Observation and Modeling of Lingual Coarticulation in the Planning Stage

Jianguo Wei[1,3], Xugang Lu[1], Jianwu Dang[1] and Pascal Perrier[2]

[1]Japan Advanced Institute of Science and Technology, Ishikawa, 923-1292, Japan
[2]ICP CNRS UMR 5009 & INPG & University Stendhal, Grenoble, France
[3]LTCI CNRS UMR 5141 & Ecole Nationale Supérieure des Télécommunications , 75634 Paris, France
{Jianguo;Xugang,Jdang}@jaist.ac.jp; perrier@icp.inpg.fr

ABSTRACT

This study investigates and models anticipatory coarticulation taking place in different portions of tongue during speech production. Based on articulatory observations, we reconstructed generalized articulatory movement from articulatory data. It is found that the movements of the tongue dorsum and tongue tip can be treated as a carrier wave and a modulation in the articulation. Accordingly, a "carrier model" is proposed to describe this mechanism of anticipatory coarticulation at the planning stage. A model based learning process was used to refine the parameters of the carrier model. The simulations using the optimized coarticulation model and the learned typical phonetic targets were consistent with the observations within an average error of 0.18 cm in the articulatory space. The listening test showed that the synthesized sounds were improved by implementing the carrier model.

1. INTRODUCTION

Coarticulation in speech production has been studied for a long time, trying to account for its origin, nature and functions. Coarticulation models are expected to predict the details of the process bridging the invariant and discrete units of representation of articulation and acoustics[1].

Many researches on this issue mainly focus on coarticulation between different speech organs like lips vs. jaw, lip vs. tongue, tongue vs. velum, and so on. In this study, we concentrate on the coarticulation between different portions on the tongue (tongue tip and tongue dorsum), which is a more challenging issue and some researchers were working on it [2, 3]. The tongue tip and dorsum perform different actions during articulation, notwithstanding they belong to one single organ and have a high correlation. For instance, the tongue tip and dorsum constrict alterative articulatory places in many utterances such as /ata/. To well understand and model the coarticulation, we must investigate how the tongue tip and tongue dorsum coarticulate with each other in speech and how to model coarticulation between them.

In order to reveal the underlying mechanism of coarticulation, a novel observation method was introduced to reconstruct a general tongue movement. A "carrier model" , which take the advantages of Henke's look-ahead model [4] and Öhman's perturbation model[5], is introduced to model the coarticulation in the planning stage based on the finding in the observations. A model based learning process is used to refine the parameters of the carrier model.

2. OBSERVATIONS OF COARTICULATION

Focusing on speech organs, Öhman proposed a principal-subordinate structure based on spectrogram analyses and X-ray data, namely Öhman's model[5]. This section attempts to investigate whether this structure exists between articulation of the tongue tip and tongue dorsum.

To uncover the intrinsic properties of the coarticulation in real speech, we used continuous speech as the speech materials in this study, but not a vowel-consonant sequence embedded in a carrier sentence which is used in most of the studies. The question arrived at how to generate a context-independent coarticulatory environment to avoid the specific phoneme sequence. Since the articulatory movements are context-dependent, it is impossible to include all contexts in a single short sentence. For this reason, we analyze the movement of speech organs in the frequency domain, and reconstructed a generalized articulatory movement, which 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 midsagittal articulographic (EMMA) system in NTT, Tokyo, Japan [6]. 352 sentences of the EMMA database were selected to generate text-independent articulatory movement. Two-second segment of speech was extracted from each sentence. The short-term DFT with 256 samples (about 1 sec.) was applied to the extracted segments windowed by a hamming window, and frame shift was about 64 samples. Complex spectra for the tongue tip (T1) and the tongue dorsum (T3) were obtained respectively by averaging all the frames of the short-term DFT[7].

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, compared with vowels. Since the constriction of the apical consonants is shaped by the tongue tip, T1 is roughly considered as a representative point for consonants (C), while T3 represents vowels (V). Because a CV syllable is the basic unit in Japanese, we can reasonably suppose that the reconstructed articulatory movement corresponds to a phoneme sequence of CVCV...CV for the generalized utterance. According to the above analysis, the tongue dorsum (T3) mainly concerns the vocalic stream of V_V_V_V excluding the consonants, while the tongue tip (T1) corresponds to both C and V in CVCVCVCV. If this speculation is correct, the movement of tongue tip should have about twice as many stable points as that of tongue dorsum in the same period. To verify this hypothesis, velocities of T1 and T3 were calculated based on the frequency analysis, which are shown in Figure 1. In the central point of a phoneme, the articulators are in a steady-state position, where velocity is equal to zero. In Figure 1, there are 14 zeros for T1 and 8 zeros for T3. The number for the tongue tip is about twice as that of the dorsum. As may be noticed, the number of steady states of the tongue tip is slightly less than twice of the dorsum. There are two factors responsible for this phenomenon. One is that some vowel-to-vowel sequences exist in the utterances. The other is that the palatal consonant contributes to the constriction of the tongue dorsum rather than the tongue tip. Altogether, the results indicate that articulation can be separated into vocalic movement and consonantal movement, and the latter is superposed on the former. This finding indicates that the tongue tip and tongue dorsum can be treated as two independent parts in producing speech, which coarticulate each other obeying the principal-subordinate structure.

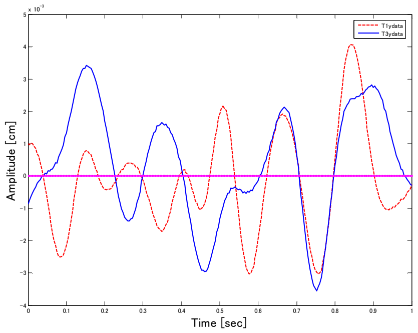
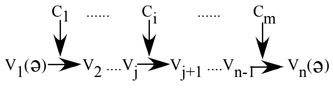


Figure 1. Velocity of reconstructed waveforms of T1&T3.

3. FORMULATION OF COARTICULATION

Based on the above analysis, an utterance in general can be considered as a stream consisting of vowels and consonants, in which an utterance can be illustrated as a principal-subordinate structure. The look-ahead mechanism is applied to realize the interaction of adjacent phonemes of inner- and inter- components. During this processing, a given utterance can be separated into two phoneme sequences as (1), where i and j are the indices of the consonants and vowels.



(1)

To realize the modulation processing, the first step is to construct the carrier wave, in which articulatory movement is considered as a continuous movement from one vowel to another. To construct the carrier wave, if the first and/or the last phonemes of the utterance are not vowels, the target vector of a neutral vowel is added preceding the first phoneme and/or following the last phoneme in the utterance. In this study, we used a degree of articulatory constraint (DAC) [2, 3] to describe the effect of each phonemes on its neighbor phonemes, so that the effects of vowel V_j on the adjacent ones depend on its DAC, denoted by d_{v_j}. Since the resultant target of consonant C_i is affected by a "tug-of-war" effect from its bilateral vocalic targets, a virtual target G_i would be introduced at the position of C_i using (2).

$$G_i = (\alpha d_{v_j} V_j + \beta d_{v_{j+1}} V_{j+1}) / (\alpha d_{v_j} + \beta d_{v_{j+1}}) \tag{2}$$

where i and j are the indices of the consonants and vowels respectively, and α and β are the weighting 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 typical phonetic target C_i and virtual target G_i according to the formula (3). Note that at this step only the crucial feature, for instance, the target of the tongue tip for /t/, is reconstructed, where no change happens in indecisive features since they depend on the coarticulation caused by the adjacent vowels.

$$C_i' = (r_{c_i} C_i + G_i) / (r_{c_i} + 1) \tag{3}$$

where r_{c_i} is a coefficient of articulatory resistance for the crucial feature of C_i. This coefficient reflects the capability of the consonants against the effects of neighbour vowels while DAC describes the capability that the concerned phoneme affects its neighbours. The larger value r_{c_i} the weaker effects are accepted from the neighboring vowels.

The effects of the consonants on the vowels are taken into account via the anticipation mechanism as:

$$V_j' = (d_{c_i} C_i' + d_{v_j} V_j) / (d_{c_i} + d_{v_j}) \tag{4}$$

where i and j are the same as those of (2), and d_{c_i} 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 $\{\{V_j'\} \cup \{C_i'\}\}$.

4. OPTIMIZATION FRAMEWORK

A brief flowchart of the procedure used in human speech production is shown in the left panel of Figure 2. Here, we suppose that there is a unique spatial target corresponding to each phonetic unit of speech, referred to as typical phonetic target. In the speech production process, there is a series of commands corresponding to each phoneme in different context. We defined the context dependent commands as planned targets, which reflect the variations of the typical phonetic target with its environments. In the flowchart, the planned targets are generated from the typical phonetic targets of a phoneme sequence in the planning stage based on the anticipation mechanism.

In the simulation, the carrier model transforms the typical phonetic targets to planned targets based on contextual information. Unfortunately, the typical phonetic targets of phonemes in the phonetic planning level can not be observed directly. In addition, we also can not observe the planned targets directly. Consequently, questions arise as to obtain the parameters of the carrier model and to estimate the typical phonetic targets. In the current situation, we can obtain articulatory movements of human from the EMMA system. If there is a physiological articulatory model that has identical functions as human at the physiological and kinematical levels, it is possible to obtain reliable planned targets by tuning the inputs of the model to match the observations. Based on this consideration, we propose a physiological model based learning process to acquire the planned targets. The physiological articulatory model adopted in this research is a partial 3D physiological articulatory model that was constructed based on volumetric MRI data obtained from a male Japanese speaker [8]. This model consists of the tongue, jaw, hyoid bone and vocal-tract wall. The muscular structure of the model was designed based on MRI measurements and anatomical literature. On the physiological level, this model has a high consistency with human.

The model based learning process is shown in the right panel of Figure 2. , which has each counterparts corresponding to the human speech production procedure. In this learning process, the physiological articulatory model and the carrier model both are base on the mechanism of human speech production and/or the observations of articulatory movements. Therefore, the learned parameters have certain physical meanings.

The main focus of this study is on the coarticulation involved in the anticipation. Actually, the observed articulatory data contained both the effects of carryover coarticulation and anticipatory coarticulation. In order to separate the anticipatory coarticulation from the carryover effect, we split the learning framework into two layers, a low layer concerning with the carryover effect and a high layer associating with anticipatory effect.

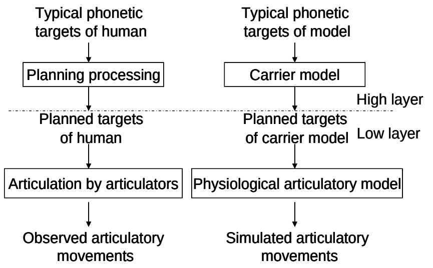

Typical phonetic targets of human
Typical phonetic targets of model
Planning processing
Carrier model
High layer
Planned targets of human
Planned targets of carrier model
Low layer
Articulation by articulators
Physiological articulatory model
Observed articulatory movements
Simulated articulatory movements

Figure2. Speech production procedure of human & model

3.1. Learning planned targets in the low layer

In the low layer, the "true" planned targets can be expected if the differences between the model simulations and observations are reduced, which is formulated as (5):

$$\{T^*_{p_v}, T^*_{p_c}\} = \arg\min_{T_{p_v}, T_{p_c}} [\upsilon(M_{o_v} - M_{s_v})^2 + (1-\upsilon)(M_{o_c} - M_{s_c})^2] \qquad (5)$$

where T_{p_v} and T_{p_c} denote the planned target of preceding vowel and central consonant, respectively, in vowel-consonant-vowel (VCV) sequences. M_{o_v} and M_{o_c} are the observed movements obtained from EMMA data, while M_{s_v} and M_{s_c} are the simulated movements of vowels and consonants. Since consonants are more sensitive to articulation places than vowels, we used the weighting coefficient υ to emphasize the locations of consonants.

3.2. Learning for the carrier model and the typical phonetic targets

The objective functions of high layer are described in (6) and (7), where C_i'' and V_j'' denote the planned targets obtained from the low layer for consonants and vowels, respectively. C_i' and V_j' are the planned targets derived from the carrier model. K is the number of VCV combinations used in the learning process. The parameters of the carrier model and the typical phonetic targets can be learned by minimizing the objective functions.

$$l(V_j, d_{v_j}, d_{c_i}, C_i') = \sum_{k=1}^{K} (V_k'(V_j, d_{v_j}, d_{c_i}, C_i') - V_k'')^2 \qquad (6)$$

$$f(r_{c_i}, C_i, d_{v_j}, \alpha, V_j) = \sum_{k=1}^{K} (C_k'(r_{c_i}, C_i, d_{v_j}, V_j) - C_k'')^2 \qquad (7)$$

The optimization processing is depicted as (8):

$$\min_{r_{c_i}, d_{c_i}, d_{v_j}, \alpha, V_j, C_i} \gamma l + (1-\gamma) f \qquad (8)$$

where γ are the weighting coefficients of the $l(\cdot)$ and $f(\cdot)$.

5. EXPERIMENTS OF THE LEARNING PROCESS

153 VCV combinations were extracted from the database, which consisted of five Japanese vowels /a/, /i/, /u/, /e/, /o/ and eight consonants /d/, /g/, /k/, /n/, /r/, /s/, /t/, /w/. To evaluate the whole learning process, speech production procedure is simulated from the learned typical phonetic targets to articulatory movements using the optimized carrier model and the physiological articulatory model. The distributions of the simulations of consonants are shown in Figure 3, where the diamonds denote the simulations and the crosses for the observations. The letters in each panel denote the learned typical phonetic targets. One can see that the learned phonetic targets for the consonants with a closure between the tongue and the palate such as /d/, /t/, /n/,/r/,/k/ and /g/ were beyond the hard palate, while the targets of fricative /s/ and semivowel /w/ were located inside the vocal tract. This implies that to form a closure between the tongue and the palate for those consonants the phonetic targets should be beyond the hard palate. These results confirmed the hypothesis that such consonants usually have virtual targets over the hard palate [9, 10].

Subjective evaluations of the optimization results were conducted using A-B comparison listening test. In which, speech sounds were synthesized using the physiological articulatory model based synthesizer under three conditions: Condition 1 is based on the targets observed from EMMA data without the carrier model, Condition 2 is based on the learned phonetic target without the carrier model, and Condition 3 is synthesized from the learned typical phonetic targets with the optimized carrier model. The 153 VCV combinations were synthesized under each condition, in which 40 VCV combinations was randomly extracted as the speech materials. Eighteen volunteers evaluated three groups using the paired A-B comparison listening test method[11]. The results have been shown in Table 1 and Table 2. In the A-B comparison listening test, a speech material pair from two different speech groups was listened by the subjects, and choose the better one from the two speech samples, or choose “unknown” if no prefer. These results showed that the naturalness of synthesized speech sound improved when the carrier model was implemented.

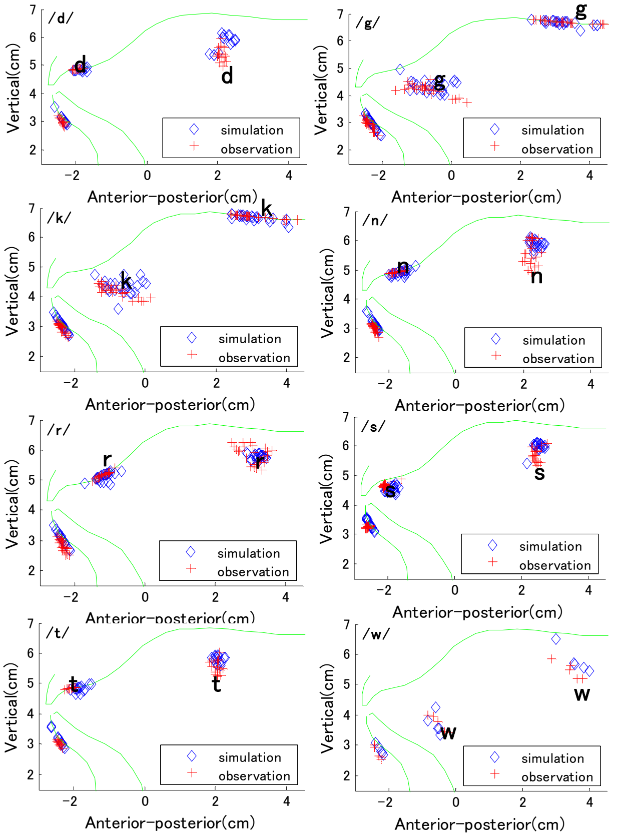


Figure 3. Articulatory movements from observations vs. simulations via the whole framework

Table 1 : The average choice rate of trial 1 and trail 2

	Trial1	Trial2	Unknown
Average percentage (%)	**10.97**	**76.11**	**12.92**
Stander deviation of percentage (%)	**2.99**	**4.64**	**3.12**

Table 2 : The average choice rate of trial 2 and trail 3.

	Trial2	Trial3	Unknown
Average percentage (%)	**15.56**	**68.61**	**15.83**
Stander deviation of percentage (%)	**2.79**	**3.45**	**2.27**

6. SUMMARY

In this paper we introduced the formulations of the carrier model that realizes the computational function for anticipatory coarticulation between tongue tip and tongue dorsum. The carrier-modulation structure in the articulatory domain was verified by reconstructing a generalized articulatory movement for the speech organs. A physiological articulatory model-based optimization framework was proposed, to obtain the typical phonetic targets in the planning stage, and refine the parameters of the carrier model. The learned typical phonetic targets of the consonants with closure were located beyond the hard palate, which is consistent with the common hypothesis that such consonants usually have overshot targets. The listening test results showed that the carrier model was confirmed by the synthesized sound.

Acknowledgements

The authors especially thank NTT Communication Laboratories for the permission to use the articulatory data. This study was supported in part by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan and in part by Grant-in-Aid for Scientific Research of Japan (No. 17300182).

BIBLIOGRAPHY

[1] E. Farnetani and D. Recasens, "Coarticulation models in recent speech production theories," in *Coarticulation, theory, data and techniques*, H. W. J. H. N, Ed. Cambridge: CUP, 1999, pp. 31-65.

[2] D. Recasens, "An EMA study of VCV coarticulatory direction," *Journal of the Acoustical Society of America,* vol. 111, pp. 2828-2841, 2002.

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

[4] L. Henke, "Dynamic articulatory model of speech production using computer simulation." vol. MIT, 1966.

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

[6] T. Okadome and M. Honda, "Generation of articulatory movements by using a kinematic triphone model," *J. Acoust. Soc. Am,* pp. 453-463, 2001.

[7] J. Dang, J. Wei, T. Suzuki, and P. Perrier, "Investigation and Modeling of Coarticulation during Speech," in *Eurospeech2005*, Lisbon, Portugal, 2005, pp. 1025-1028.

[8] J. Dang and K. Honda, "Construction and control of a physiological articulatory model," *J. Acoust. Soc. Am.,* vol. 115, pp. 853-870, 2004a.

[9] A. Löfqvist and V. L. Gracco, "Control of oral closure in lingual stop consonant production," *J. Acoust. Soc. Am,* vol. 111, pp. 2811–2827, 2002.

[10]S. Fuchs, P. Perrier, and C. Mooshammer, "The role of the palate in tongue kinematics: An experimental assessment in VC sequences," in *Eurospeech*, 2001.

[11]R. Cox, J. Snyder, R. Crochiere, D. Bock, and J. Johnston, " Testing of wideband digital coders," in *IEEE International Conference on ICASSP*, 1984.