Proceedings of
The XIII\textsuperscript{th} International Congress
of Phonetic Sciences

ICPhS 95

Stockholm, Sweden
13 - 19 August, 1995

Editors
Kjell Elenius & Peter Branderud

Volume 2 of 4

The XIII\textsuperscript{th} International Congress of Phonetic Sciences (ICPhS 95) has been organised by the Department of Speech Communication and Music Acoustics, KTH (Royal Institute of Technology), and the Department of Linguistics, Stockholm University.
INTERGESTURAL TIMING IN SPEECH PRODUCTION:
DATA AND MODELING

Elliot Saltzman
Haskins Laboratories, 270 Crown Street, New Haven, CT USA; Center for the
Ecological Study of Perception and Action, University of Connecticut; Storrs, CT USA

ABSTRACT
The task-dynamic model of speech production is reviewed, and current
developments that incorporate a recurrent network dynamics of inter-
igestural timing are described. Four sets of data are considered that provide a set
of temporal patterning benchmarks for the evaluation of any model of inter-
igestural dynamics. These benchmarks are being used to evaluate and to guide
the development of the extended task-
dynamic model.

INTRODUCTION
This paper examines the nature of intergestural timing in speech production
from both a theoretical and an empirical perspective. The theoretical focus is on
how a model developed previously in our laboratories, the task-dynamic
model, is being extended to incorporate a dynamics of intergestural timing. In the
extended model, a recurrent, connec-
tionist network is used to simulate these
dynamics. This network plays the role of
a utterance-specific central timing
network for speech that is bidirectionally
coupled to a simplified set of speech
articulators. The outputs of the network
serve to parameterize the dynamics of
the articulators, whose ongoing state is
fed back to the network in a manner that
modulates its timing. As this work is in
its preliminary stages (Saltzman, Mitra,
Levy, & Hogden, in progress), dis-

cussion will be limited to the directions
being taken in these developments.

The empirical focus of this paper is a
review of data from four classes of intergestural timing phenomena that
provide a significant set of constraints
that must be satisfied by this or any
model of speech intergestural timing.
These data thus serve to establish a set of
benchmark criteria in the temporal
domain, according to which different
models can be evaluated.

THE TASK-DYNAMIC MODEL
Work on the task-dynamic model of
speech production has been conducted in
collaboration with several of our
colleagues at Haskins Laboratories as
part of an ongoing project focused on the
development of a gesturally-based,
computational model of linguistic
structures (e.g., [1], [2], [3], [4], [5], [6]).
The focus of this model is on the control
schemes that underlie gestural pat-
terning, and represents an attempt to
reconcile the linguistic hypothesis that
speech involves an underlying
sequencing of abstract, context-inde-
dependent units, with the empirical
observation of context-dependent inter-
leaving of articulatory movements.

The central thesis of the model is that
the spatiotemporal patterns of speech are
shaped by a dynamical system with two
functionally distinct but interacting
levels. The interarticulator level is
defined according to both model
articulator (e.g., lips and jaw) and tract-
variable (e.g., lip aperture and protrusion) coordinates; the intergestural
level is defined according to a set of
activation coordinates. Invariant gestural
units are posited in the form of context-

independent sets of dynamical
parameters (e.g., lip protrusion target,
stiffness, and damping coefficients), and
are associated with corresponding
subsets of these coordinates. Each unit’s
activation coordinate reflects the
strength with which the associated
gesture “attempts” to shape vocal tract
movements at any given point in time.
The tract-variable and model articulator
coordinates of each unit specify,
respectively, the particular vocal-tract
constriction (e.g., bilabial) and articu-
latory synergy whose behaviors are
affected directly by the associated unit’s
activation. The formation and release of
such constrictions are governed
according to an overall control regime
defined as a damped, second-order
dynamical system that spans tract-variable and model articulator coordinates, and whose parameters are functions of the current set of gestural activations.

The interarticulator level accounts for the coordination among articulators at a given point in time due to the currently active gesture set. At present, the dynamics of this level are sufficiently well developed to offer promising accounts of spatially goal-directed, multiarticulator movement patterns observed during unperturbed and mechanically perturbed gestures, and during periods of coproduction. In all cases, the evolving configuration of the articulators during a given utterance results from the gesturally- and posturally-specific way that driving influences generated in the tract-variable space are distributed across the associated articulatory synergies.

**Intergestural Timing and Sequential Networks**

The intergestural level accounts for the relative timing of activation intervals for the gestural units participating in a given utterance, e.g., for vocalic and bilabial gestures in a vowel-bilabial-vowel sequence. At present, this relative timing is accomplished with reference to a *gestural score* that specifies the activation of gestural units over time across parallel tract-variable output channels. Currently, gestural scores are not derived from an underlying implicit dynamics. Rather, they are derived explicitly either "by hand" or according to the rules of Brown and Goldstein's *articulatory phonology* ([11], [2], [3]).

We have adopted the recurrent, sequential network architecture of Jordan ([7], [8], [9], [10]; see also [11], [12], [13], [14]) at the intergestural level, as a means of patterning the gestural activation trajectories for the task-dynamic model. In the network, a separate output unit exists for each distinct gesture in a sequence, with the values of the output units corresponding to the activation values of the associated gestures. The network additionally includes a set of *state units* that determine among themselves a dynamical flow with an intrinsic time scale specific to the intended sequence. The dynamics of activity among the state units are defined by weighted recurrent connections: a) among the state units themselves, and b) from the output units to the state units. A set of intermediate *hidden* units are connected to both the state and output units by two respective layers of weighted paths. Temporal ordering among the output units is an implicit consequence of the network architecture and the sequence-specific set of weight values associated with each connection path.

Using *back-propagation* training methods ([e.g., 15]), sequential networks can learn single sets of weights that will allow them to produce several different sequences. The information that is used to define the *teaching vector* sequences applied to the network during training will be obtained from the same articulatory phonology rules that are presently used to generate gestural scores. Such information will include gestural targets stiffness and damping coefficients, and the required times or phases of target attainment. Each simulated sequence is produced (and learned) in the presence of a corresponding constant activation pattern in a set of *plan units*. These units provide a second, external set of inputs to the network's hidden layer, in addition to the internal inputs provided by the state units.

**A simplified pilot network**

Pilot investigation now under way is focused on the behavior of a simplified model system. This system consists of a sequential network whose outputs represent the activations of gestures that are defined in a small set of "tract variables," whose dynamics are first order. The activation nodes act only to insert their associated target values into the tract-variable dynamics; the tract-variable stiffness coefficients are fixed. There are no model articulators at this point, equivalent to the simplifying assumption that tract variables and model articulators are defined in a one-to-one manner. These simplifications were adopted in order to gain familiarity and to sharpen intuitions regarding the types of intergestural timing phenomena that are intrinsic to such coupled dynamical systems, while keeping the
“controlled” articulatory system as simple as possible.

FOUR BENCHMARK CLASSES OF TEMPORAL PHENOMENA IN SPEECH

This section focuses on four empirical classes of phenomena related to intergestural temporal patterns: 1) the temporal extent of anticipatory coarticulation; 2) the durational changes that result when mechanical perturbations are introduced to the articulators during speaking; 3) the effects of linguistic boundaries on the relative timing of gestures associated with the same phoneme; 4) the continuous sliding and aggregation effects, and discontinuous phase transition effects, that accompany increases of speaking rate.

Emphasis is placed on the first two classes of phenomena, since they are being addressed currently in our pilot modeling efforts. The remaining two classes will be reviewed more briefly, since they are not being addressed presently in our model, although they will provide significant constraints in guiding its future development.

Temporal Extents of Anticipatory Coarticulatory Fields

The definition adopted in this paper of a gesture’s anticipatory field of coarticulation is the time from gesture onset to the time of target attainment. In a review of the literature on anticipatory lip rounding (e.g., [17], [18], [19]) and velum lowering (e.g., [20]), and on transconsonantal vowel-to-vowel coarticulation (e.g., [21]), Fowler and Saltzman [22] concluded that anticipatory coarticulation fields are temporally constrained, and that they do not typically extend very far backward in time from the time of target attainment. This interpretation is consistent with that provided by Bell-Berti & Harris’ [23] frame model of coarticulation, and contrasts with the extensive degrees of anticipation that are possible in look-ahead models (e.g., [24]; in fairness, however, it should be noted that the anticipatory feature-spreading used in Henke’s [24] model looked ahead only to the immediately following segment, although unlimited anticipation was allowed in principle).

In the task-dynamic model, this definition corresponds (roughly) to the interval from the time of gestural activation onset to the time of target attainment. Currently, the activation trajectories are specified in the gestural score as simple step functions, whose values change discretely between zero (the gesture is inactive) and one (the gesture is maximally active). Consequently, a gesture’s anticipatory field is simply proportional to the gesture’s intrinsic time constant, which is itself a function of the gesture’s fixed stiffness and damping parameters. The onset of a given activation wave is specified according to the rules of articulatory phonology so that the gesture attains its target at the appropriate point in the simulated utterance (see also [16]).

Anticipatory behavior in the pilot network

In the pilot network, each first-order tract-variable is represented by a linear unit whose inputs are current target value and tract-variable position. The unit’s output is current tract-variable velocity, which is fed into a linear, self-recurrent unit that provides a discrete time, Euler integration to generate the next tract-variable position. Taken together, therefore, the tract-variable and integrator units represent a model of the forward dynamics from current tract-variable state and target inputs to the next tract-variable state. Additionally, a delay line from the integrator unit is used to feed back current tract-variable state to the hidden units of the sequential network (see [8], [10], [25], [26] for related treatments).

Teaching vectors are applied intermittently at the tract-variable integrator units, and output errors are measured. At all other times, “don’t care” conditions exist and no errors are defined. When errors are defined, however, they are propagated backward through the fixed tract-variable forward dynamics, and applied to the sequential net’s output units. From this point on, these backpropagated errors are used in the usual manner to train the weights inside the sequential net. Because of the downstream recurrence implicit in the tract-variable dynamics, and the delayed tract-variable feedback into the sequen-
tial network, the back-propagation through time (e.g., [15]) training procedure is used.

It is known from previous work [7] that sequential nets whose outputs directly represent the distinctive features of speech will display relatively unconstrained temporal fields of anticipatory behavior. Hence, we hypothesize that "frame-model-like" behavior will require the addition of appropriate sets of side constraints, perhaps to the sequential net's output units, during the network's training phase (see [8] for detailed discussion of such constrained optimization methods). Under such training conditions, the anticipatory fields of activation waves should be limited to durations proportional to the tract-variable time constants of the activated gestures.

**Effects of Mechanical Perturbations: Phase Resetting**

Transient mechanical perturbations delivered to the speech articulators during repetitive and discrete speech sequences can alter the underlying trajectory structure of the ongoing sequence, and induce systematic shifts in the timing of subsequent movement elements [27], [28], [29], [30], [31], [32], [33]. We have recently used such methods to examine the sequential dynamics governing bilabial and laryngeal gestures, both within segments and between successive syllables, in repetitive [... p e p a ...] and discrete [p a m p a] utterances ([34], [35], [36]).

Analyses of the repetitive utterances indicated that steady-state shifts occurred for both lips and larynx, but that no steady-state shifts occurred in the relative phasing of these articulators. These effects occurred only when the perturbation was delivered within a "sensitive phase" of the cycle. During this period, the downwardly directed lower lip perturbation opposed the just-initiated, actively controlled bilabial closing gesture for /p/. Thus, the sensitive period corresponded (roughly) to the acceleration portion of the closing gesture (Kawato, personal communication). When perturbations were delivered at the system's sensitive phase, the bilabial and laryngeal gestures were phase-advanced as a relatively coherent unit, maintaining their relative phasing as they were advanced in absolute time. Durational changes observed in other phases were systematic yet transient peripheral responses to the perturbation, and did not indicate central phase-resetting. Finally, the patterns observed in the discrete utterances resembled the transient effects seen in the repetitive utterances, indicating that similar dynamics underlie the production of both types of utterance.

**Phase Resetting: Implications for the Pilot Network**

The fact that steady-state phase shifts exist supports the hypothesis that central intergestural dynamics can be "permanently" reset by peripheral articulatory events. Thus, any hypothesized central timing network for speech cannot not unidirectionally drive the articulatory periphery; rather, central and peripheral dynamics must be coupled bidirectionally, so that feedback information from the biomechanical periphery also can influence the state of the central "clock." In this regard, it is interesting to recall that feedback connections from the interarticulator level (tract-variable state) to the intergestural level (sequential network hidden units) were required in designing the pilot network used to produce unperturbed gestural sequences (see above section Anticipatory behavior in the pilot network).

It is unlikely, however, that this architecture will be sufficient to simulate the phase-resetting results. One reason is that such resetting occurred only during the syllable's sensitive phase, when the lip-opening perturbations opposed the actions of the just-initiated bilabial closing gesture. Additionally, although changes in syllable duration were found for other perturbed phases, these changes were simply transient effects, and did not indicate resetting of the central "clock". If one hypothesizes that efferent commands to the periphery are strongest near gestural initiation, then this pattern of results implies that the timeframe of afferent sensitivity mirrors that of efferent strength. In turn, such an hypothesis suggests a modeling possibility that we will investigate, namely, that "sensory" information regarding a perturbed tract-variable trajectory,
perhaps in the form of a tract-variable error signal (i.e., current target - current tract-variable position), might be multiplicatively gated into the sequential network as a function of the activation state of gestures associated with the tract-variable.

**Effects of Linguistic Boundaries**

Many so-called phonemes are produced as coordinated constellations [1] of gestures, with characteristically different patterns of intergestural phasing depending on the constellation’s location relative to a linguistic (e.g., syllable) boundary [13], [37], [38], [39], [40]). For example, /l/ is produced with two tongue gestures — tongue tip raising and tongue body retraction. In certain dialects of English, word initial /l/-’s ("light" /l/-’s) are produced using roughly synchronous gestures; in word final /l/-’s ("dark" /l/-’s), the gestures are produced asynchronously, with gestural onset for the tongue tip aligned roughly with the time of target attainment for the tongue body. Similary, the nasal consonant /m/ is produced word-initially with roughly synchronous velic lowering and bilabial closing gestures; in word final position, these gestures are asynchronous, with gestural onset for velic lowering aligned roughly with the bilabial closing gesture’s time of target attainment.

**Linguistic boundaries: Constraints on modeling**

It is possible that these timing differences are incorporated explicitly into the phonological rules that, in our modeling scheme, are used to generate the set of tract-variable-target teaching vectors for a given sequence. It is also possible, however, that the boundaries themselves are represented as non-tract-variable elements in the teaching vectors. For example, one possibility is that such boundary elements might serve to dampen the activations (i.e., drive the output units of the sequential network toward zero) of all the gestures in the simulated sequence in proportion to the strength (e.g., [41], [40]) of the associated boundary. Such a boundary element would serve to reduce the magnitudes of tract-variable gestures within the boundary "gesture’s" temporal field. If the anticipatory field is much larger than the carryover field, this could produce pre-boundary gestural reduction similar to that reported in the literature. Whether such boundary elements could also alter the relative timing of nearby gestures is an intriguing possibility that will be explored in future developments of our model.

**Effects of Increased Speaking Rate**

A striking phenomenon that accompanies increases in speaking rate is that the gestures associated with temporally adjacent phonemes tend to "slide" relatively continuously into one another with a resultant increase in temporal overlap (e.g., [42], [3], [43], [44], [45], [46], [47], [48]). For example, Hardcastle [43] has shown with electropalatographic data that the tongue gestures associated with producing the (British English) consonant sequence /kl/ tend to slide into one another with experimentally manipulated increases in speaking rate.

Continuous increases in speaking rate can also produce discontinuous transitions of intergestural phasing ([49], [50], [51]). In these studies, when subjects spoke the syllable /pi/ repetitively at increasing rates, the relative phasing of the bilabial and laryngeal gestures associated with the /p/ did not change from the pattern observed at a self-elected, comfortable rate. However, when the repeated syllable /pi/ was similarly increased in rate, its relative phasing pattern switched relatively abruptly at a critical speed, from that observed for a self-elected, comfortable rate to the pattern observed for the /pi/ sequences.

**Speaking rate: Constraints on modeling**

At this point, we can only speculate as to the dynamical underpinnings of these continuous and discontinuous changes as a function of speaking rate. For example, it is possible that the continuous patterns of intergestural sliding might result from relatively simple changes in the values of a control parameter or parameter set at the intergestural level; e.g., the increased parallelism might be simulated by having rate increases serve to decrease the effective inhibition among the sequential net's output nodes.
The discontinuous, intergestural phase transitions may be viewed as nonequilibrium phase transitions of the type seen in other physical and biological systems (e.g., [52]). In such a framework, the rhythmic units in question are characterized as nonlinearly coupled, limit cycle oscillators; the transition is characterized as a bifurcation from a modal pattern that becomes unstable with increasing rate to another modal pattern that retains its stability. The obvious implications for model development are that when the model system is trained to perform extended repetitive /pl/ or /fl/ sequences, the rhythmically active gestures should be similarly characterizable, and should display corresponding intergestural phase transitions.

ACKNOWLEDGEMENTS
This work was supported by NIH Grant DC-00121 and NSF Grant DBS-9112198 to Haskins Laboratories. I am grateful to Dani Byrd and Philip Rubin for discussion and helpful comments on an earlier draft of this paper.

REFERENCES


