Implement of Coarticulation in Physiological Articulatory Model

Jianwu Dang[1,2]*, Kiyoshi Honda*[2]*, and Pascal Perrier*[3]

[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[2]ATR Human Information Science, Kyoto, Japan
[3]ICP CNRS UMR 5009 & INPG & University Stendhal, Grenoble, France
jdang@jaist.ac.jp; honda@atr.co.jp; perrier@icp.inpg.fr

Abstract

One of the fundamental issues to achieve a speech synthesizer using a physiological articulatory model is the implementation of coarticulation. This study focuses on the coarticulation occurred in the target planning stage, and proposes a "carrier model" to realize the coarticulation mechanism. Simulation results show that the carrier model demonstrates a good performance to account for the coarticulation.

1. Introduction

A physiological articulatory model has been constructed based on the displacement-based finite element method to replicates midsagittal regions of the speech organs [1]. To achieve a speech synthesizer using such an articulatory model, the important issue is to implement coarticulation in the model. This study adopts spatial targets in the control strategy. According to a commonly accepted concept, the speech production system is controlled by relative (dynamic) spatial targets, which give the best combination of the articulators for configuring a desired vocal tract shape [2]. As a first step, we used absolute (static) spatial targets instead of relative spatial targets, while such an absolute-target method can be extended to a relative-target method by using a sequence of time-varying or scattered absolute targets. The modeling of the coarticulation in this study is focused on the coarticulation taking place in the target planning stage while the coarticulation in the physiological level is naturally realized by the physiological articulatory model itself.

2. A "Carrier Model" for Coarticulation

Generally speaking, there are two types of coarticulatory overlap during natural speech: the carryover and anticipatory processes. In carryover coarticulation, the paths taken by the tongue, jaw, and lips as they move to a given target depend on the preceding target. Anticipatory coarticulation can occur only if a speaker can "look-ahead" in time and anticipate oncoming sounds. Carryover coarticulation must therefore reflect a high-level central type of phonological-phonetic processing, since an entire utterance must be scanned [3]. To describe this process, Henke proposed a phonemic-segment model [4]. Each segment was described in a matrix of articulatory target features, in which some features changed abruptly as the target shifted. Öhman [5] proposed another model to describe the mechanism of coarticulation, in which the articulation was represented by a basic diphthongal vowel gesture with an independent consonant gesture that is superimposed on its transitional portion during a vowel-consonant-vowel sequence.

Essentially, 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 [4]. In an utterance stream, therefore, there are a vocalic "component" with strong but sustaining effects, and a consonantal "component" with relative weak and rapid effects. Accordingly, we can use a concept of "carrier model" to reconsider such a mechanism. The vocal tract moves slowly and continuously from vowel to vowel, with periodic interference from rapidly amplitude- and frequency-modulated effects by consonants. As a result, coarticulation, the overlap properties of speech sounds, is built into the varying nature of the articulation process. The carrier model focuses on the principal-subordinate relation between vowels and consonants, while the "look-ahead" model [4] pays particular attention to the time order.

3. Target Planning for an Utterance

To plan articulation based on the conception of a "carrier" movement plus "modulation" movement, it requires us to separate a given utterance into vowel and consonant sequences. Therefore, for a given utterance, the vowels and consonants are separated into two phoneme sequences as the following.

$$\begin{tikzcd}[column sep=small] & C_1 \arrow[d] & \cdots\cdots & C_i \arrow[d] & \cdots\cdots & C_m \arrow[d] & \\ V_1(\text{ə}) \arrow[r, dash] & {} \arrow[r] & V_2 \cdots V_j \arrow[r, dash] & {} \arrow[r] & V_{j+1} \cdots V_{n-1} \arrow[r, dash] & {} \arrow[r] & V_n(\text{ə}) \end{tikzcd} \qquad (1)$$

To construct the "carrier wave", articulatory movement is considered as a continuous movement from one vowel to another. Thus, a virtual vocalic target G_i is created in the position of the consonant C_i by an interpolation between the neighboring vocalic targets. This procedure can be 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 phoneme is not a vowel, the target vector of the neutral vowel is added for the interpolation, as shown in (1). Based on the look-ahead mechanism, the following target has stronger effects on the virtual target than the preceding one in the target planning stage, where the carryover effects are realized by properties of the articulatory model in a low level. Therefore, the coefficient β should be larger than α. In our experiment, β is set to 0.7 and α to 0.3.

The second process accounts for the effect of the consonants on the vowels, where the effects between vowels and consonants are considered only for the immediately adjacent phonemes. This processing is carried out by the following formula.

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

where d_{ci} is coefficient based on the degree of the articulatory constraint of consonant C_i [6], and i is the index of consonants. One more processing is to account for the effect of the following consonant on the preceding vowel via the look-ahead mechanism.

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

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 indecisive 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 $\{\{V_j\} \cup \{C_i^{'}\}\}$.

4. Simulations

For a given phoneme sequence, all the targets of the phonemes used in this simulation are out of context-independent typical target codebook, which were average positions obtained from the X-ray microbeam data. The proposed model is implemented in the phoneme sequence with the typical target to take the coarticulation into account in the target planning stage. Figure 1 shows the results for phoneme sequences of /aka/ and /aki/ simulated by the physiological articulatory model [1]. The upper panel shows the horizontal component of the trajectory for tongue dorsum, and the middle panel shows the vertical component. The trajectories involved both of the coarticulation components, where the carryover effects are automatically accounted for by the model properties. One can see that the closure of the palatal consonant /k/ has different locations for the sequences. The location for /aki/ is anterior to that for /aka/. The preceding /a/ shows some differences in the given utterances. The lower panel shows the velocity of the dorsum during the articulation. The velocity of the dorsum for /aka/ has two outstanding peaks from /a/ to /k/ and from /k/ to /a/, while there is only one major peak for /aki/. These results are consistent with articulatory observation.

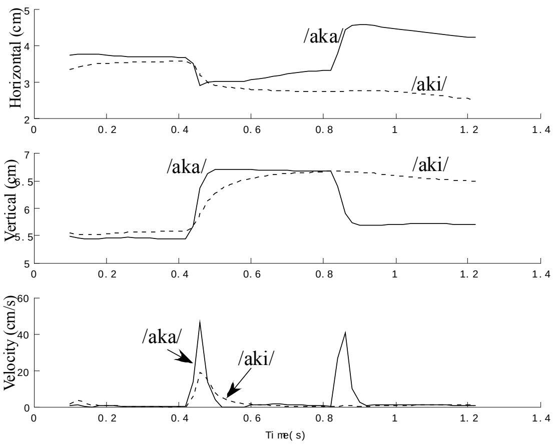

Figure 1. Trajectories of the horizontal (upper) and vertical (middle) components of the tongue dorsum. Velocities of the dorsum during the articulations.

5. Conclusions

In this study, we proposed a "carrier model" to deal with the coarticulation during speech, which treats a phoneme sequence as a carrier wave (vowel articulation) and a modulation signal (consonant articulation). Simulation showed that this model could account for the coarticulation in the target planning stage.

6. Acknowledgment

This research has been supported in part by the Telecommunications Advancement Organization of Japan.

7. References

[1] Dang, J. and Honda, K. "Construction and control of a physiological articulatory model," J. Acoust. Soc. Am. **115 (3),** (2004), in press

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

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

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

[5] Öham, S. (1966). "Coarticulation in VCV utterance: Spectrographic measurements," J. Acoust. Soc. Am. **39**, 151-168.

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