Feedforward Control of A 3D Physiological Articulatory Model for Vowel Production

Qiang Fang[1], *Akikazu Nishikido*[2], *Jianwu Dang*[2], *Aijun Li*[1]

[1] Phonetics Lab., Institute of Linguistics, Chinese Academy of Social Sciences
[2] IIPL, School of Information Science, Japan Advanced Institute of Science and Technology
fq0237@gmail.com, a-nishi@jaist.ac.jp, jdang@jaist.ac.jp, liaj@cass.org.cn

Abstract

A 3D Physiological articulatory model has been developed to account for the biomechanical properties of speech organs in speech production. To control the model for investigating the mechanism of speech production, a feedforward control strategy is necessary to generate proper muscle activations according to desired articulatory targets. In this paper, we elaborated a feedforward control module for the 3D physiological articulatory model. In the feedforward control process, an input articulatory target, specified by articulatory parameters, is transformed to intrinsic representation of articulation; then, a muscle activation pattern by a proposed mapping function. The results show that the proposed feedforward control strategy is able to control the proposed 3D physiological articulatory model with high accuracy both acoustically and articulatorily.

Index Terms: articulatory model, feedforward control, vowel production

10.21437/Interspeech.2009-11

1. Introduction

Physiological articulatory models are expected to help speech scientists understand the biomechanical characteristics of articulators, and further, understand coarticulation in speech planning, which are difficult to infer from observed articulation directly [1].

In previous study, we developed the partial-3D model [4] into a full 3D model [2]. To shed light on coarticulation at the planning level by using the proposed 3D model, it is necessary to elaborate an efficient control strategy that could generate proper muscle activations according to input articulatory targets. In this paper, we address the preliminary effort to elaborate a feedforward control module to drive the proposed 3D physiological articulatory model for vowel production.

Dang *et al.* proposed the muscle workspace [3] and Ep-map [4] methods for model control. The former one dynamically generates muscle activations by reducing the distance between current position and the target position of control points, while the latter one generates muscle activation according to the target position of the control points of the partial 3D model. Nonetheless, those two control methods are essentially multi-point control strategy, with which the control points of the model are independently controlled. This may cause contradiction for the control of each control point while achieving some articulatory targets. In addition, those control methods only concern the manipulation of the articulators in the sagittal plane alone without accounting for the active control of transversal deformation of the tongue. In this paper, we constructed a mapping from articulatory posture to muscle activation to realize feedforward control of the 3D behavior of the articulatory model for vowel production.

For this purpose, at first, we conduct model simulation to provide paired articulatory posture-muscle activation data; then, the articulatory parameters are extracted by Linear Component Analysis (LCA) to specify articulatory targets (articulatory postures). Finally, the mapping from articulatory targets to muscle activations is elaborated by using General Regression Neural Network (GRNN) via intrinsic representation of vowel articulation.

2. Model simulation

Since it is difficult to collect empirical data that provide both muscle activations and articulatory postures (AP) for elaborating the functional relationship between AP and muscle activation, we conduct model simulation to obtain the posture-muscle activation data pairs.

The simulations are conducted by accounting for the functions of tongue muscles [5]. According to Fang *et al.* [6], in simulation, each muscle group consists of 6 tongue muscles. Among those muscles, transversus together with verticalis manipulate the width and length of tongue. As for the other 4 muscles, two of them are antagonist pair for tongue tip/dorsum, and the other two muscles are chosen to enlarge the region that the tongue covers by activating the antagonist pair. Eight-level muscle forces (0.0, 0.1, 0.2, 0.4, 1.0, 2.0, 4.0, and 6.0) are assigned to each tongue muscle in each muscle combination; three-level forces (0.5, 1.5, 3.0) are assigned to jawCl; and six-level forces (0.0, 0.5, 1.2, 2.4, 4.0, 6.0) are assigned to jawOp.

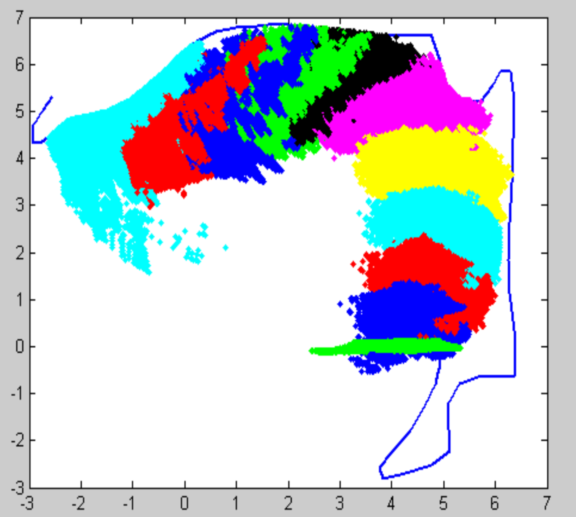


Figure 1: the distribution of the 11 nodes along the tongue surface in the midsagittal plane of part of the simulation results.

Figure 1 gives the distribution of the 11 nodes along the tongue surface in the midsagittal plane of the simulation results. The area with different colors corresponds to different tongue nodes. It shows that the simulation is able to cover a large variety of vocal tract shapes. Moreover, the tongue can form constrictions and full closure with the anterior and posterior part of the palate, which are important for apical and dorsal consonants.

For acoustic evaluation, the corresponding acoustic response (F1, F2) of the obtained vocal tract in the simulation are calculated. Figure 2 depicts the dispersion of the first and second resonance peaks (F1 and F2) of the vocal-tracts obtained from simulations. The ellipses delimit the regions of the acoustic targets of 5 Japanese vowels within the limen of 5% for F1 and 10% for F2, respectively. It shows that the simulations can cover the sustained vowels of the prototype subject.

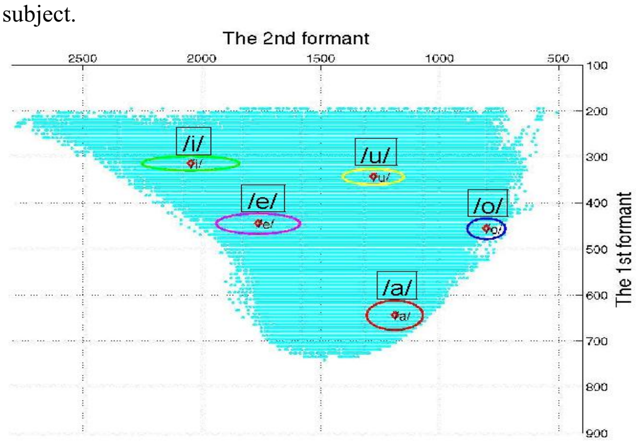


Figure 2: The distribution of the acoustic responds in the F1-F2 plane.

As mentioned by Atal *et al.* [7], many different articulatory configurations can generate acoustic signals with similar spectral characteristics. To exclude the simulations corresponding to unexpected vocal tract shapes, the vocal-tract configurations, whose acoustic response (F1 and F2) locate in the ellipsis of Figure 2, are constrained by observed articulatory data of the prototype subject. Finally, 364, 312, 573, 437, and 423 samples are obtained for vowel /a/, /i/, /u/, /e/ and /o/, respectively.

3. Extraction of articulatory parameters

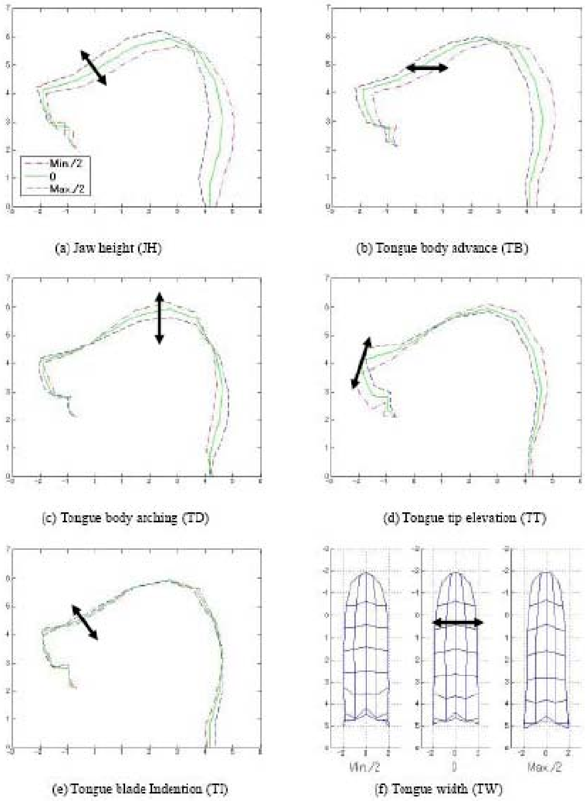
The original AP vectors are represented by the coordinates of the nodes on tongue surface. The dimensions of original AP are difficult to be interpreted from articulatory perspective directly. This brings difficulties for the control of the physiological articulatory model. Hence, we adopt LCA to extract the linear components in vowel production. The advantage of using LCA is that each extracted parameter has a well-defined articulatory interpretation.

As shown in Eq. (1), **s** is the AP vector, $\mathbf{s}_m$ is the average AP vector of the data for analysis, ***M*** is the transformation matrix, each column of ***M*** corresponds to one basic AP vector, and a is a vector of loadings factors.

$$\boldsymbol{s} = \boldsymbol{s}_m + \boldsymbol{Ma} \qquad (1)$$

In contrast to principle component analysis, in LCA, the correlation between the columns of ***M*** is allowed.

The loading factors of the basic AP vectors are determined in light of the following procedure: i) the data that describe the jaw movement are feed to PCA. Then the corresponding loading factor is subject to linear regression to extract the first basic shape vector; ii) the residue are obtained by subtracting the influence of jaw, and are feed to PCA to extract the other component of tongue movements; iii) the rest column vectors of ***M*** are determined by linear regression on the extracted factor loading.



(a) Jaw height (JH)

(b) Tongue body advance (TB)

(c) Tongue body arching (TD)

(d) Tongue tip elevation (TT)

(e) Tongue blade Indention (TI)

(f) Tongue width (TW)

Figure 3: The articulatory parameters extracted by linear component analysis: a). JH; b). TA; c). TD; d). TT; e) TI; f) TW.

Through above procedure, six linear components are extracted. They are Jaw height (JH), Tongue body advancing (TB), Tongue body arching (TD), Tongue tip elevation (TT), Tongue width (TW), and Tongue blade indention (TI). The impact of each component on the tongue shape is demonstrated in Figure 3. It shows that: a). With the increasing of JH, the jaw moves upward, and cause some tongue deformation at the tongue tip and rear part of the tongue body; b). With the increasing of TB, the tongue moves forward and upward; c). With the decreasing of TD, the tongue body moves upward and backward, while the tongue tip moves backward and downward; d). With the decreasing of TT, the tongue tip moves upward; e). With the increasing of TW, the tongue body is narrowed; and f). With the increasing of TI, the indention pattern at the tongue blade is more and more clear.

Table 1. Components extracted by LCA procedure and their contributions.

	Explanation	Var.
JH	Jaw Height	0.36
TB	Tongue body advance	0.42
TD	Tongue body arching	0.12
TT	Tongue tip elevation	0.04
TW	Tongue width	0.02
TI	Tongue blade Indention	0.01

The extracted six linear components explain 97% variance of the data of vowel production. The details are shown in Table 1.

The components (JH, TB, TD, TT) extracted in this study resemble the components extracted from X-ray [8] and MRI measurement [9] using similar methods, although the source of data and the AP vectors for analysis show some difference. A novel component TW proposed here can be used to control the deformation of tongue in the transversal dimension.

When comparing the original articulatory postures and the postures reconstructed from the extracted parameters, it is found that the average difference between them is 0.07cm, and the standard deviation is 0.03cm. This indicates that the tongue shape can be represented by these six linear components with high accuracy. Accordingly, these six components are used as articulatory parameters hereafter.

4. Feedforward control for vowel production

In vowel production, the tongue contacts with different parts of the vocal tract wall according to the position of the required constrictions, which makes the boundary constraints different from vowel to vowel. For this reason, it is necessary to cluster the articulatory postures and construct function from articulatory posture to muscle activation for each cluster. It requires finding an appropriate transformation that keeps the similarity relationship among the original articulatory postures.

Our experiments show that Euclidian distance is an appropriate measure of the similarity for the original APs. The similarity of the articulatory postures can be derived in terms of the extracted articulatory parameter from Eq. (1).

$$d_{ij} = \| \boldsymbol{s}_i - \boldsymbol{s}_j \|^2 = (\boldsymbol{a}_i - \boldsymbol{a}_j)^T \boldsymbol{M}^T \boldsymbol{M} (\boldsymbol{a}_i - \boldsymbol{a}_j) \quad (2)$$

where $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are original articulatory posture vectors, $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ are corresponding articulatory parameter vectors.

Because the columns in $\boldsymbol{M}$ are not orthogonal to each other, the matrix $\boldsymbol{M}^T\boldsymbol{M}$ is not diagonal. This makes it difficult to define a simple similarity measure that keeps the similarity property of the original APs in terms of the articulatory parameters directly.

For this reason, we first look for a low dimensional space (intrinsic representation space, see Section 4.1) in which the similarity of the original APs can be kept by using the Euclidean distance between the representations of the APs in the low dimensional space. Then, the articulatory parameters are transformed into the low dimension representation, and then projected from the low dimensional representation to muscle activation. The problem arrived at establish a mapping from articulatory parameters to muscle activation, which is the essential part of feedforward control for the 3D model. In the rest part of this section, we will introduce them in detail.

4.1. Intrinsic representation

Multidimensional scaling (MDS) is a technique to transform high dimensional data into a low dimensional space. In such a space, the similarity between samples in the original space is kept. Here, we use the method proposed by Webb, in which every sample in the shape space can be transformed into the lower dimensional space with a combination of a set of radial basis functions [10] by using Eq. (3).

$$f(\boldsymbol{s}) = \boldsymbol{W}\boldsymbol{\phi}(s) \quad (3)$$

where $\boldsymbol{W}$ is the weighting coefficient matrix, $\boldsymbol{\phi}(\boldsymbol{s})$ is vector with *jth* element $\phi_j(\boldsymbol{s})$. The coefficients in $\boldsymbol{W}$ can be obtained by minimizing the cost function defined in Eq. (4).

$$\boldsymbol{W} = \arg\min_{W} \sum_{i=1}^{N} \sum_{i=1}^{N} \left(q_{ij}(\boldsymbol{W}) - d_{ij} \right)^2 \quad (4)$$

where $\boldsymbol{q}_{\mathrm{ij}}$ is the Euclidian distance between the *ith* and *jth* samples in the transformed low dimensional space, $\boldsymbol{d}_{\mathrm{ij}}$ is the similarity measure defined in Eq. (2).

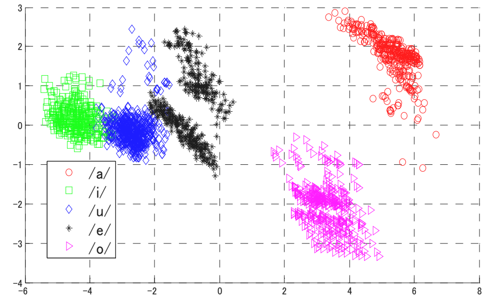


Figure 4: The dispersion of the vowels in low dimensional space obtained by MDS analysis.

Figure 4 shows the distribution of the five Japanese vowels (/a/, /i/, /u/, /e/ and /o/) in the MDS space. It shows that the APs for different vowels congregate into different clouds in the low dimensional space, and the clouds are well separated. This result is similar to the intrinsic structure of vowel production obtained by using manifold learning method [11]. Hence, we refer to the low dimensional representation as intrinsic representation of vowels.

4.2. Mapping from articulatory posture to muscle activation

The mapping from articulatory parameters to muscle activation consists of two steps: a). mapping articulatory postures specified by articulatory parameter to intrinsic representation; b). mapping intrinsic representation to muscle activations. They will be address in the following part.

4.2.1. From articulatory parameter to intrinsic representation

According to Eq. (1) and Eq. (3), articulatory parameters can be transformed into corresponding intrinsic representation by using Eq. (5) directly.

$$\boldsymbol{f}(\boldsymbol{a}) = \boldsymbol{W}\boldsymbol{\phi}(\boldsymbol{M}\boldsymbol{a} + \boldsymbol{s}_m) \quad (5)$$

where matrix $\boldsymbol{M}$, $\boldsymbol{s}_{\mathrm{m}}$, $\boldsymbol{W}$, and $\boldsymbol{\phi}(\boldsymbol{s})$ are already obtained in previous sections.

4.2.2. From intrinsic representation to muscle activation

To account for the different boundary constraints resulted from contact between tongue and surrounding structures, the k-means method is applied to clustering the APs based on Euclidian distance. The 'Elbow' criterion is used to determine the number of clusters in the data.

Then, GRNN is applied to elaborate the mapping from intrinsic representation to muscle activation for each cluster, of which 80% and the left 20%are chosen as the training set and testing set, respectively.

The average differences between output of GRNN and target muscle activation are shown in Figure 5, where the abscissa represents the cluster indices, and the ordinate denotes the average difference between the target muscle activation and muscle activation estimated by the trained

GRNN for the test set of each cluster. The horizontal line indicates the minimal nonzero muscle activations for each muscle in simulation. It shows that, for most of the clusters, the difference between the target and estimation is less than of the minimal nonzero muscle activation for each muscle in our simulation.

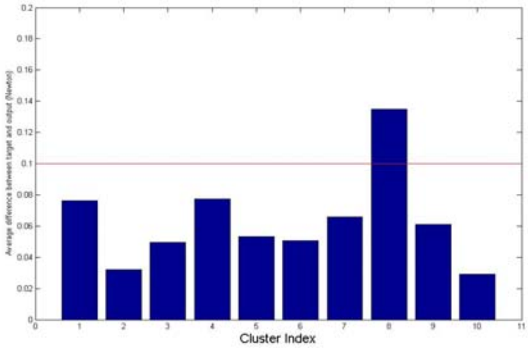


Figure 5: The average difference between the target muscle activation and that estimated by the GRNN

We also run the whole control process from input articulatory target to generate the muscle activations, and then drive the 3D physiological articulatory model. The intermediate output of the muscle activations and final output of acoustic responses are used to evaluate the differences between the target articulatory postures and the ones obtained using feedforward control process, and the differences between their corresponding acoustic consequences. The results are shown in Table 2. It indicates that, in most cases, the articulatory posture and corresponding acoustic consequences obtained by activating the proposed physiological articulatory model are close to the required articulatory and acoustic target.

Table 2. The average and standard deviation of difference between the postures of the target articulation and the ones obtained by proposed method, and average and standard deviation of between the corresponding acoustic consequences (F1 and F2 in Hz)

	Shape Dev.	Shape Std.	Acoustic Dev. (F1/F2)	Acoustic Std. (F1/F2)
Cluster 1	0.03	0.06	6.9/14.6	6.2/14.5
Cluster 2	0.13	0.09	9.5/14.5	5.1/12.1
Cluster 3	0.07	0.07	20.8/24.3	6.4/21.1
Cluster 4	0.07	0.06	5.2/26.2	7.4/23.1
Cluster 5	0.06	0.07	24.4/10.4	9.6/7.9
Cluster 6	0.08	0.08	68.0/65.0	28.7/25.4
Cluster 7	0.04	0.05	9.8/55.0	9.5/49.8
Cluster 8	0.02	0.05	9.8/20.5	7.9/20.8
Cluster 9	0.03	0.06	9.1/20.7	33.2/15.7
Cluster 10	0.03	0.06	18.7/9.5	5.0/9.5

5. Summary

In this study, we proposed a feedforward control strategy for the 3D model to obtain muscle activations from the given articulatory targets. The results show that this method can control the proposed 3D physiological articulatory model with high accuracy.

In the current stage, we focus on developing a method that can estimate appropriate muscle activation according to the input articulatory targets. In the future, we will develop a feedforward mapping for vowel production by fully considering the human mechanism. A feed forward control will be introduced for consonant production and the coarticulation process in the planning level of speech production to shed light on the detail mechanism of speech production.

6. Acknowledgements

This research is conducted as a program for the "21st Century COE Program" for promoting Science and Technology by Ministry of Education, Culture, Sports, Science and Technology, and also supported partly by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC), Japan

7. References

1. Perrier, P., *et al. Modeling the production of VCV sequences via the inversion if a biomechanical model of the tongue*. in *INTERSPEECH 2005*. 2005. Lisbon, Portugal.
2. Fang, Q., *et al. A model based investigation of activation patterns of the tongue muscles for vowel production*. in *InterSpeech2008*. 2008. Brisbane, Australia.
3.Dang, J. and Honda, K., *Estimation of vocal tract shape from sounds via a physiological articulatory model.* Journal of Phonetics, 2002. **30**: p. 511-532
4. Dang, J., Honda, K., *Construction and control of a physiological articulatory model.* J. Acoust. Soc. Am., 2004. **115**(2): p. 853-870.
5. Fang, Q., *et al.*, *Investigation of functions of tongue muscles for model control.* Chinese Journal of Phonetics (in press), 2008.
6. Fang, Q., *et al.*, *A model-based investigation on activation of tongue muscles in vowel production.* Acoustical Science & Technology, 2008. (to appear)
7.Atal, B. S., *et al.*, *Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique.* J. Acoust. Soc. Am., 1978. **63**(5): p. 1535-1555.
8.Maeda, S., *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model.* Speech production and modeling. 1990: Kluwer Academic Publishers.
9.Badin, P., *et al.*, *Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images.* Journal of Phonetics, 2002. **30**(3): p. 533-553.
10.Webb, A. R., *Multidimensional scaling by iterative majorization using radial basis functions.* Pattern Recognition, 1995. **28**(5): p. 753-759.
11.Dang, J. and Lu, X. A perspective on the relation between speech production and perception based on a vowel study. in The 8th Phonetic Conference of China (PCC2008) and the International Symposium on Phonetic Frontiers (ISPF2008). 2008. Beijing.