Generating Articulator Motion from Muscle Activity Using Artificial Neural Networks

Eric Vatikiotis-Bateson, Makoto Hirayama, Yasuhiro Wada
and Mitsuo Kawato

ATR Human Information Processing Research Laboratories,
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Introduction

During the past two years, we have modeled speech production computationally using artificial neural networks that learn associations between physiological, kinematic, and acoustic data. Our aim has been to identify and emulate various stages of the production process as plausibly yet simplistically as possible. As schematized in Fig. 1, linguistic intentions are realized neurophysiologically by combining phoneme-specific motor command sequences and global factors conditioning overall performance (e.g., speaking rate, style, speaker mood), which together determine what muscles are activated and to what extent (Dornay, Uno, Kawato, & Suzuki, 1992). They also affect articulator behavior through adjustment of dynamical parameters of the musculoskeletal system. The ensuing articulator kinematics effect changes in vocal tract shape and consequently the acoustic output when accompanied by appropriately coupled periodic and non periodic sources.

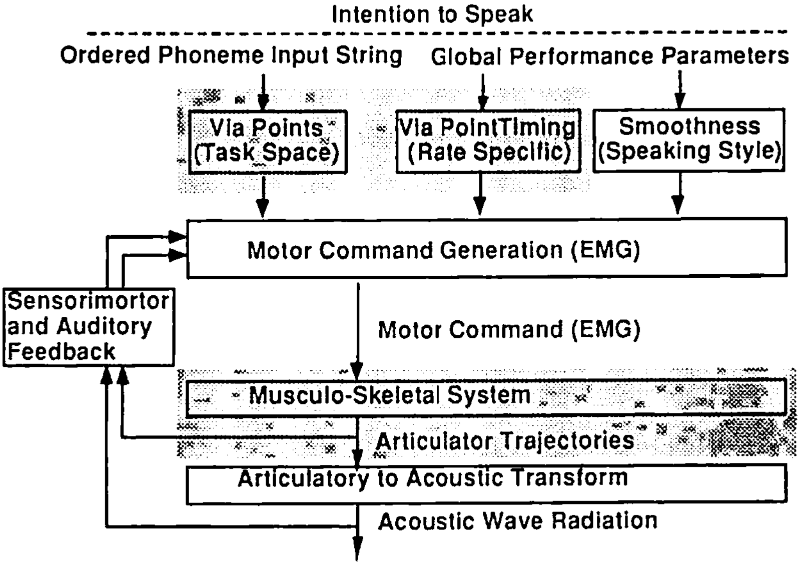
Intention to Speak
Ordered Phoneme Input String
Global Performance Parameters
Via Points (Task Space)
Via PointTiming (Rate Specific)
Smoothness (Speaking Style)
Motor Command Generation (EMG)
Sensorimortor and Auditory Feedback
Motor Command (EMG)
Musculo-Skeletal System
Articulator Trajectories
Articulatory to Acoustic Transform
Acoustic Wave Radiation

Fig. 1 Overview of Speech Production. Shading denotes topics of this paper.

Initially, the basic formal and conceptual aspects of the model were adapted from those applied earlier to the motor control of discrete arm movements ((Kawato, Maeda, Uno, & Suzuki, 1990; Uno, Kawato, & Suzuki, 1989), and were successfully extended to the generation of reiterant speech sentences (Vatikiotis-Bateson, 1988) composed of *ba* or *bo* sequences (Hirayama, Vatikiotis-Bateson, Kawato, & Jordan, 1992b). Using kinematic data (horizontal and vertical position) from the lips and jaw, physiological data (EMG) from several relevant muscles, and the speech acoustics, artificial neural networks learned aspects of the dynamical mapping between EMG activity and articulator motion (Forward Dynamics) and of the mapping between articulator motion and the acoustics (Forward Acoustics). After training, the acquired forward model of the musculoskeletal system was incorporated into a cascade network that generated plausible EMG signals from the sequence of phoneme-specific targets (via points corresponding to /b/ and /a/ or /o/) specified in task space (lip aperture), as shown in Fig. 2. This network employed a constant smoothness constraint on muscle activity, whose setting can be adjusted for differences in speaking rate and style. Via point specification and setting of the smoothness constraint interact in such a way that estimated (model) trajectories corresponding to fast casual speech will be more smooth and undershoot via point targets more than those corresponding to slower more precise productions. The EMG signals generated by the cascade network then drove the forward dynamics model recurrently connected to produce continuous estimates of articulator position. Finally, articulator positions were input to the Forward Acoustics network to generate PARCOR parameters for speech synthesis. (Hirayama, Vatikiotis-Bateson, Kawato, & Honda, 1992a).

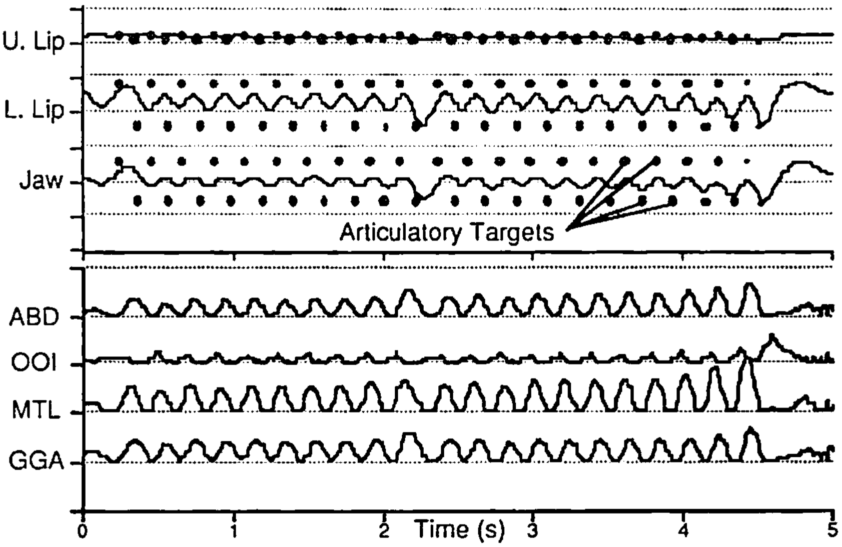

Fig. 2. Via point assignment (top) and EMG (bottom) generated by the Cascade neural network for a reiterant speech production of *When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow*, using *ba*.

Recently, we have applied this modeling scheme to more natural utterances. Compared to discrete arm movements or reiterant speech, the greater complexity of real speech has proved to be technically challenging and has led to modifications and extensions of the model at various levels of the scheme shown in Fig. 1. In this paper, we discuss two areas that are particularly interesting because, while crucial to our specific effort to model natural utterances, they address issues of general interest. The first concerns changes in the way the forward mapping between EMG and articulator motion is obtained. We are now using separate networks, whose architecture is being extended to include boundary constraints such as the shape of the hard palate, to model different articulator components and functional groupings; specifically, the jaw, the jaw+lower lip, and the jaw+tongue. The second development is automated via point estimation, which makes it possible to assign phoneme-specific articulatory targets to complex articulator configurations in a principled way.

Acquiring The Forward Dynamics

Compared to reiterant speech, real speech entails activity of many more articulators and muscles, particularly those of the tongue. The increased complexity in articulatory patterning and added data channels makes the computational task substantially more difficult. For example, in a recent experiment using the magnetometer at Haskins Laboratories (Perkell, Cohen, Svirsky, Matthies, Garabieta, & Jackson, 1992), we recorded nine muscle EMG channels, horizontal and vertical position channels for the lips, jaw, tongue tip and tongue blade during production of fairly natural sentences. In modeling the data, we first tried to use all EMG, 2D position, and corresponding velocity signals for each articulator dimension as input vectors to the training network (multilayer perceptron). The network task was to compute a unique mapping between the 39 inputs and 10 acceleration outputs for each time sample of the 8 second training utterances (1600 samples per utterance).

In theory, even such a large network can be trained successfully, provided the data contain sufficient information to compute unique solutions between muscle EMG and articulator acceleration. In practice, however, performance of the fully-connected network was poor in three respects: slow training time, coincidental correlations confusing structure and function (described below), and failure to converge on a solution due to insufficient data for the tongue. The latter two reasons in particular led us to split the network training task into sub networks for the jaw alone, the jaw+lips, and the jaw+tongue.

At present, the sub networks are independent and are not incorporated into a modular architecture of articulator-specific expert systems whose activity is coordinated by a gating network (Jacobs, Jordan, Nowlan, & Hinton, 1991). Therefore, the apparent redundancy of the jaw in each sub network is necessary. Since the analyzed motion of the tongue and the lips includes the jaw component, the desired mapping between muscle EMG and tongue or lip acceleration must include the kinematic and physiological contributions of the jaw.

Obviously, size and complexity of these sub networks is much smaller than the fully connected network, so training time is reduced substantially. For example, the smallest network for the jaw alone has as input only an antagonist muscle pair — ABD and MPT (anterior belly of the digastric for opening, medial pterygoid for closing) — and position and velocity for horizontal and vertical motion. Thus, there are only six input vectors for acquiring the sample-step mapping between muscle EMG and two outputs for horizontal and vertical jaw acceleration. Using continuous EMG input and the forward dynamics acquired from this network, jaw position was estimated recurrently as schematized in Fig. 3 for the sentence, *Sam sat on top of the potato cooker and Tommy cut up a bag of tiny potatoes and popped the beet tips into the pot.* EMG input to the recurrent network and the match up between experimental and estimated jaw position are shown in Fig. 4. Smaller network size is useful not only because it simplifies and speeds up the training process, but it also makes it easier to see the results of the inevitable adjustments in parameter values needed to improve the results.

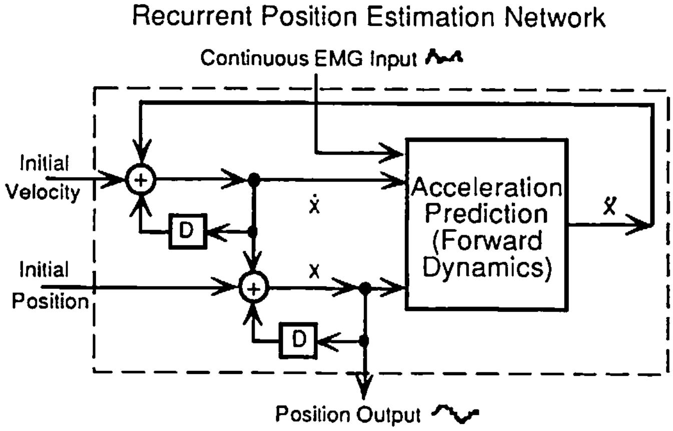

Fig. 3 Recurrent position estimation network. The network estimates articulator position from initial values for position and velocity and continuous EMG. For each EMG sample, the acquired model of the forward dynamics calculates an acceleration value which is then summed and integrated with the preceding values of position and velocity.

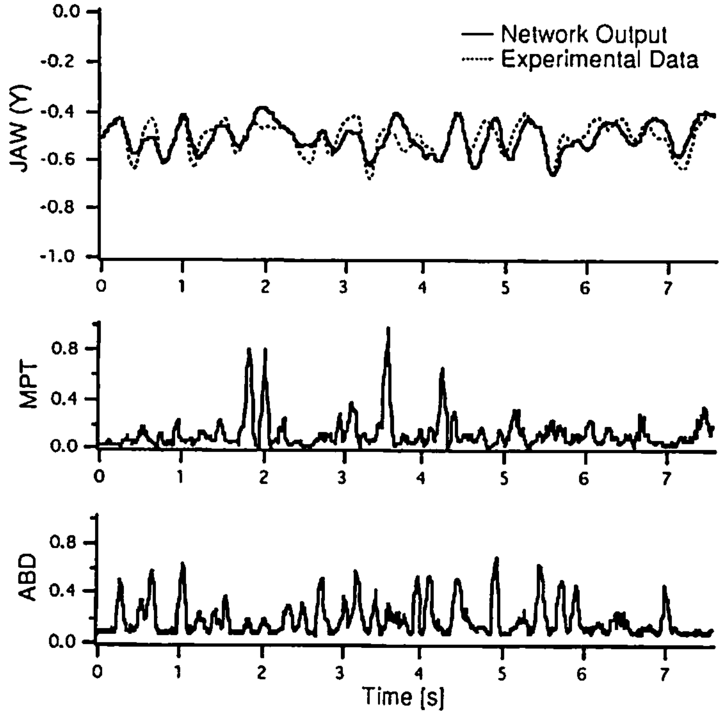

Fig. 4 Jaw position estimated by the recurrent network is compared with the original data. EMG signals from ABD (opening) and MPT (closing) were used to train the network.

In addition to faster training, use of specific sub networks reduces the confusion of coincidental functional aspects of articulatory patterning and structural components of the musculoskeletal dynamics. This problem first appeared in modeling reiterant speech whose reduced phonetic variability, inherent cyclicity and simplified interarticulator coordination made discrimination between dynamical and functional aspects of the musculoskeletal system quite difficult (Vatikiotis-Bateson, Hirayama, & Kawato, 1991). Network training of the forward dynamics for reiterant speech data associated activity of GGA (genioglossus anterior) muscle, active during tongue lowering, with jaw opening (see Fig. 2). While tongue lowering may co-occur with jaw opening during production of /ba/ sequences, GGA does not 'cause' the jaw to open. At best, tongue and jaw are functionally coupled during the vowel portion of these productions. Use of sub networks that isolate tongue-jaw from lip-jaw articulations cannot prevent such "errors" in mapping from occurring, unless the perceptron is explicitly blocked from learning specific associations between correlated physiological and kinematic events; but their incidence can be greatly reduced. Also, comparison of common components such as the jaw across sub networks allows functional and structural components to be distinguished empirically.

A second subnet was used for acquiring the mapping between EMG and acceleration for the lips. Since lip motion is coupled to jaw motion, jaw EMG was included in addition to the muscles related to lower lip raising (OOI — orbicularis inferior), protrusion (MTL — mentalis), and lowering (DLI — depressor labii inferior). Since no upper lip EMG, such as OOS (orbicularis oris superior), was available for network training, horizontal and vertical position of the upper lip was treated as an implicit boundary constraint.

Both the jaw and the lip-jaw sub networks gave better estimation results for jaw position than did the fully-connected network (in which jaw position estimation failed entirely), demonstrating that inclusion of the tongue data in the fully-connected network degraded overall performance, not just performance specific to the tongue. The problem of the tongue is not trivial and not one we are likely to overcome soon. While it is the most important speech articulator, ultimately responsible for shaping most of the vocal tract, it is also the most difficult to observe and has extremely complex physiology and anatomy (Miyawaki, 1974; Smith & Kier, 1989). Since it is not a rigid body, its overall shape (hence vocal tract shape) cannot be reliably inferred from observation of only a few points along its midsagittal surface (Kaburagi & Honda, 1993). By the same token, EMG activity was recorded for only a few muscles (genioglossus anterior [GGA] and posterior [GGP], and hyoglossus [HG]), whose correlation to the recorded tongue positions may not be very high.

However, the success of the other sub networks does lead us to believe that we may be able to address tongue behavior explicitly and incorporate the modeling piecemeal over the course of a number of experiments using the same subject. For example, data for more points on the anterior tongue surface as well as the midsagittal outline of the hard palate are needed to adequately describe the kinematic behavior of the fore tongue. In order to obtain reasonable EMG-to-kinematic mappings for continuous estimation, EMG activity is needed for enough tongue muscles to approximate agonist-antagonist pairs. Although such pairs are not as easily defined for the tongue as for the jaw, addition of styloglossus (SG) to the three muscles previously recorded would provide better antagonist activity for two dimensions of tongue motion: GGA *vs.* SG for front-lowering <> back-raising; and GGP *vs.* HG for front-raising <> back-lowering (Baer, Alfonso, & Honda, 1988; Honda, Kusakawa, & Kakita, 1992; Maeda, 1992).

Assigning Phoneme-Specific Articulatory Targets

The second area in which modeling real speech data has necessitated interesting modification is in the assignment of phoneme-specific via point targets. Within a phrase, via point assignment for reiterant speech (see Fig. 2) required only a simple alternation of two target positions evenly spaced in time (except at phrase boundaries) — one via point for each phoneme. The assignment task was unnaturally easy because both targets could be specified within a single, highly coupled articulator group consisting of the lips and jaw. Thus, the via point target could be defined either in articulator space using the positions of the lips and jaw or in task space as lip aperture. Speaking rate differences could be controlled by specifying the frequency of a simple oscillator and by adjusting the constant smoothness constraint in the cascade network, responsible for generating motor commands (EMG). Each oscillator cycle includes one vowel (V) and the intervocalic string of consonants ($C_{1,2,3}$), phased around 180 degrees. This results in different degrees of approximation to the via point target, as illustrated in Fig. 5 for the simple reiterant speech case.

Real speech is both temporally and spatially more complex. Spatially, we have found that there is a particularly strong linear relation between jaw position and the first two formants (F1, F2) only for /a/, probably because the vocal tract is shaped as a wide front cavity connected to a narrow back cavity. Thus, the jaw is a suitable primary articulator for /a/ and, being a rigid structure, may be measured at a single point. For other vowels, however, tongue shape and position determine more complex vocal tract shapes. Formant values for mid and high vowels are at best weakly correlated with jaw position. Therefore, target vocal tract configurations must be derived from measurement of the tongue. But, unlike the jaw or even lip aperture, there is no one point on the tongue surface, at least not one that anyone has found, whose position can be used to uniquely specify different vowels. Furthermore, the correlation among markers placed along the midsagittal tongue decreases with distance (Kaburagi & Honda, 1993). Thus, for specification of a particular vowel phoneme, via points may need to be assigned to multiple articulator components, from whose individually weak correlations specific patterns may emerge computationally.

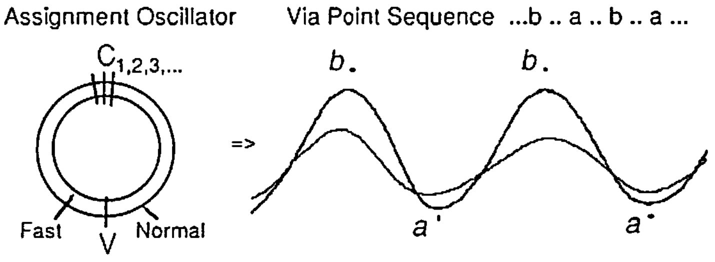

Fig. 5. Oscillator assignment of via points for consonant-vowel (CV) sequences and normalized effects of smoothness on faster rate gestures (lighter trace at right).

Similarly, while stop consonants have theoretically clear places of articulation, they are susceptible to extensive contextual variability both acoustically and articulatorily, not all of which can be accounted for by classical theories of coarticulatory undershoot (Lindblom, 1963). Note that, as described for speaking rate differences, the combination of via points and a smoothness constraint on muscle activity in our model of motor command generation should account for neuromuscular undershoot quite nicely, while biomechanical properties contributing to apparent "articulator sluggishness" are inherent to the acquired model of the forward dynamics. Other consonants, such as fricatives may involve numerous articulator complexes — e.g., the alveolar fricative /s/ depends on tongue tip, blade, jaw, and probably both lips (Faber, 1989).

Temporally, the difficulty with assigning phoneme-specific targets is well-known, and has led many people to abandon hope of finding them in the articulatory stream in anything like the beads-on-a-string representation to which we still adhere (cf. (Browman & Goldstein, 1986; Browman & Goldstein, 1990)). That is, the timing of phoneme-specific articulatory events may violate the serial order of phonemes in the string.

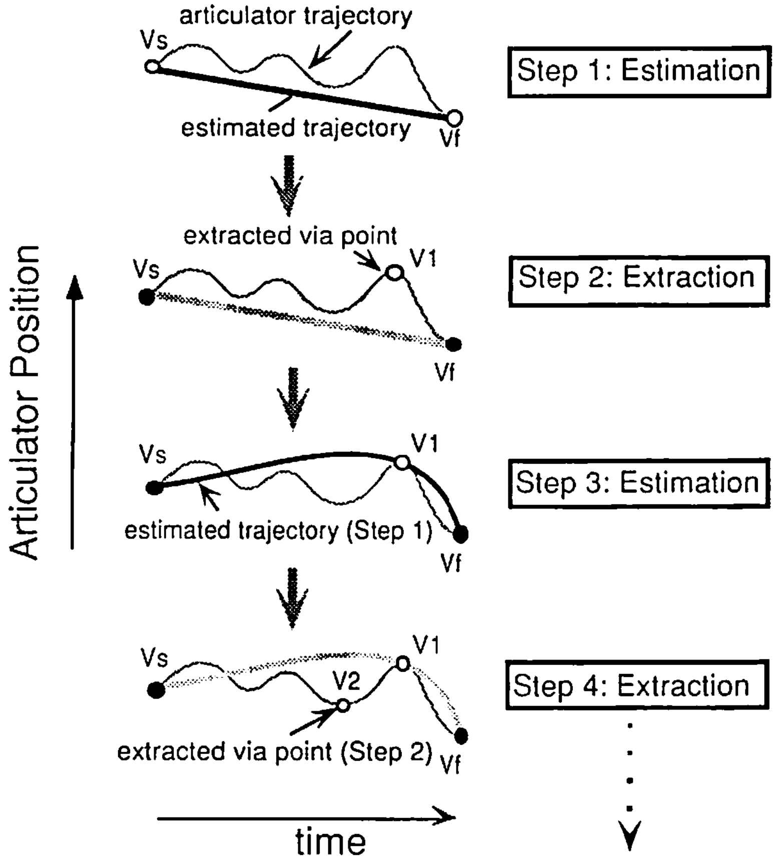

Fig. 6. Via point assignment by minimum jerk (MJ)estimation. Step 1: Minimum jerk is estimated for phrase or sentence between known start (Vs) and end (Vf) points (i.e., velocity and acceleration are zero). Step 2: Via point (V1) is assigned to data point at maximum distance from MJ trajectory, if summed error threshold (S) for all trajectories (jaw + tongue tip+tongue blade + lower lip) is reached and if the error threshold for the specific articulator (e.g., Ejaw) is reached. Step 3: a new MJ trajectory is calculated through Vs, V1, and Vf. Step 4: the next via point (V2) is assigned by repeating Step 2. Steps 3-4 are repeated until S cannot be reached.

The problem therefore is to find a principled means for assigning via point targets to complex sound sequences in which phoneme-specific articulatory or task correlates are likely to be obscured due to coarticulation and other allophonic processes. Furthermore, since we lack firm *a priori* knowledge of these correlates, via point assignment should be empirical and automatic. To address this, we have adapted the via point assignment scheme currently being used for automatic character recognition based on the dynamics of cursive handwriting (Wada, Koike, & Kawato, 1993). While the handwriting scheme integrates via point estimation with computation of the inverse and forward dynamics, the current scheme entails only calculating minimum jerk trajectories. As outlined in Fig. 6, minimum jerk trajectories are initially calculated between rest positions (i.e., where tangential velocity and acceleration are zero) at the beginning and end of each phrase of the utterance. Via points are then assigned to articulatory trajectories by successively applying the procedure shown in the figure. The number of via points assigned to a given utterance is controlled by setting a limit on the cumulative error between estimated and real trajectories.

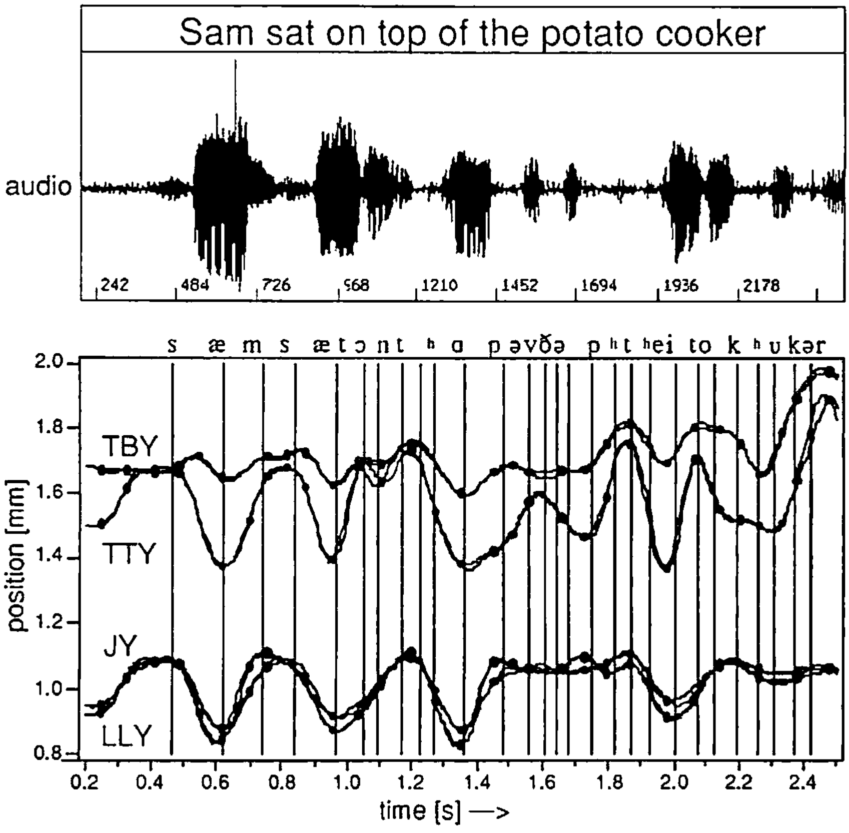

Fig. 7. Temporal acoustics and vertical positions for tongue blade (TBY), tongue tip (TTY), jaw (JY), and lower lip (LLY) are shown with overlaid via point trajectories. Vertical lines correspond to acoustic segment centers; thick dots denote via points. Speaking rate was normal.

Wada has experimented with various forms of the procedure in order to find the best fit for the data with the number of via points per articulator most closely matching the number of phonemes in the utterance. A one-one phoneme-to-via point mapping would be desirable for historical reasons, i.e., the notion of one primary target (acoustic or articulatory) per phoneme, and might allow us to posit one abstract motor command per phoneme. However, despite the conceptual and computational advantages of such an isomorphism, it is unlikely that the model components responsible for converting motor commands to muscle activity and subsequent articulator movements can cope with the range of coarticulatory and coproduction effects induced by contextual variability. Also, there is no reason to insist that "targets" be articulatory entities rather than acoustic (as commonly believed) or some higher-level equivocation of the two entities (suggested by Kiyoshi Honda).

As can be seen in Fig. 7, the number of via points required for adequate trajectory estimation is about the same as the number of acoustically derived segment labels. Notice that via points are not always assigned to each articulator at a given point in time. This situation arises when the error threshold for an articulator is not reached. Also, note the tendency for via points to be placed at peaks and valleys, similar to our original oscillator assignment scheme used for reiterant speech and specified previously in coupled oscillator models of handwriting (Hollerbach, 1981). Furthermore, when extra points are assigned, they tend to fall near the midpoint of the trajectory. This result corroborates Rumelhart's intuition (1993) that trajectory midpoints are necessary for handwriting recognition, but is not based on any *a priori* target specification. We assume therefore that multiple via points may be required for adequately generalized specification across the range phoneme transitions, and that template extraction will be necessary. To this end, data for subsets of English and Japanese diphones are being collected for tongue, lip, and jaw kinematics using real speech with which we hope to refine and verify the phoneme-specific via point estimation scheme outlined here.

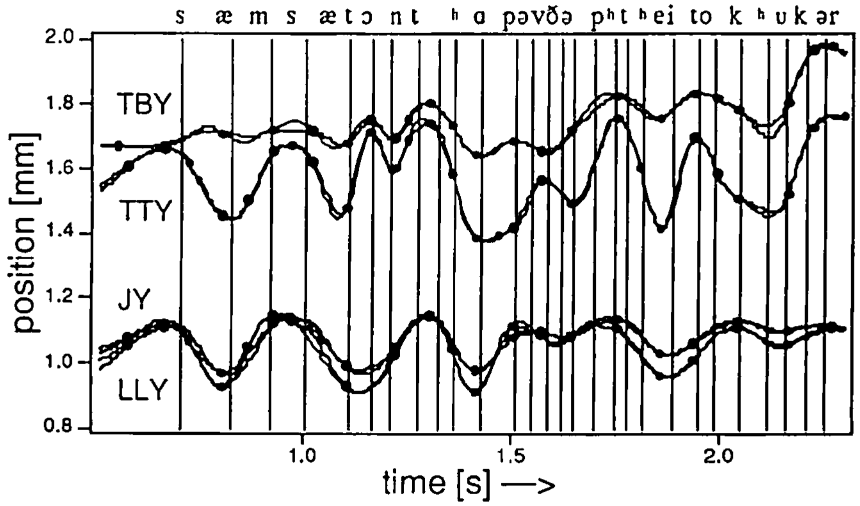

Fig. 8. Articulator trajectories and via point assignments produced at the faster speaking rate for the phrase shown in Fig. 7.

It has been observed for reiterant speech (Vatikiotis-Bateson & Kelso, 1993) that the kinematics of lip and jaw better fit the constraints of a simple second-order system such as a linear mass-spring when produced at faster speaking rates. In real speech, fast, casual productions are marked by increased coarticulation resulting in loss of phoneme-specific distinctions both kinematically and acoustically. This is born out here as well. In Fig. 8, via point assignments using the same error threshold are shown for the same utterance phrase produced at the faster speaking rate. Comparing the trajectories in Figs. 7 and 8, we see that they are quite similar but fewer points tend to be assigned at the faster rate. Presumably, phoneme-specific templates consisting of at least one, but probably no more than three, via points per primary articulator can be derived and the differences observed here between the two rates can be ascribed to rate-specific settings of the smoothness constraint (e.g., minimum change of EMG) at the motor command generation stage (Figs. 1, 5).

Summary

We have described two areas in which our modeling of speech production has progressed. Addressing the problem of determining the forward dynamics of the system in terms of separate articulatory structures has improved our ability to evaluate the quality of our approach and the data needed. It is clear that Hirayama's computational approach is adequate for modeling the lips and jaw, where adequate muscle and kinematic data have been acquired. It is equally clear that the model fails for the tongue and that we need to ascertain whether this is due to inadequate data, as we hope, inadequate modeling, or some combination of the two. At another level, we have shown that lip, jaw, and tongue position trajectories can be recovered by estimating via points, roughly equivalent in number and distribution to the phoneme string. Furthermore, the effect of speaking rate differences on the estimation results is small and in the direction we would predict for increased smoothness of faster productions. Finally, we infer from the success of via point estimation for the kinematics of the tongue tip and blade that what is most needed for better forward modeling of the tongue is higher quality EMG from a wider variety of tongue muscles.

Acknowledgment

We thank Shinji Maeda for helpful comments and suggestions.

References

Baer, T., Alfonso, P. J., & Honda, K. (1988). Electromyography of the tongue muscles during vowels in /əpVp/ environment. *Ann. Bull. (RILP)*, **17**, 7-18.

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

Browman, C. P., & Goldstein, L. (1990). Representation and reality: physical systems and phonological structure. *Jour. Phon.*, **18**, 411-424.

Dornay, M., Uno, Y., Kawato, M., & Suzuki, R. (1992). Simulation of optimal Movements using the minimum-muscle-tension-change model. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems 4* . San Mateo, CA: Morgan Kaufmann Publishers.

Faber, A. (1989). Lip protrusion in sibilant production. *JASA*, 86(Suppl. 1), S113.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Honda, K. (1992a). Neural network modeling of speech motor control. In *Proceedings of the International Conference on Spoken Language Processing-1992*, Banff:.

Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. (1992b). 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* (pp. 191-198). San Mateo, CA: Morgan Kaufmann Publishers.

Hollerbach, J. M. (1981). An oscillation theory of handwriting. *Biol.Cyber.*, 39, 139-156.

Honda, K., Kusakawa, N., & Kakita, Y. (1992). An EMG analysis of sequential control cycles of articulatory activity during /əpVp/ utterances. *Jour. Phon.*, 20, 53-63.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3, 79-87.

Kaburagi, T., & Honda, M. (1993). Prediction of the tongue shape by using the magnetic multiple-point position sensing (in Japanese). *Technical Report of IEICE*, SP92-141, 33-40.

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. Cyber.*, 62, 275-288.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *JASA*, 35, 1773-1781.

Maeda, S. (1992). From EMG to sound patterns of vowels: Software. *ATR Technical Report*, TR-H-005.

Miyawaki, K. (1974). A study on the musculature of the human tongue. *Ann. Bull. (RILP*, 8, 23-50.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *JASA*, 92, 3078-3096.

Rumelhart, D. E. (1993). Computational learning and cognition: Supervised learning for coordinative motor control. In E. B. Baum (Eds.), *SIAM Frontier Series* (pp. 177-196). Philadelphia: Society for Industrial & Applied Mathematics

Smith, K. K., & Kier, W. M. (1989). Trunks, tongues, tentacles: Moving with skeletons of muscle. *American Scientist*, 77, 29-35.

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

Vatikiotis-Bateson, E. (1988). *Linguistic structure and articulatory dynamics*. Bloomington, IN: Indiana University Linguistics Club.

Vatikiotis-Bateson, E., Hirayama, M., & Kawato, M. (1991). Neural network modeling of speech motor control using physiological data, *Perilus* (Stockholm Univ. Linguistics Dept.), XIV, 63-67.

Vatikiotis-Bateson, E., & Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Jour. Phon.*, 21(No. 3), 231-265.

Wada, Y., Koike, Y., & Kawato, M. (1993). A cursive handwriting model based on the minimization principle (in Japanese). *Technical Report of IEICE*, NC92-124, 229-236.