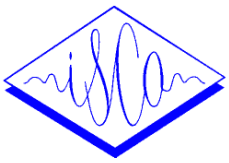

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

NEURAL NETWORK MODELING OF SPEECH MOTOR CONTROL

*Makoto Hirayama** *Eric Vatikiotis-Bateson*** *Mitsuo Kawato** *Kiyoshi Honda***

*ATR Human Information Processing Research Laboratories
**ATR Auditory and Visual Perception Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Tel. +81-7749-5-1042 Fax. +81-7749-5-1008

ABSTRACT

An artificial neural network approach to modeling speech production is presented. The model uses simultaneously recorded data for neuromuscular activity (EMG), articulator motion, and the speech acoustics and based on dynamic optimization principle based on forward dynamics of articulators. The experimental configuration and data processing are explained, then the simulation result of forward dynamics modeling, which is the essential part of our modeling approach, is presented. Preliminary result of the forward acoustic modeling, which learns the correlation between articulator trajectories for the lips & jaw and PARCOR parameters of acoustic waveform, is presented.

I. INTRODUCTION

Speech production entails extraordinary coordination among diverse neurophysiological and anatomical structures from which unfolds through time a complex acoustic signal that conveys to listeners something of the speaker's intention. Analysis of the speech acoustics has not revealed the encoding of these intentions, generally conceived to be ordered strings of some basic unit, e.g., the phoneme. Nor has analysis of the articulatory system provided an answer, although recent pioneering work by Jordan [1][2], Saltzman [3], Laboissière [4] and others [5][6] have brought us closer to an understanding of the articulatory-to-acoustic transform and have demonstrated the importance of modeling the articulatory system's temporal properties. However, these efforts have been limited to kinematic modeling because they have not had access to the neuromuscular activity of the articulatory structures.

In this paper, we present an artificial neural network approach to modeling speech production. The model uses simultaneously recorded data for neuromuscular activity (EMG), articulator motion, and the speech acoustics and consists of four information processing units: 1) the transformation from linguistic information to phoneme-specific articulatory targets and setting of a biologically plausible smoothness constraint; 2) the generation of appropriate neuromuscular activity (EMG) to satisfy the constraints; 3) the transformation from EMG activity to articulatory movements; and 4) the transformation from articulation to acoustic output.

The model has several important features: First, real rather than simulated physiological and kinematic data are used during network training to acquire the forward dynamics relating neuromotor activity and musculo-skeletal constraints to the ensuing articulatory behavior. Second, continuous movement behavior is modulated by serially-ordered spatial targets corresponding to the utterance-specific phoneme string. Third, global performance parameters, such as speaking rate and style, determine the relative strengths of the smoothness and spatial target constraints. This enables examination of the effects of different performance constraints on various aspects of interarticulator coordination such as articulatory undershoot and blending [7].

In previous work [8], using physiological data (articulator movement and EMG activity of relevant muscles), a neural network learned the forward dynamics relating motor commands to muscles and the ensuing articulator behavior. Then, a cascade neural network containing the forward dynamics model along with a suitable smoothness criterion was used to produce continuous motor commands from a sequence of discrete articulatory targets corresponding to the phoneme input string.

Although a promising beginning, the earlier work [8] was limited by the simplicity of the reiterant speech paradigm used to elicit movements whose primary articulators were the lips and jaw. This technique was useful for recording motion optoelectronically, but it resulted in alternating sequences of just two phonemes with a high degree of interarticulator coupling. Another limitation was the small number of physiological channels recorded; four muscles and generally only one dimension of articulator motion. For any articulator, then, EMG activity could be recorded for only one direction of motion (e.g., jaw opening and lip closing). These limitations are remedied in the present study by using more natural utterances, produced in more than one speaking style, and by recording three-dimensional movement and EMG activity for 10 muscles. Finally, this study extends the scope of our previous modeling effort to the transformation of the model-generated articulator trajectories to acoustic output. A neural network is used to acquire the mapping between articulation and acoustics and assigns PARCOR parameters [9], which are correlated with vocal tract area functions [10].

II. DATA COLLECTION

Movements of articulators, EMG from 10 orofacial muscles and acoustic data were recorded while a subject produced multiple repetition of sentences. Figure 1 shows an example of the processed data.

2.1 Test Utterance and Speech

One speaker produced multiple repetitions of five sentences, including a reiterant copy (using */baba/*) of one sentence for each of two speaking rate/styles: fast/casual and slow/precise. Test sentences were:

1. When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow.
2. Reiterant copy of sentence 1), "babababa....."
3. Pam put the bobbin in the frying pan and added more puppy parts to the boiling potato soup.
4. The pope put on a purple robe to appease the people who prefer popular opinion to polite praise of papal posture.
5. After Papa beamed aboard the love boat, Mama popped their baby into the bubbling mud bath.

Sentence duration was between 5 and 7 seconds depending on sentence and speaking style. The speech acoustics were sampled at 10 kHz using a 12 bit A-D converter.

2.2 Articulator Movement Measurement and Processing

Three-dimensional movements of the lips and jaw were recorded at 200 Hz using a Northern Digital OPTOTRAK position sensing system. Infrared LED markers were put on the center of upper and lower lips. For jaw movement, 2 markers were attached to a rigid jaw splint. Missing data points, which occurred about 10 times per trial, were linearly interpolated. Movement data were corrected for head movement via rigid body reconstruction for three reference markers on the subject's head (see Vatikiotis-Bateson & Ostry [11] for details). Velocity was derived from the smoothed (at 40Hz) position data. The velocity data were then smoothed and derived to get acceleration.

2.3 EMG Recording and Processing

Using surface electrodes, EMG activity was recorded for 10 orofacial muscles at 2000 Hz, then rectified and integrated at 200 Hz, and finally smoothed at 40 Hz. The 10 muscles used were: orbicularis oris superior (OOS), levator labii superioris (LLS), orbicularis oris inferior (OOI), depressor labii inferioris (DLI), mentalis (MTL), medial pteragoid (PT), anterior belly of the digastric (ABD), geniohyoid (GH), sternohyoid (SH), and masseter (MAS).

III. NEURAL NETWORK MODELING

We are interested in how discrete linguistic intentions become continuous and complex speech signals, through continuous articulator trajectories that are driven by continuous neuromuscular activity. How does the brain control this process? What is the mechanism for generating motor commands? It is thought that the

10.21437/ICSLP.1992-268

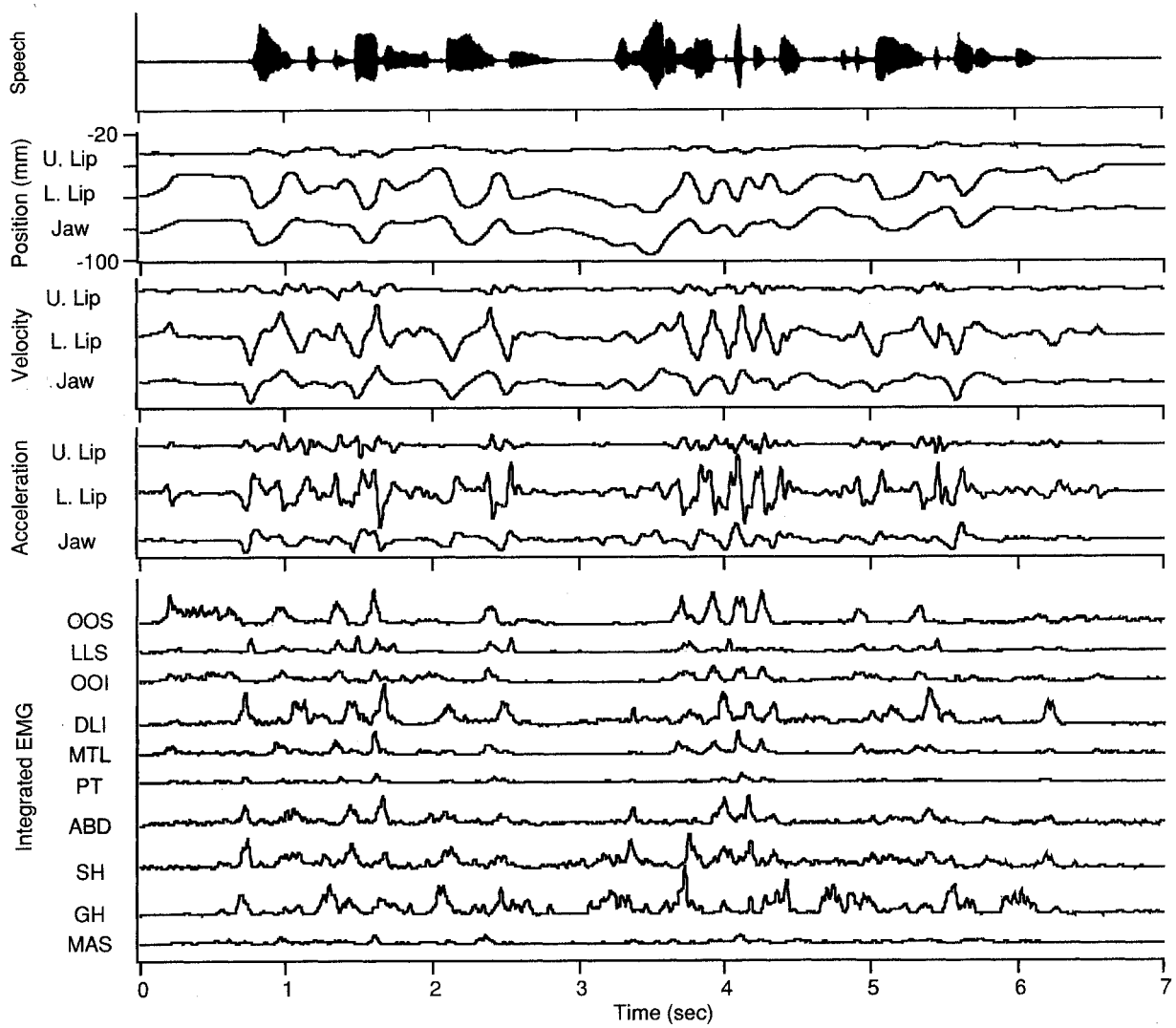

Fig. 1. Experimental & processed data. The audio waveform, kinematics for vertical (head-corrected) motion, and rectified, integrated, and smoothed EMG are shown for the sentence: "Pam put the bobbin in the frying pan and added more puppy parts to the boiling potato soup". The 10 muscles used are: orbicularis oris superior (OOS), levator labii superioris (LLS), orbicularis oris inferior (OOI), depressor labii inferioris (DLI), mentalis (MTL), medial pteragoid (PT), anterior belly of the digastric (ABD), geniohyoid (GH), sternohyoid (SH), and masseter (MAS).

mechanism of continuous speech generation is mainly effected by feedforward control. Auditory feedback and somatosensory feedback also play a role, but feedback delay makes it difficult to control articulators in real time. If feedforward control is the means of control, an internal model of the controlled object, i.e., a dynamics model of the articulator system's musculo-skeletal properties, is required. Therefore, we made a forward dynamics model of the articulators using an artificial neural network. Then, we can use a dynamic optimization process based on the learned forward dynamics and a biologically plausible smoothness constraint to generate motor commands.

The model uses simultaneously recorded data for neuromuscular activity (EMG), articulator motion, and the speech acoustics and consists of four information processing units: 1) task specification networks for the transformation from linguistic information to phoneme-specific articulatory targets and setting of the smoothness constraint; 2) a motor command generation network for producing appropriate neuromuscular activity (EMG) to satisfy the constraints; 3) a musculo-skeletal system network for the transformation from EMG activity to articulatory movements; and 4) a forward acoustic network for the transformation from articulation to acoustic output. Networks (2) and (3) contain forward dynamics models.

3.1 Forward Dynamics Modeling of Speech Articulators

Modeling the forward dynamics of speech articulation is essential to our approach. The network learns the correlation between position, velocity, EMG at time t_n and the acceleration for all articulators at the next time sample t_{n+1}. Figure 2 shows the network architecture of the forward dynamics model. Inputs to the network are position and velocity for each dimension of motion for the lips and jaw, and the 10 EMG signals; outputs are accelerations. Using real physiological data as teacher signals, the error back-propagation learning algorithm [12] updates the weights of input-hidden & hidden-output layers to acquire the forward dynamics of articulators.

Learning was done using 10 trials of collected data from various sentences as the teacher signal. Figure 3 shows the generalization results comparing predicted vertical acceleration of lips and jaw with the experimentally obtained acceleration for a test trial which was not included in the training (teacher) set. A concern about our previous work [8], in which a network succeeded in learning the forward dynamics relating EMG from 4 muscles and lip and jaw motion was that the model's success was due to the simplicity of the data set. We were also concerned that we might not be able to collect data from enough channels for the model to successfully generalize. As shown in Figure 3, the acquired model produces appropriate acceleration trajectories for real speech utterances, suggesting that utterance complexity is not a limiting factor in this approach.

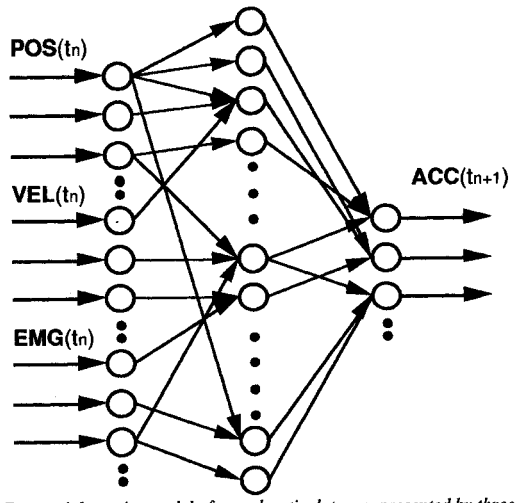

Fig. 2. Forward dynamics model of speech articulators represented by three-layer perceptron. Network inputs are articulator positions, velocities and EMGs, outputs are accelerations. Using real phisiological data as the teacher signals, error back propagation learning algorithm aquire the relation between input and output by updating weights between input and output through hidden layer neurons.

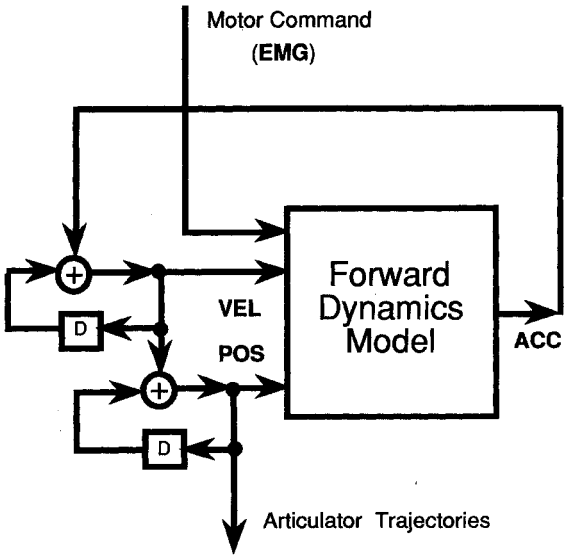

*Fig. 4. Network for transforming motor commands to articulator trajectories. The forward dynamics model, acquired by a three-layer perceptron, is incorporated into the recurrent network shown. Continuous motor command (EMG) input drives the network, which uses estimated acceleration at time **tn**, to predict new velocity (integration) and position (double integration) values at the next time step **tn+1**. **D** is a 1 sample (5 ms) delay unit. The network is initialized with position and velocity values taken from the test utterance at **t0**.*

3.2 Musculo-skeletal System Network

Using the learned model, trajectory prediction from the motor command input (rectified and integrated EMG) can be done by connecting the forward dynamics model recurrently as shown in Figure 4. The network uses only the initial articulator position and velocity values and the continuous EMG "motor command" input to generate predicted trajectories. The Forward Dynamics Model estimates the acceleration at time t_n, to predict new velocity (integration) and position (double integration) values at the next time step t_{n+1}.

3.3 Motor Command Generation

To generate continuous motor commands from discrete linguistic information, an optimization process is necessary. Phoneme-specific articulatory targets and global performance parameters such as speaking rate and style are given to the motor command generation network from the task specification networks (refer to Vatikiotis-Bateson et al., this volume, for the conceptual scheme of the task specification networks). In order to solve the one-to-many mapping problem that exists between discrete linguistic information and motor command generation, another constraint is necessary. Observed articulator movements are smooth. Their smoothness is due partly to physical dynamic properties (inertia, viscosity). Furthermore, smoothness may be an attribute of the motor command itself, thereby resolving the ill-posed computational problem of generating

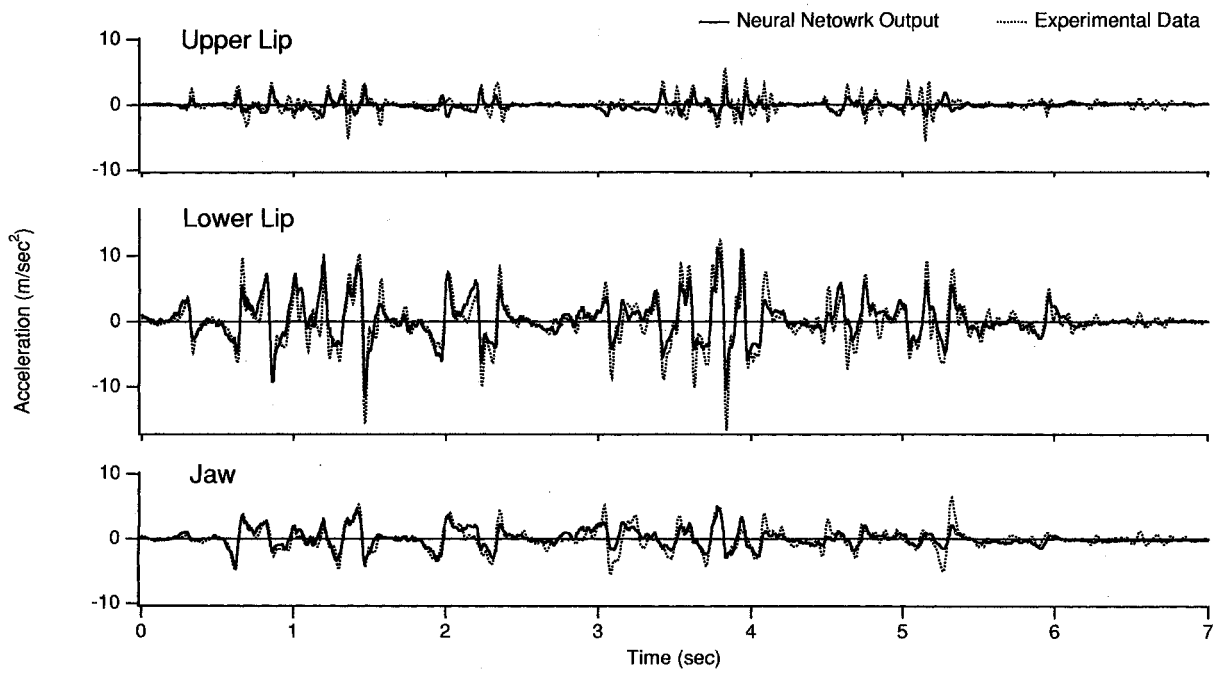

Fig. 3. Network predicted acceleration vs. experimental data. Estimated acceleration for vertical motion of the three articulators is compared to that of the test sentence: "Pam put the bobbin in the frying pan and added more puppy parts to the boiling potato soup".

continuous motor commands from a small number of discrete articulatory targets. Optimization based on objective functions have been used to model arm movement dynamics (e.g., Uno et al. [13][14]). Here, a biologically plausible smoothness constraint is used — the minimization of the motor command change.

Currently, a cascade neural network proposed by Kawato et al. [15] is used to generate motor commands and articulator trajectories from phoneme-specific articulatory targets and speaking rate, which is used to determine the appropriate setting of the smoothness constraint (minimum motor command change). This model calculates a motor command sequence to satisfy 2 types of constraints — articulatory target and minimum motor command change — using relaxation calculation with the internal model of the forward dynamics of articulators. For reiterant speech [8], this network generated smooth EMG and articulator trajectories whose spatiotemporal asymmetry approximated the prosodic patterning of the natural test utterances. Although this was only a preliminary implementation of via-point and smoothness constraints, the model's ability to generate trajectories of appropriate spatiotemporal complexity from a series of alternating via-point inputs is encouraging.

3.4 Forward Acoustic Network

The final stage of our speech production model entails using a neural network to acquire a model of the relation between articulator motion and the ensuing acoustics. Using articulator position as input, a 3-layer perceptron was used to learn PARCOR analysis and to generate appropriate PARCOR parameters [9] for subsequent speech synthesis. We chose PARCOR parameters, rather than more commonly used formant values, because the parameters have some relation to specific sections of the vocal tract — e.g., the first PARCOR corresponds to the cross-sectional area closest to the lips [10]. Also, PARCOR estimation errors have less radical consequences than formant estimation errors. Finally, there is a unique mapping from PARCOR to formant values, but not the reverse [9]. Without tongue data, accurate prediction of all 16 PARCOR parameters needed to model the vocal tract is not possible. However, the network does recover a surprising degree of correlation from only the "visible" lip and jaw data (Figures 5). When we finish analysis of tongue data for these utterances and apply an appropriate glottal source, this neural network model can be used for articulatory speech synthesis.

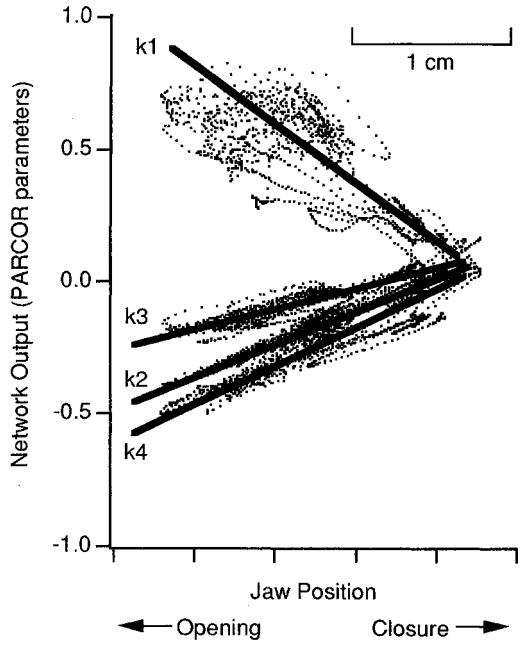

Fig. 5. Generalized relation between articulator position and PARCOR parameters. The first four PARCOR parameters corresponding to the anterior sectors of the vocal tract are plotted as a function of jaw position and show reasonable correspondence to changes in vocal tract area. Smaller values of jaw position correspond to lip-jaw opening for vowel production, while higher values correspond to lip-jaw closure for consonant constrictions.

IV. CONCLUSIONS

Our model's characteristics and virtues are 1) real rather than simulated physiological and kinematic data are used during network training to acquire the forward dynamics relating neuromotor activity and musculo-skeletal constraints on the ensuing articulatory behavior, 2) continuous movement behavior is modulated by serially-ordered spatial targets corresponding to the utterance-specific phoneme string, and 3) global performance parameters, such as speaking rate and style, determine the relative strengths of the smoothness and spatial target constraints. This enables examination of the effects of different performance constraints on various aspects of interarticulator coordination such as articulatory undershoot and blending.

Although our model is still preliminary and the training data are very limited, it is already capable of rudimentary articulatory synthesis from phoneme input strings via the motor command. Furthermore, the simulation results of PARCOR parameter generation from only lip and jaw movements suggests that such networks may be useful in audiovisual speech recognition tasks. As the data are extended to tongue motion and more appropriate muscles, we are confident that the quality of synthesized speech will improve. Ultimately, however, and critical to their use in recognition tasks, such models must resolved the difficulties posed by coarticulation, segmentation, prosody, and speaking style. These issues are of mutual importance to engineers and speech scientists, and our hope is that neural networks of the sort outlined here can help us better understand the computational physiological aspects of speech motor control.

Acknowledgment

We thank Yoh'ichi Toh'kura, Philip Rubin, Elliot Saltzman, Vincent Gracco, Michael I. Jordan for insightful discussion; Haskins Laboratories for use of their facilities (NIH grant DC-00121); Further support was provided by HFSP grants to M. Kawato.

References

[1] Jordan, M. I., "Serial order: a parallel distributed processing approach", *ICS (Institute for Cognitive Sciences, University of California) Report*, **8604**, 1986.

[2] Jordan, M. I., "Motor Learning and the Degrees of Freedom Problem", (M. Jeannerod editor) *Attention and Performance XIII*, pp. 796-836, Hillsdale, NJ, Erlbaum, 1990.

[3] Saltzman, E. L., "Task dynamics coordination of the speech articulators: A preliminary model", *Experimental Brain Research*, Series 15, pp. 129-144, 1986.

[4] Laboissière, R., Schwarz, J. L. & Bailly, G. "Motor Control for Speech Skills: a Connectionist Approach", *Proceeding of the 1990 Summer School*, Morgan Kaufmann Publishers, pp. 319-327, 1990.

[5] Bailly, G., Laboissière, R. & Schwarz, J. L., "Formant trajectories as audible gestures: an alternative for speech synthesis, *Journal of Phonetics*, **19**, pp. 9-23, 1991.

[6] Bengio, Y., Houde, J., & Jordan, M. I., "Representations Based on Articulatory Dynamics for Speech Recognition", Presented at Neural Networks for Computing, Snowbird, Utah, 1992.

[7] Saltzman, E. L., & Munhall, K. G., "A dynamic approach to gestural patterning in speech production", *Ecological Psychology*, **1**, pp. 333-382, 1989.

[8] Hirayama, M., Vatikiotis-Bateson, E., Kawato, M., & Jordan, M. I., "Forward dynamics modeling of speech motor control using physiological data", In Moody, J. E., Hanson, S. J., and Lippmann, R. P. (eds.) *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann Publishers, 1992.

[9] Itakura, F., Saito, S., "Speech analysis and synthesis by partial correlation parameters", Proceeding of Japan Acoust. Soc., **2-2-6**, 1969.

[10] Wakita, H., "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", *IEEE Trans. Audio Electroacoust.*, vol. **AU-21**, pp. 417-427. 1973.

[11] Vatikiotis-Bateson, E., & Ostry, D. J., "Rigid body reconstruction of jaw motion in speech", To be presented at 124th meeting of the Acoustic Society of America, New Orleans, LA, 1992.

[12] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. "Learning Internal Representation by Error Propagation", *Parallel Distributed Processing* Chap. 8, MIT Press, 1986.

[13] Uno, Y., Kawato, M., & Suzuki, R. "Formation and Control of Optimal Trajectory in Human Multijoint Arm Movement", *Biol. Cybern.* **61**, pp. 89-101, 1989.

[14] Uno, Y., Suzuki, R. & Kawato, M. "Minimum muscle-tension-change model which reproduces human arm movement", *Proceedings of the 4th symposium on Biological and Physiological Engineering*, pp. 299-302, 1989, in Japanese.

[15] Kawato, M., Maeda, M., Uno, Y. & Suzuki, R. "Trajectory Formation of Arm Movement by Cascade Neural Network Model Based on Minimum Torque-change Criterion", *Biol. Cybern.* **62**, pp. 275-288, 1990.