Acoust. Sci. & Tech. **25**, 5 (2004)

PAPER

Investigation of coarticulation in continuous speech of Japanese

Jianwu Dang[1,3,4,*], Masaaki Honda[2,†] and Kiyoshi Honda[3,‡]

[1]*Japan Advanced Institute of Science and Technology, Japan*
[2]*School of Sport Science, Waseda University, Japan*
[3]*ATR Human Information Science Laboratories, Japan*
[4]*ICP CNRS UMR 5009 & INPG & University Stendhal, France*

(*Received 31 July 2003, Accepted for publication 16 February 2004*)

Abstract: This study analyzed coarticulation involved in continuous speech based on articulographic data obtained from three Japanese male speakers. Distribution of the articulation points of vowels and consonants revealed that speakers might compensate for morphological differences in the hard palate by adjusting the location of vowel articulation points. To evaluate the effects of the left and right phonemes on the target phoneme, three-phoneme sequences, consisting of five Japanese vowels (V) and ten apical and two palatal consonants (C), were extracted from read sentences and used in this analysis. A stepwise multiple regression method was used to analyze the phoneme sequences, in order to evaluate the "contributions" of the surrounding phonemes to the central one. The results showed that the horizontal component of the articulatory movement had a dominant function during articulation. The movement of the tongue tip was highly correlated to the tongue dorsum movement in the horizontal dimension, but was almost independent in the vertical dimension. This phenomenon suggested that coarticulation in VCV sequences can be considered as an independent consonantal gesture superimposed on a transitional portion between vowels. For apical-vowel combinations, the preceding consonant in CVC had a stronger effect on the vowel than the following one, but there was no dominance caused by the positions of the vowels in VCV sequences. For palatal-vowel combinations, the following phoneme showed a greater effect than the preceding phoneme in CVC sequences. In VCV sequences, the open and closed vowels showed different behaviors.

Keywords: Speech production, Coarticulation, Speech dynamics, Articulatory movement, Speech analysis

PACS number: 43.70.Bk, 43.72.Ar, 43.70.Fq [DOI: 10.1250/ast.25.318]

1. INTRODUCTION

Coarticulation is a natural phenomenon found in human speech, and is affected both at the physiological level by the properties of the speech organs, and in the motor planning stage by the look-ahead mechanism. Coarticulation generates the characteristics of natural sounding speech, and at the same time introduces so-called supra-segmental characteristics that degenerate phonemic boundaries. It is therefore a crucial issue in both speech synthesis and speech recognition to deal with coarticulation when performing human speech behaviors.

In general, there are two types of coarticulatory overlaps occurring during natural speech: the left-to-right (LR, carryover) and right-to-left (RL, anticipatory). The former reflects the sequence as the tongue, mandible, and lips move from the preceding phoneme toward the following one. In this case, the target of this phoneme is reached in different ways, depending on the status of the previous phoneme. Anticipatory (RL) coarticulation can occur only if the speaker can "look ahead" in time and anticipate oncoming sounds. While LR coarticulation occurs at the physiological level, RL coarticulation reflects a high-level (central) type of phonological-phonetic processing, since an entire segment must be scanned before it is articulated [1]. To describe this process, Henke proposed a phonemic-segment model [2]. Each segment with multiple phonemes was described in a matrix of articulatory target features, in which some features change abruptly as the targets vary from one to another.

Öhman investigated coarticulation in VCV utterances using spectrographical measurements, in which three stop

*e-mail: jdang@jaist.ac.jp
†e-mail: hon@waseda.jp
‡e-mail: honda@atr.jp

consonants (C) and five vowels (V) were employed [3]. He concluded that VCV articulation could be represented by a basic diphthongal gesture with an independent consonantal gesture superimposed on its transitional portion. Kiritani observed articulatory movements for a Japanese speaker using the X-ray microbeam system and investigated coarticulation in VCV and CVC sequences within a carry sentence [4]. He found that in CVC, the effect of the preceding consonant on the vowel articulation was larger than that of the following one, regardless of the vowel types, while the following vowels in VCV sequences showed larger effects than the preceding ones. Recasens *et al.* employed electropalatographic data and the trajectory of the second formant to investigate anticipatory and carryover effects within VCV sequences, consisting of seven consonants and two vowels /a/ and /i/ [5]. They found that the effects were dependent on the degree of articulatory constraint (DAC) of two adjacent phonemes, where the DAC was defined as having three levels. They also tried to model coarticulation between the consonants and vowels using the DAC.

The dynamic characteristics of the speech organs have been observed using X-ray cinematography, X-ray microbeam system, and electromagnetic sensing systems, in order to systematically analyze and model speech production mechanisms. Shirai and Honda, for example, analyzed X-ray data obtained from five Japanese male speakers, and built a statistical model with five parameters, including the jaw and tongue [6]. Maeda analyzed articulatory movements obtained from an X-ray film of a French female speaker and constructed a statistical articulatory model with eight parameters [7]. Dang and Honda have developed a physiological articulatory model with a target-based control strategy [8–10]. The model was constructed based on volumetric MR images and anatomic data, and is driven by physiological mechanisms. The model is expected to naturally realize the coarticulation caused by physiological factors. However, the coarticulation derived from high-level processing, such as anticipation in the target planning stage, has not yet been taken into account since there is insufficient knowledge for quantifying such effects.

In this study, we attempt to clarify and quantify the coarticulation involved in continuous speech based on articulatory data obtained from the electromagnetic articulographic system. Distribution of the articulatory points is first investigated for vowels and consonants, to obtain a uniform articulatory coordinate. We then evaluate contributions of observation points to a target phoneme using a stepwise multiple regression analysis. Finally, we investigate the anticipatory (RL) and carryover (LR) effects of the coarticulation separately by measuring changes in the tongue shapes that resulted from the left and right phonemes.

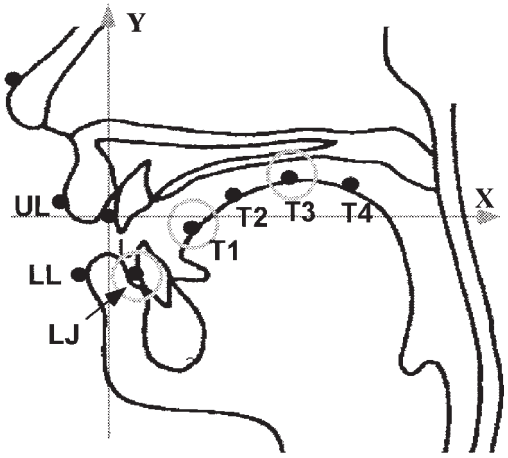

Fig. 1 Placement of the receiver coils in the electromagnetic articulographic experiment, and the coordinate system used in this study. The gray circles show the observation points in the target vector.

2. ARTICULATORY DATA AND LABELING

The articulatory data used in this study were collected using the electromagnetic articulographic (EMA) system at NTT Communication Science Laboratories, Japan [11]. Figure 1 shows the placement of the receiver coils in the experiment. Four receiver coils were placed on the tongue surface in the midsagittal plane, from the tongue tip named $T1$ through $T4$, and one coil each was placed on the upper lip, lower lip, maxillary incisor, and mandible incisor (LJ). The sampling rate was 250 Hz for the articulatory channels and 16 kHz for the acoustic channel. Speech materials comprised 360 Japanese sentences read by three adult male Japanese speakers at a normal speech rate. The acoustic data and the articulatory data were recorded simultaneously.

The central point of the phonemes was first labeled manually, cross-referencing the acoustic cues to the articulatory cues. The label location was then refined automatically by finding the steady point of the articulators: $T1$ was used for labeling apical consonants, and $T3$ for the others [11].

3. ANALYSIS OF SPEECH DYNAMICS

We first performed a preliminary analysis to investigate effects of the individual morphological differences on the speech dynamics. In the articulatory observation, each phoneme was represented by a set of vectors with eight points corresponding to the receiver coils. Phonetically, a phoneme can be roughly described by a constriction location in the vocal tract, which is usually caused by a part of the articulators (*cf.* [12]). Therefore, it is not always necessary to account for all the observation points. With this in mind, we focus on three points in this study: the jaw (LJ), tongue tip ($T1$), and tongue dorsum ($T3$), which were

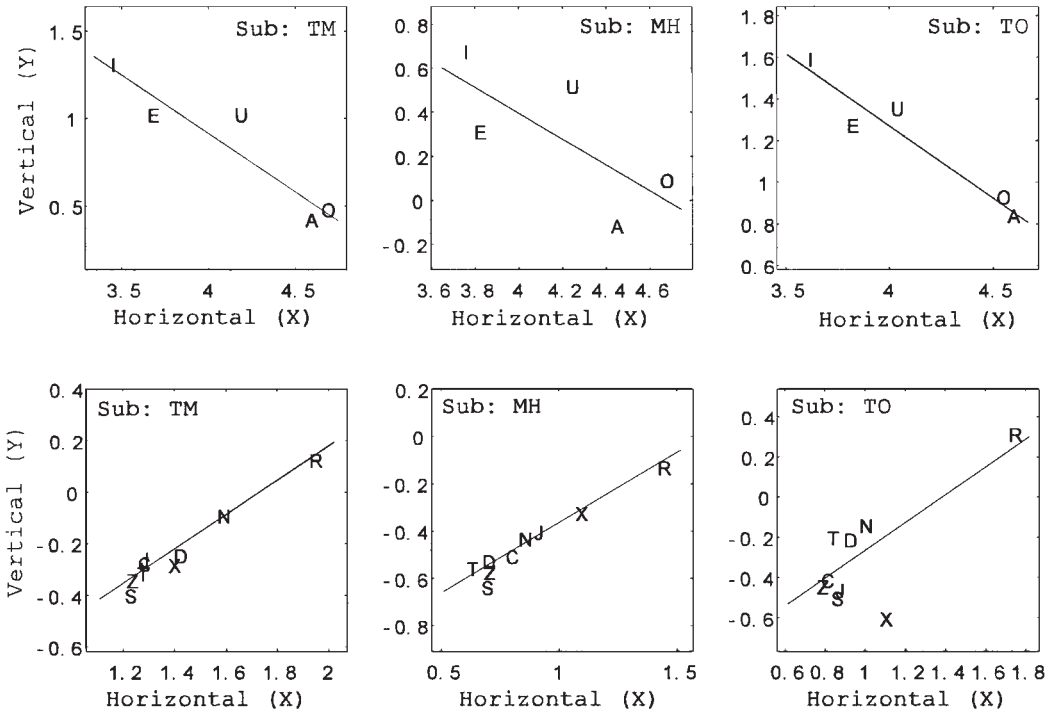

Fig. 2 Articulatory coordinates of vowels and consonants. The upper panels show the mean location of the crucial points ($T3$) for vowels using upper case characters; the flowerer panels are for the apical consonants ($T1$). The lines are the optimal fitting to the scattering distributions based on the least mean square method (Dimension: cm).

used to control our physiological articulatory model [8–10]. These three points are represented in Fig. 1 by the circles.

3.1. Articulatory Coordinate for Vowels and Consonants

Among the above three points, usually only one is crucial for constructing a particular phoneme; the other points are indecisive. Thus, the relation among the phonemes can be analyzed using a distribution of the crucial points (CP). In phonetic literature, distribution of the CPs is commonly described based on isolated utterances. However, CP distribution can vary during continuous speech because of coarticulation. This study begins with an investigation into the distribution of the CPs of the tongue in continuous speech. Here, $T1$ is used to define the CP for ten apical Japanese consonants (including alveolar and post-alveolar consonants), and $T3$ is used for vowels and the palatal consonants /k/ and /g/. The bilabial and nasal consonants are not dealt with in this paper because they are anatomically independent of the tongue. The locations of the CPs were calculated by averaging the coordinates over a segment of 20 ms, 6 samples of EMA data surrounding the central point of the phoneme. For each speaker, the dorsal CP was collected from about 1,600 segments for /a/, 1,200 for /i/, 900 for /u/, 800 for /e/ and 1,200 for /o/. The phoneme number included in this data set varied from 50 to 300 for the consonants.

Figure 2 shows the mean locations of CP for the vowel ($T3$) in the upper panels and for the apical CP ($T1$) in the lower panels for three speakers. The coordinates correspond to those in Fig. 1. The upper case characters denote the mean location of the CP for vowels and consonants, which is the average over the selected segments. The main axis of the scattered CPs, denoted by the lines, was calculated using a least mean square fitting approach. As seen in the plots, the mean CPs for both vowels and consonants have a roughly linear distribution, although the scattering of the dorsal CPs is wider for speaker MH than the others, and the scattering of the apical CPs is wider for TO. The in-line distribution of the CPs in continuous speech differs significantly from the common distributions for vowel systems, which have been plotted as a triangular or trapezoidal shape. This difference can be explained by the fact that in continuous speech the tongue body moves in a shortened path to achieve a minimal energy, rather than a path with an exaggerated triangular distribution seen in isolated words.

The fitting lines show the main direction of the movements of the tongue tip and the dorsum. The lines of the consonantal and vocalic movements have an intersection with an acute angle by lengthening them in articulatory space, which is named an *articulatory coordinate*. It is interesting to find that the angles between the directions of the consonantal and vocalic movements are consistent with one another for all three speakers; angles

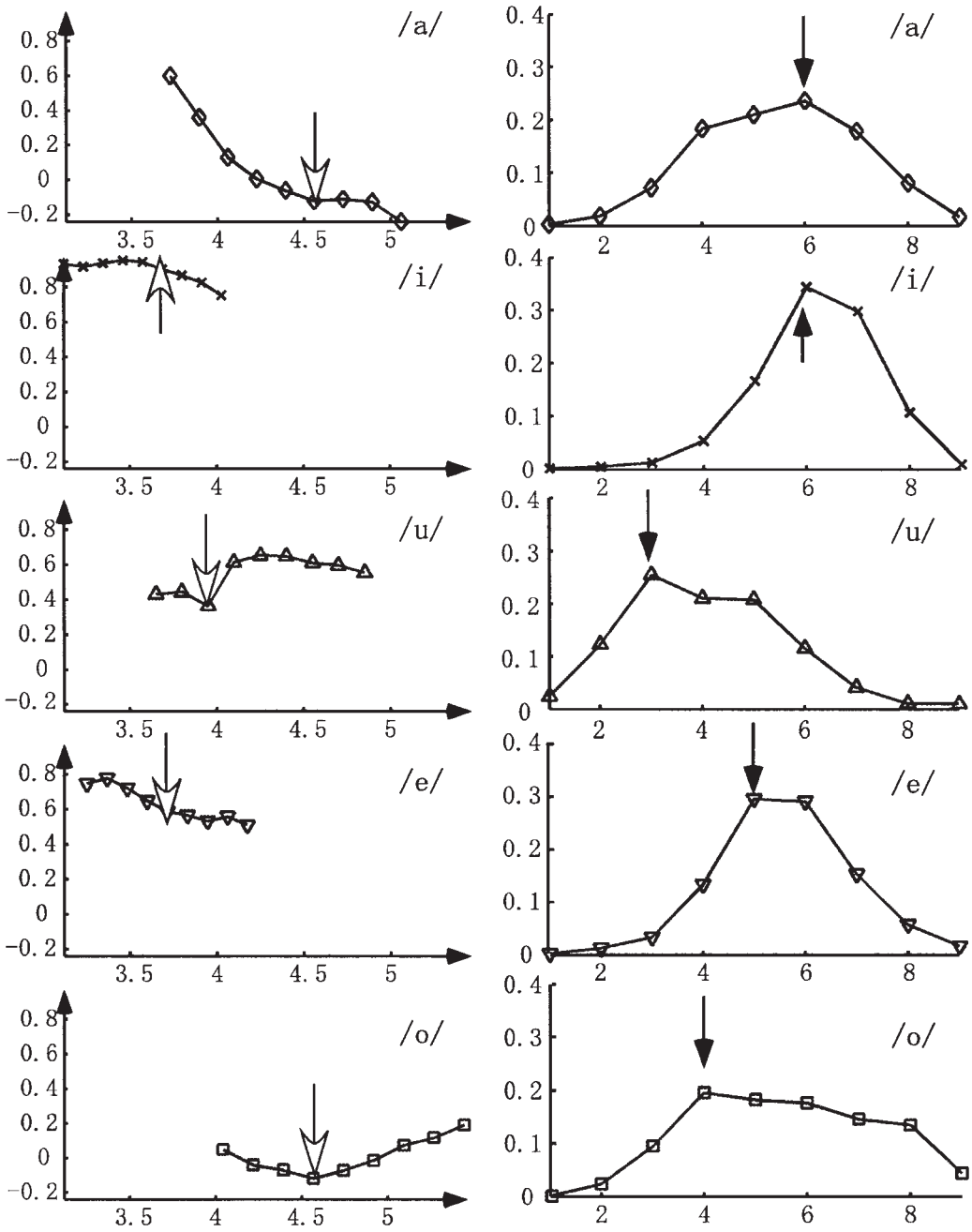

Fig. 3 Distribution of the CP (*T*3) of five Japanese vowels from one speaker. The left panels show the location of the CPs in the articulatory space (dimensions in cm); the right panels show the relative frequency of the CP's appearance in each location (section number).

were 61, 67, and 68 degrees, respectively. The articulatory movement of the tongue tip to form an apical CP is heavily dependent on the individual shape of the hard palate, where slopes of the anterior portion of the hard palate for the three speakers ranged from about 10 to 30 degrees referring to the bite-plate plane. The presence of consistent angles demonstrates that for a speaker who has a flatter hard palate, the movement direction of the tongue body is more tilted in vocalic articulation. This reveals the fact that in articulation, speakers adjust the position of the CPs to compensate for morphological differences in their hard palate shapes. After this compensation, an optimal articulatory space (the consistent coordinate) is achieved.

In processing such data, the head angle is usually calibrated using the bite plate to obtain a compatible coordinate between speakers. Since the shape of the hard palate was not considered, that calibration is not always optimal for constructing a compatible coordinate. For example, the main direction of the CP of the tongue tip lay almost horizontally for one of the speakers. In that situation, it was difficult to obtain a reliable correlation of the vertical and horizontal directions. In contrast, the proposed articulatory coordinate provides a proper reference for calibrating the speakers' head during the EMA

experiments. The calibration using the proposed coordinate derives a maximum projection on the horizontal axis for both the vocalic and consonantal movements, as shown in Fig. 2. In the following part of the paper, "horizontal" and "*X*" indicate the same axis, where the former is mainly used to describe the spatial relationship and the latter mainly for the data component. The same distinction is used for "vertical" and "*Y*."

3.2. Variation of Articulation and Acoustics

Because of coarticulation, both the left and right phonemes affect movements of the articulators. For this reason, phonemes in continuous speech usually show larger variations in articulatory and acoustic characteristics than in isolated words. This section investigates the variations in articulatory and acoustic characteristics of the vowels. Figure 3 shows the distribution of the tongue dorsum CP (*T*3) of five Japanese vowels spoken by one speaker. As seen in the left panels, the CP during continuous speech shows a wide distribution in the horizontal dimension, ranging from 0.9 to 1.4 cm for the five vowels. The vertical distribution spans less than 0.3 cm for most vowels, while it is 0.8 cm for /a/. This implies that the tongue moves more widely in the horizontal dimension than in the vertical.

To examine the relationship between articulatory and acoustic characteristics, the distribution of the CP is segmented into nine equal sections in the horizontal dimension. The appearance frequency of the CPs in each section is defined by the ratio of the appearance number in this section to the total number for the vowels, and shown in the right panels. The black arrows indicate the section with the highest appearance frequency, and the white arrows denote the corresponding location of the CPs.

Figure 4 shows the first three formants (F_1–F_3) of the vowels, which correspond to the nine sections of the CPs. The second formant (F_2) varies largely with the CP locations while the first formant (F_1) shows little change.

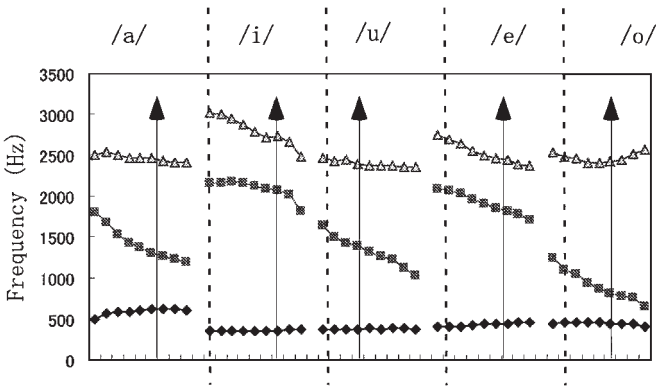

Fig. 4 Variation of the formants with the CP location for five Japanese vowels.

Consistent results were confirmed for the other speakers. Roughly speaking, F_1 mainly corresponds to the degree of mouth opening, and F_2 heavily depends on the anterior-posterior location of the tongue. The F_1 without obvious changes implies that there are no significant changes in the degree of mouth opening. This is consistent with one of the articulatory characteristics for Japanese speakers, where movements of the lips are relatively small. On the other hand, the large horizontal movement of the tongue results in an F_2 with a large variation.

4. REGRESSION ANALYSIS OF COARTICULATION

According to previous studies [1,13], the anticipatory effects sometimes can span up to three phonemes. For example, the velum possibly starts moving toward its target at the third vocalic gesture before a nasal sound. To our knowledge, there has been no report on the spanning range of the carryover effects. In past studies [3–5], the coarticulation was focused on the interval between the immediately adjacent phonemes. Following those studies, we consider the effects of vowels on consonants and of consonants on vowels only for the immediately adjacent (preceding and following) phonemes. Accordingly, VCV and CVC sequences were adopted in this study. The consonants were classified as two groups according to their CPs: the apical consonants and palatal consonants.

4.1. Analysis Method

From the articulation point of view, the tongue tip and dorsum can be considered as two independent CPs for consonants and vowels, respectively. In fact, the movement region of the tongue tip is constrained by the position of the dorsum (tongue body), while the dorsum would follow the tongue tip to construct a constriction at the apex. This physical connection between the tongue tip and the dorsum cannot be ignored when examining the relation between the observation points. For this reason, we consider the spatial (physiological) constraints of the articulators as well as the temporal effects of the phoneme sequences in the coarticulation analysis.

Since there is a high correlation between the jaw and tongue, and between the tongue tip and dorsum, one observation point can be described as a function of the others. Accordingly, the CP coordinates are assumed to be a linear function of the other factors, and the multiple regression method is employed to evaluate the contribution of each factor to the resultant articulatory movement of the CP. Formulas (1) and (2) describe the *X*- and *Y*-components of the CP for the tongue dorsum of the central phoneme, respectively,

$$
\begin{aligned}
T3_x(i) = c &+ \sum_{n=i-1}^{n=i+1} w_j(n) LJ_x(n) + \sum_{n=i-1}^{n=i+1} w_t(n) T1_x(n) + \sum_{\substack{n=i-1 \\ n\neq i}}^{n=i+1} w_d(n) T3_x(n) \\
&+ \sum_{n=i-1}^{n=i+1} q_j(n) LJ_y(n) + \sum_{n=i-1}^{n=i+1} q_t(n) T1_y(n) + \sum_{n=i-1}^{n=i+1} q_d(n) T3_y(n)
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
T3_y(i) = c &+ \sum_{n=i-1}^{n=i+1} q_j(n) LJ_y(n) + \sum_{n=i-1}^{n=i+1} q_t(n) T1_y(n) + \sum_{\substack{n=i-1 \\ n\neq i}}^{n=i+1} q_d(n) T3_y(n) \\
&+ \sum_{n=i-1}^{n=i+1} w_j(n) LJ_x(n) + \sum_{n=i-1}^{n=i+1} w_t(n) T1_x(n) + \sum_{n=i-1}^{n=i+1} w_d(n) T3_x(n),
\end{aligned} \tag{2}
$$

where i is the phoneme index, which was set as 2 since we focus on the central phoneme in the three-phoneme sequence. Here, w is the partial multiple regression coefficients for X-components, q is for Y-components, and c is a constant term. The subscripts j, t, and d denote the coefficients for the jaw, tongue tip, and tongue dorsum, respectively. The first line of the formulas is the factors in the same dimension, and the second line is the factors from the perpendicular dimension. Factors with the same phoneme index ($n = i$) describe the spatial constraints between the different observation points.

For the apical consonants in VCV, the dependent variable is $T1$. The description for $T1$ is similar to the above formulas; Eq. (3) is for $T1_x$ and (4) for $T1_y$.

$$
\begin{aligned}
T1_x(i) = c &+ \sum_{n=i-1}^{n=i+1} w_j(n) LJ_x(n) + \sum_{\substack{n=i-1 \\ n\neq i}}^{n=i+1} w_t(n) T1_x(n) + \sum_{n=i-1}^{n=i+1} w_d(n) T3_x(n) \\
&+ \sum_{n=i-1}^{n=i+1} q_j(n) LJ_y(n) + \sum_{n=i-1}^{n=i+1} q_t(n) T1_y(n) + \sum_{n=i-1}^{n=i+1} q_d(n) T3_y(n)
\end{aligned} \tag{3}
$$

$$
\begin{aligned}
T1_y(i) = c &+ \sum_{n=i-1}^{n=i+1} q_j(n) LJ_y(n) + \sum_{\substack{n=i-1 \\ n\neq i}}^{n=i+1} q_t(n) T1_y(n) + \sum_{n=i-1}^{n=i+1} q_d(n) T3_y(n) \\
&+ \sum_{n=i-1}^{n=i+1} w_j(n) LJ_x(n) + \sum_{n=i-1}^{n=i+1} w_t(n) T1_x(n) + \sum_{n=i-1}^{n=i+1} w_d(n) T3_x(n)
\end{aligned} \tag{4}
$$

In this analysis, the Z-score is used to standardize all the variables. A backward stepwise regression is employed to focus on the most significant factors [14]. The processing starts with all factors and removes the least significant terms in a stepwise fashion until all the remaining terms are statistically significant. The probability of the rejection region $F > F\alpha$ was set $P(F > F\alpha) = 0.001$. The multiple correlation coefficients with the partial regression coefficients (PRC) reflect the "contribution" of each factor to the target. The PRCs of the formulas (1)–(4) are used to evaluate the effects of each factor on the CP of the central phoneme. Note that after using the Z-score the partial correlation coefficients (PCC) are proportional to the partial regression coefficients. Therefore, this paper does not calculate the PCCs directly. The following discussions concerned with the correlation in the following section reference to the PRCs.

4.2. Coarticulations within Different Combinations

Three-phoneme sequences of VCV and CVC were segmented from read sentences. Four combinations were used in this analysis: apical-vowel-apical (CVC), vowel-apical-vowel (VCV), palatal-vowel-palatal (CVC), and vowel-palatal-vowel (VCV). The preceding and following consonants in CVC come out of the same consonant group. To obtain a reliable relation, the phrase boundaries were marked and excluded from the VCV and CVC sequences. The sequences with a pause were also excluded from the materials for analysis. The data size for each speaker is about 1,500 sequences for the combinations of vowels and apical consonants, and about 90 for the combinations of vowels and palatal consonants. The analysis was first carried out for three speakers individually. The results indicated that two speakers had a consistent tendency, but one speaker showed some differences, especially for the

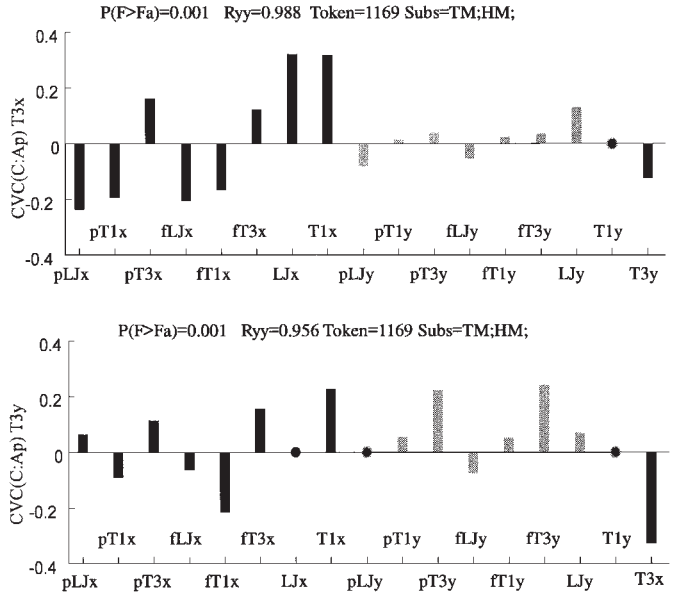

Fig. 5 PRCs for vocalic target $T3_x$ (upper) and $T3_y$ (lower) in apical-vowel-apical target. Dots denote the removal of variables without statistical significance.

VCV sequence with apical consonants. To increase data size and obtain reliable results, the data from the two consistent speakers were put together and used in the following analyses.

4.2.1. Results for apical-vowel-apical sequences

Figure 5 shows the PRCs obtained from 1,169 apical-vowel-apical (CVC) sequences, where $T3$ of the central vowels was described by Eqs. (1) and (2). The multiple correlation coefficient (MCC) was 0.988 for $T3_x$ and 0.956 for $T3_y$, and the average prediction errors were 0.069 cm and 0.141 cm, respectively. The first eight variables are concerned with X-components, shown in a dark color, while the following eight are concerned with Y-components, shown in a pale color, and the last one shows the effects of the perpendicular dimension of the same observation point. The dots indicate that the factors were removed in the stepwise regression since they were statistically less significant. The variables starting with "p" and "f" denote the components of the preceding and the following phonemes, respectively, and the others show the components out of the same phoneme to account for physiological constraints inter and/or intra the articulators.

For $T3_x$ on the upper panel, X-components in general show a definite contribution to $T3_x$, but Y-components do not show any significant contribution. Among the X-components, the tongue tip and jaw shows large positive PRCs. Effects of the preceding phoneme on $T3_x$ are slightly greater than those of the following one. For $T3_y$ on the lower panel, the Y-components show strong effects on $T3_y$, while the X-components also have comparable effects with the Y-components. This implies that the dorsum movement in the Y-dimension strongly depends on the X-components while the movement in the X-dimension is almost independent of the Y-components.

During production of the central vowel, $T1$ and $T3$ have a higher correlation in the X-dimension but have no significant correlation in the Y-dimension, which was removed in the stepwise processing. This implies that the tongue tip moves consistently with the dorsum in the X-dimension but independently in the Y-dimension. Moreover, $T3_y$ has a larger negative correlation with $T3_x$, indicating that the tongue moves in an oblique direction from front-upper to back-lower.

It is natural to expect that the dorsum may have a higher correlation to the lower jaw in the Y-dimension than in the X-dimension. In this figure, however, the LJ_x has an extremely high positive correlation with the $T3_x$, which implies that the movement of the dorsum is constrained by the jaw in the X-direction more than in the Y-direction. For the surrounding apical consonants, the PRCs of the jaw to $T3_x$ are negative because the jaw moves in the opposite directions during the vocalic gesture and the consonantal gesture.

Both the preceding and following $T3$s of the apicals positively contribute to the vocalic target $T3$. In contrast, their $T1_x$s negatively contribute to the target $T3_x$. The positive PRCs reflect the fact that the tongue dorsum positions of the consonants are strongly dependent on the adjacent vowels because the apical consonants have no crucial target for the dorsum. The negative PRCs are concerned with the movement of the tongue tip for the apical target, which is opposite to the dorsal movement for the vocalic target. The opposite motions in the X-dimension during the apical consonants stretch the tongue surface to make it flat. This is consistent with the articulatory observation.

4.2.2. Results for vowel-apical-vowel sequences

Figure 6 shows the PRCs obtained from 3,226 vowel-apical-vowel (VCV) sequences. The crucial point of the apical consonants, $T1$, was described by Eqs. (3) and (4). The MCC was 0.990 for $T1_x$ and 0.961 for $T1_y$, and the prediction errors were 0.064 cm and 0.066 cm, respectively. As seen in the case of CVC, the X-components play a dominant role in forming the consonantal target.

To form an apical constriction, the tongue tip moves forward at the same time the dorsum moves in the same direction to synergize the tongue tip. This analysis consistently shows that $T3_x$ contributes largely to $T1_x$, as shown in the upper panel. It is interesting to find that $T3_y$ contributes more to $T1_x$ than $T1_y$ does. This may be because in most cases an upward movement of the dorsum induces a forward movement of the tongue tip, especially in the vowel-apical-vowel context.

During both preceding and following vowels, the jaw actively participates in the synergic coarticulation to form

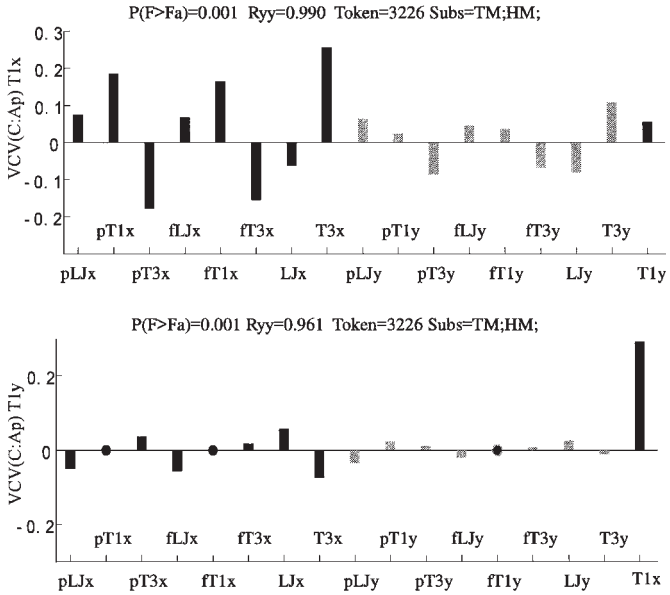

Fig. 6 PRCs for consonantal target $T1_x$ (upper) and $T1_y$ (lower) in vowel-apical-vowel sequences. Dots denote the removal of variables without statistical significance.

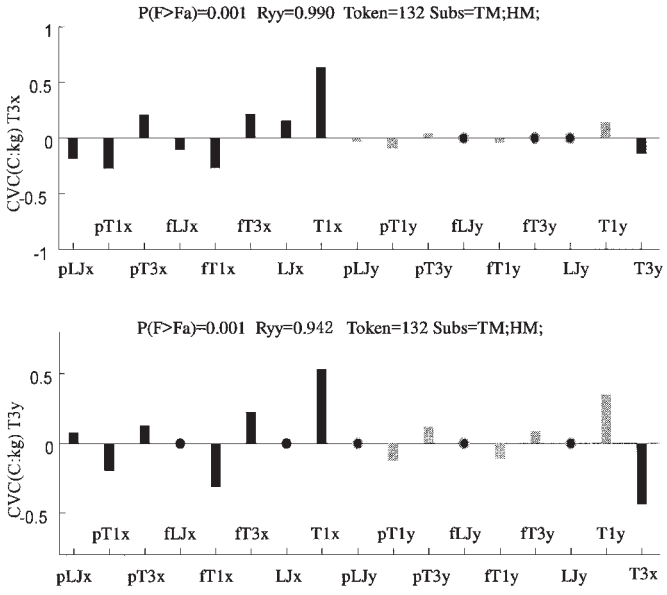

Fig. 7 PRCs for vocalic target $T3_x$ (upper) and $T3_y$ (lower) in palatal-vowel-palatal sequences. Dots denote the removal of variables without statistical significance.

the apical gesture since both the *X*- and *Y*-dimensions of the jaw show the positive contribution to $T1_x$. In contrast, the jaw has negative PRCs during the consonants. The opposite movement of the jaw to the tongue tip might be required for forming a wide space in front of the apical constriction.

Differing from Fig. 5, $T1_x$ shows an extremely large PRC for $T1_y$, while the other variables definitely have small PRCs. This phenomenon indicates that the vertical position of the tongue tip depends solely on its horizontal position. Since the tongue tip contacts the hard palate in the midsagittal or bilateral part in producing the apicals, for a given *X*-position, the *Y*-position of the tongue is uniquely determined according to the shape of the hard palate. This is why the $T1_y$ of the apical target depends on its $T1_x$ alone. In contrast, $T1_y$ did not show any significant contribution to $T1_x$, as seen in the upper panel.

In addition, $pT1_x$ and $fT1_x$ contribute positively to $T1_x$ of the apical consonants, while $pT3_x$ and $fT3_x$ contribute negatively. The positive contribution of the tongue tip results from a synergic action of the surrounding vowels for forming the apical targets of the consonants. At the same time, the tongue dorsum moves backward to approach the vocalic targets. This is why the $pT3_x$ and $fT3_x$ contribute negatively to the forward apical movement. The result of such an opposite motion is that the tongue surface becomes flatter.

4.2.3. Results for palatal-vowel-palatal sequences

Figure 7 shows the PRCs obtained from 132 palatal-vowel-palatal (CVC) sequences, where the vocalic target *T*3 was described by Eqs. (1) and (2). The MCC was 0.990 for $T3_x$ and 0.942 for $T3_y$, and the prediction errors were 0.077 cm and 0.109 cm, respectively.

In this combination, vowels and consonants have the same crucial point, *T*3. Since the tongue tip does not have a crucial target, its movements show a good consistency with those of the tongue dorsum. The result demonstrates that *T*1 contributes heavily to *T*3 in both the *X*- and *Y*-dimensions. As seen in the other cases, the *X*-components of the independent variables contribute more to $T3_x$ than the *Y*-components do. The preceding and following phonemes did not show any superiority of one over another in their contribution.

Differing from the cases of the apical consonants, the *Y*-component of the surrounding palatals does not show any significant contribution to $T3_y$. This is because the dorsum always rises at about the same height to contact the palate during production of the palatal consonants, irrespective of the height of the dorsum during the vowel. $T1_y$ gives a relatively large contribution to $T3_y$ since the tongue tip and the dorsum drop during the release of the palatal closure. The contribution of $T3_x$ to $T3_y$ is greater than that of $T3_y$ to $T3_x$. Again, this implies that the tongue dorsum moves more actively in the *X*-dimension than in the *Y*-dimension. As shown in Fig. 5, the following phoneme shows stronger effects on the $T3_y$ than the preceding one.

4.2.4. Results for vowel-palatal-vowel sequences

Figure 8 shows the PRCs obtained from 772 vowel-palatal-vowel (VCV) sequences. The MCC was 0.989 for $T3_x$ and 0.921 for $T3_y$, and the prediction errors were 0.076 cm and 0.100 cm, respectively. As seen in Sec. 4.2.3, *T*1 has large contributions to the dorsum in both *X*- and *Y*-

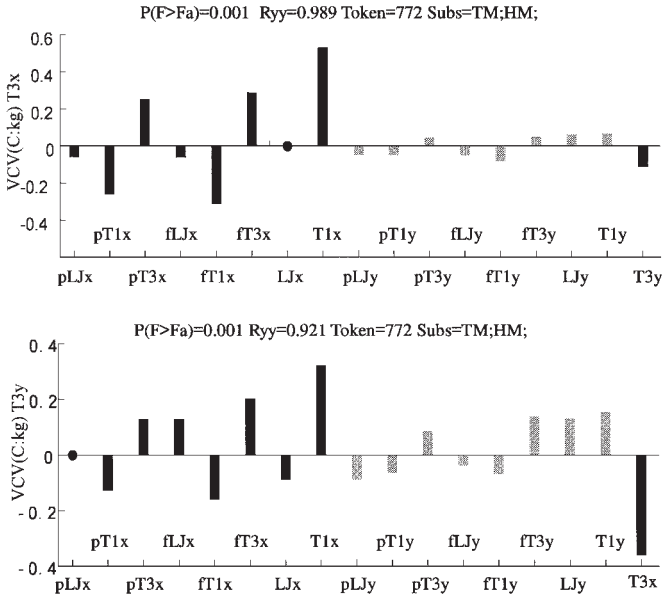

Fig. 8 PRCs for palatal target $T3_x$ (upper) and $T3_y$ (lower) in vowel-palatal-vowel sequences. Dots denote the removal of variables without statistical significance.

dimensions.

In general, *X*-components of the variables make greater contributions to $T3_x$ than *Y*-components, which is a common phenomenon for all the combinations. Both vocalic $T3_x$s represent positive PRCs for the palatal $T3_x$ because the crucial point of each vowel affects the location of the dorsal-palatal constriction. The negative PRCs of vocalic $T1_x$ result from the opposite motion of the tongue tip, which retracts backward during the palatals and extends forward during the vowels. Retraction of the tongue tip might be necessary for raising the tongue dorsum by compressing the tongue body in the anterior-posterior dimension. The following phoneme obviously makes larger contributions to the $T3_y$.

In this combination, the jaw and tongue tip work together to synergize the movement of the dorsum in the *Y*-dimension. The jaw shows a clear contribution to $T3_y$ during the palatal consonants, which was not seen in the other combinations. This might suggest that there is a special requirement for the jaw to generate a synergic movement in forming the palatal-dorsal closure.

5. EVALUATION OF LR AND RL EFFECTS

To realize coarticulation in an articulatory model, the left-to-right (LR) effect and the right-to-left (RL) effect should be investigated separately. The former reflects the carryover effects of the preceding phoneme on the following one, whereas the latter corresponds to the anticipatory effects of the following phoneme on the preceding one. In this section, we evaluate the LR (carryover) effect and RL (anticipatory) effect by examining their relative influences of the preceding and following phonemes on the central phoneme using the three-phoneme sequences.

5.1. Method of Evaluation

The basic idea for the evaluation is to quantify the effects of the lateral phonemes of the three-phoneme sequences on the central phoneme, and then make a comparison between the relative effects of the two sides. To do this, we fix a selected phoneme on a given side of the three-phoneme sequences and change the phonemes on the opposite side, so that variations in the tongue shape would be observed. The variation reflects the effects of the side with varying phonemes. In the same way, we can measure the effects of the other side by switching the sides and fixing the same selected phoneme. The standard deviation (STD) is used to evaluate variations in the tongue shape resulting from this manipulation. For a given phoneme sequence C_1VC_2, for example, we fix C_2 and then replace C_1 by all possible consonants in the observation data one by one. This replacement induces some variations in the tongue shapes of the central vowel. The STD of the tongue shapes from the average shape corresponds to effects of C_1. The STD value resulted from C_1 is referred to as Std1. Similarly, fixing C_1 and replacing C_2 by all the possible consonants, we can obtain another STD for tongue shape changes, named Std2. If Std1 is larger than Std2, it means the preceding phoneme has greater effect than the following one, and vice versa. Note that in this processing, all four points on the tongue surface were used to describe tongue shape, regardless of the definition of the CP. The fixed phoneme was included from the varying phonemes in the palatal-vowel-palatal sequence to increase data size, while it was excluded in the other sequences.

5.2. LR Effect vs RL Effect

Table 1 shows the STD of tongue shapes during CVC sequences, where the palatal and apical consonants were treated as two separate groups. For a sequence of C_1VC_2, Std1 is the value obtained by fixing C_2 and altering C_1, and Std2 is the value obtained by the opposite order. The dark shadow indicates that such combinations do not exist symmetrically. The pale shadow denotes the case in which the following phonemes have larger effects than the preceding ones. The part without any shadow is for the case in which the preceding phonemes have greater effects than the following ones. In the results, the preceding consonants show larger effects on the vowel /a/ while the following ones have larger effects on /o/. For the other vowels, there is no significant superiority shown between the vowel positions.

For the selected sequences, the number of combinations beginning with a given phoneme may be different from that

Table 1 The STD (cm) of tongue shape variations for C_1VC_2. Std1 is the value obtained by fixing C_2 and replacing C_1 by all the consonants, and Std2 is the value obtained in the opposite order. * denotes the average values over all the vowels. Note that the calculation was carried out for the palatal and apical consonants separately.

		t	d	s	z	sh	Ts	ch	J	k	g
a	Std1	0.162	0.186	0.106	0.108	0.108	0.102	0.106		0.118	0.095
	Std2	0.129	0.135	0.064						0.119	0.097
i	Std1	0.158	0.076	0.198	0.084	0.151	0.103	0.062	0.070	0.025	0.058
	Std2					0.157		0.076	0.091	0.025	0.058
u	Std1	0.108	0.096	0.061	0.100	0.139	0.053	0.029	0.107	0.099	
	Std2			0.112	0.095		0.128			0.099	
e	Std1	0.040	0.065	0.077		0.077	0.059			0.212	
	Std2	0.070	0.092	0.056	0.067					0.212	
o	Std1	0.170	0.191	0.170		0.203	0.126	0.145		0.136	0.130
	Std2	0.168	0.236	0.211						0.151	0.110
*	Std1	0.226	0.229	0.213	0.166	0.122	0.130	0.212	0.167	0.136	0.128
	Std2	0.199	0.187	0.199	0.134	0.084	0.098	0.072	0.084	0.140	0.134

of combinations terminated by the same phoneme. To reduce some potential influences caused by the asymmetry of the data, an identical number was chosen from the two combinations in calculating the average value. The last two rows, indicated by *, show the average values over all the vowels. The apical consonants show that the preceding phonemes clearly tend to have larger effects on the vowels than the following ones. For the palatal consonants, shown in the last two columns, the following phonemes show greater effects than the preceding ones. The two consonant groups, meanwhile, show a different tendency to one another.

Table 2 shows the STD (Std1 and Std2) of tongue shapes during VCV sequences, where C denotes the apical consonants. The shading has the same meaning as in Table 1. In most cases, for individual consonants, the preceding phoneme has a slightly greater effect than the following one. For the average values, however, there is no clear tendency corresponding to the vowels; both the preceding and following vowels show about the same magnitude of effect on the central apicals.

Table 3 shows the results for VCV with the palatal consonants. For most combinations, the following vowel has a greater effect on the consonant than the preceding one. On average, for vowels /i/, /u/, and /e/, the following vowel shows stronger effects on the palatals than the preceding one, while for /a/ and /o/ the preceding vowel shows larger effects. The opposite tendencies seem to be attributed to closed vowels and open vowels. As shown in Fig. 3, the CP is relatively loose (scattering more widely) for the open vowels of /a/ and /o/, and is strict for the closed vowels. The articulation points for both the vowels and the palatals are on the dorsum, and they affect each other during articulation. It can be expected that the phonemes (vowels and consonants) with a strict CP have

Table 2 The STD (cm) of tongue shape variations for V_1CV_2, where C is the apical consonant; Std1 is the value obtained by fixing V_2 and replacing V_1 using all the vowels, and Std2 is the value for the opposite order. * indicates the average values over the consonants.

		a	i	U	e	o
t	Std1	0.106			0.129	0.117
	Std2	0.105	0.107	0.064	0.088	0.125
d	Std1	0.134			0.140	0.127
	Std2	0.140	0.118	0.089	0.099	0.138
s	Std1	0.070		0.084	0.098	0.078
	Std2	0.059	0.105	0.065	0.066	0.066
z	Std1	0.076		0.075	0.140	0.086
	Std2	0.069	0.101	0.060	0.047	0.093
sh	Std1		0.078			
	Std2	0.068	0.100	0.077	0.105	0.056
Ts	Std1			0.082		
	Std2	0.063	0.097	0.070	0.059	0.070
ch	Std1		0.074			
	Std2	0.081	0.071	0.041		0.074
j	Std1		0.068			
	Std2	0.073	0.057	0.063	0.058	0.068
*	Std1	0.109	0.090	0.063	0.095	0.107
	Std2	0.110	0.090	0.066	0.085	0.103

Table 3 The STD (cm) of tongue shape variations for V_1CV_2, where C is the palatal consonant. The rest is the same as that in Table 2.

		a	i	U	e	o
k	Std1	0.299	0.158	0.137	0.204	0.322
	Std2	0.288	0.179	0.192	0.236	0.214
g	Std1	0.313	0.149	0.162	0.209	0.320
	Std2	0.292	0.237	0.161	0.215	0.265
*	Std1	0.304	0.153	0.147	0.206	0.322
	Std2	0.289	0.214	0.184	0.228	0.242

stronger effects on the others.

6. DISCUSSION AND CONCLUSION

In this study, we investigated coarticulation within three-phoneme sequences from Japanese read sentences, where pauses and phrase boundaries were excluded. Our analysis focused on the effects of coarticulations on phoneme targets. For this purpose, each phoneme was represented using three observation points in the stationary period: the lower jaw, tongue tip, and tongue dorsum. In continuous speech, distribution of the average crucial points for both vowels (the tongue dorsum) and the apical consonants (the tongue tip) was nearly linear. These two distribution lines constructed an articulatory coordinate for vowels and consonants, which was consistent for all three speakers. Since the distribution of the apical CPs depends heavily on the palate shape, the main direction of the distribution is speaker-dependent. To construct a constant

angle between the vocalic and consonantal distributions, it seems that the speakers adjust the movement directions of the dorsum while producing vowels to fit the consonant distribution. This implies that the speakers might compensate for the morphological difference of the hard palate by adjusting the main moving direction of the tongue for the vowels.

6.1. Coarticulation and Context

Four combinations of three-phoneme sequences were used in this study, which consist of vowels, apical consonants, and palatal consonants. The CP of the central phoneme was represented by a linear function of the observation vectors with 17 elements. A stepwise multiple regression method was adopted to focus the analysis on the factors with statistical significance. For this linear description, multiple correlations were about 0.99 for the *X*-dimension and 0.92 for the *Y*-dimension, and the prediction errors were about 0.07 cm for the *X*-dimension and about 0.1 cm for the *Y*-dimension. Because of the high correlations and small prediction errors, we can expect this analysis to give reliable results.

The *X*-component of the independent variables showed dominant contributions to both *X*- and *Y*-components of the CP of the central phoneme in all the combinations. This result suggests that Japanese speakers may control tongue movement more actively in the horizontal dimension than that in the vertical dimension. Our model simulation confirmed this suggestion to some extent, in which for a given force the tongue muscles generated a larger movement for the tongue dorsum in the horizontal dimension than that in the vertical dimension [10].

Using the above result, we may explain the opposite behaviors of the preceding and following phonemes in CVC sequence. There, the preceding phoneme had larger PRCs in the *X*-dimension than the following one, whereas the following phoneme showed larger PRCs in the *Y*-dimension than the preceding one. The former phenomenon may be related to the basic CV structure of Japanese, in which the vowel has a closer relationship with the preceding phoneme than the following one. The latter phenomenon may be explained as a passive component of movement if Japanese speakers truly control movements in the *X*-dimension more actively as mentioned above. In other words, the *Y*-movement of the dorsum behaves as a relaxation motion (carryover effects), which depends on the past deformation. Thus, it is reasonable to state that the movement for the right phoneme in the *Y*-dimension depends more on the central vowel. Similar tendency can be seen in palatal-vowel combinations.

In V-apical-V sequences, $T1_x$ of both preceding and following vowels contributes positively to apical $T1_x$ while the $T3_x$ of both the vowels contributes negatively. The positive contribution results from a synergic action of the surrounding vowels for forming the apical targets of the consonants. The negative contribution implies that the dorsum moves in the opposite directions during the vocalic gesture and consonantal gesture. The opposite motions in *X*-dimension during the apical consonants stretch the tongue surface to make it flat. This result is consistent with the articulatory observation.

The tongue tip and the dorsum have a high correlation in the *X*-direction but they almost have no correlation in the *Y*-direction. According to the results from VCV with apical consonants, the *X*-positions are the dominant factor for movements of the tongue tip. The *Y*-movement of the tongue tip seems to be an independent motion superimposed on a continuous movement of the body (see Fig. 6 for details). This observation supports Öhman's conclusion that VCV articulation could be represented by a basic diphthongal gesture with an independent consonantal gesture superimposed on its transitional portion [3].

The majority of the combinations showed the dorsum moves more consistently with the jaw in the *X*-direction than in the *Y*-direction. This is somewhat different from the conventional wisdom, and might imply that the translation movement of the jaw may positively contribute to tongue movement during speech. In the combination of V-palatal-V, however, the jaw shows a clear contribution to $T3_y$ during the palatal consonants, and the jaw and tongue tip work together to synergize the movement of the dorsum in the *Y*-dimension. It might suggest that there is a special requirement for the jaw to generate a synergic movement in forming the palatal-dorsal closure.

6.2. LR Effect and RL Effects

We used the standard deviation of the tongue shape of the central phoneme to evaluate the relative effects from the phonemes on the left and right sides. The deviation was measured by fixing one phoneme on a given side and varying the phonemes on the other side. The results showed that in apical-V-apical (CVC) sequences the preceding consonant affected the vowel more strongly than the following one. This result is the same as the one obtained by the regression analysis, and is also consistent with the one observed in [4]. Unlike CVC, in V-apical-V sequences, the effects of the preceding and the following vowels on the central consonant did not show any significant positional difference. This observation was not consistent with that in [4].

In [5], Recasens *et al.* investigated the LR and RL effects by examining the transition between the consonant and vowel when fixing a phoneme at one end. They found that some coarticulations were vowel-dependent and others were consonant-dependent. The dependences were concerned with the degree of articulation constraint (DAC),

where the DAC scale ranks phonetic sound categories from maximally to minimally constrained as follows: /ʃ/, /ɲ/, /i/, /k/, dark /l/, (/s/) (DAC = 3) > /n/, /a/, (/s/) (DAC = 2) > /p/, /ə/ (DAC = 1) [5]. Referring to the DAC, the apical consonants in general have a larger effect on vowels at the crucial point because the consonants have stricter constraints than vowels. This may be the reason why vowels do not show any significant positional difference on the apicals.

For vowel-palatal combinations, clear tendencies were shown in both CVC and VCV. For CVC sequences, the following palatal generally showed larger effects on the vowels. For VCV, the preceding vowel showed a greater effect than the following one when fixing open vowels, while the opposite tendency was shown when fixing closed vowels. How should we explain this phenomenon? Here, we suppose that the look-ahead mechanism works on the vowel sequence over the palatal consonant as observed in [2], and the DAC plays an important role in coarticulation [5]. Accordingly, the CP of the right vowel affects that of the left vowels, and the extent of the effects depends on the DAC of the vowels. When fixing /i/ on the right side, for example, the replaceable vowels on the left side are /a,u,e,o/. According to the look-ahead mechanism, the CP of the left vowels approaches that of /i/ to some extent, so that the difference between the left and right vowels gets smaller. If fixing /i/ on the left side, in contrast, the CP of /i/ does not approach the CP of /a,u,e,o/ so much because /i/ has a higher DAC, and thus the difference between the left and right vowels is larger than in the former case. Therefore, the combination of /a,u,e,o/-palatal-/i/ has a smaller STD than /i/-palatal-/a,u,e,o/. When fixing /a/, the replaceable vowels are /i,u,e,o/, whose DACs are higher than or the same as that of /a/. The CP of /a/ approaches the CP of /i,u,e,o/ if /a/ is fixed on the left side so that the difference between the left and right vowels becomes smaller. If /a/ is on the right side, the difference is not reduced since /a/ cannot strongly affect /i,u,e,o/. As a result, /a/-palatal-/i,u,e,o/ has a smaller STD than /i,u,e,o/-palatal-/a/. From the above, one can see that all the results of the vowel-palatal combinations can be explained using the look-ahead mechanism and the DAC.

It has been observed that the articulation point for the palatal consonants is easily affected by vowels (*cf.* [12]). This analysis for vowel-palatal combinations showed a consistent result with the observation. The results for the vowel-palatal combination can be explained by using the look-ahead mechanism. However, the results from the apical-V-apical sequences showed that the preceding consonant had greater effects than the following one. A possible reason for this difference is that in vowel-apical combinations the CPs switch between the tongue tip and the dorsum, but the CPs in vowel-palatal combination are always located in the same place, the dorsum. In vowel-apical combinations, the targets for the tongue tip and dorsum can overlap temporally, and/or reach optimal positions for both targets by a deformation of the tongue, for example, a stretch of the tongue surface. In this case, effects of carryover are also playing an important role. In contrast, in vowel-palatal combinations, the CP for the palatals tends to approach the following vocalic target for the coarticulation as well as accept some effects from the previous vocalic target. Accordingly, both phenomena are not contradictory to the look-ahead mechanism. This hypothesis will be verified in future studies, using simulations with our physiological articulatory model [10].

ACKNOWLEDGMENTS

This research has been supported in part by CREST of Japan Science and Technology. The authors would like to thank Pascal Perrier and Donna Erickson for their beneficial comments on this work.

REFERENCES

[1] D. Raymond, *The Physiology of Speech and Hearing* (Prentice-Hall, Englewood Cliffs, 1980).

[2] L. Henke, "Dynamic articulatory model of speech production using computer simulation," Doctoral Dissertation, MIT, (1966).

[3] S. Öhman, "Coarticulation in VCV utterances: Spectrographic measurements," *J. Acoust. Soc. Am.*, **39**, 151–168 (1966).

[4] S. Kiritani, "Perturbation of the consonant and vowel articulation by a adjacent segments," *J. Acoust. Soc. Jpn.* (*J*), **34**, 132–139 (1978).

[5] D. Recasens, M. Pallares and J. Fontdevila, "A model of lingual coarticulation based on articulatory constraints," *J. Acoust. Soc. Am.*, **102**, 544–561 (1997).

[6] K. Shirai and M. Honda, "Estimation of articulatory parameters from speech sound," *Trans. IECE*, **61**, 409–416 (1978).

[7] S. Maeda, "Compensatory articulation during speech: evidence from the analysis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. (Kluwer Academic Publishers, Dordrecht, 1990), pp. 131–149.

[8] J. Dang and K. Honda, "A physiological model of a dynamic vocal tract for speech production," *Acoust. Sci. & Tech.*, **22**, 415–425 (2001).

[9] J. Dang and K. Honda, " Estimation of vocal tract shape from sounds via a physiological articulatory model," *J. Phonet.*, **30**, 511–532 (2002).

[10] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.*, **115**, 853–870 (2004).

[11] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," *J. Acoust. Soc. Am.*, **110**, 453–463 (2001).

[12] K. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, Mass., 2000).

[13] K. Moll and R. Daniloff, "Investigation of timing of velar movement during speech," *J. Acoust. Soc. Am.*, **50**, 678–684 (1971).

[14] N. Draper and H. Smith, *Applied Regression Analysis*, 2nd Ed. (John Wiley & Sons, New York, 1981), pp. 307–312.