CASY AND EXTENSIONS TO THE TASK-DYNAMIC MODEL

1026

Rubin, P.[1,2], Saltzman, E.[1,3], Goldstein, L.[1,4], McGowan, R.[1,2], Tiede, M.[5] & Browman, C.[1] ([1]Haskins Lab.; [2]Yale School of Medicine; [3]Univ. of Connecticut; [4]Yale Univ.; [5]ATR HIP Res. Lab.)

Abstract. This paper focuses on selected changes in the computational model of speech production being developed at Haskins Laboratories. The overall structure of the model's several components and the manner in which the components communicate with each other is reviewed. Emphasis is placed on recent developments to the task-dynamic component and to CASY (Configurable Articulatory SYnthesizer), the vocal tract model/ synthesizer component.

Résumé. Cet article porte sur certains changements du modèle computationnel de production de la parole développé aux Laboratoires Haskins. La structure générale des différentes parties du modèle et la manière dont les parties communiquent entre elles est discutée. Nous soulignons les développements récents de la partie *task-dynamics* et de la partie conduit vocal/synthétiseur, appelée CASY (Configurable Articulatory SYnthesizer).

Introduction. This paper is focused on selected developments in the computational model of speech production being used at Haskins Laboratories. There are four major components in the overall model (Fig. 1): a) the *linguistic gestural model* embodies Browman and Goldstein's (1986, 1992) Articulatory Phonology and translates a phonetically specified input utterance into a *gestural score* that specifies the relative timing and the set of dynamic parameters for the speech gestures corresponding to the utterance; b) the *task-dynamic model* (Saltzman & Munhall, 1989) generates coordinated patterns of articulator movements from the gestural score; c) the *articulatory synthesizer* (Rubin, Baer, & Mermelstein, 1981) is based on a midsagittal model of the vocal tract and is used to compute the vocal tract's changing cross-sectional area function, resonance characteristics, and acoustic signal from these articulatory movement patterns; and d) the *recovery procedure* component uses an analysis-by-synthesis procedure incorporating genetic algorithms to recover gestural scores from the acoustic outputs of the overall model (McGowan, 1994).

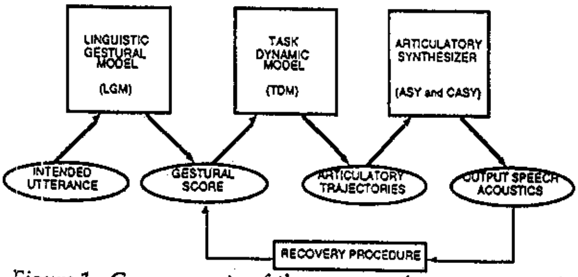


Figure 1. Components of the proposed computational model of speech production.

In what follows, we describe recent developments to the task-dynamic and articulatory synthesizer (CASY; Configurable Articulatory SYnthesizer) components of the overall model.

The task-dynamic model. The central thesis of the task dynamic model of speech production is that the spatiotemporal patterns of speech emerge as behaviors implicit in a dynamical system with two functionally distinct but interacting levels (Saltzman & Munhall, 1989). 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 (Fig. 2); 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 articulatory synergy whose behaviors are affected directly by the associated unit's activation. The formation and release of such constrictions are governed by 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. The intergestural level accounts for the relative timing of activation intervals for the gestural units participating in a given utterance. Until recently, this relative timing has been accomplished with reference to gestural scores that explicitly specify the activation of gestural units over time.

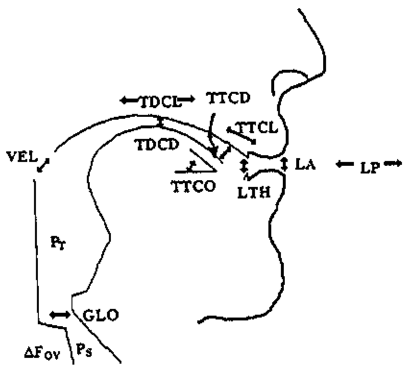
Recent progress on the task-dynamic model has been made in the following areas:

Aerodynamics—The task dynamic model was generalized to include control of aerodynamic flow variables in addition to its previous repertoire of kinematic constriction shape variables (McGowan & Saltzman, 1995; Fig. 2). The new tract variables are: subglottal pressure, P_s, to simulate long-term declination or decay in F_0 over the course of an utterance; transglottal pressure, P_t, to allow continuation of voicing during oral closure; and change in virtual fundamental frequency, ΔF_{0V}, to simulate stress and intonational contours. The associated set of new articulator variables are: total lung force, upper vocal tract volume, and laryngeal tension.

Gestural inventory—An additional kinematic tract-variable, tongue-tip constriction orientation, was defined to supplement our previous inventory of tongue-tip gestures (Fig. 2). The new tract-variable is defined by the orientation of the tongue surface just behind an alveolar constriction, and uses the same articulatory synergy as is used for the tract-variables of tongue-tip constriction location and degree.

Intergestural dynamics—An important step in modeling the relative timing of gestural activations is to derive these temporal patterns as implicit consequences of an intrinsic *serial dynamics*, and not as extrinsically imposed, explicit gestural scores. We are incorporating the dynamics of recurrent, connectionist networks at the intergestural level as a means of patterning gestural activation trajectories.

Specifically, we have adopted the sequential network architecture of Jordan (Jordan, 1992; see also Bailly, Laboissiere, & Schwartz, 1991) in current simulations focusing on the behavior of a simplified model system for producing unperturbed gestural sequences (Fig. 3). This system consists of a sequential network whose outputs represent the activations of gestures defined in a small set of "tract variables" with first-order dynamics. These activation nodes act only to insert target values into the tract-variable dynamics; the tract-variable stiffness coefficients are fixed. 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 the current tract-variable state to the hidden units of the sequential network. During training, teaching vector targets are applied at the tract-variable integrator units for those time intervals of desired target attainment ("care" intervals), 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 back-propagated errors are used 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



Tract variables		Model articulators
LP	lip protrusion	upper & lower lips
LA	lip aperture	upper & lower lips, jaw
TDCL	tongue dorsum constrict location	tongue body, jaw
TDCD	tongue dorsum constrict degree	tongue body, jaw
LTH	lower tooth height	jaw
TTCL	tongue tip constrict location	tongue tip, body, jaw
TTCD	tongue tip constrict degree	tongue tip, body, jaw
TTCO	tongue tip constrict orientation	tongue tip, body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottal width
P_S	subglottal pressure	total lung force
P_T	transglottal pressure	supralaryngeal vocal tract volume
ΔF_{OV}	delta virtual fundamental frequency	vocal fold tension, total lung force glottal width

Figure 2. Left: Schematic midsagittal vocal tract outline, with tract-variable degrees of freedom indicated by arrows; Right: Table listing the model's tract variables and associated model articulators.

sequential network, the *back-propagation through time* training procedure is used.

Our initial efforts focused on modeling a given gesture's *anticipatory field of coarticulation*, defined operationally as the time from motion onset to the time of required target attainment. In a review of the literature, Fowler and Saltzman (1993) 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 (cf., Abry & Lallouache, 1995). This interpretation is consistent with that provided by Bell-Berti and Harris' (1981) *frame* model of coarticulation, and contrasts with the extensive degrees of anticipation that are possible in *look-ahead* models (e.g., Henke, 1966).

In our model this definition of anticipatory field corresponds to the interval from the time of gestural activation onset to the time of required target attainment. Using this model, we have generated both unconstrained look-ahead and appropriately constrained, frame-like fields of anticipatory coarticulation. Jordan (1986) showed that sequential networks whose outputs directly represent the distinctive features of speech displayed relatively unconstrained temporal fields of anticipatory behavior. Hence, we tested the hypothesis in our pilot network that "frame-model-like" behavior requires the addition of *side constraints* to the sequential net's output units, uniformly applied during both the "care" and "don't-care" intervals of the network's training phase. Using side constraints during training that penalized activity of the sequential network's output units (i.e., gestural activation units), the anticipatory fields were constrained temporally in a manner consistent with the frame model (see Fig. 4).

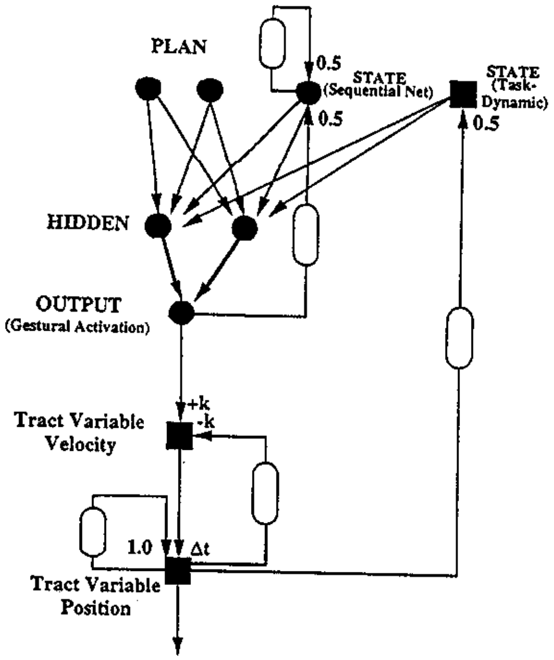


Figure 3. Architecture of the pilot hybrid network. Filled circles = sequential network elements; filled squares = task-dynamic elements; open "lozenges" label delay lines; arrowheads = inter-element synapses; numbers and symbols denote the fixed weights assigned to some synapses.

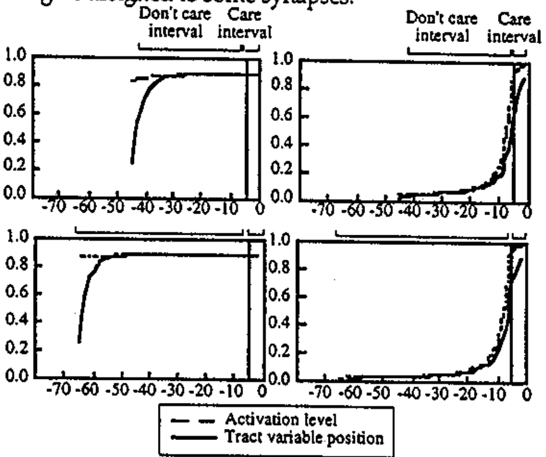


Figure 4. Anticipatory behavior of the hybrid network. Left column: look-ahead behavior occurs when side constraints are absent, and gestural onset occurs near the beginning of don't-care interval regardless of its length; Right column: frame model behavior occurs when side constraints are present, regardless of the don't-care interval's length, and gestural onsets are approximately time-locked to the care interval.

Configurable Articulatory Synthesizer (CASY). Over the past several years, we have been designing and implementing a new approach to articulatory synthesis at Haskins Laboratories that will move us beyond the restrictions inherent in the Mermelstein model (Mermelstein, 1973) that underlies our present articulatory synthesizer, *ASY* (Rubin, Baer & Mermelstein, 1981). The major goals of the endeavor include the following: design and incorporate more realistic source and aerodynamic simulations; provide for greater flexibility in the specification of midsagittal control parameters; remove the fixed articulatory dependencies of the Mermelstein model; provide for better tongue surface fitting; improve the sagittal distance to area calculations; provide for a better representation of 3D tract shape.

A preliminary version of this new Configurable Articulatory Synthesizer (CASY) has been developed by Goldstein, Tiede, and Rubin, with assistance from Browman and McGowan. CASY parameters are a superset of those in ASY and include values that were in Mermelstein's (1973) original model, but which could not be adjusted by the user. In addition, the fixed surfaces of the vocal tract are represented parametrically (2nd-order Bezier curves, with a user-specified number of Bezier segments), so that they can be adjusted to match any arbitrary speaker's vocal tract by aligning the model's representation with previously acquired MR images. CASY embodies several other improvements over the Mermelstein model, including the following: surfaces are defined as shapes, independent of any set of gridlines that is superimposed on them; the outline of the soft palate and uvula is improved by the addition of an outline parameter; the relationship between the soft palate position and nasal coupling is made easier to control by the addition of parameters; there is a new model for the tongue tip; model parameters can be adjusted and re-defined, to fit different speaker's heads, and different tract configurations; a new algorithm for computing the area function from a sagittal outline is used; finally, the new algorithm provides a display of how area is computed at each tract section, letting the user flexibly change the parameter values of the equations that relate sagittal distance to area in various parts of the tract.

The CASY system has been expanded to include dynamic display of pellet data tracked by x-ray microbeam (XRMB) or electromagnetic midsagittal articulometer (EMMA) techniques. The horizontal and vertical positions of the pellets can be displayed as time functions; a point in time can then be graphically selected, and the locations of the pellets at that time can be displayed in the midsagittal plane, superimposed on the model vocal tract, and/or any currently displayed image. This capability allows us to begin to compare dynamic pellet data with movements of the model vocal tract, which will be particularly useful for our "designated talker," for whom we have collected XRMB, EMMA, and MRI data. Future work on CASY will focus on the following areas: specifying the output of the task-dynamic model in a form that is appropriate for providing input to CASY; establishing a procedure for choosing among specific parameter sets that are associated with known speakers, so that the appropriate CASY outline can be selected; incorporating into CASY the improvements in transfer function and sound source computation that have already been developed to supercede the original ASY model; and upgrading CASY to an *X-Windows/Motif* graphical standard.

References.

Abry, C., & Lallouache, T. (1995). Modeling lip constriction anticipatory behavior for rounding in French with the MEM (Movement Expansion Model). In K. Elenius & P. Branderud, (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences.* (Pp. 152-155). Stockholm: KTH and Stockholm University

Bailly, G., Laboissière, R., & Schwartz, J. (1991). Formant trajectories as audible gestures: An alternative for speech synthesis. *Journal of Phonetics, 19,* 9-23.

Bell-Berti, F., & Harris, K. (1981). A temporal model of speech production. *Phonetica, 38,* 9-20.

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

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

Fowler, C., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36,* 171-195.

Henke, W. (1966). *Dynamic articulatory models of speech production using computer simulation.* Unpublished Ph.D. diss., MIT, Cambridge, MA.

Jordan, M. (1986). *Serial order in behavior: A parallel distributed processing approach* (Technical Report No. 8604). San Diego: University of California, Institute for Cognitive Science.

Jordan, M. (1992). Constrained supervised learning. *Jounal of Mathematical Psychology, 36,* 396-425.

McGowan, R. (1994). Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication, 14,* 19-48.

McGowan, R. & Saltzman, E. (1995). Incorporating aerodynamic and laryngeal components into task dynamics. *Journal of Phonetics, 23,* 255-269.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53,* 1070-1082.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 321-328.

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1,* 333-382.

Acknowledgements. Work described in this paper was supported by NIH Grants HD-01994, DC-00121, DC-01247, and NSF Grant DBS-9112198.