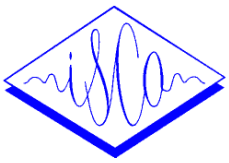

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

THE ARTICULATORY DYNAMICS OF RUNNING SPEECH: GESTURES FROM PHONEMES?

Eric Vatikiotis-Bateson, Makoto Hirayama, Kiyoshi Honda, and Mitsuo Kawato

ATR Auditory and Visual Perception Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

As demonstrated by Hirayama, V.-Bateson, Kawato, & Honda [1], we are attempting to model speech production by means of neural networks. At one stage, a 3-layer perceptron learns the dynamics relating muscle activity and articulator motion and, at a later stage, another perceptron learns the PARCOR parameters relating the effect of articulator motion on vocal tract shape and the speech acoustics. After learning, motor commands to the musculo-skeletal system are generated through time by performing the trajectory formation and inverse dynamics using a cascade neural network, parametrized by the via point and smoothness constraints imposed by the phoneme input string and global performance factors, such as speaking rate and speaking style. Articulator trajectories are then generated which serve as input to the PARCOR synthesizer that produces the speech acoustics. Although this effort is still in its infancy, it is proving to be a fairly successful piece of engineering and we think a discussion of the attendant speech science and motor control issues is warranted. Therefore, in this paper, we discuss our modeling effort in terms of well-known problems common to both computational modeling and speech motor control, such as excess degrees-of-freedom in the mapping between different levels, coordinate transformation between articulator and task space variables, and extrinsic versus intrinsic timing. Problems of less common concern, such as biological and cognitive plausibility, are also considered. For the most part, we focus on issues leading up to the generation of articulator motion and leave discussion of the articulatory-to-acoustic transform to a later date.

DIMENSIONALITY AND DEGREES-OF-FREEDOM

We are continually reminded that speech communication requires the extraordinary ability of the human to coordinate and/or recover excess degrees of freedom in the time-bound processes of speech production and perception. For example, the speech production process could entail successive reductive mappings from central nervous system to periphery. Higher level cortical behavior leads to relatively fewer motor commands which activate a relatively small number of muscles that result in movement of an even smaller number of articulator structures. The movement of these structures then gives rise to an acoustic signal that unfolds unidimensionally through time. Actually, we do not have a precise idea of what this means, because we do not know exactly what the relevant components of the system are, how they are represented either cognitively or neurophysically, or what their temporal constraints are. So, we cannot be entirely sure what constitutes a degree-of-freedom. Are membership in a particular set of cortical motor neurons or neuronal firing rates important control parameters? Does it matter exactly which muscle fibers are activated or the degree of activation in a particular articulator component? We tend to test these questions parametrically, but the answers are often equivocated by seemingly trivial differences in experimental conditions.

One encouraging aspect of the speech production process is that it has temporal structure at all levels. Neural activity associated with cognitive and motor behavior occurs in time. Events in the motor cortex are realized musculo-skeletally after some delay, both neuromotor and biomechanical. Finally, the generation and propagation of the acoustic signal takes time. The fact that these events occur in time greatly reduces the dimensionality of the behavior through successive processing stages and insures ultimately that the system is deterministic. But even at the culmination of the complex processes we try to address here, when the dimensionality is at its lowest, namely, the acoustic signal, there appear to be excess degrees-of-freedom. The acoustic signal contains a wealth of information from which a listener may simultaneously infer the speaker's mood, gender, intended message, unintended message, the phoneme string, diagnostic aspects of articulation, and so on. The fact that no two acoustic signals are the same, although their informational content may be, implies that the information contained within the speech signal is redundantly specified. This redundancy is the basis for categorical perception and frees the listener from strict adherence to the acoustic time course.

The loose mapping between recoverable information and acoustic carrier signal has well-known advantages in environmentally conditioned behavior. In the context of this discussion, however, we use it to support the hypothesis that excess degrees-of-freedom is a systemic necessity for the generation and interpretation of purposive behavior. If we take at all seriously the notion that speech motor behavior is regulated in part by acoustic and/or sensorimotor targets, mapped cognitively the way a listener does, then it is unlikely that speech production is carried out with much precision at any level we care to observe. Indeed, evidence of imprecision abounds at every level of observation: e.g., diffuse activation over multiple cortical areas prior to movement of discrete movements of fingers and toes [2, 3]; equifinality between muscle activity and articulator position [4]; free variation of coupled articulators, such as the lips and jaw during bilabial production [4]; and the loose mapping between critical dimensions of articulator motion and acoustic consequences, e.g., [5] supporting the notion of quantal states [6].

In recent years there have been several approaches to modeling the degrees-of-freedom problem in biological behaviors. One has been to identify functionally coupled structures or synergies specified at physiological levels [7, 8] and acting in physically lawful ways, e.g., [9, 10]. Other approaches grounded more firmly in engineering and computation have sought unique or "optimal" solutions to the formally ill-posed problems that arise when a system must reduce or expand the number of conditioning variables between the input and output sides of some mathematical function [11]. Artificial neural networks are an example of this latter approach and as such must constantly be helped over the hurdles imposed by apparent many-to-one and, even worse, possibly one-to-many mappings relating degrees-of-freedom at different levels of description (or representation). A commonly cited example for the control of limb movement is the mismatch of degrees-of-freedom between joint space (e.g., shoulder, elbow, and wrist) and the task space of the end-effector (e.g., planar motion of the hand).

SMOOTHNESS

One way to reduce the effect of variable input to a temporal system which shows less variable output is to impose an objective function that minimizes change over time of some criterial aspect of the behavior being modeled. In modeling the control of movement behavior, this has taken the form of a smoothness constraint. That is, if we are interested in predicting successive positions of a moving articulator, a smoothness constraint greatly reduces the range of possible next positions in a sequence. The most widely used form of such a constraint has been to minimize the movement jerk (rate of change of acceleration), as originally proposed for planar arm movement by Flash and Hogan [12]:

$$C_J = \frac{1}{2}\int_0^{t_f}\left\{\left(\frac{d^3X}{dt^3}\right)^2 + \left(\frac{d^3Y}{dt^3}\right)^2\right\}dt. \qquad (1)$$

Here, (X, Y) are Cartesian coordinates of the hand, and t_f is the movement duration. This objective function has proved useful in predicting point-to-point movements as well as a range of via-point movements (i.e., through some target specified between movement endpoints). Formally, it provides unique solutions to trajectory control from knowledge of only the initial, final, and via-point positions and the movement duration.

It is readily observed that biological movements are quite smooth kinematically, across a wide range of structures and conditions that cannot be accounted for solely by passive biomechanical factors, such as inertia and viscosity. For example, tongue tip and jaw

10.21437/ICSLP.1992-269

movements show similar smoothness despite differences in mass and structure. Smoothness could function not only to reduce the complexity of movement control, but also to minimize physiological wear and tear and prevent damage to joints and muscles [13]. This latter possibility, along with the minimum jerk function's independence of the underlying dynamics of the system, makes its biological relevance questionable. Being strictly kinematic and dependent only on the start, end, and via points of the planned trajectory, minimum jerk cannot account for abrupt changes in either the organism, the task, or the environment (e.g., adaptation to perturbation or sustained external force [14, 15, 16]) that impinge on the system's dynamics. For this reason, a number of alternative smoothness constraints, which are formally quite similar to minimum jerk, but which make direct use of the dynamics, have been proposed. In succession, these include minimum torque change [16], minimum muscle-tension change [17], and finally, minimum motor-command change shown as equation 2 [13]:

$$C_M = \frac{1}{2}\int_0^{t_f} \sum_{i=1}^{n} \left(\frac{dM_i}{dt} \right)^2 dt, \qquad (2)$$

where, M_i is the motor command sent to the ith of n muscles.

Empirically, the first two functions predict curved movement paths appropriate to arm movements, where the minimum jerk function always predicts straight line paths. Conceptually, the third function, which we use in our speech production model, addresses the insight that smoothness might serve to reduce the enormous degrees-of-freedom problem believed to exist within the central nervous system itself rather than only at the periphery. That is, motor commands are constrained to evolve smoothly over time, thus greatly reducing the possible responses to higher order neural activity. Observed smoothness at the periphery will thus combine constraints on neuromotor behavior as well as the passive biomechanics (e.g., inertia and viscosity).

VIA POINTS

Via-points have been used to model limb movement control and have proved to be more effective in guiding trajectory-formation algorithms than simple end-point control [18]. Basically, a via point is an intermediate spatial target specified somewhere between a movement's start and end points that allows a recurrent neural network, for example, to generate more complex and realistic movement trajectories than those resulting from the step functions characteristic of end-point control models [19]. The via point is an optimal target that does not actually have to lie on the realized trajectory. In our model, the extent to which trajectories approach the via point is modulated primarily by the strength of the smoothness constraint. In modeling continuous movement behavior, via points are specified sequentially, and adjacent via points are formally similar to the end points specified in point-to-point movements. Initially, via points were specified temporally as well as spatially, but we have shown that exact temporal specification is not necessary in generating plausible articulator trajectories in simple tasks such as reiterant speech [20]. We return to this very important feature of via points below.

Specifying targets, either acoustic or articulatory, is not new to speech production [21]. Typically, they have been of the end-point control type, as exemplified by Masaki *et al.* [22], and more recently by Browman and Goldstein [23]. For example, target position for the production of a phoneme or "gesture" is specified either in Cartesian articulator coordinates or in task coordinates such as location and degree of tongue constriction [24]. Motion toward such targets is usually controlled by specifying parameter values, such as spring stiffness and damping, for a second-order equation of motion. A problem for this approach is that the step function parameters and the timing of the function's activation relative to all other gestures must be precisely determined to achieve the observed behavior. That is, precise values must be found for the second-order movement equation, and the timing (onset and offset) and degree (strength) of activation must be determined. Formerly, this has been done in an informed, yet *ad hoc*, manner, e.g., [23]. However, to determine these parameters empirically from articulatory and physiological data is a formidable task requiring analysis of enormous amounts of data, not to mention identification of the relevant events.

Up until now we have specified via points spatially in articulator coordinates for each phoneme of the input string. Values are set *a priori* by the researcher to represent extreme values, which either are observable (vowels) or could be observed in the absence of boundary constraints (consonants). For example, lower lip height for bilabial closure is specified higher than presence of the upper lip would allow. Although lip compression is easily observed, it is difficult to measure instrumentally. Specification of via points in task coordinates, such as lip aperture, has the same problem. In this case, the boundary constraint can be resolved either by specifying negative aperture values or an offset. Similarly, negative values must be specified for tongue constriction degree, although it is possible to monitor and measure the area of tongue-palate contact using dynamic electropalatography.

There are, however, advantages to specifying via points in task rather than articulator coordinates. First, as much as possible, cognitively linked control parameters should be specified in cognitively plausible terms. It may or may not be the case that speech structures are represented cognitively in terms of task-specific vocal tract configurations or gestures [3, 25], but it is highly unlikely that position of individual articulators is represented. Indeed, it seems that many basic synergies are specified innately, as exemplified by the birth cry and supine locomotory behavior in neonates [26]. This suggests that speech acquisition entails fine-tuning of existing articulatory synergies rather than creation of new ones. Second, in the near future, we will be able to replace the homuncular role of the researcher, who at present must specify the via points, with network acquisition of the phoneme-specific via point values as well as the

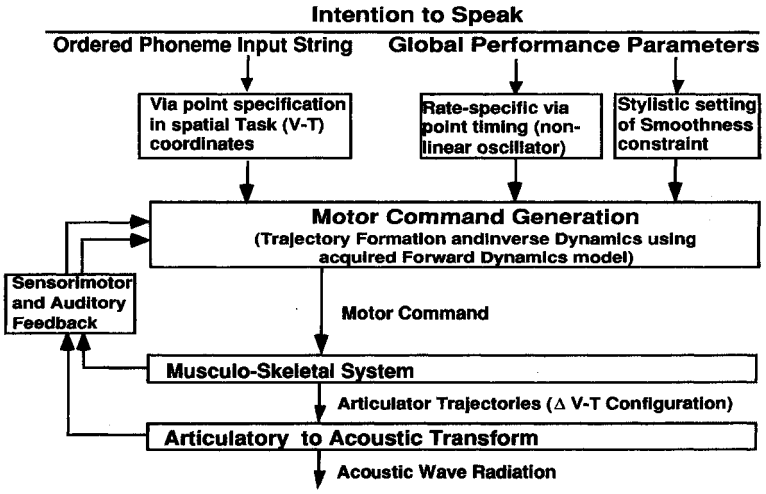

Figure 1. Overview of Speech Production Model

basic and acquired articulator coupling constraints that will solve the one-to-many relation between coordinate structure and articulator components. The result will be a table-lookup of phoneme-specific via point values in task coordinates.

GLOBAL PERFORMANCE PARAMETERS

By global performance parameters we mean those constraints on an utterance which are not themselves subject to constant revision during utterance production. Examples for which we have collected data and have done some modeling include speaking rate (e.g., normal *vs.* rapid) and speaking style (e.g., precise *vs.* casual, loud *vs.* soft). Other, as yet unexplored parameters are intonation and prosody. Although these parameters are not independent of one another — e.g., precise speech tends to be produced at slower syllable rates than casual speech — their input to the model is being treated quite separately (see Figure 1); their interdependency reappears in their effects on the motor command generation network (see next section).

Speaking rate serves to parametrize a nonlinear (limit cycle) oscillator, which is constrained to contain no more than one vowel per cycle. Conceptually, this is nothing more than a tunable central pattern generator driving the network in a rhythmical fashion, a need previously recognized in Grenoble [27]. The oscillator governs syllable production rates and is plausibly anchored to the vowel. Results of both tapping and "P-Center" studies have shown the rhythmic beat of English to be very near (within 5 ms) of English acoustic vowel onset [28, 29]. Intervening consonants are specified 180 degrees anti-phase to the vowel, their order determined by the phoneme input string. This results in a loose, sub-optimally timed succession of via points. Fine-tuning of via point timing is achieved by interaction in the motor command generation network with the smoothness constraint and interarticulator couplings reflecting the acquired dynamical and language-specific phonotactic constraints. In this way, language-specific differences in coarticulation (or co-production) need not be under active control, and context-specific timing differences need not be specified extrinsically.

All other things being equal, an increase of speaking rate results in smaller trajectories that undershoot the via points more. This is shown schematically in Figure 2. Speaking style, on the other hand, affects the smoothness constraint to allow closer articulatory approximation of the via point constraints as, for example, in the case of very precise speaking.

NETWORK MODELING

As described in detail elsewhere [1, 20], a 3-layer perceptron is used to learn the forward dynamics linking muscle activity and the resulting articulatory movement behavior. Briefly, we have used EMG activity from as many as ten orofacial muscles and the motion of the lips, jaw, and recently the tongue to train the network during reiterant and real speech tasks. After training, a cascade neural network containing the acquired dynamics model is used to estimate motor commands and articulator trajectories for test data by specifying the via points associated with the phoneme input string and setting the smoothness constraint.

What is particularly interesting about the learning phase is that the acquired forward dynamics is not pure, because the network learns the functional couplings between articulators in addition to the dynamics governing their motion. If the coupling constraints can be separated from the dynamics, then this impurity will be very useful to our modeling effort. This has been done for the simple case of reiterant speech where interarticulator couplings are very clear, and suggests that the model reduces the degrees of freedom by identifying functional synergies, such as the coordinated action of lips and jaw in producing a bilabial [20]. These synergies amount to task-specific constraints governing interarticulator coordination and support the proposal that, for production, via points can be specified

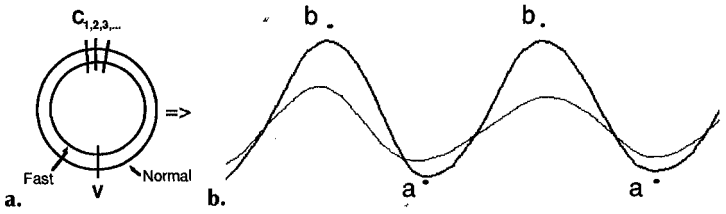

Figure 2. a. Limit-cycle oscillator determining sub-optimal timing of via points shown for two speaking rates. **b.** Hypothesized interaction of rate and via points (dots) for lip aperture during reiterant speech production. Trajectories are time-normalizedspeaking ; darker trace denotes normal rate.

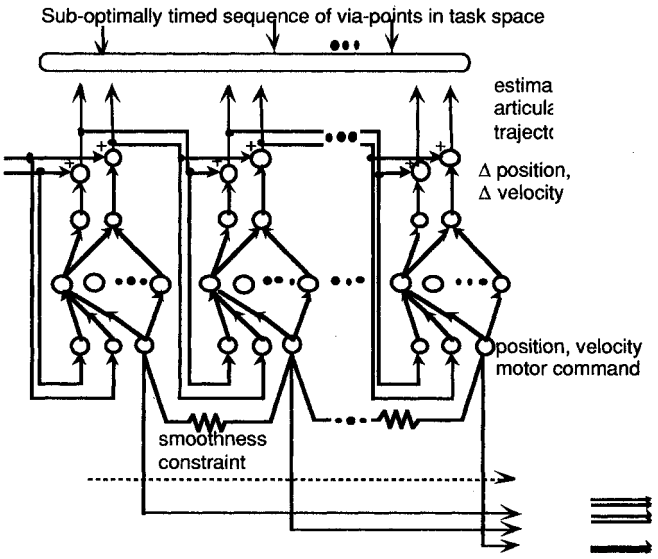

Figure 3. Cascade neural network for motor command generation.

in task space and sub-optimally in time (see Figure 1). The cascade network (Figure 3) acquires the coordinate mapping between via points and articulators and, as shown for arm movement [19], it can fine-tune via point timing.

Thus, the motor command generation network receives as input via points specified in task space, e.g., lip aperture, degree and location of tongue constriction, their sub-optimal temporal specification, and information about smoothness requirements (Figure 3). All three of these inputs interact in the network: The strength of the smoothness constraint limits the rate of change in the motor command and affects the degree to which resulting movement trajectories approach the spatial via point. Smoothness will be greater for casual than for precise speech, therefore, casual speech will show more undershoot. Fast speech also will show more undershoot, unless the smoothness constraint is relaxed, simply because of the reduced time available to approach via points. Thus, greater articulatory precision is achieved at the expense of smoothness and even more so if precision is required at a fast speaking rate.

The interarticulator coupling constraints acquired during network training will determine the extent to which adjacent phonemes may overlap and hence their relative timing. The network can then use these adjacency constraints together with frequency of the nonlinear oscillator and the setting of the smoothness constraint to fine-tune via point timing. Indeed, the model's ability to make appropriate adjustments in via point timing has been demonstrated for arm movement control [19]. Timing, then, is intrinsically controlled through the combination of rhythmic priming provided by the nonlinear oscillator, hardware constraints on production such as smoothness, task-specific and general dynamic constraints on coproduction, and the sequence of task coordinates specified by the phoneme input string.

The network produces smooth motor command output to the musculo-skeletal system, which generates articulator trajectories. These trajectories then serve as input to the PARCOR synthesis network for acoustic wave generation [1]. The PARCOR synthesis network currently learns the correlation between estimated articulator trajectories and vocal tract area functions. However, learning might be improved if the task-to-articulator mappings acquired in the motor command generation network could be converted directly into changes of vocal tract configuration, which might be more appropriate correlated to the PARCOR coefficients.

SUMMARY

We have outlined our approach to modeling speech production and discussed some of the speech motor control issues that must be addressed along the way. As usual, the ideas are a step or two ahead of the data. Even now, as we evaluate the current proposal with EMG and movement data for the tongue, lips, and jaw, other ideas emerge. For example, we know the optimization window of the current scheme is too large (i.e., the entire 7-8 second utterance) and the relaxation time too long. Addition of the rhythmic component should reduce the temporal scope of network convergence significantly. A more realistic window size would be 3-4 oscillator cycles, which would better match the physiological scope of look-ahead processes such as anticipatory coarticulation [30]. We are also considering replacing the cascade network with a more online network that combines forward and inverse dynamics in such a way that convergence is achieved within 25-50 ms [31].

REFERENCES

[1] Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Honda, K. "Neural Network Modeling of Speech Motor Control," *Proceedinsgs of the International Conference on Spoken Language Processing - 2*, Banff, Canada. 1992.

[2] Boschert, J., Hink, R.F., & Deecke, L.. "Finger Versus Toe Movement-Related Potentials: Further Evidence for Supplementary Motor Area (SMA) Participation Prior to Voluntary Action," *Experimental Brain Research*, **52**, 73-80, 1983.

[3] Gracco, V. L. "Characteristics of Speech as a Motor Control System." In G.R. Hammond (ed.), *Cerebral Control of Seech and Limb Movements (Advances in Psychology* Series). Amsterdam: Elsevier. 1991.

[4] Gracco, V.L., & Abbs, J.H. "Variant and Invariant Aspects of Speech Movements," *Experimental Brain Research*, **65**, 156-166, 1986.

[5] Perkell, J.S., & Nelson, W.L. "Variability in Production of the Vowels /i/ and /a/," *Jour.of the Acoustic Society of America*, **77**, 1889-1895, 1985.

[6] Stevens, K.N. "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data." In E.E. David & P.B. Denes (Eds.), *Human Communication: A Unified View*. New York: McGraw-Hill. 1972.

[7] Bernstein, N. *The coordination and Regulation of Movements*. New York: Pergamon Press. 1967.

[8] Greene, P.H. "Problems of Organization of Motor Systems." In R. Rosen and F. Snell (Eds.), *Progress in Theoretical Biology*. New York: Academic Press. 1972.

[9] Turvey, M.T. Preliminaries to a Theory of Action with Reference to Vision." In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1977.

[10] Kelso, J.A.S., & Tuller, B. "Converging Evidence in Support of Common Dynamic Principles for Speech and Movement Coordination," *American Journal of Psychology*, **15**, R928-R935, 1984.

[11] Marr, D. *Vision*. New York: Freeman. 1982.

[12] Flash, T., & Hogan, N. "The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model," *Journal of Neuroscience*, **5**, 1688-1703, 1985.

[13] Kawato, M. "Trajectory Formation in Arm Movements: Minimization Principles and Procedures." In H.N. Zelaznik (Ed.), *Advance in Motor Learning and Control*. Human Kinetics Publishers, in press.

[14] Atkeson, C.G., & Hollerbach, J.M. "Kinematic Features of Unrestrained Vertical Arm Movements," *Jour.of Neuroscience*, **5**, 2318-2330, 1985.

[15] Kelso, J.A.S, Tuller, B., Vatikiotis-Bateson, E., & Fowler, C.A. "Functionally Specific Articulatory Cooperation Following Jaw Perturbations During Speech: Evidence for Coordinative Structures," *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 812-832, 1984.

[16] Uno, Y., Kawato, M., & Suzuki, R. "Formation and Control of Optimal Trajectory in Human Multijoint Arm Movement — Minimum Torque-Change Model," *Biological Cybernetics*, **61**, 89-101, 1989.

[17] Dornay, M., Uno, Y., Kawato, M., & Suzuki, R. "Simulation of Optimal Movements Using the Minimum-Muscle-Tension-Change Model." In R.P. Lippmann, J.E. Moody, & D.S. Touretzky, *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann Publishers, in press.

[18] Bizzi, E., Accornero, N., Chapple, W., Hogan, N. "Posture Control and Trajectory Formation During Arm Movement," *Journal of Neuroscience*, **4**, 2738-2744, 1984.

[19] Kawato, M., Maeda, Y., Uno, Y., & Suzuki, R.. "Trajectory Formation of Arm Movement by Cascade Neural Network Model Based on Minimum Torque-Change Criterion," *Biological Cybernetics*, **62**, 275-288, 1990.

[20] Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. "Forward Dynamics Modeling of Speech Motor Control Using Physiological Data." In R.P. Lippmann, J.E. Moody, & D.S. Touretzky, *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann Publishers, in press.

[21] Lindblom, B. "Spectrographic Study of Vowel Reduction," *Journal of the Acoustical Society of America*, **35**, 1773-1781, 1963.

[22] Masaki, S., Shirai, K., Imagawa, H., & Kiritani, S. "Differences in Jaw Opening for Vowels Due to Speaking Rate and Word-Internal Position in the Production of Vowel Sequence Words," *Annual Bulletin* (RILP, Tokyo), **19**, 29-46, 1985.

[23] Browman, C.P., & Goldstein, L. "Gestural Specification Using Dynamically-Defined Articulatory Structures," *Journal of Phonetics*, **18**, 299-320, 1990.

[24] Saltzman, E.L. "Task Dynamic Coordination of the Speech Articulators: A Preliminary Model." In H. Heuer & C. Fromm (Eds.), *Generation and Modulation of Action Patterns*. Berlin: Springer-Verlag. 1986.

[25] Liberman, A.M., & Mattingly, I.G. "The Motor Theory of Speech Perception Revised," *Cognition*, **21**, 1-36, 1985.

[26] Thelen, E. "Developmental Origins of Motor Coordination: Leg Movement in Human Infants," *Developmental Psychology*, **18**, 1-22, 1985.

[27] Bailly, G., Laboissière, R., & Schwartz, J.L. "Formant Trajectories as Audible Gestures: An Alternative for Speech Synthesis," *Journal of Phonetics* **19**, 9-24, 1991.

[28] Allen, G.D. "Speech Rhythm: Its Relation to Performance Universals and Articulatory Timing," *Journal of Phonetics*, **3**, 75-86, 1975.

[29] Fowler, C.A. "Converging Sources of Evidence on Spoken and Perceived Rhythms of Speech: Cyclic Productions of Vowels in Sequences of Monosyllabic Stress Feet," *Journal of Experimental Psychology: Gerneral*, **112**, 386-412.

[30] MacNeilage, P.F. "Motor Control of Serial Ordering of Speech," *Psychological Review*, **77**,182-196, 1970.

[31] Wada, Y., & Kawato, M. "A Neural Network Model for Arm Trajectory Formation Using Forward and Inverse Dynamics Models," *Neural Networks*, submitted