds as follows (Fig. 3). On a given learning iteration, the child selects a relative phase value, $\psi_{\text{SEL}}$, to produce from its evolving distribution of relative phases, and a single-well potential function with a minimum at that value is added to the double-well (HKB) intrinsic potential function to create a resultant composite potential function.

![Figure 3. Self-organization learning model for emergence of CV and VC structures](image_url)
The agent then plans the production of a pair of gestures by using the composite potential to specify the coupling function between for a pair of corresponding planning oscillators. Oscillator motion is initialized with a random pair of initial phases, and oscillator motions settle to a stabilized relative phase, $\psi_{\text{OUT}}$, in accordance with the shape of the composite potential. The child then compares $\psi_{\text{OUT}}$ to the (veridically) perceived relative phase of an utterance token produced by the adult, $\psi_{\text{ADULT}}$. If the difference between these two relative phases meets a given criterion, the child tunes its relative phase representation to increase the likelihood of producing that phase again. The details of the model are now described.

### 3.1.1. Phase Representation and Selection Model

The relative phase of the oscillators responsible for triggering C and V gestures is represented (for both the child and the adult agents) by a set of virtual ("neural") units, $\psi_i$, each of which represents some value of the relative phase parameter. At the outset of the simulation, values of relative phase are assigned to these neural units in one degree increments from $-179^\circ$ to $360^\circ$. Thus, $\psi_1 = -179^\circ$, $\psi_2 = -178^\circ$, ..., $\psi_{540} = 360^\circ$, for a total of 540 units. On a given learning trial, one of the 540 units is selected at random and its relative phase value, $\psi_{\text{SEL}}$, is used as the agent’s intended relative phase. Since at the beginning of the simulation, the units’ relative phase values are uniformly distributed across the phase continuum, the value of $\psi_{\text{SEL}}$ will be completely random. As learning progresses through attunement (section 3.1.3 below), the values of the neural units will come to be clustered, with most units having values near $0^\circ$ or $180^\circ$. We will represent this clustering at various points in the simulation by plotting a frequency histogram, showing the number of units as a function of relative phase. We will refer to this distribution as the density distribution, and the number of units sharing a value of relative phase as its density. The density distribution will develop peaks at $0^\circ$ and $180^\circ$; therefore, since $\psi_{\text{SEL}}$ values are chosen by randomly sampling the density distribution, the values of $\psi_{\text{SEL}}$ will also tend to be either $0^\circ$ or $180^\circ$ as the distribution develops over the course of learning.

### 3.1.2 Planning and Production Model

Once an intended relative phase, $\psi_{\text{SEL}}$, is selected by the child, it is used to specify a corresponding relative phase target value, $\psi_{\text{SEL}}$, that will be
used by the coupled oscillator model to plan the production of a corresponding pair of relatively phased C and V gestures. The value of $\psi^{\text{SEL}}$ helps to shape the coupling forces between these oscillators through its contribution to the composite potential function for relative phase. To model the fact that not all relative phases patterns can be equally easily learned, the intended relative phase contributes to the shape of this potential, but so does the HKB potential function (see section 2.2) that represents the intrinsic modes of coupling two oscillators. The intended relative phase is modeled by a cosine-shaped potential function with a single attractor at the value of $\psi^{\text{SEL}}$. This potential function is added to the HKB potential function to build a composite potential function. The relative weight of the intended potential in the composite should depend on how well learned, or well-practiced, the intended pattern is. To implement this, the coupling strength of the single attractor, represented by the height and steepness of the intended potential well, is scaled according to the density defined for the intended relative phase, $\psi^{\text{SEL}}$, in the child’s evolving, experience-dependent phase representation (section 3.1.1). The squashing function relating density to coupling strength shown in Fig. 4.

![Figure 4. Function relating Coupling Strength to mode Density. ($a=-0.5$, $g=0.15$, $x_0=70$ and $b=0.5$)](image)

At the beginning of a planning simulation, the phases of the planning oscillators are chosen at random; over the course of the simulation, they settle to a stable relative phase under the control of the composite potential function. This final, steady-state relative phase value, $\psi^{\text{OUT}}$, is thus determined both by the landscape of the composite potential and the basin selected by the randomly chosen initial conditions. As a result, the produced $\psi^{\text{OUT}}$ may correspond to neither the intended relative phase, $\psi^{\text{SEL}}$, nor to either the in-phase or anti-phase modes intrinsic to the HKB potential. This
Self-organization of Syllable Structure

final relative phase of the planning oscillators could then be used to trigger activation of C and V gestures but, in this agent model, the simulation simply stops with the oscillators settling at their final, steady-state pattern.

3.1.3 Attunement Model

The attunement of the child to the language environment is modeled by comparing the child’s produced relative phase ($\psi^{\text{OUT}}$) on a given learning cycle to a randomly chosen relative phase from the adult density function, $\psi^{\text{ADULT}}$. If the child’s produced value matches the randomly chosen adult value such that $|\psi^{\text{OUT}} - \psi^{\text{ADULT}}| < 5^\circ$, then the intended phase used by the child on that trial ($\psi^{\text{SEL}}$) is gated into the tuning (or learning) process. Tuning occurs as the units of phase representation that have values ($\psi_i$) similar to $\psi^{\text{SEL}}$ respond by increasing their level of activation as a function of the distance between $\psi_i$ and $\psi^{\text{SEL}}$ in phase space, $\psi_i - \psi^{\text{SEL}}$. The “receptive field” of each unit ($i$) is Gaussian-shaped and defines the unit’s activation level, $G_i$, as a function of this distance, as in (1):

\[
G_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{\psi_i - \psi^{\text{SEL}}}{\sigma} \right)^2}
\]

The value of $\sigma$ employed in the simulations was $40^\circ$. The values of the all units, $\psi_i$, are then attracted to $\psi^{\text{SEL}}$ in proportion to their activation levels, according to the parameter-dynamics equation in (2):

\[
\psi_i' = \psi_i - r G_i (\psi_i - \psi^{\text{SEL}}), \text{ where}
\]

$\psi_i'$ is the new unit value and $r$ is a learning rate parameter (equal to 1 in this simulation). The result of this parameter-dynamic tuning process is an evolution of the density distribution of units along the relative phase continuum. The example in Fig. 5 shows the initial uniform state of the child’s density distribution and the effect on the units of gating a value of $\psi^{\text{SEL}} = 2^\circ$ into the tuning process. Tuning ends one cycle of the phase learning process.
The adult’s phase distribution was varied across simulations to model ambient languages with different properties. For example, both English and Spanish exhibit preference for onsets (CV) over coda (VC) in production but coda consonants are more frequently produced in English than Spanish. This kind of asymmetry can be expressed by the difference between the phase distributions of English and Spanish adult speakers, with the density of units clustered in the anti-phase region being higher for English than Spanish. In the simulation presented here, three different hypothetical languages with different densities of in-phase and anti-phase modes were tested: (a) CV>VC (in-phase = 290 units/deg; anti-phase = 250 units/deg), (b) CV=VC (in-phase = anti-phase = 270 units/deg) and (c) CV<VC (in-phase = 250 units/deg; anti-phase = 290 units/deg). The first two types can be thought of as modeling Spanish and English, respectively. The third might be an Australian language, many of which have a preference for Coda Cs (e.g., Tabain et al, 2004), the extreme case being Arrernte, for which the claim of no onsets has been made (Breen & Pensalfini, 1999). These three different types of adult distributions are shown in the upper left insets in Fig. 6a, b, and c, respectively.

Figure 5. Visualization of tuning(learning) process in self-organization model
3.1.2. Results

Results of the three simulations are summarized in Fig. 6. For each simulation, the figure shows the density of in-phase and anti-phase modes as a function of iteration, or learning cycle, for 1000 iterations (main panels), the adult density (upper left insets), and the child’s agent-density distribution after 200, 400, 600, and 800 iterations (bottom row of insets).

(a) CV>VC  
(b) CV=VC  
(c) CV<VC

Figure 6. The learning process in three hypothetical language environments with different corresponding densities of in-phase and anti-phase modes

The results show that in-phase (CV) mode is stabilizes more quickly, than anti-phase (VC) mode regardless of different distribution of these modes in the adult agent. The number of trials required to achieve a stabilized density (180) is: (a) CV 369, VC 880, (b) CV 407, VC 765, (c) CV 468, VC 625. If we assume that adult-like production of CV or VC depends on development of the corresponding mode, the advantage of CV in acquisition is
predicted. In addition, as the data from different languages shows, the lag between production of onsets and codas is less in languages with a higher frequency of coda consonants (condition c). Finally, as the simulations continue (beyond the 1000 iterations shown in the graphs), the mode densities of the child agent come to match that of the adult. Thus, the adult pattern is learned, but en route, CVs are always acquired first.

3.2. Emergence of CCV vs. VCC structures

Now let’s consider a child who has begun to learn the distinct in-phase and anti-phase modes of coordinating C and V, and who now intends to produce a structure (s)he perceives as having two prevocalic Cs (CCV) or two postvocalic Cs (VCC). We assume that (s)he doesn’t know anything about coordinating the two Cs, so when (s)he attempts to produce the CCV, for example, she uses the learned in-phase CV mode to coordinate both Cs to the V. This will have the effect of synchronizing the two Cs, since they will both be coordinated in-phase with the V. The output, therefore, will be a form that does not typically match the adult form. By hypothesizing an additional (evolving) CC coordination pattern that can compete with the with CV pattern, the child can begin to produce adult-like structures once the CC pattern is well established enough to push the Cs apart, despite the synchronizing pressures of the CV pattern.

If we embed this scenario in an agent model like that section 3.1, and treat VCCs as completely parallel in structure to CCVs, then we would predict that the child’s CC (anti-phase) mode would emerge more quickly for VCC than for CVV. This would be the case because coupling strength in the model for an intended relative phase is dependent on mode density, which will typically be less for the anti-phase VC than for the in-phase CV early in the child’s experience. Thus, it would be easier to pull the Cs apart for the relatively less stable anti-phase VC couplings of VCC than for the stronger in-phase CV couplings of CCV. In real languages, the structure of VCC is not always parallel to CCV, as discussed in section 2.3, and it is not clear whether there is any cross-language difference in acquisition of coda clusters as a function of moraic status. However, if the child never even attempts to produce coda clusters with a competitive structure, then we would also expect VCC to develop more rapidly than CCV, since there would be no synchronizing force working against learning to produce the sequence of Cs in VCC. It seemed to us more sound methodologically to test the more challenging case in which the structures of CCV and VCC are
completely parallel and differences are in the strength of CV vs. VC coupling alone. We tested this prediction by performing additional agent simulations, one for CCV structures and one for VCC structures.

3.2.1. Extending the model to clusters

Both CCV and VCC simulations began identically to that in section 3.1, with adult input on each trial restricted to CV or VCs. These simulations assume that the frequencies of adult CV and VC are equal (see Fig. 6b). By iteration 200, there are already two established modes of C^V coordination (see the leftmost density distribution in the bottom row of Fig. 6b), though the frequencies associated with those modes continue to grow. (We use the term C^V to refer to the coordination of V and C gestures, without regard to whether they correspond to CV or VC structures). Thus, the child can effectively choose to produce either a CV or a VC by this time, depending on the part of the C^V distribution the child samples from (near 0° or near 180°). In this simulation, the C^V distribution is frozen at this iteration, and the child partitions the C^V distribution into CV and VC subparts (Fig. 7) and begins to attempt to produce CCVs or VCCs using the appropriate subpart for the simulation being run. At this point (iteration 200), frequencies associated with the CV and VC modes have not reached their adult levels, and all tested language types show a stronger mode for CV.

![Figure 7. Child agent’s developed CV (left) and VC (right) modes captured at iteration 200.](image)

**Selection and Planning.** From this point on, the child selects three intended phases on each trial: two from the CV sub-distribution (or VC sub-distribution, depending on simulation), specifying the CV (or VC) phase for each of the two Cs, $\psi_{C1V_{\text{SEL}}}$ (or $\psi_{V_{\text{SEL}}}$) and $\psi_{C2V_{\text{SEL}}}$ (or $\psi_{V_{\text{SEL}}}$), and one from a new CC distribution (which begins completely flat), $\psi_{CC_{\text{SEL}}}$. The adult output is also now restricted to CCVs or VCCs (depending on simula-
tion): two CV (or VC) phases, and also a C-C phase, drawn from a distribution around 120°, which represents the result of competition of the adult CV (or VC) mode with an anti-phase mode between Cs.

Thus, starting from iteration 200, the child’s coupling graph input to the planning model includes three coupling links with selected target relative phases. For the CCV simulation, these are $\psi_{C1V}^{SEL}$, $\psi_{C2V}^{SEL}$ and $\psi_{CC}^{SEL}$ and for VCC, these are $\psi_{VC1}^{SEL}$, $\psi_{VC2}^{SEL}$ and $\psi_{CC}^{SEL}$. There is no addition of the intrinsic HKB potential in this simulation\(^2\). The final set of relative phases is computed from the coupled oscillator model, and the final relative phase of the consonant cluster $\psi_{CC}^{OUT}$ is compared to the value selected from the adult distribution with a mode at 120°, $\psi_{CC}^{ADULT}$.

**Attunement.** When first attempting CCV or VCC outputs, the CV and VC modes already developed by the child both tend to foster CC synchronization. Hence, early cluster productions will be nowhere near the adult phase values (which, we would explain is why in real life we do not hear the child producing any clusters). Thus, tuning of the CC mode would never get off the ground if it depended on a child-adult matching criterion as stringent as that used in simulation in 3.1. Consequently, we hypothesize a different form of attunement here. A selected value ($\psi_{CC}^{SEL}$) is gated into the learning process whenever $\psi_{CC}^{OUT}$ is such that the Cs are planned to be triggered in the same temporal order as the adult Cs (C₁ before C₂). Thus, the condition for gating is that the final relative phase is positive: $\psi_{CC}^{OUT} > 0$.

Then, in order to make the simulation somewhat dependent on the goodness of the match between the correctly ordered CCs, the learning rate (in equation 2) was set to be proportional to the inverse of the relative phase mismatch between output and adult:

\[
(3) \quad r = \min (a / | \psi_{CC}^{ADULT} - \psi_{CC}^{OUT} |, 3), \quad \text{where } a = 20.
\]

### 3.2.2. Results

Results of the CCV and the VCC simulations are shown in Fig. 8. The density of the CC mode grows much more quickly in the VCC simulation than the CCV simulation. We assume that until stable sequential coupling of CC is acquired, phasing to the vowel will result in multiple Cs being produced synchronously. Therefore, they will not be readily perceivable in the child’s output, and the child would be described as not producing clusters of the relevant type. Thus, the model predicts that we should perceive children as
producing VCC structures before CCV structures, because CC coupling stabilizes earlier in VCC.

![CCV and VCC simulations](image)

Figure 8. Density of CC anti-phase mode, as a function of iteration number for CCV and VCC simulations.

4. Discussion

In summary, results presented here show that it is possible to model the course of acquisition of CV vs. VC and CCV vs. VCC structures as emerging from a self-organized process, if we make three basic hypotheses that form the boundary conditions for the process: 1) syllable structure can be modeled in terms of modes of coupling in an ensemble of gestural planning oscillators; 2) infants come to the learning process with very generic constraints that predispose them toward producing in-phase and anti-phase coordinations between pairs of gestures; and 3) infants attune their action patterns to those they perceive in the ambient language environment. Our results are striking in that the seemingly contradictory acquisition trends in the emergence of onsets and codas with single Cs vs. C clusters follow in this model from the same principle—the relatively greater strength of in-phase than anti-phase coupling.

There are, of course, many limitations to the type of modeling presented here. One major limitation is that we do not provide an explicit account of how the child agent is able to extract relevant phase information from the articulatory and acoustic patterns that result from adults’ phasing patterns. Behavioral evidence across several domains shows that sensory informa-
tion must make contact in some common form with motor plans (evidence for ‘common currency’ in speech gestures—Goldstein & Fowler, 2003; more generally a ‘common coding’ principle in action systems, Prinz, 1997; Galantucci, et al, 2006), and the discovery of mirror neurons (Rizzolatti et al, 1988) has made this notion seem more biologically tractable. However, it would certainly strengthen the kind of simulations presented here if we could show how that is accomplished in the case of a relatively abstract property like phase3.

The cluster simulation has a more specific limitation. For reasons discussed in section 3.2, we assumed in both CCV and VCC learning simulations that both Cs are identically coupled to the V, in-phase in CCV, and anti-phase in VCC. This was appropriate to do, we argued, as we wanted to assume complete parallelism between onsets and codas, and show that the observed differences could emerge from differential coupling strength of in-phase vs. anti-phase alone. However, in our model, the child acquires the coupling associated with a language in which coda Cs are not moraic, and we are left with the problem of how a child might acquire the pattern exhibited by English and other languages in which coda Cs are moraic. One possible answer is that in this case, adult CC phasing in VCCs would presumably be 180° (as opposed to the 120° employed here), so perhaps the infant would never, in this case, attempt to use the VC coordination to produce the final coda C. However, the implications of this would have to be tested in further simulations.

There are also predictions made by the model that could, in theory, be tested, although the relevant data are not yet available. One prediction is that CV structures should appear earlier than VC even in a language like Arrernte, in which the CV structures would presumably be phonologically ill-formed, as Arrernte has been argued to have no onsets. Such developmental data are not presently available. A second is that careful analysis of children’s early productions of intended CCVs that are perceived by adult transcribers as CV should reveal cases in which both Cs are being produced, but synchronously. Testing this would require articulatory data from children that is also not presently available.

Finally, we should consider alternative accounts for distributional and developmental regularities of syllable structure. Ohala (1996) attributes the CV preference in languages to the perceptual robustness of initial versus final Cs, because particularly in the case of stops, acoustic information that affords perceptual recovery is more salient in CV (e.g., the intensity of release bursts). While it is plausible that such differences form part of the
Self-organization of Syllable Structure

explanation for the CV preference, it is not clear how this explanation could account for the developmental lag of CCV compared to VCC. Rather, this view would presumably hypothesize no difference between CCV and VCC, since only the C adjacent to the V would be associated with positional asymmetries in perceptual robustness.

The other major alternative model of the development of syllable structure is the frame-content model (e.g., MacNeilage, 1998; MacNeilage & Davis, 2000). The model hypothesizes that a syllable structure ‘frame,’ based on mandibular oscillation, develops before the ‘content’ provided by individual C or V gestures. While the model has some plausibility, arguments have been raised that MacNeilage and Davis’ evidence for jaw-only oscillations in children’s babbling (a preponderance of certain CV combinations) cannot by itself be used as evidence for the jaw-only strategy (Goldstein et al, 2006; Giulivi et al, 2006). Regardless of how that issue is resolved, it is not clear how the frame-content model would account for the pattern of results predicted by the coupled oscillator model: the earlier acquisition of CV compared to VC, but the earlier acquisition of VCC compared to CCV.

References

Bender, B. 1999 Marshallese grammar (Chapter 1, 2). Ms., University of Hawai’i.
Chervela, N.


Demuth, K. & Fee, E.J. 1995 Minimal prosodic words in manuscript. Brown University and Dalhousie University.


Löfqvist, A. & V. L. Gracco

Macken, M. A.

MacNeilage, P. F.

MacNeilage, P.F. & B.L. Davis


McLeod, S., J. van Doorn, & V. A. Reed

Nam, H.
in press A competitive, coupled oscillator model of moraic structure: Split-gesture dynamics focusing on positional asymmetry. *Laboratory phonology VIII*

Nam, H., L. Goldstein, & E. Saltzman
submitted a Intergestural timing in speech production: the role of graph structure.
submitted b A dynamical model of gestural coordination.

Nam, H. & E. Saltzman

Ohala, J.

Oudeyer, P-Y.


Self-organization of Syllable Structure 27

Paul, R. & Jennings, P.

Pouplier, M.
in press Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech*.

Pouplier, M. & L. Goldstein
2005 Asymmetries in the perception of speech production errors. *Journal of Phonetics* 33, 47-75.

Prinz, W

Rizzolatti, G., R. Camarda, L. Fogassi, M. Gentilucci, G. Luppino, & M. Matelli

Roark, B. & K. Demuth

Salidis, J. & J.S. Johnson

Saltzman, E. & D. Byrd

Saltzman, E., A. Löfqvist, B. Kay, J. Kinsella-Shaw & P. Rubin

Saltzman, E. & K. Munhall

Stetson, R. H.
1951 *Motor Phonetics*. Amsterdam: North-Holland

Templin, M.

Tuller, B. & J.A.S. Kelso

Turvey, M.

Vihman, M.M. & Ferguson, C.A.


Watson, M. M. & G. P. Scukanec

Notes

* We gratefully acknowledge support of the following NIH grants: DC-00403, DC-03663, and DC-03782.

1 Since a cycle of tuning is done by attracting neural units to an experienced stimulus, the density of the units can grow differently at the ends of the range covered by units from. Thus, to prevent mode growths from emerging at the boundaries, 180 units are added beyond in-phase (0°) and anti-phase (180°) respectively, which are the predicted modes in this simulation. In addition, phases employed as input to tuning process are wrapped between -90° and 270°, which is medially positioned within the range of the units.

2 We assumed that, at this stage of the simulations, the intrinsic HKB function is so weak relative to the intended $\psi_{SEL}$ potentials that it could be ignored. When we experimented by adding a relatively strong intrinsic potential function to each of the $\psi_{SEL}$ intended potential functions, the result was to create multiple attractors in a given coupling. This led to unpredictable or chaotic sets of resultant relative phases.

3 It is encouraging to note that some progress has been made along these lines in extracting *syllabic phase* (a continuously varying, normalized measure of temporal position within syllables) from speech acoustics (Hartley, 2002).