A task-dynamic approach to gestural anticipation in speech: A hybrid recurrent network

Elliot Saltzman* and Suvobrata Mitra

*Department of Physical Therapy, Boston University, 635 Commonwealth Ave., Boston MA 02215; and Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695; 1Department of Psychology, University of Connecticut, Storrs CT 06269

Abstract: Recent simulations of the anticipatory timing of speech gestures within the task-dynamic model of gestural patterning are described. To accomplish this, a sequential neural network architecture is used in a simplified dynamical system in which the network’s outputs define the activations of gestures within a set of vocal tract constrictions (“tract variables”) whose dynamics are second-order. These activations act only to insert target values into the tract-variable dynamics; the stiffness and damping coefficients are fixed. The forward dynamics, from current tract-variable state and target inputs to the next tract-variable state, are then modeled by a second recurrent network whose outputs are fed back to the sequential network. In this manner the simulated articulator timing patterns are sensitive both to the intended linguistic sequence and to the articulator dynamics themselves. Simulations suggest that anticipatory gestural timing is defined along a simply implemented continuum between rigid time-locking and unrestricted lookahead behavior. Temporally elastic anticipatory behavior such as that reported for French lip protrusion falls naturally within this continuum. These results provide a promising account of interspeaker and interlanguage differences in anticipatory gestural timing.

INTRODUCTION

In the task dynamic model of speech production (1), there are two functionally distinct but interacting levels: the interarticulator level, which is defined by both model articulator (e.g., lips and jaw) and tract-variable (e.g., lip aperture and protrusion) coordinates; and the intergestural level, which is defined by a set of activation coordinates. Invariant gestural units are defined as context-independent sets of dynamical parameters (e.g., lip protrusion target and stiffness) and are associated with corresponding subsets of all three coordinates. Each gesture’s activation coordinate reflects the strength with which it “attempts” to shape vocal tract movements at any point in time, while the tract-variable and model articulator coordinates specify the vocal-tract constriction (e.g., bilabial) and articulator synergy, respectively, whose behaviors are directly affected by the gesture’s activation. Constriction formation and release are governed by 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 due to the currently active gesture set. The intergestural level accounts for the relative timing of activation intervals for the gestures in a given utterance. Until recently, this relative timing was controlled solely by a gestural score that was specified explicitly either “by hand” or by the phasing rules of Browman & Goldstein’s Articulator Phonology (2). Here, we report recent work in which the temporal patterning of activation intervals is derived implicitly according to an intrinsic serial dynamics of intergestural-level timing. In particular, we focus on modeling the phenomenon of anticipatory coarticulation.

A DYNAMICS OF ANTICIPATORY GESTURAL BEHAVIOR

When does gestural motion begin relative to its required time of target attainment? This has been a controversial question in speech science whose answer has usually been couched in terms of either lookahead models (3), in which gestures begin as early as possible during a preceding “don’t care” interval where there are no other conflicting demands, or time-locking models (4), in which gestural onsets are time-locked to the time of target attainment regardless of the length of the “don’t care” interval (cf., 5). Recently, data on anticipatory lip-protrusion in French (6) suggested that gestural anticipation is neither rigidly time-locked nor totally unconstrained, but shows a speaker-specific degree of temporal elasticity, i.e., the anticipation interval lengthens with the preceding “don’t care” interval, but only fractionally. We have modeled this constrained temporal elasticity on a continuum that includes look-ahead and time-locking behavior as polar extremes, using a hybrid model whose intergestural dynamics are those of a recurrent sequential network (7) that is bidirectionally coupled to an interarticulator level with simplified task-dynamics (Figure 1). The sequential net’s output nodes represent gestural activations and insert target values into the second-order tract-variable dynamics; stiffness and damping are fixed. Each tract-variable is represented by a linear unit whose inputs are current target value, position, and velocity. It’s output is acceleration which is fed into a cascade of two linear, self-recurrent units that provide discrete time, Euler integrations to generate the next tract-variable velocity and position. Thus, the tract-variable and integrator units model the forward dynamics from current tract-variable state and target inputs to the next tract-variable state. Additionally, delay lines from the integrator units feed back the current tract-variable state to the network’s hidden units. Teaching vectors are used during backpropagation-through-time (8) training to apply targets to the tract-variable position unit for the
time intervals of desired target attainment ("care" intervals), and output errors (target position minus tract-variable position) are measured. At all other times, "don’t care" conditions exist and no errors are defined. During the "care" intervals distal supervised learning procedures (9) are used to propagate the errors backward through the fixed tract-variable forward dynamics to the sequential net’s output units, where the backpropagated errors are used to train the weights inside the sequential net. 

In our model, a gesture’s anticipation interval is defined as the interval, after a short startup time (5 time-ticks), from the time it leaves a criterion baseline (−20% of the distance from initial position to target) to the time of target attainment. A continuum of anticipatory behavior was generated using a corresponding continuum of side constraint, c, values (7) during network training. These constraints were applied to the sequential net’s output (gestural activation) units uniformly during both the "care" (5 ticks) and "don’t-care" (40, 50, or 60 ticks) training intervals, and acted to penalize unnecessary output unit activity. The results (Figure 2) were that pure lookahead behavior was obtained without side constraints (c = 0), pure time-locking occurred with the strongest side constraints (c = .5), and intermediate, temporally elastic behavior resulted from intermediate side constraint values. These results appear to provide a promising account of interspeaker and interlanguage differences in anticipatory gestural timing.

ACKNOWLEDGMENTS

This work was supported by NIH grants DC-03663 and DC-00121

REFERENCES