Investigation and Modeling of Coarticulation during Speech

Jianwu Dang1,2, Jianguo Wei1, Takeharu Suzuki1, and Pascal Perrier3

1Japan Advanced Institute of Science and Technology, Ishikawa, Japan
2ATR Human Information Science Lab, Kyoto, Japan
3ICP CNRS UMR 5009 & INPG & University Stendhal, Grenoble, France
{jw.dang, jianguo}@jaist.ac.jp; perrier@icp.inpg.fr

Abstract
Coarticulation is an important phenomenon of speech production, which involves in both the physical level concerned with articulators’ properties, and the planning stage for generating motor commands. The authors have proposed a model for coarticulation, named “carrier model”, to imitate the coarticulation mechanism. The carrier model is built on an assumption that articulation can be separated into a vocalic movement and a consonantal movement, and that coarticulation takes place within and between the movements as a modulation process. This study first verifies the assumption using a reconstructed text-independent articulatory movement based on observed articulatory data. The carrier model is elaborated based on those two movements and the look-ahead mechanism is employed to materialize the coarticulation. Coefficients of the carrier model were quantified via statistical analysis of articulatory data. A numerical experiment was conducted by implementing the carrier model in a physiological articulatory model (Dang and Honda, 2004). The results showed that the carrier model gave a good performance for simulating coarticulation.

1. Introduction
Coarticulation is a basic phenomenon of speech, which originates from motor planning strategies and physical interactions of the speech articulators. Since coarticulation is a factor to bring naturalness to speech sounds, it is necessary to be taken into account to attain a high-quality synthetic speech sound.

To investigate 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]. Kiritan used the X-ray microbeam system to investigate coarticulation in VCV and CVC sequences [2]. Perkell et al. [3] and Matthies et al. [4] 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 [5]. 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, 8-9, 12]. However, there is no general agreement among the models since they resulted from different experimental data and conditions.

The authors proposed a model for coarticulation, named “carrier model” [7], based on two well-known models of the literature, the “perturbation model” [1] and the “look-ahead model” [9]. The former mainly focused on the principal-subordinate relation between vowels and consonants, while the latter paid a particular attention to the time order and anticipation. The carrier model takes the advantages of those two models. This study is to refine the carrier model based on the analysis of articulatory observations which were obtained from the electromagnetic articulographic system. Performance of the carrier model is evaluated by simulation using a physiological speech production model proposed by Dang and Honda [16].

2. Verification of Principal-Subordinate Structure
Dang et al. have proposed a model to describe the coarticulation during speech [7, 15]. The basic idea for the model is that an utterance in general can be considered as a stream consisting of consonants and vowels. In the stream, there are a vocalic “component” with strong and sustaining effects, and a consonantal “component” with relative weak and rapid effects. The look-ahead mechanism is applied to realize the interaction of adjacent phonemes of inner- and inter- components. By treating the vocalic component as “carrier wave” and the consonantal part as a “modulation signal”, the coarticulation can be considered as a modulation process. Therefore, the proposed model is referred to as a “carrier model”. The principal-subordinate speculate of the carrier model is similar to that used in Öhman’s model [1], which was derived from a spectrogram analysis. This section attempts to verify the carrier-modulation structure using articulatory data, which correspond to the articulatory movements directly.

2.1. Generalization of articulatory movements
It is difficult to verify the principle-subordinate structure speculated in the carrier model by using a specific phoneme sequence or contextual environment since it is context-dependent. For this reason, we analyze movement components of speech organs in frequency domain, and reconstruct a mean articulatory movement for the speech organs by averaging their frequency components. Such a mean characteristics is expected to reflect the inherent property of the speech organs in a general contextual environment.

The articulatory data used in this study were collected using the electromagnetic mid sagittal articulographic (EMMA) system in NTT, Japan [10]. Four receiver 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, 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 is located in the maxilla incisor, 0.5 cm upper its 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.

In this analysis, the 352 sentences were selected from the EMMA database, and used to generate text-independent articulatory movement. A two-second segment of speech was extracted from each sentence. The short-term DFT with 256 samples (about 1 sec.) was applied on the extracted segment, and frame shift was about 0.25 sec (64 samples). Complex spectra for T1, T2, and T3 were obtained respectively by averaging on all the frames of the short-term DFT. Figure 1 shows the average amplitude spectra of T1 and T3 for vertical (Y) dimension. The results indicate that the generalized movement of the tongue tip has stronger component in higher frequency region than the tongue dorsum.

To clarify the relation of the tongue tip and tongue dorsum, subtractions of T2 from T1 and T3 respectively are calculated and shown in Figure 3. The results show that the tongue tip and tongue dorsum have opposite phases one another. This implies that when the tongue tip rising to form an apical target, the tongue dorsum generally lowers its position for a synergic movement, and vice versa. Such an interaction is also a kind of coarticulation, which may concern with the physiological constraints of the volume.

2.2. Carrier and modulation waveforms in articulation

Generally speaking, vowel production has a strong and relatively slow movement that governs the whole tongue, while a consonantal movement is relative weak and rapid, which usually shows a local effect comparing with vowels. Since the constriction for the majority of the consonants is shaped by the tongue tip, here, T1 is roughly considered as a representative point for consonants (C), while T3 represents vowels (V). Because CV syllable is the basic unit in Japanese, we can reasonably suppose that the reconstructed articulatory movement corresponds to a phoneme sequence of CVCVCVCV for the one-second generalized utterance.

According to the above analysis, the tongue dorsum (T3) mainly concerns with vocalic stream of V_V_V_V excluding the consonants, while the tongue tip (T1) corresponds to CVCVCVCV. If this speculation is correct, the movement of the tongue tip should have about twice maximum (or minimum) peaks as that the tongue dorsum in the same period. To verify this hypothesis, velocities of T1 and T3 were calculated and are shown in Figure 4.
articulation can be separated into consonantal movement and vocalic movement, and the former is superposed on the latter. Note that the “vocalic component” here indicates the phoneme whose articulation constriction is formed with the tongue dorsum, so that the palatal consonants are classified into the vocalic component in this description.

3. Formulation of Coarticulation

According to the above analysis, an utterance can be formulated and quantified of the carrier model. In the processing, a given utterance is separated into two phoneme sequences as (2), where \(i\) and \(j\) are the indices of the consonants and vowels. If the first and/or the last phonemes in the utterance are not a vowel, the target vector of the neutral vowel is added. This section focuses on the formulation and quantification of the carrier model.

\[
V_1(\Theta) \rightarrow V_2 \rightarrow \cdots \rightarrow V_j \rightarrow V_{j+1} \rightarrow \cdots \rightarrow V_n(\Theta) \tag{2}
\]

3.1. Formulation of coarticulation

In the formulation, the first step is to construct the carrier wave. Articulatory movement is considered as a continuous movement from one vowel to another, where effects of one vowel to another depend on a degree of articulatory constraint (DAC). The DAC is denoted by \(d_{ij}\), which is a function of \(V_j\). 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 \(C_i\), as described in (3).

\[
G_i = (a_{ij} V_j + \beta d_{ij}, j+1 \gamma_{ij}) (a_{ij} V_j + \beta d_{ij+1}) \tag{3}
\]

where \(i\) and \(j\) are the indices of the consonants and vowels respectively, and \(a\) and \(\beta\) are the weight coefficients concerned with the tug-of-war in the look-ahead process.

The second process is to construct a resultant consonantal target \(C_i\) according to the “abstract” target \(C_i\) and virtual target \(G_i\) according to the following formula (4). Note that at this step only the specified feature is reconstructed, where no change happens in unspecified features since they depend on the coarticulation caused by the adjacent vowels.

\[
C_i = (r_{ci} C_i + G_i) (r_{ci} + 1) \tag{4}
\]

where \(r_{ci}\) is a coefficient of articulatory resistance for the specified feature of \(C_i\). When \(r_{ci}\) is sufficiently large, the effects of right to left will stop at this phoneme. In other words, this coefficient reflects how much the consonantal target can be affected by the virtual target.

The effects of consonants on vowels are taken into account via the look-ahead mechanism.

\[
V_j = (d_{ij} C_i + d_{ij} V_j)/(d_{ij} + d_{ij}) \tag{5}
\]

where \(i\) and \(j\) are the same as those of (3), \(d_{ij}\) is the DAC of consonant \(C_i\). Finally, the planned target sequence is obtained by the summation set of the principal and subordinate components of \(\{V_j\} \cup \{C_i\}\).

3.2. Quantification of the carrier model

In formulas (3)-(5), there are a number of undetermined coefficients in the carrier model. Those coefficients are determined in this part based on a statistical analysis of the articulatory data.

3.2.1. Estimation of DAC

To estimate the degree of articulatory constraint (DAC) for vowels, the phoneme sequences of \(V_j C_j V_i\) were segmented out of the read sentences. Here, \(V_j\) indicates the vowels on both sides, \(V_j\) is the central vowel, and \(C_i\) is a consonant. The distribution of the average location was investigated for each phoneme with a variety of contexts. The action of bilateral phonemes is to pull the central phoneme away from its average position. To evaluate the interaction between the adjacent phonemes, the displacement of the central phoneme apart from its average position is calculated (see [7] for details). The displacement of the tongue dorsum is calculated using (6) for five Japanese vowels, and treated as a measure of the DAC.

\[
d_{i} = \frac{1}{M} \sum_{m=0}^{M} \sum_{k=0}^{K} \mu (V_j, V_j | V_i) \tag{6}
\]

where \(\mu\) is a function to calculate the distance between the two locations, \(V_j\) is the average locations of vowel \(V_j\), \(V_j\) is the location depending on the bilateral vowel \(V_i\), \(C_i\) is any possible consonant existing in Japanese context.

Similarly, \(V_j C_i\) sequence is used to estimate the DAC for consonants by (7).

\[
d_{i} = \frac{1}{M} \sum_{m=0}^{M} \sum_{k=0}^{K} \mu (V_j, V_j | C_i) \tag{7}
\]

Table 1: The DAC of vowels and consonants.

<table>
<thead>
<tr>
<th>a</th>
<th>i</th>
<th>u</th>
<th>e</th>
<th>o</th>
<th>c</th>
<th>d</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dv</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Dv</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Dv</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Dv</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

3.2.2. Weighting coefficient for look-ahead mechanism

\(\alpha\) and \(\beta\) in formula (3) concern with the look-ahead mechanism to quantify those values, we calculated the displacement of the central consonant caused by the preceding and following vowels in \(V_j, C_j V_i\) sequence, respectively, using the following formulas.

\[
L_{\alpha} = \frac{1}{N} \sum_{m=0}^{N} \sum_{k=0}^{K} \mu (C_i, C_j | V_{ij}), \quad L_{\beta} = \frac{1}{N} \sum_{m=0}^{N} \sum_{k=0}^{K} \mu (C_i, C_j | V_{ij}) \tag{8}
\]

where \(N\) is the number of consonants. \(m\) and \(k\) are the combinations of the preceding and following vowels with consonant \(i\), respectively. The weighting coefficient for the look-ahead mechanism is calculated using (9).

\[
\alpha = L_{\alpha}, \quad \beta = L_{\beta} \tag{9}
\]

Applying (9) on the X- and Y-dimension respectively, the coefficients of \(\alpha\) and \(\beta\) for X-dimension are 0.33 and 0.67 respectively, while \(\alpha\) and \(\beta\) for Y-dimension are 0.46 and 0.54, respectively.

3.2.3. Coarticulation resistance coefficient

A coarticulation resistance coefficient \(r_{ci}\) is used in (4), which
reflects how tough a consonant resists vowels’ effects in the tug of war. VCVCV sequences are employed in the estimation of this coefficient, in which the bilateral vowels are the same. Based on the EMMA data, the distance between the average positions of vowels and consonants was calculated, and the displacement of the consonant induced by the bilateral vowels was measured. The articulatory resistance of a consonant reflects the ratio of the distance between the consonant and a vowel to the displacement induced by the vowel. Equation (10) gives the coefficient \( r_{ci} \) a description.

\[
 r_{ci} = \frac{1}{M} \sum_{j=1}^{M} \frac{L(C_i, C_j)}{L(C_j, V_j)}
\]

where \( M \) is the number of vowels concerned with consonant \( C_i \). The results are listed in Table 2, the \( r_{ci} \) for apical consonants is corresponding to the tongue tip, while it is corresponding to the tongue dorsum for the palatal consonants of /g/ and /k/.

<table>
<thead>
<tr>
<th>( r_{ci} )</th>
<th>d</th>
<th>n</th>
<th>s</th>
<th>l</th>
<th>z</th>
<th>g</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>k + d + b</td>
<td>9.0</td>
<td>2.1</td>
<td>21.4</td>
<td>11.3</td>
<td>11.9</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>k - d + b</td>
<td>10.3</td>
<td>10.9</td>
<td>11.9</td>
<td>13.5</td>
<td>16.9</td>
<td>2.0</td>
<td>2.8</td>
</tr>
</tbody>
</table>

4. Simulation

The carrier model is implemented in the physiological articulatory model [16] to take coarticulation into account in the target planning stage. VCVCV phoneme sequences were used in the simulation, in which the “abstract” target was out of context-independent typical target codebook. The coefficients obtained above were implemented in the formulas.

In the simulation, behaviors of the carrier model are evaluated by investigating of distributions of the tongue tip during the central vowel. VCVCV sequence used in the simulation consists of five Japanese vowels, apical consonants and two palatal consonants /k/ and /g/. Figure 5 shows the simulation result with the distribution obtained from EMMA data. The contours demonstrate the distribution of the tongue tip during the central vowels from EMMA data, while the circles illustrate the one obtained from simulation. The simulation shows a consistent distribution for vowel /e/, /i/ and /a/, while some differences are seen for vowel /u/ and /o/.

5. Conclusions

Adopting “carrier” concept, a phoneme sequence is treated as a carrier wave (vowel-to-vowel articulation) and a modulation signal (superimposed consonantal articulation). The concept of the carrier model is confirmed by means of reconstructed context-independent articulatory movements using complex Fourier series analysis. Based on the principal-subordinate structure, the carrier model is elaborated with the look-ahead mechanism. The parameters of the carrier model are estimated based on articulatory observations. After implementing the carrier model in the physiological articulatory model, distributions of the tongue tip and dorsum were simulated and compared with articulatory observations. As the result, the carrier model showed a good performance for the anticipatory function of coarticulation.

6. Acknowledgements

The authors especially thank NTT Communication Laboratories for permitting us to share the articulatory data. This research is mainly supported by a program for the Fostering Talent in Emergent Research Fields in Special Coordination Funds. This study is also supported in part by the National Institute of Information and Communications Technology and in part by Grant-in-Aid for Scientific Research of Japan (No. 17300182).

7. References