Neural Network Modeling of Speech Motor Control Using Physiological Data*

Eric Vatikiotis-Bateson, Makoto Hirayama, and Mitsuo Kawato
ATR Auditory and Visual Perception Research Laboratories
Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Introduction

In this paper, we focus on the dynamical relation between observed elements of neuromuscular activity, indicative of neuromotor commands, and the ensuing articulatory movement behavior during reiterant speech production. Although progress has been made modeling the articulatory kinematics from which inferences can be made about the underlying dynamics of the musculoskeletal system (for review, Saltzman & Munhall, 1989), the mapping between actual neuromuscular and articulatory events - in which the system's dynamical properties could be directly assessed - still remains a mystery. Ultimately, we hope this computational approach to speech production can be used to model speech perception and recognition, and to elucidate structural and functional aspects of the central nervous system's orchestration of motor behavior.

Specifically, we use a neural network scheme developed previously to characterize the motor control of human arm movement (Kawato *et al.*, 1990). The task of the network is to learn the forward dynamics linking neuromotor commands and their articulatory consequences. Once learned, the forward dynamics model (FDM) is used both to generate appropriate phoneme- and context-specific movement trajectories and to quantify viscoelastic properties such as damping and stiffness. As far as we know, this is the first time such modeling has been based on real, rather than simulated, physiological data (cf., Flash & Hogan, 1985; Jordan, 1990; Laboissière *et al.*, 1990; Ostry *et al.*, in press; Uno *et al.*, 1989).

This study is a first step in implementing the scheme for speech production shown in Figure 1. We would like to assume that inputs to the model are phoneme strings where each phoneme has an associated target or 'via point' specified spatially in the relevant task domain(s) - e.g., lip aperture for bilabials, or lip aperture, protrusion, and the location and degree of tongue body constriction for /o/ (Saltzman, 1986). Smoothness, an observed attribute of biological movement behavior, undoubtedly arises from active control at a variety of neural and physiological levels as well as from the passive biomechanics. As has been done for modeling arm motion (Kawato *et al.*, 1991), we plan to subsume smoothness under one hard-wired constraint whose influence is constrained by global performance criteria such as speaking rate and style. Thus, via point and smoothness constraints are functionally opposed, since the relative weight of the smoothness constraint determines the degree to which movement trajectories approach phoneme-specific via points.[1]

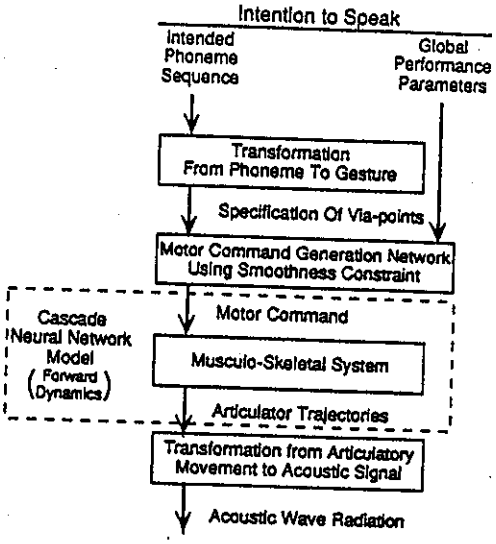

Fig. 1. Forward model of speech production. The dashed-line box indicates the scope of this study.

* We thank Vincent Gracco and Kiyoshi Oshima for nearly painless muscle insertions; Haskins Laboratories for use of their facilities (under the auspices of NIH grant DC-00121); and to Kiyoshi Honda, Michael Jordan, Philip Rubin, Elliot Saltzman and Yoh'ichi Toh'kura for insightful discussion.

[1] For the preliminary modeling presented in this paper, via-points are not specified and the need for a smoothness constraint is eliminated by using the EMG signals themselves to drive the CNN (see Fig. 5).

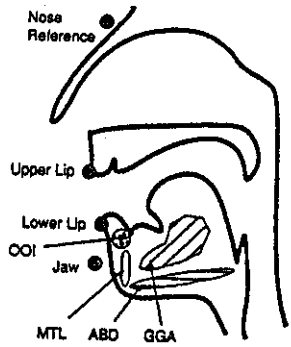

Fig. 2. Approximate positions of infrared markers and muscle insertions for recording movement and EMG.

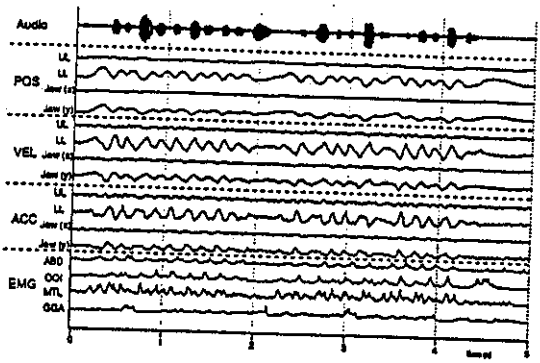

Fig. 3. Time series representations for all channels of one reiterant rendition of "When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow", using *ba*.

The Experiment

Movement, EMG, and acoustic data were recorded for one speaker who produced reiterant verisons of two sentences from the Rainbow Passage (as in Kelso *et al.*, 1985). Speaking rate was fast and the reiterant syllables were *ba, bo*. Each sentence-syllable condition was produced 10 times in a randomized order. Figure 2 shows approximate marker positions for tracking (Haskins Huntspot) position of the jaw (horizontal and vertical) and lips (vertical only) and muscle insertion points for hooked-wire, bipolar EMG recording from four muscles: ABD (anterior belly of the digastric) for jaw lowering, OOI (orbicularis oris inferior) and MTL (mentalis) for lower lip raising and protrusion, and GGA (genioglossus anterior) for tongue tip lowering.

All movement and EMG (rectified and integrated) signals were digitized (12 bit) at 200 Hz and then numerically smoothed at 40 Hz. Position signals were differentiated to obtain velocity and then, after smoothing at 22 Hz, differentiated again to get acceleration. (for processing details, see Kay *et al.*, 1985).[2] Results are shown in Figure 3 for one reiterant utterance using *ba*.

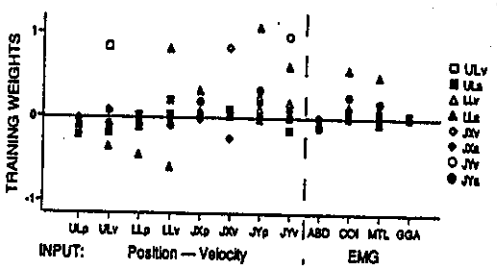

Fig. 4. Learned weights correlating each input value (position, velocity, and EMG) and the change in position (velocity) and velocity (acceleration) from sample *t* to t+1.

Network Training

The forward dynamics model for *ba, bo* production was obtained using a three-layer perceptron (a component of the cascade neural network discussed below) with back propagation. Briefly, the network learns the correlations between the values for all signals at time *t* (e.g., upper lip position or OOI) and the changes of position and velocity for all articulators at the next time sample t+1. The first four tokens of the four conditions using *ba, bo* were used for network training. The learned correlations, expressed as weights, are shown in Figure 4. There are two points of particular interest: First, the positive weightings near +1 for velocity (unfilled symbols for horizontal and vertical jaw, upper and lower lip) indicate that the velocity was nearly constant from one sample to the next. Second, the articulator weightings for the four muscle inputs are sensible: lower lip (which includes vertical jaw) and jaw acceleration are positively weighted for lip raising muscles, OOI and MTL, but negatively (albeit slightly) weighted for the jaw lowering muscle, ABD; and uncorrelated entirely for the tongue muscle, GGA.

[2] Originally, velocity signals were not smoothed prior to obtaining acceleration. Vincent Gracco suggested that the high frequency noise induced by multiple derivatives might un-necessarily hinder network learning, since the process of converting neuromuscular activity to articulator motion acts as a kind of low pass filter. Although secondary smoothing of the training data had no "visible" effect on the very smooth trajectories estimated by the network, error signals between estimated and desired output were lower.

Trajectory Prediction from Muscle EMG

After learning, the acquired forward dynamics model (FDM) was tested on data not included in the training set. First, we used the perceptron to estimate changes of articulator position and velocity from the EMG input over a single timestep. Predicted velocity (Δposition) values matched desired values almost perfectly. Acceleration (Δvelocity) values were consistently underestimated, resulting in a slightly smoother-than-desired output. More important, however, was to test the FDM's ability, when used in conjunction with a cascade neural network (CNN), to generate continuous articulator trajectories over a large number (n=1000) of time steps. A partial representation of this network is shown for the first three sample steps in Figure 5.

The idea of the CNN is that the kinematic consequences of trajectory estimation are chained together from the first sample onward. Only the articulator position and velocity values for the very first sample of the test utterance are given to the FDM. Using the EMG "motor command" input for each sample step, the FDM estimates the changes of position and velocity and sums them with the position and velocity values of the previous sample to obtain estimated values at the next time sample. Thus, error in trajectory estimation is cumulative from the first sample onward.

Figure 6 shows a comparison of actual vs predicted trajectories. In general, their spatiotemporal characteristics are very similar, e.g., amplitude, frequency, and phase, indicative of the generally good performance of the model. There is, however, a tendency towards negative DC offset in the predicted position signals. For upper lip position, the offset persists from the third peak (bilabial closure, counting from the left). For lip and jaw, offset occurs primarily around phrase breaks, although the trajectories tend to converge again by the third movement cycle after a pause. Noting that offsets occur at specific extrema (peaks and valleys), where the lower lip-jaw motion of the test utterance is particularly large, and that the network learned the functional coupling between the lips and jaw for bilabial closure (see below), there are two important limitations that reduce the current model's ability to compensate for position shifts in the test utterance. First, there is no specified equilibrium or rest position, either in articulator or task space, towards which articulators might tend in the absence of EMG activity. Second, the acquired FDM is based on very limited EMG; at most, there is correlated EMG for only one direction of motion per articulator. Addition of antagonist EMG and/or an estimate of equilibrium position should greatly increase the model's generalization capability.

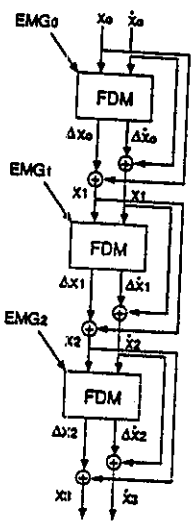

Fig. 5. Cascade Neural Network (CNN) for generating estimated movement trajectories from EMG. Input is shown for one muscle and articulator.

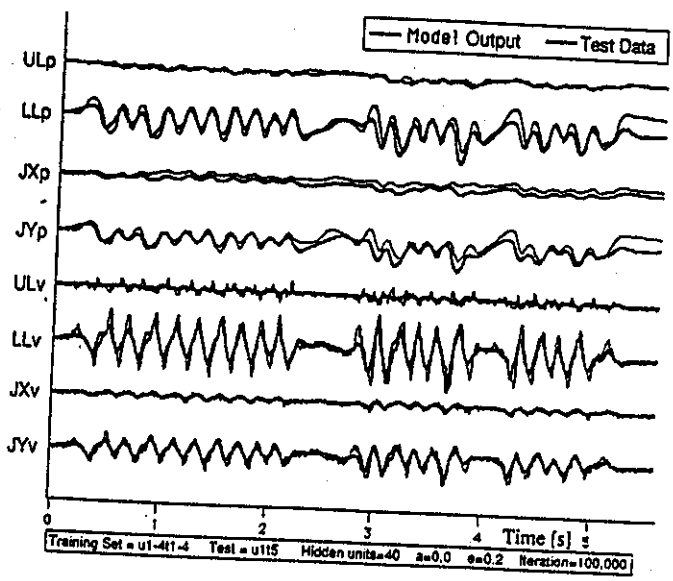

Fig. 6. Comparison of test data and estimated trajectories generated by the CNN.

Estimation of Dynamic Parameters

It is clear from the preceding that the neural network learned to generate good movement trajectories from EMG input; but did it learn the forward dynamics? To show that it did, the model system's responses to two types of simulated perturbation were observed (for details, see Hirayama *et al.*, in press). The first, shown in Figure 7, confirmed that the model indeed learned an appropriate nonlinear dynamics and affords rough estimation of the system's viscoelastic properties, such as natural frequency (1.0-1.3 Hz) and damping (0.24). Simulated release of the lower lip at various distances from rest revealed underdamped, though stable behavior (as shown) within a certain range of values and unstable behavior outside that range. The simulated perturbation also revealed the learned functional coupling among articulators as shown by the in-phase lower lip/jaw trajectories and the rapid 180 degree phase shift of the upper lip appropriate to bilabial production.

The second simulated perturbation entailed observing articulator response to a step increase (50% of full-scale) in EMG activity for each muscle. The results shown in Figure 8 demonstrate that the learned relation between EMG input and articulator movement output is dynamical rather than mechanical. Furthermore, the direction of movement response is appropriate for ABD (jaw lowering) and OOI (lower lip raising). However, responses to increased MTL (lip raising and protrusion) and GGA (tongue lowering) are less

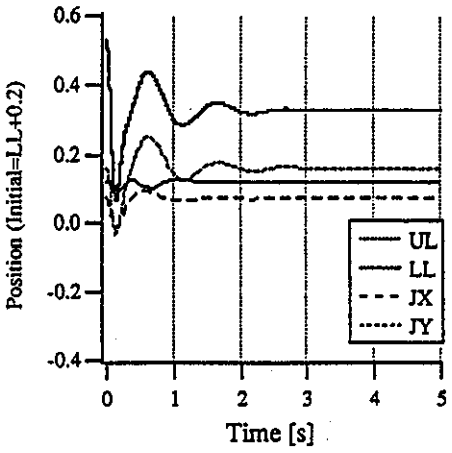

Fig. 7. Response of model system to simulated perturbation of lower lip from estimated rest values of position and EMG (continuously delivered for 5 sec.).

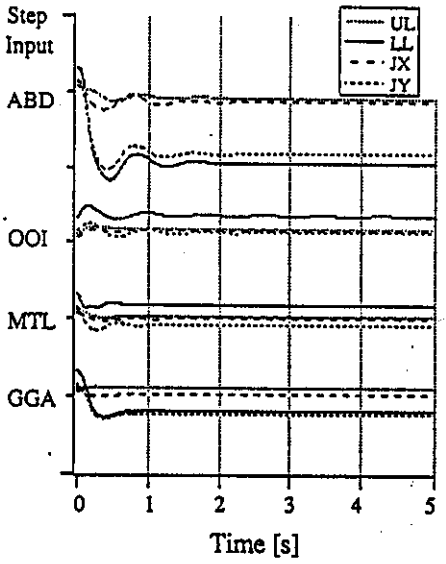

Fig. 8. Response of model system to simulated step increase in EMG from estimated rest values of position and EMG.

straightforward. Step increase of MTL activity has essentially no effect on articulator motion, yet fairly strong weightings were learned for lower lip and jaw acceleration (Fig. 4). A possible explanation is that, compared to this muscle's consistently high level of activity (Fig. 3), the step increase used in the simulation is too small to generate a response. A more likely possibility stems from the observation (due to Kiyoshi Honda) that the MTL signal contains two bursts per movement cycle and, thus, may include lip depressor activity associated with lip-jaw lowering. Closer inspection shows the two bursts are highly correlated with the onsets of both the raising and lowering phases of the lower lip (and jaw) movement cycle. This would account for the strong positive weightings between MTL input and lip/jaw acceleration, and would result in no motion during a continuous step response test. Finally, GGA activity was observed only during large opening movements associated with heavily stressed and phrase-final gestures. The absence of learned weights from only several bursts per utterance is therefore not surprising. However, that the network appears to have learned the functional coupling between lip/ jaw motion and tongue EMG is due to the use of reiterant speech. The repetition of the same consonant-vowel sequence results in

predictable, but misleading couplings between jaw and tongue - e.g., it is not the case that GGA causes jaw lowering.

Summary Discussion

In this paper, we outlined a computational approach to motor control in which a neural network was used to model the nonlinear dynamics and functional constraints causally linking muscular activity and ensuing articulator motion. Simulated perturbations showed that the model has research potential beyond its ability to synthesize reasonable articulatory trajectories from muscular input; since the dynamic parameters of the acquired model can be quantified, they can be tested empirically against our knowledge of this and other motor control systems. Use of real data at both levels of observation does not insure that the acquired dynamical model is biologically real, but it does make prior knowledge of the system being studied unnecessary. Finally, while useful and encouraging as a first test, reiterant speech is far too limited with its phonemic simplicity and the unnaturally cyclical motion of a small set of primary articulators and narrowly-defined interarticulator couplings (Vatikiotis-Bateson, 1988). The validity and practicality of the approach must be demonstrated using real speech. This will require data from many more muscles and articulators whose motions will not necessarily be rhythmically cyclic or easy to specify in terms of phoneme-specific tasks.

References

Flash, T. & Hogan, N. (1985). The coordination of arm movements: An experimentally confirmed mathematical model, *Jour. of Neuroscience*, 5, 1688-1703.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. (in press). Forward dynamics modeling of speech motor control using physiological data. In: R.P. Lippmann, J.E. Moody, & D.S. Touretzky (Eds), *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann Publishers.

Jordan, M.I. (1990). Motor learning and the degrees of freedom problem, In: M. Jeannerod (ed.) *Attention and Performance XIII*, 796-836. Hillsdale, NJ: Erlbaum.

Kawato, M., Maeda, Y., Uno, Y., & Suzuki, R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion, *Biol. Cybern.*, 62, 275-288.

Kawato, M. (1991). Optimization and learning in neural networks for formation and control of coordinated movement. In: D. Meyer (Ed.), *Attention and Performance, XIV*. Hillsdale, NJ: Lawrence Erlbaum.

Kay, B., Munhall, K., Vatikiotis-Bateson, E., & Kelso, J.A.S. (1985). A note on processing kinematic data: Sampling, filtering, and differentiation. *Haskins Laboratories Status Report on Speech Research*, SR-81, 291-303.

Kelso, J.A.S., Vatikiotis-Bateson, E., Saltzman, E., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Jour. of the Acous. Soc. of Am.*, 77, 266-280.

Laboissière, R., Schwartz, J.L., & Bailly, G. (1990). Motor control for speech skills: A connectionist approach, In: D.S. Touretzky, J.L. Elman, T.J. Sejnowski, & G.E. Hinton (eds), *Connectionist Models: Proceedings of the 1990 Summer School*, 319-327. San Mateo, CA: Morgan Kaufmann Publishers.

Ostry, D.J., Flanagan, J.R., Feldman, A.G., & Munhall, K.G. (in press). Human jaw motion control in mastication and speech. In: J. Requin & G.E. Stelmach (Eds), *Tutorials in Motor Neuroscience*. Kluwer Academic Publishers.

Saltzman, E.L. (1986). Task dynamics coordination of the speech articulators: A preliminary model, *Exp. Brain Res.*, Series 15, 129-144.

Saltzman, E.L. & Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4), 333-382.

Uno Y., Kawato M., & Suzuki R. (1989). Formation and control of optimal trajectory in human multijoint arm movement, *Biol. Cybern.*, 61, 89-101.

Vatikiotis-Bateson, E. (1988). *Linguistic Structure and Articulatory Dynamics*. Bloomington, IN: Indiana University Linguistics Club.