TRAJECTORY FORMATION OF ARTICULATORY MOVEMENTS FOR A GIVEN SEQUENCE OF PHONEMES

OKADOME, Takesi *Tokihiko Kaburagi* *Masaaki Honda*
NTT Basic Research Laboratories
3-1 Morinosato Wakamiya, Atugi-Si, Kanagawa, 243-0198 Japan
e-mail: houmi@idea.brl.ntt.co.jp

ABSTRACT

The method proposed here produces trajectories of articulatory movements based on a kinematic triphone model and the mimimum-jerk model. The kinematic triphone model, which is constructed from articulatory data obtained in the experiments through the use of a magnetic sensor system, is characterized by three kinematic features for a triphone and intervals between two successive phonemes in the triphone. After extracting a kinematic feature for a phoneme in a given sentence, for each point on the articulator, the minimum-jerk trajectory which coincides with the extremum of the time integral of the square of the magnitude of jerk of the point is formulated, which requires only linear computation. The method predicts both the qualitative features and the quantitative details experimentally observed.

1. INTRODUCTION

Articulatory-based speech-synthesis requires high-fidelity generation of articulatory behavior. The articulatory organs constitute a multiple degrees-of-freedom system and thus the phoneme-specific tasks related to the production of vocal-tract shapes can be shared by different articulators. Furthermore, an infinite number of trajectories of the articulator can achieve phoneme-specific tasks aligned in order of time. To cope with these redundancies and to determine the articulator movements uniquely, we require additional constraints.

This article proposes a method for forming trajectories of articulatory movements, where phoneme-specific tasks are specified by using a classical context-sensitive coding method (for example, Wickelgren, 1969) and trajectories are uniquely determined by minimizing a cost function. As the classical context-sensitive coding method, we develop a kinematic triphone model which is described in Section 2. As a cost function, we adopt the time integral of the square of the magnitude of jerk (the time derivative of acceleration) of each point on the articulator. This article deals only with speech production in normal speed.

2. A TRIPHONE MODEL

The triphone model presented here, which is called a *kinematic triphone model*, is characterized by intervals between two successive phonemes in a triphone and three kinematic features for the triphone; each kinematic feature is defined for each phoneme contained in the triphone. A *kinematic feature* for a phoneme is represented by the position, velocity, and acceleration of each point on an articulator.

To construct the kinematic triphone model, we used articulatory data obtained in the experiments through the use of a magnetic sensor system in which a single subject read 354 sentences. In the experiments, we observed 9 points on the articulator with 250Hz sampling in both the vertical and horizontal orientations.

For the observed data, we first did the time alignment for each phoneme. The time alignment was done by putting a marker to the time at which the kinematic feature of each phoneme was most remarkably seen. For example, we put the marker for /b/ to the time at which the lips are closed. We call the time aligned for a phoneme the *articulation time for the phoneme*. Using 338 sentences of the 354, we calculated the position, velocity, and acceleration of the 9 points on an articulator for each phoneme of triphones. Then, for each triphone and for each phoneme contained in the triphone, we calculated the average of positions and the median values of velocities and accelerations of the 9 points on the articulator.

As phonemic symbols of Japanese, we used 40 kinds of phonemes and two special symbols which represent the articulation start and end, respectively. The 338 training sentences contained 11154 phonemes and 2460 triphones in all. To evaluate our method, we used the remaining 16 sentences which had 507 triphones. The 338 training sentences did not contain 31 triphones of the 507 triphones in the test sentences (the coverage rate: 93.89%).

3. PRODUCING TRAJECTORIES

3.1. Kinematic Feature Extraction

Our method for producing trajectories extracts three kinematic features for each phoneme in a given sequence of phonemes on the basis of the triphone model because each phoneme in the sequence is contained in three successive triphones. The kinematic feature for each phoneme in the sequence is determined to be the weighted average of the three kinematic features. Let (x_j^*, v_j^*, a_j^*) be the triple of the position, velocity, and acceleration of the jth point on the articulator. For a sequence of phonemes $p_1 p_2 \cdots p_n$, the kinematic feature of a phoneme

Table 1. Distances between predicted and observed trajectories (mm).
The times aligned for observed data are used as the articulation time for each phoneme in the sentences.

	s1	s2	s3	s4	s5	s6	s7	s8
average	1.67	1.66	1.44	1.44	1.52	1.63	1.44	1.99
maximum	7.51	7.71	6.51	7.57	5.31	6.85	8.10	9.47
	s9	s10	s11	s12	s13	s14	s15	s16
average	1.38	1.50	1.44	1.65	1.60	1.67	1.60	1.46
maximum	7.50	7.08	8.05	8.04	6.96	7.68	9.80	8.66

Table 2. Distances between predicted and observed trajectories (mm),
in which we specify only the position of each phoneme as a kinematic feature.

	s1	s2	s3	s4	s5	s6	s7	s8
average	2.09	1.76	1.62	1.63	1.62	2.08	1.73	2.20
maximum	12.70	9.04	7.47	7.45	7.13	12.03	8.13	11.39
	s9	s10	s11	s12	s13	s14	s15	s16
average	1.43	1.79	1.73	1.80	1.69	1.82	1.94	1.57
maximum	5.25	9.90	8.71	7.89	7.68	7.62	20.86	6.91

p_i, $1 \leq i \leq n$, is given by the following expression:

$$(x_j^*, v_j^*, a_j^*) = \frac{(x_j, v_j, a_j) + 4 \cdot (x_j', v_j', a_j') + (x_j'', v_j'', a_j'')}{6},$$

where (x_j, v_j, a_j) is the kinematic feature of p_i for triphone $p_{i-2}p_{i-1}p_i$ in the kinematic triphone model, (x_j', v_j', a_j') is that for triphone $p_{i-1}p_ip_{i+1}$, and (x_j'', v_j'', a_j'') is that for triphone $p_ip_{i+1}p_{i+2}$.

Likewise, the interval times between two successive phonemes in the sequence are also determined on the basis of the triphone model. That is, the interval time between p_i and p_{i+1}, $1 \leq i \leq n-1$, is given by the following expression: $t_i^* = \frac{(t_i + t_i')}{2}$, where t_i is the interval time between p_i and p_{i+1} for triphone $p_{i-1}p_ip_{i+1}$ in the kinematic triphone model; and t_i' is the interval time between p_i and p_{i+1} for triphone $p_ip_{i+1}p_{i+2}$ in the kinematic triphone model.

3.2. Minimum-Jerk Trajectories

Using each kinematic feature as a constraint, we can formulate the trajectory by calculating the *minimum-jerk trajectory* (Flash & Hogan, 1985) for each point on the articulator which coincides with the extremum of the following cost function:

$$\frac{1}{2}\int_0^{t_f} \left(\left(\frac{d^3x}{dt^3}\right)^2 + \left(\frac{d^3y}{dt^3}\right)^2\right)dt, \qquad (1)$$

where (x, y) are the time-varying Cartesian coordinates on the sagittal plane of the point on the articulator. To find the trajectory that optimizes the cost function, we use a variational calculus method and a dynamic optimization theory (Pontryagin, et al., 1962) which allows us to obtain a set of linear differential equations. Solving the set of linear differential equations gives us a piecewise polynomial function of time (for details, see the Appendix). The kinematic features are used in the linear computation, which determines the coefficients of the piecewise polynomial function. Thus, the trajectory formation method produces trajectories simply by extracting kinematic features and by linear computation.

4. EVALUATION OF THE METHOD

The trajectory formation method was evaluated for 16 test sentences. Table 1 shows the average and maximum distances between predicted and observed trajectories in which the times aligned for observed data were used as the articulation time for each phoneme in the sentences. The average distances between the observed and predicted trajectories were 1.38 to 1.99 mm, which are compatible with the average distances between trajectories of articulatory movements such as when the subject reads a sentence twice. Figure 1 shows an example of the predicted and observed trajectories of the articulatory organs, where the times aligned for observed data were used as the articulation time for each phoneme in the sentences. Incidentally, Table 2 shows the average and maximum distances between the predicted and observed trajectories in which, as another triphone model, we specify only the position of each phoneme as a kinematic feature. The average distances between the minimum-jerk and observed trajectories are larger than those for the original triphone model.

Table 3 lists the distances between the predicted and observed trajectories, where. the estimated times were used as the articulation time for each phoneme in the sentences. We calculated the distances after carrying out a time adjustment by using a DP matching technique. Again, we can see that the average distances between the

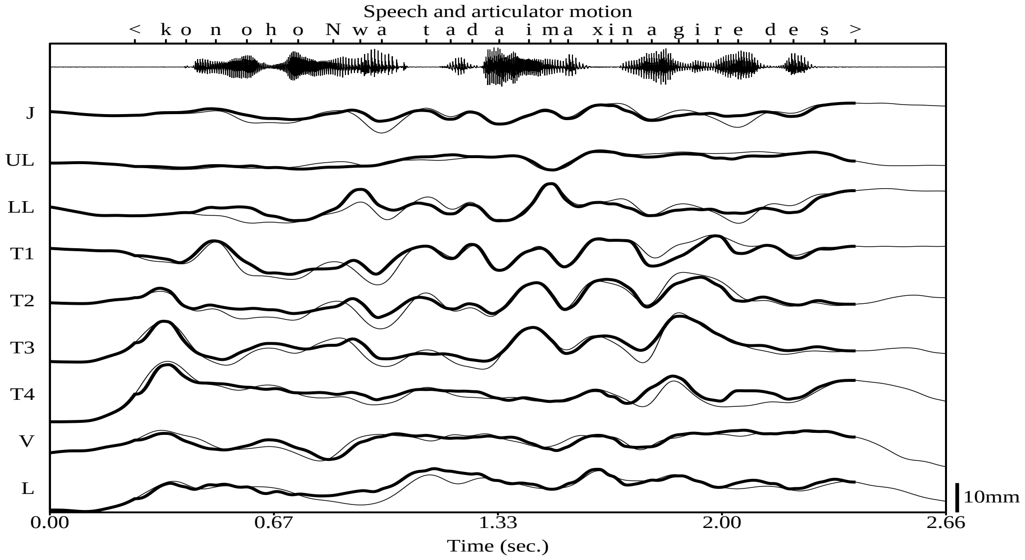

Figure 1. Predicted (thick lines) and observed (thin lines) trajectories. The times aligned for the observed data are used as the articulation time for each phoneme in the sentences. The top trace is the speech waveform, the others are the movements of the jaw (J), the upper lip (UL), the lower lip (LL), the tongue points (T1, T2, T3, and T4), the velum (V), and the larynx (L).

observed and predicted trajectories are compatible with the average distances between trajectories of articulatory movements such as when the subject reads a sentence twice. Figure 2 shows an example of the predicted and observed trajectories of the articulator, in which the estimated times were used as the articulation time for each phoneme in the sentences. We can see that the method predicts the quantitative details experimentally observed.

Furthermore, we see that the method predicts the qualitative features of the observed trajectories. For example, the method is good for predicting the characteristics of the articulator that are consistent for each consonant because of the small variability of the articulatory configuration. See, for example, the position of the tongue tip (T1) for the consonants /*t*/ and /*d*/ and that of the dorsum (T3 and T4) for the consonats /*k*/ and /*g*/. in Figure 1. The method is also particularly good for predicting the fast motion in the release of occlusion for stop consonants. Again see the position of the tongue tip (T1) for the consonants /*t*/ and /*d*/.

To evaluate the interval time estimation, we conducted another experiment in which the subject read 16 sentences, each of them 15 times. For each diphone, we calculated the observed range of interval times between two succesive phonemes in the diphone as a reference of error estimation of interval times. The results of the experiments show that 70 percent of the predicted interval times were inside the ranges. The predicted interval times for diphones, including long vowels or geminative consonants, are relatively out of range.

5. CONCLUSION

This article has presented a method for producing articulator movements for continuous speech utterances at a normal speed. In the method, each phoneme-specific task is specified by a kinematic triphone model constructed on the basis of the experimental data using a magnetic sensor system and the trajectories of articulator movements are determened by minimizing the jerk of each point on articulator organs. The method predicts both the qualitative features and the quantitative details experimentally observed.

REFERENCES

1. Flash, T. and N. Hogan (1985). The coordination of arm movements: an experimentally confirmed mathematical model. *The Journal of Neuroscience*, **5**, 1688-1703.
2. Pontryagin, L. S., V. G. Boltyanski, R. V. Gamkrelidze, and E. F. Mischenko (1962). *The Mathematical Theory of Optimal Processes*, Interscience Publishers, Inc., New York.
3. Wickelgren, W. A. (1969). Contex-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, **76**, 1-15.

APPENDIX

Let $\xi = (x, y)$ be the time-varying Cartesian coordinates of a point p on a system. We determine the trajectory which minimizes

$$\frac{1}{2}\int_0^{t_f}\left(\left(\frac{d^3x}{dt^3}\right)^2+\left(\frac{d^3y}{dt^3}\right)^2\right)dt,$$

Table 3. Distances between predicted and observed trajectories (mm). The estimated times are used as the articulation time for each phoneme in the sentences.

	s1	s2	s3	s4	s5	s6	s7	s8
average	1.75	1.83	1.35	1.50	1.53	1.67	1.53	1.99
maximum	7.64	7.15	4.52	7.43	5.30	7.11	8.03	7.13
	s9	s10	s11	s12	s13	s14	s15	s16
average	1.49	1.77	1.82	1.84	1.75	1.83	1.35	1.51
maximum	5.50	6.76	8.57	8.15	7.64	7.15	4.52	7.16

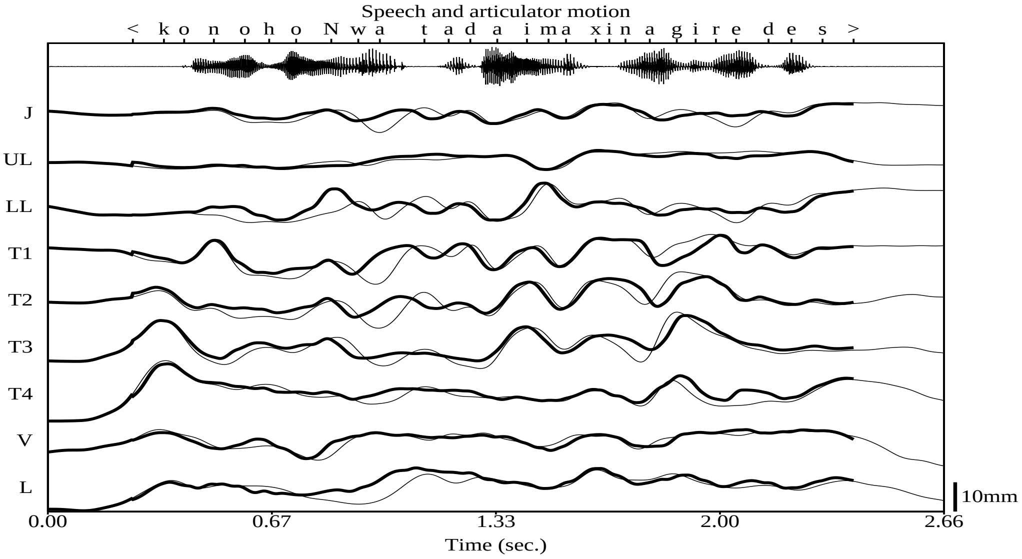

Figure 2. Predicted (thick lines) and observed (thin lines) trajectories. The estimated times are used as the articulation time for each phoneme in the sentences. The top trace is the speech waveform, the others are the movements of the jaw (J), the upper lip (UL), the lower lip (LL), the tongue points (T1, T2, T3, and T4), the velum (V), and the larynx (L).

where the time interval $[0, t_f]$ is divided into $t = t_0$, t_1, t_2, ..., and $t_n = t_f$ and at each t_i, $i = 0$, ..., n, the position (x_i, y_i), the velocity $(\dot{x}_i, \dot{y}_i)$, and the acceleration $(\ddot{x}_i, \ddot{y}_i)$ of p are given. In general, for a cost function $L[t,\ \xi,\ \dot{\xi},\ \ldots,\ d^n\xi/dt^n]$, the trajectory $\xi(t)$ which minimizes $\int_{T_1}^{T_2} L\left[t, \xi, \dot{\xi} \ldots, \frac{d^n\xi}{dt^n}\right] dt$ satisfies the following Euler-Poisson equation:

$$\frac{\partial L}{\partial \xi} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\xi}}\right) + \cdots + (-1)^n \frac{d^n}{dt^n}\left(\frac{\partial L}{\partial \xi^{(n)}}\right) = 0,$$

where $\xi^{(n)} = d^n\xi/dt^n$. In case of $L = \frac{1}{2}((d^3x/dt^3)^2 + (d^3y/dt^3)^2)$, we obtain

$$\frac{d^3}{dt^3}\left(\frac{\partial (x^{(3)})^2}{\partial x^{(3)}}\right) + \left(\frac{\partial (y^{(3)})^2}{\partial y^{(3)}}\right) = 0$$

and thus

$$\frac{d^6 x}{dt^6} = 0, \qquad \frac{d^6 y}{dt^6} = 0.$$

Solving this equation brings to us the following time-varing functions:

$$\begin{aligned} x(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_6 t^5, \\ y(t) &= b_0 + b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_6 t^5. \end{aligned}$$

If, as constraints, we give $x(T_1)$, $\dot{x}(T_1)$, $\ddot{x}(T_1)$, $x(T_2)$, $\dot{x}(T_2)$, $\ddot{x}(T_2)$, $y(T_1)$, $\dot{y}(T_1)$, $\ddot{y}(T_1)$, $y(T_2)$, $\dot{y}(T_2)$, and $\ddot{y}(T_2)$, then we can determine the coefficients a_0, ..., a_5, b_0, ..., and b_5 uniquely. Hence, for each interval $[t_i, t_{i+1}]$, a trajectory which satisfies (x_i, y_i), $(\dot{x}_i, \dot{y}_i)$, and $(\ddot{x}_i, \ddot{y}_i)$, at each t_i, $i = 0, \ldots, n$, and minimizes $L = \frac{1}{2}((d^3x/dt^3)^2 + (d^3y/dt^3)^2)$ is uniquely determined. Clearly, the trajectory constructed by piece-wisely joining these trajectoies is the solution to our problem.