社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

Optimization and Evaluation of a Coarticulation Model based on Observation and Simulation

Jianguo Wei† Xugang Lu† and Jianwu Dang †

†Japan Advanced Institute of Science and Technology 1-1, Asahidai, Nomi-shi, Ishikawa, 923-1292, Japan

E-mail: †{jianguo, xugang, jdang}@jaist.ac.jp

Abstract: A coarticulation model, namely 'carrier model', has been proposed previously by Dang et al. to improve the performance of a physiological articulatory model based speech synthesizer. The carrier model offers a good framework to account for coarticulation in the planning stage, while its parameters need to be refined for improving the performance of the model. This study is to refine the parameters of the carrier model and estimate typical phonetic targets by minimizing the differences between model simulations and observations. A simulation based optimization framework is proposed for this purpose. The framework consists of two layers: obtaining planned targets in a low layer; estimating phonetic targets and optimizing the parameters in a high layer. A direct search method was applied to the low layer due to the non-analytic nature of the articulation model, while the high layer adopts bilevel optimization strategy to decompose the complicated problem into a set of subproblems. Objective and subjective evaluation were conducted by combining the refined carrier model and the learned phonetic targets together using the physiological articulatory model and the average error between observations and simulations was 0.15 cm over 153 VCV combinations on the jaw, tongue tip and tongue dorsum, meanwhile mean opinion score(MOS) were improved about 0.28 compared with the sound synthesized by averaged target obtained from electromagnetic midsagittal articulographic (EMMA) data through the physiological articulatory model.

Keyword coarticulation, physiological articulatory model, bilevel optimization, simulation

1. Introduction

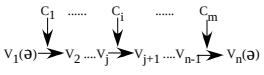
Coarticulation is a longstanding issue, which brings naturalness to speech sounds, since it is necessary to be taken into account for high-quality synthetic speech sound. Dang et al. proposed a computable model for coarticulation, named "carrier model" [1,2], based on two well-known models; the "perturbation model" [3] and the "look-ahead model" [4]. The former mainly focused on the principal-subordinate relation between vowels and consonants, while the latter paid particular attention to time order and anticipation. The carrier model takes advantages of those two models so as to provide a good framework to account for coarticulation in the planning stage. The initial parameters of the carrier model came from EMMA data by means of statistical method [2]. There, however, is no guarantee that those parameters reached the optimal values. The typical phonetic targets of phonemes in the phonetic planning level need to be identified based on observations.

In order to refine the parameters we implemented the carrier model in a physiological articulatory model to construct an optimization framework based on the simulation method. To reduce the complexity and computational cost of the simulation procedure, we divided the optimization procedure into two parts: a high layer and a low layer. Different optimization strategies are adopted in the different layers according to their own properties. In order to assess the optimization performance, objective and subjective evaluation were conducted. We calculated the averaged distance between observation and simulation to objectively assess the optimization result in spatial space and listening test was served as the subjective evaluation in perception space.

2. The optimization framework

First, we briefly describe the idea and concept of the carrier model. During speech production, two types of coarticulation can be identified: carryover and look-ahead. The carrier model mainly focuses on look-ahead coarticulation in the planning stage while

the carryover is supposed to be realized by the physiological properties. Articulatory movement for speech can be considered consisting of a principal (vocalic) component and subordinate (consonantal) component. Thus, a given utterance can be separated into two phoneme streams as shown in (1), where i and j are the indices of the consonants and vowels respectively.



(1)

Based on this process, the planned targets are obtained by applying the carrier model to the typical phonetic targets, which are supposed to be constant for each phoneme in the phonetic planning level. The planned targets are used to drive the physiological articulatory model to produce articulatory movements and speech sound.

If we have a physiological articulatory model that can model human mechanism at the physiological and kinematical levels, the objective of the above processing arrived at how to obtain the typical phonetic targets and how to refine the parameters of the carrier model. Actually, the observed articulatory data reflect both the effects of carryover coarticulation caused by physiological properties of the articulators, and the look-ahead coarticulation in the high level. We propose a physiological articulatory model based optimization by analogizing speech production processing in humans and in simulation, which is shown in Fig. 1. The left panel of the figure shows the speech production procedure of humans, where the planned target is generated from typical phonetic targets based on the look-ahead mechanism in the planning stage, and the planned targets are applied to drive the articulators to produce speech movements and then speech sound. The right part of the figure shows the proposed simulation framework which has every counterpart corresponding to the human speech production procedure.

In this simulation based optimization framework, the computation cost is mainly due to the computation of the physiological articulatory model. If simulation based optimization is carried out according to a flowchart step by step, processing is time consuming. To reduce complexities of the simulation, we divided it into a low layer and a high layer. The low layer mainly focuses on the physiological process, in which planned targets drive a physiological articulatory model to produce articulatory movements. In contrast, the high layer mainly focuses on optimizing the coefficients of the carrier model and learning typical phonetic targets by minimizing the distance of the output of the carrier model and the planned targets obtained from the lower layer. This two-layer simulation framework is of great benefit in the reduction of computation cost.

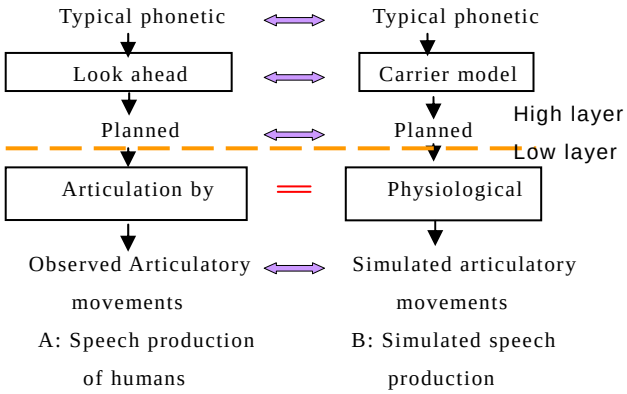


Figure 1 Comparison of speech production processes of human and in the simulations

3. Optimization in the low layer

This section describes the optimization procedure in the low layer.

3.1 Strategy of optimization used in the low layer

The optimization in the low layer from the articulatory movements to planned targets is shown in Fig. 1. The difference between observations and model simulations can be considered to be caused by the planned targets if the model can perform the identical functions as humans. Since it was proved that the physiological articulatory model realizes human articulation very well [5], we can say that “true” planned targets can be obtained if the differences between the simulations and observations are reduced. Therefore, the planned targets can be obtained by minimizing the distance between simulated articulatory movements and observed articulatory movements.

Since the physiological articulatory model is constructed by the finite element method, there is no analytic formula to directly describe the relation between the planned targets and the simulation outputs. According to the nature of the problem, a mesh adaptive direct search algorithm (MADS) is

adopted to serve as the optimizer, which is designed to adapt to derivative free optimization problems [7]. In contrast to traditional optimization methods using the gradient or higher derivatives to search an optimal point, the direct search algorithm searches an area around the current optimal point, looking for any point whose objective value is lower than that of the current point.

3.2 Formulation of objective function in low layer

The purpose of this layer is to obtain the planned targets by minimizing the distance between observed articulatory movements and the simulated articulatory movements.

$$\{T_{p_v}^*, T_{p_c}^*\} = \arg\min_{T_{p_v}, T_{p_c}} [(M_{o_v} - M_{s_v})^2 + (M_{o_c} - M_{s_c})^2] \quad (2)$$

where T_{p_v} and T_{p_c} denote the planned target of preceding vowel and central consonant respectively in VCV tokens, which are six dimensional vectors including x and y dimensions of the jaw, tongue tip and tongue dorsum. M_{o_v} and M_{o_c} are the observed movements obtained from EMMA data, while M_{s_v} and M_{s_c} correspond to simulated movements of vowels and consonants respectively.

4. Optimization in the high layer

This section describes the optimization procedure in the high layer.

4.1 The carrier model

The basis of the carrier model is an assumption that articulatory movement for speech consists of a principal (vocalic) component and subordinate (consonantal) component. As illustrated in (1), the planned target of consonant C_i is affected by a "tug-of-war" of the adjacent vocalic targets, while the vocalic target is also affected by adjacent consonants. Two steps are employed in constructing the planned target. The first step is to construct a virtual target G_i in the position of C_i as shown in (3).

$$G_i = (\alpha d_{v_j} V_j + \beta d_{v_{j+1}} V_{j+1}) / (\alpha d_{v_j} + \beta d_{v_{j+1}}) \quad (3)$$

where i and j are the same as in (1), α and β are the weighting coefficients concerned with the tug-of-war in the look-ahead process. d_{v_j} is the degree of articulatory constraint (DAC)[6] of vowel V_j. The second step is to construct a planned consonantal target $C_i^{'}$ according to the phonetic target C_i and virtual target G_i according to formula (4).

$$C_i^{'} = (r_{c_i} C_i + G_i) / (r_{c_i} + 1) \quad (4)$$

where r_{c_i} is the coefficient of articulatory resistance for the crucial feature of C_i.

The effects of consonants on vowels are taken into account via the look-ahead mechanism in (5).

$$V_j^{'} = (d_{c_i} C_i^{'} + d_{v_j} V_j) / (d_{c_i} + d_{v_j}) \quad (5)$$

where i and j are the same as in (1), and d_{c_i} is the DAC of consonant C_i. Finally, the planned target sequence is obtained by the summation sets of the principal and subordinate components of $\{\{V_j^{'}\} \cup \{C_i^{'}\}\}$ [2].

4.2 Strategy of optimization in the high layer

4.2.1 Objective function

Objective function of this part is to measure the difference between the learned planned target and the one calculated based on the phonetic target and the carrier model. The parameters of the carrier model and the "true" phonetic target are learned by minimizing the objective function. Here, C_i" and V_j" denote the planned targets of consonants and vowels respectively, which were obtained from the low layer. C_i' and V_j ' are the output of the carrier model. The errors for vowels and consonants are defined by (6) and (7) respectively.

$$F(\mathrm{V_j}, \mathrm{d_{c_i}}, \mathrm{d_{v_j}} \mathrm{C_k^{'}}) = \sum_{\mathrm{j}=1}^{\mathrm{K}} (V_{\mathrm{j}}^{'} - \mathrm{V_j^{''}})^2 \quad (6)$$

$$f(\mathrm{C_i}, \mathrm{r_{c_i}}, \mathrm{d_{v_j}}, \mathrm{V_j}) = \sum_{\mathrm{i}=1}^{K} (C_{\mathrm{i}}^{'} - \mathrm{C_i^{''}})^2 \quad (7)$$

Then the task is to minimize the objective function as follow (8):

$$\min_{\mathrm{d_{v_j}}, \mathrm{C_i}, d_{\mathrm{c_i}}, \mathrm{V_j}, \mathrm{r_{c_i}}} \gamma F + \eta f \qquad (\gamma + \eta = 1) \quad (8)$$

where γ and η are the weighting coefficients of the $F(.)$ and $f(.)$. When this objective function is applied to the tongue tip, η=0.8, since the tongue tip is the crucial feature of apical consonants, while $\gamma = 0.8$ is adopted in the tongue dorsum, because the dorsum is the crucial feature for vowels. K is the number of VCV combinations.

4.2.2 Bilevel optimization

From the idea of the carrier model, we can see that the planning procedure of vowels and consonants can be considered as a tug-of-war condition, the positions and features of vowels and consonants

affect each other. In light of this effect, bilevel optimization method was adopted in the high layer, which is suitable for this kind of problem.

The bilevel programming problem (BLPP) is an optimization problem stemming from the Stackelberg game. In a Stackelberg game, the leader knows that the follower will respond to any decision he makes, but the leader can not control the follower's responses [8]. At each level the decision makers can optimize its variables for reaching their objective, but may be partially affected by variables controlled by others. BLPP is often used in decomposition procedures [8]. In the optimization function (8), the first part focuses on optimizing the objectives of vowels and the second part on the objective of consonants, and each part affects the opposite part by shared variables. In our case, the formulation of (6) and (7) can be described as a Stackelberg game. So we can decompose this problem into 2 subproblems based on the bilevel method shown in (9).

The MADS optimization method serves as the optimizer at both levels. Because we can not guarantee that the follower level is in a convex region, the traditional bilevel optimizer can not work well. The vectors V_j and d_{c_i} are dealt with in the leader level, while the vectors r_{c_i}, d_{v_j}, and C_i are treated in the follower level.

$$\min_{V_j, d_{c_i}} \sum_{k=1}^{K} (\gamma l + \eta f) \tag{9a}$$

where C' and d_{v_j} solve $\min_{C_i, r_{c_i}, d_{v_j}} \sum_{k=1}^{K} (\gamma l + \eta f)$ (9b)

$$0 < d_{v_j}, d_{c_i}, r_{c_i} \tag{9c}$$

The initial values of C_i and V_j are in the rational region as decided by the physiological articulatory model empirically.

5. Experiment and results of this simulation based optimization.

5.1 Observation data

In this numerical experiment, the NTT EMMA observations were employed [9]. 153 VCV combinations were extracted from the EMMA data, they consist of five Japanese vowels /a/,/i/,/u/,/e/,/o/ and eight consonants /d/,/g/,/k/,/n/,/r/,/s/,/t/,/w/ which were used in the learning process. Each phoneme is represented by a vector of the positions of the jaw, tongue tip and tongue dorsum.

5.2 Experiment in the low layer

In the low layer we obtained planned targets for VCV combinations consisting of five Japanese vowels and eight consonants by means of the optimization method. For each token, 90 iterations were carried out in the optimization. The convergence curve of the optimization is shown in Figure 2. One can see that the optimization error is reduced as the iteration times increases.

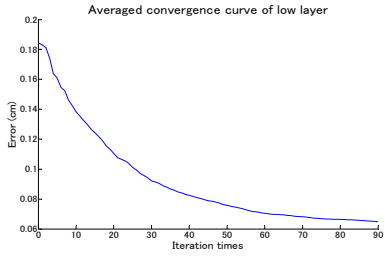


Figure 2 Convergence curve of low layer

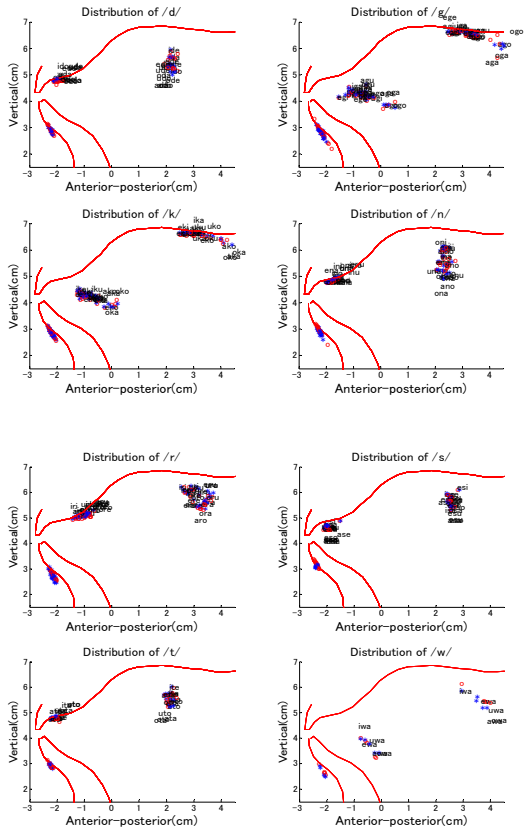


Figure 3 Distribution of observed and simulated articulatory movements of 8 consonants in the low layer. The circles denote the simulations while stars show the observed data. The VCV tokens are the learned planned

targets.

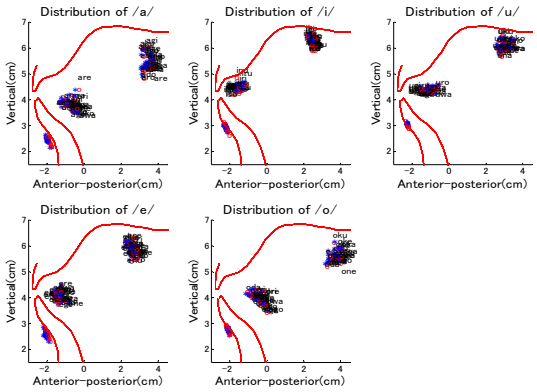


Figure 4 Distribution of observed and simulated articulatory movements of 5 vowels in the low layer. The symbols represent the same meaning as used in Figure 3.

Figures 3 and 4 show the jaw, tongue tip and tongue dorsum distribution of simulated results using learned planned targets through a physiological articulatory model. One can see that the articulators' movements of simulations and observations have almost identical distributions. The average error between them is 0.065 cm. Figure 3 shows the distribution of 8 consonants in the low layer. Most of the planned targets for the consonants with closure are over the hard palate. Figure 4 shows that the planned targets of vowels for tongue tip have a movement tendency towards the following consonants.

5.3 Experiments in the high layer

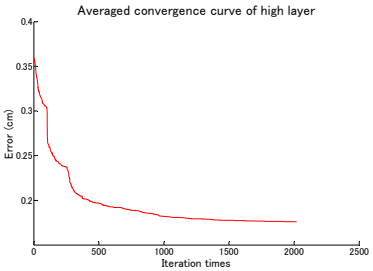


Figure 5 Convergence curve of high layer

The bilevel decomposition optimization method was used in the high layer, in which the loop includes leader and follower levels. The loop was run 10 times, each time the MADS ran 100 iterations for leader and follower levels, respectively. Figure 5 shows the averaged error over x-dimension and y-dimension of the tongue tip and dorsum in the high layer. The optimization error curve becomes flat when the iterations are over 2000 times. The unsmooth features in the convergence curve are caused by switching levels in the bilevel method. The averaged error was 0.178 cm between the planned targets learned in the low layer and ones calculated by the optimized carrier model using the learned phonetic targets.

6. Evaluation

6.1 Objective evaluation

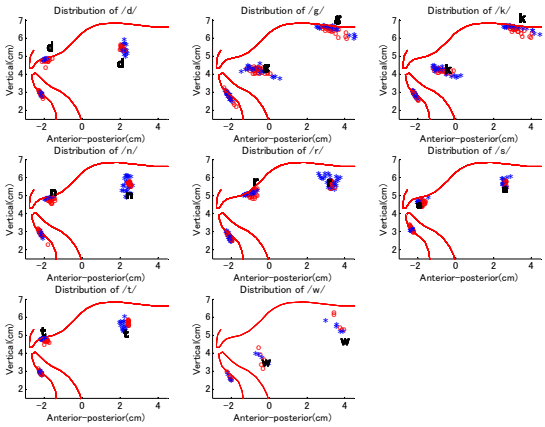


Figure 6 Distribution of observed and simulated articulatory movements of 8 consonants through whole framework. The circles denote the simulations, stars show the observations. Each black phoneme is a typical phonetic target.

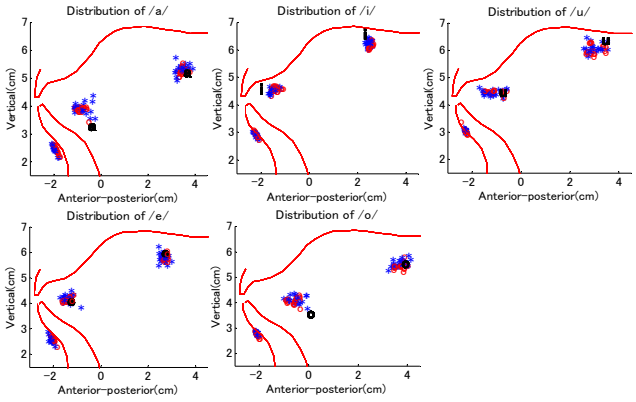


Figure 7 Distribution of observed and simulated articulatory movements of 5 vowels. The symbols represent the same meaning as used in Figure 6.

The distributions of simulations from the phonetic targets obtained by the carrier model and the physiological articulatory model are shown in Fig. 6 and 7, where the observations were plotted using different symbols for comparison. The average error

between the simulations and observations was 0.15cm. One can see that the phonetic targets for the apical consonants with a closure such as /d/, /g/,/k/,/t/, /n/, and /r/ were beyond the hard palate, while fricative /s/ and semivowel /w/ were located inside the vocal tract. This implies that the phonetic targets should be virtual one beyond the hard palate to form closure with the apex.

6.2 Subjective evaluations

Subjective evaluations of optimization result were performed by a listening test. In the listening test, there are two groups of synthesized sounds using the physiological articulatory model based synthesizer; one group is synthesized from the learned typical phonetic targets with the carrier model, and the other group is based on averaged targets observed from EMMA data. 40 combinations out of each group were used as listening speech database in the learning test. Seven volunteers evaluated each speech by mean opinion score (MOS)[12] method on a scale of 1 to 5(1-very unnaturalness, 2-unnaturalness, 3-neutral, 4-naturalness, 5-very naturalness) through binaural headphones at a comfortable loudness level in a sound-proof room. The speeches were played using a laptop, which was controlled by the subject to repeat the speech until the score can be given, but it did not permit to rescore. The results are shown in Table 1. One can see that the MOS score gets higher when the learned typical targets and carrier model were applied. The T-test results showed that there are significant differences between these two groups at 1% significance level.

Table 1. The value of MOS obtained in two groups

	subjects	Mean	SD	SE
Group A*	7	2.879	0.333	0.1258
Group B*	7	3.181	0.424	0.1803
Difference	7	0.282	0.143	0.0542

Group A: the sound synthesized by averaged target from EMMA with the physiological articulatory model.

Group B: synthesized by carrier model with typical phonetic target using the physiological articulatory model

7. Conclusion

In this paper we proposed a simulation-based optimization framework for obtaining the typical phonetic targets in the phonetic planning stage, and for refining the parameters of the carrier model simultaneously. The distributions of simulated articulatory movements are consistent with the EMMA-based observations with an average error of 0.15 cm. The MOS has been improved 0.28, which implies that the naturalness of synthesized speech sound is improved by using the carrier model. The learned typical phonetic targets of the apical consonants with closure showed the overshot properties beyond the hard palate, which is consistent with the hypothesis that such consonants usually have virtual targets [10,11]. These results indicated that the optimal values of coefficients of carrier model have been obtained for eight consonants and five vowels.

8. 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 Grant-in-Aid for Scientific Research of Japan (No. 17300182).

9. Reference

[1]Dang, J., Honda, K., and Perrier, P., “Implement of Coarticulation in Physiological Articulatory Model,” International Congress of Acoustics 2004, 1335-1338.

[2]Dang, J., Wei, J., Suzuki, T., Perrier, P., “Investigation and modeling of Coarticulation during Speech,” Europeech2005, 1025-1028.

[3]Öhman, S., “Coarticulation in VCV utterance: Spectrographic measurements,” *J. Acoust. Soc. Am*, 39, 151-188, 1988.

[4]Henke, L., “Dynamic articulatory model of speech production using computer simulation.” Doctoral dissertation, MIT, 1988.

[5]Dang, J., Honda, K., “Construction and control of a physiological articulatory model,” *J. Acoust. Soc. Am*,, 115(2), 853-870, 2004.

[8]Recasens, D., Pallares, M., and Fontdevila, J., “A model of lingual coarticulation based on articulatory constraints,” *J. Acoust. Soc. Am,* 102, 544-581, 1997.

[7]Audet, C.,Orban, D., “Finding optimal algorithmic parameters using a mesh adaptive direct search”, *Les Cahiers du GERAD* G-2004-98, Montreal.

[8]J.F. Bard, Practical Bilevel Optimization: Algorithms and Applications, Kluwer Academic Publishers, 1998.

[9]Okadome, T. and Honda, M., “Generation of articulatory movements by using a kinematic triphone model,”*J. Acoust. Soc. Am*453-483 ,2001.

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

[11]Löfqvist, A. & Gracco, V. L.,“Control of oral closure in lingual stop consonant production” *J. Acoust. Soc. Am,* 111 (8) 2811–2827, 2002.

[12] ITU-T(1993) “A method for subjective performance assessment of the quality of speech voice output devices”, Draft ITU-T Recommendation P.85, COM 12-R 8.