INVESTIGATION AND MODELING OF COARTICULATION IN SPEECH PRODUCTION

Jianwu Dang[1,2],Jianguo Wei[1],Takeharu Suzuki[1],Kiyoshi Honda[2],Pascal Perrier[3] and Masaaki Honda[4]

[1]Japan Advanced Institute of Science and Technology, Ishikawa
[2]ATR Human Information Science Lab, Kyoto
[3]ICP CNRS UMR 5009 & INPG & University Stendhal, Grenoble
[4]School of Sport Science, Waseda University, Tokyo
{jdang, jianguo, takeharu}@jaist.ac.jp; honda@atr.jp; perrier@icp.inpg.fr; hon@waseda.jp

ABSTRACT

Coarticulation during speech production takes place at both the physiological level concerned with articulators' properties and the planning stage for elaborating motor commands. This study focuses on the investigation and modeling of planning aspects of coarticulation with the ultimate objective to implement human mechanism in controlling a physiological speech production model. A "carrier model" was derived from articulatory data to describe mechanisms of the coarticulation, in which the vocalic movement is considered to be the primary component as a "carrier wave" and the consonantal movement as a "modulation wave". Interactions between the carrier and modulation waves were evaluated using phoneme sequences of $V_bCV_cCV_b$ (V_c: the central vowel; V_b: the bilateral vowel; C: consonants) out of articulatory data that was obtained from the electromagnetic articulographic system. The analysis of the articulatory data showed that the articulatory position of the central vowel tended to be assimilated towards that of the bilateral vowels.

1. INTRODUCTION

Coarticulation is a natural phenomenon involved in human speech, which originates from movement planning strategies and from physical interactions among speech articulators. Since coarticulation is a factor to bring natureness to speech sounds, it is necessary to be taken into account to attain high-quality synthetic speech sound.

To clarify and model coarticulation, a large number of experiments have been carried out [cf. 1-6]. Öhman used spectrographical measurements to describe coarticulation in VCV utterances, where he paid more attentions to the relation between vowels (V) and consonants (C) [1]. Kiritani used the X-ray microbeam system to investigate coarticulation in VCV and CVC sequences [2]. Perkell *et al*. [3] and Matthies *et al*. [5] used the electromagnetic articulographic system to investigate coarticulation mainly concerned with lip movements. Recasens *et al*. employed electropalatographic data and trajectories of the second formant to investigate anticipatory and carryover effects within VCV sequences [4]. Dang *et al.* conducted statistical analysis on articulatory data to clarify the effects of the pre- and post-phonemes on the central phoneme [6]. Those studies resulted in a number of models to account for the mechanism of coarticulation [3-5, 7-8, 12]. However, there is no general agreement among the models since they resulted from different experimental data and conditions.

This study aims to derive a more plausible model of coarticulation than the previous models by using articulatory observations to quantify the parameters of the model. Our analysis of coarticulation has two distinctive characteristics. First, the speech material used in this study consists of continuous utterances of read sentences while almost former studies employed a carrier sentence with nonsense sequences of vowels and consonants. The continuous utterances are adopted here because they are of benefit to investigate effects of coarticulation on the nature of speech sounds. Second, while many studies focused on the coarticulation involved in labial movements that is to some extent independent of the other articulators, the results for lip data may not be adequate for tongue movements. Therefore, we focused on coarticulation of the tongue movements, which is related to contextual and physiological factors. This report consists of two parts. The first part proposes a model for coarticulation based on the previous studies [1, 4, 6-8, 12]. The second part is to evaluate the parameters of the model using articulatory observations.

2. "CARRIER MODEL" OF COARTICULATION

There are a large number of models of coarticulation. However, those models are lack of generality since they were derived from difference data sets for different purposes. This study aims to derive a model of coarticulation for the purpose of speech synthesis with reference to the past studies and models.

2.1. Derivation of the "carrier model"

During natural speech, two types of coarticulatory overlaps can be identified: 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 phonemes toward a given phoneme. In this case, the target of the phoneme is reached in different ways, depending on the status of the previous phoneme(s). Anticipatory (RL) coarticulation occurs when the speaker can "look ahead" in time and anticipate oncoming sounds. While LR coarticulation occurs mainly at the physiological level, RL coarticulation reflects essentially a high-level phonological-phonetic processing, since the entire utterance must be scanned before it is articulated [9]. To describe this process, Henke proposed a phonemic-segment model [8]. Each utterance was described by a matrix of articulatory targets with distinctive features, in which some features change abruptly as the target switches from one to another. Öhman proposed another model to describe the mechanism of coarticulation, in which the articulation was represented by a basic diphthongal vocalic gesture with independent consonantal gestures that are superimposed on the vocalic transition in a vowel-consonant-vowel sequence [1, 7, 12].

As the preceding work, Dang *et al.* analyzed coarticulation in continuous speech based on articulographic data using a stepwise multiple regression method [6]. It was found that the movement of the tongue tip was highly correlated to that of the tongue dorsum in the horizontal dimension, while they were almost independent of one another in the vertical dimension. This observation supports the idea that coarticulation in VCV sequences can be considered as an independent consonantal gesture superimposed on a transitional portion between vowels, which is consistent with the Öhman's proposal [1].

To quantify coarticulation between the adjacent phonemes, Recasens *et al*. employed electro-palatographic data and the trajectory of the second formant to investigate anticipatory and carryover effects within VCV sequences, which consist of seven consonants and two vowels /a/ and /i/ [4]. They found phoneme-dependent differences and used a phoneme specific degree of articulatory constraint (DAC) to explain the differences, where the DAC was defined as having three levels. They tried to account for coarticulation between the consonants and vowels using the concept of DAC.

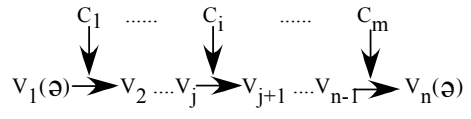
In general, a spoken utterance can be considered as a stream consisting of consonants and vowels. The coarticulatory effect of a vowel on consonants is generally greater than that of a consonant on vowels [9]. Therefore an utterance stream can be considered to consist of a vocalic "component" with strong and sustaining effects, and an additional consonantal "component" with relative weak and rapid effects. Accordingly, we used the concept of "carrier and modulation" to reconsider such a mechanism with following underlying assumptions: during speech, the articulators move slowly and continuously from vowel to vowel, with periodic interference from rapidly amplitude- and frequency-modulated effects caused by consonants. As a result, coarticulation, the overlapping properties of speech sounds, is built into the varying nature of the articulation process. Therefore, the interaction between adjacent phonemes shows both the principal-subordinate property and time-varying property. A "carrier model" is proposed to describe these properties of coarticulations.

Comparing with the previous models, Öhman's model mainly focused on the principal-subordinate relation between vowels and consonants, while the "look-ahead" model paid a particular attention to the serial order. The proposed model takes the advantages of those two models.

2.2. Description of the carrier model

This study adopts spatial targets in the control strategy. According to a classical assumption, the speech production system is controlled by dynamic spatial targets, which give the proper combination of the articulators to configure a desired vocal tract shape [11, 13]. Modeling of coarticulation in this study focuses on the coarticulation taking place at the target planning stage, while the coarticulation at the physiological level is supposed to be realized by the physiological properties in a physiological articulatory model.

Planning an articulation based on the concept that a "carrier" movement superimposed by a "modulation" movement requires us to separate a given utterance into vowel and consonant sequences. To do so, the vowels and consonants are sorted out from the given utterance, and then built into two phoneme sequences as the following.



(1)

To construct the "carrier wave", articulatory movement is considered as a continuous movement from one vowel to another. Since the consonantal movement is superimposed on the "carrier wave", the resultant target of consonant C_i is dependent on a "tug-of-war" of the adjacent vocalic targets, and thus a virtual target G_i is generated in the position of C_i. This procedure is described by formula (2).

$$G_i = \alpha V_j + \beta V_{j+1} \qquad \alpha < \beta \tag{2}$$

where i and j are the indices of the consonants and vowels, and α and β are the weight coefficients. If the first and/or the last phonemes in the utterance are not a vowel, the target vector of the neutral vowel is added, as shown in (1), for the interpolation of (2). Note that if the carryover effect was completely realized by the physiological properties of an articulatory model, ideally, it would need to consider the effect of look-ahead mechanism alone, and thus α in (2) should be zero. At the planning stage, however, the virtual target for a consonant cannot be generated only referencing one of the bilateral vowels, since the both sides are necessary for the tug-of-war. For this reason, the weight coefficient α is nonzero in our treatment. Based on the look-ahead mechanism, the following target has stronger effects on the virtual target

than the preceding one at the target planning stage. Therefore, the coefficient β should be larger than α somewhat.

The second process is to deal with the effects of consonants on vowels, where only the immediately adjacent phonemes are taken into account. Here, the consonantal target $C_i^{'}$ is first constructed according to the "abstract" feature C_i and virtual feature G_i. Note that at this step only the crucial feature is modified, no change happens in the indecisive features since they depend on the adjacent vowels. This processing is carried out by the following formula:

$$C_i^{'} = (d_{ci}C_i + G_i)/(d_{ci} + 1) \qquad (3)$$

where d_{ci} is a weight coefficient depending on the degree of the articulatory constraint of the crucial feature of C_i. Second processing in this step is to account for the effects of the following consonant on the preceding vowel via the look-ahead mechanism.

$$V_j^{'} = \gamma d_{ci}C_i^{'} + \tau d_{vj}V_j/(\gamma d_{ci} + \tau d_{vj}) \qquad \gamma \approx \tau \qquad (4)$$

where i and j are the same as those of (2) and γ and τ are the weight coefficients for the crucial consonantal feature and the corresponding feature of the preceding vowel. In this study, the values for both γ and τ are set as 0.5. Finally, a target sequence is obtained by the summation set of the primary and subordinate components of $\{\{V_j^{'}\} \cup \{C_i^{'}\}\}$.

3. ESTIMATION OF MODEL PARAMETERS

To complete the proposed model, in this section, we employ articulatory data to estimate the parameters of the model.

3.1. Articulatory data

The articulatory data used in this study were collected using the electromagnetic midsagittal articulographic system in NTT, Japan [10]. Four receive coils were placed on the tongue surface in the midsagittal plane, named T1 through T4, and one coil on each of the upper lip, lower lip, maxilla incisor, mandible incisor (LJ), and the velum. The sampling rate was 250 Hz for the articulatory channels and 12 kHz for the acoustic channel. The origin of the coordinate system is located in the maxilla incisor, 0.5 cm upper the tip. Speech materials were about 360 Japanese sentences, and three adult male speakers read the sentences at a normal speech rate. The acoustic signal and 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 a steady point with the minimum velocity: T1 was used for labeling apical consonants and T3 for the others [10].

3.2 Distribution of the articulatory points

In this analysis, we focus on the coarticulation concerned with tongue movements. Our previous study showed that the articulation point (T3) of vowels from the continuous utterance of read sentences distributed widely [6]. Figure 1 shows the articulation point of five Japanese vowels obtained from a male Japanese speaker. The articulatory points were sorted out from read sentences, and the numbers were between 1600 and 3500 for the five vowels. The vertical axis is the density of the articulation point appearing in the articulatory space. Vowel /i/ has the smallest distribution region and the highest density among the five vowels. This implies that the articulation point of /i/ is difficult to be affected by the other vowels. Note that vowel /u/ has a relatively wide distribution. The articulatory data show that locations of T3 of /u/ have a certain correlation with the lip movements in the horizontal direction. This means that the lip protrusion may be one of accountable factors in producing Japanese /u/.

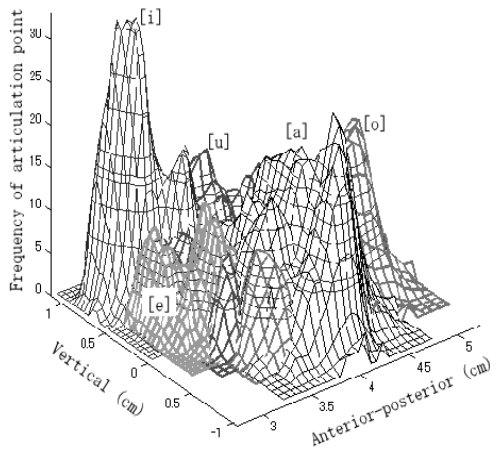


Figure 1: Distribution of articulation points for five Japanese vowels from one male Japanese speaker.

As shown in Fig. 1, there are a number of peaks in the distribution for each vowel. It does shape a single Gaussian distribution. However, if the scatter of a vowel is grouped according to the surrounding vowels respectively, each sub-scatter shows a Gaussian distribution. This indicates the surrounding vowels have distinctive effects on the central vowel.

To estimate parameters for the carrier model, phoneme sequence of $V_bCV_cCV_b$, where V_b is the vowel on both sides, V_c is the central vowel, and C is a consonant, are segmented out from the read sentences. Figure 2 shows the average location for each vowel with various surrounding vowels for three speakers. Grouping by the central vowels, vowel /i/ has the smallest distribution, while the other vowels show a wider distribution. This implies that articulatory point of vowel /i/ is more crucial than the others. The distribution for each central vowel shows a similar tendency as that seen in isolated vowels, which the closed vowels are located in front-high position and open vowel in back-low position.

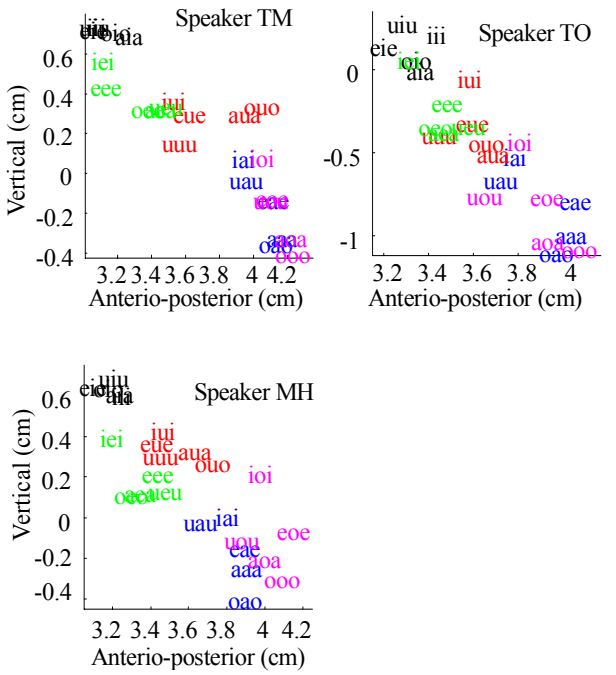


Figure 2: The average location for each vowel with various surrounding vowels in $V_bCV_cCV_b$ sequence, where V_b is the vowel on both sides, V_c is the central vowel, and C is a consonant.

Table 1: Displacement (cm) of the central vowel V_c apart from the average position (T3), which is caused by the bilateral vowels.

V_b \ V_c	a	i	u	e	o
a	0.152	0.058	0.106	0.124	0.107
i	0.288	0.059	0.236	0.281	0.389
u	0.238	0.122	0.196	0.180	0.184
e	0.085	0.130	0.117	0.029	0.101
o	0.226	0.016	0.143	0.092	0.201

As seen in Fig. 2, the bilateral vowels pull the central vowel away from its average position. To evaluate the interaction between the adjacent vowels, the displacement of the central vowel apart from its average location is calculated and shown in Table 1. As a result, vowel /i/ was displaced less than 0.13 cm from the average position by the other vowels while /i/ displaced the other vowels larger than 0.23 cm. The value of each column corresponds to the effects of the central vowel accepted from the bilateral ones. The value of each row shows the effect of the bilateral vowel on the central vowel. The values in the rows correspond to the degree of articulatory constraint (DAC). The larger the value, the stronger the DAC. We adopted the average displacement of each row as a measure for the DAC. After normalization, the DAC is 1.0 for /e/, 1.2 for /a/, 1.5 for /o/, 2.0 for /u/ and 2.7 for /i/.

4. CONCLUSIONS

Adopting "carrier" concept, a phoneme sequence is treated as a carrier wave (vowel-to-vowel articulation) and a modulation signal (overlapping consonantal articulation). Coarticulation was discussed based on the interaction of these two components. Partial parameters of the "carrier model" were quantified using phoneme sequences $V_bCV_cCV_b$ out of the read sentences obtained from the electromagnetic articulographic system. The degree of articulatory constraint (DAC) of vowels was obtained in this evaluation, while estimation of the DAC of consonants is planned for the future work.

Acknowledgment: This research has been supported in part by the National Institute of Information and Communications Technology. The authors especially thank NTT Communication Laboratories for permitting us to share the articulatory data.

5. REFERENCES

[1] Öhman, S. "Coarticulation in VCV utterance: Spectrographic measurements," *J. Acoust. Soc. Am*, 39, 151-168, 1966.

[2] Kiritani, S. "Perturbation of the consonant and vowel articulation by a adjacent segments," *J. Acoust. Soc. Jpn*, 34, 3, 132-139, 1978. (in Japanese)

[3] Perkell, J., Matthies, M. "Temporal measures of anticipatory labial coarticulation for the vowel /u/: Within- and cross-subject variability," *J. Acoust. Soc. Am.* 91, 2911-2925, 1992.

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

[5] Matthies, M., Perrier, P., Perkell, J., Zandipour, M. "Variation in articulatory Coarticulation with changes in clarity and rate," *J. Speech, Language, and Hearing Research*, 44, 340-353, 2001.

[6] Dang, J., Honda, M., Honda, K., Perrier, P. "Investigation of coarticulation in continue speech in Japanese", *Acoustical Science and Technology*, 25, 318-329, 2004.

[7] Öhman, S. "Numerical models of coarticulation," *J. Acoust. Soc. Am,* 40, 310-320,1967.

[8] Henke, L. "Dynamic articulatory model of speech production using computer simulation." Doctoral dissertation, MIT, 1966.

[9] Raymond, D. *The physiological of speech and hearing*, Prentice-Hall, Inc, Englewood Cliffs, N.J,1980.

[10] Okadome, T. and Honda, M. "Generation of articulatory movements by using a kinematic triphone model," *JASA* 453-463 ,2001.

[11] Browman, C. and Goldstein, L. "Articulatory gestures as phonological units," *Phonology*, 6, 201-251,1989.

[12] Fowler, A. "Coarticulation and theories of extrinsic timing," *J. Phonetics*, 8, 113-133, 1980.

[13] Kelso, S., Saltzman, E. and Tuller, B. "The Dynamical Perspective on Speech Production: Data and Theory," *J. Phonetics, 14*, 29-59. 1986