SELECTED PAPER

Investigation of Auditory-Guided Speech Production while Learning Unfamiliar Speech Sounds

Kazuya Fujii[1], Qiang Fang[2] and Jianwu Dang[1,3]

[1]Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa
923-1292, Japan
Phone:+81-761-51-1235
E-mail:{s1020011,jdang}@jaist.ac,jp

[2]Institute of Linguistics, Chinese Academy of Social Sciences
5# Jian Guo Men Nei Da Jie, Beijing
100732, China
Phone:+86-10-65237408
E-mail: fq0237@gmail.com

[3]Tianjin University
92 Weijin Road, Nankai District,
Tianjin 300072, China
Phone:+86-22-27406149
E-mail: dangjianwu@tju.edu,cn

Abstract

During the learning process, learners perceive the difference between a target sound and their produced sound, and manipulate speech organs to reduce the difference. Based on this fact, we proposed a framework called Auditory-Guided Speech Production (AGSP) to approximately represent the learning process for a new vowel category. Within the framework, there are two basic functions: one determines articulation according to the target sounds (function G), and the other determines articulatory increments with reference to the acoustic differences (function H). These two functions have a certain relation in acquiring new articulations, where function G is updated with reference to the output of function H. To investigate these two functions, we design an experiment to monitor the learning process during imitation, by recording the acoustic and articulatory data using electro-magnetic articulography, and we discuss the formulation of the AGSP.

1. Introduction

There are many unknown factors involved in the natural acquisition process of a new language for human beings. For instance, how do speech production and perception interact with each other during the imitation process where learning unfamiliar sounds. Previous studies on the relationship between speech production and perception focused on the effects of acoustic or articulatory perturbations [1-3], and studies revealed some of the aspects of interaction between speech production and perception. Guenther et al. modeled the relation between speech production and perception, and simulated brain activities during speech processing [4]. Kröger et al. have simulated the McGurk effect using the neurocomputational model [5]. However, there is little knowledge about the function of Auditory-Guided Speech Production (AGSP) on the learning process in the brain.

During acquisition of an unfamiliar sound, learners perceive the difference between a target speech sound and their produced sound, and then manipulate speech organs to imitate the sound by reducing the acoustic difference. Auditory-Guided articulation is the main procedure for this kind of learning process.

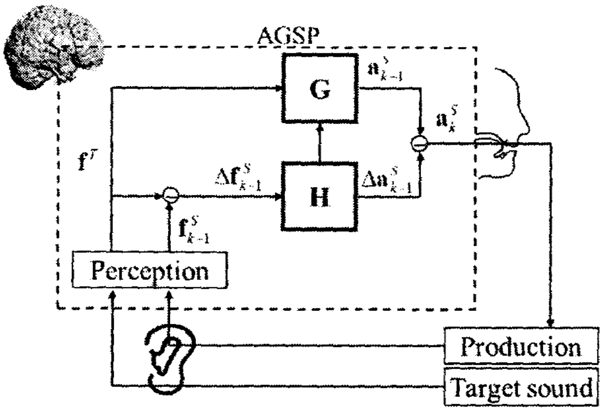

Figure 1: Schematic block diagram of AGSP function on the learning process during imitation

In previous studies [6,7], we focused on the learning process during imitation, and monitored it by means of both acoustic and articulatory data via the electro-magnetic articulography (EMA) system. Based on the assumption that the articulatory movement during imitation is modified through reference to perceived acoustic differences, we used Jacobian matrix to approximate the mapping from acoustic feature to articulatory configuration. However, we did not find a reliable mapping. The possible reasons for this problem are that the experimental data were not reliable, and/or the method used was not able to describe nonlinear relations. In this study, therefore, we use the neural network to evaluate the experimental data, and investigate the relations between acoustic difference and articulatory increment. A possible formulation of the AGSP to describe the learning process in the imitation procedure is also discussed.

2. AGSP Function in Imitative Learning

To clarify unknown factors involved in the process of imitative learning, we propose a functional model and show it in Fig. 1. During the learning process, learners perceive the difference, $\Delta\mathbf{f}_{k-1}^{S}$, between their produced sound, $\mathbf{f}_{k-1}^{S}$, and the target sound, $\mathbf{f}^{T}$, then manipulate the configuration of speech organs with an articulatory modification $\Delta\mathbf{a}_{k}^{S}$ to reduce the acoustic difference $\Delta\mathbf{f}_{k-1}^{S}$ in each utterance time $k(k=1,\cdots,K)$.

Within the model, there are two basic functions: one determines articulation of each utterance according to the target sound (function G), and the other determines modification of articulation with reference to acoustic differences (function H).

In human imitation, after perceiving a speech sound, listeners use the function G, which is a phoneme-dependent mapping, to estimate the corresponding or nearest articulation and to produce the sound. Accordingly, when learning a new sound, the listeners estimate an approximate articulation based on their existing mapping and use it to produce a closer sound. The difference between the target sound and imitated sound is used to modify the articulation via function H. At the same time, function G is updated based on the difference via function H. This learning process based on the AGSP framework can be described using a formula of the Kalman filter as follows:

$$\begin{cases}\mathbf{a}_{k}^{S}=\mathrm{G}_{k-1}\left(\mathbf{f}^{T}\right)+\mathrm{H}\left(\Delta\mathbf{f}_{k-1}^{S}\right)\\ \mathbf{G}_{k}=\mathbf{G}_{k-1}+\lambda^{k}\cdot\mathbf{H}\\ \mathbf{f}_{k}^{S}=\mathrm{g}\left(\mathbf{a}_{k}^{S}\right)\end{cases}\qquad(1)$$

Let us consider learning an unfamiliar speech sound in this learning process. At the beginning ($k=1$), the initial articulation $\mathbf{a}_{1}^{S}$ is estimated from the perceived target sound $\mathbf{f}^{T}$ using $\mathbf{G}_{0}$, which would be one of the close mapping functions in the original mapping function pool, where there is no acoustic difference for reference. Then, the learner produces a closer sound according to $\mathbf{a}_{1}^{S}$ and represents using $\mathbf{f}_{1}^{S}$. At the next step ($k=2$), the difference $\Delta\mathbf{f}_{1}^{S}$ between $\mathbf{f}_{1}^{S}$ and $\mathbf{f}^{T}$ is used to modify the articulation by function H, meanwhile $\mathbf{G}_{0}$ is updated to $\mathbf{G}_{1}$. λ is a weighting coefficient between 0 and 1 which is used to renew the function G. The learner produces a closer sound using $\mathbf{a}_{2}^{S}$. The procedure repeats according to formula (1). When the difference between target sound and imitated sound becomes small enough, the learning process is terminated and function G becomes the learned mapping function for auditory-guided speech production.

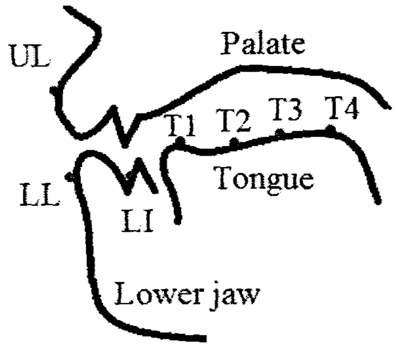

Figure 2: Schematic illustration of midsagittal seven-sensor placement constellation (UL: upper lip, LL: lower lip, LI: lower incisor, Tn: tongue No. n)

3. Experiment on the Learning Process during Imitation

In order to investigate the AGSP function, we designed an experiment on the imitation process for unfamiliar vowels. Three male Japanese speakers (Subjects-A, B, and C) participated in the experiment. English vowels /æ/, /ʊ/, /ɪ/, and /ɚ/ served as the target vowels. The acoustic data and the articulatory data of the target English vowels were selected from the X-ray microbeam database, and used in the experiments [8].

Seven sensors in the EMA experiment were set on the tongue surface along the midsagittal plane (T1–T4), upper lip (UL), lower lip (LL), and lower incisor (LI) as shown in Fig. 2, to record articulatory movement. Three more sensors were set on the upper incisor, canine tooth, and under the right ear for reference. In the experiment, the subjects were asked to imitate given vowels. The learning process was evaluated by calculating the Euclidean distance between the produced sound and the target sound, where the first and second formants were used as the acoustic features. The experiment was terminated when the distance was reduced, or no improvement is observed.

4. Investigation of the Relation between Acoustic Difference and Articulatory Modification

In our previous study [6,7], we approximated the mapping from acoustic features to articulatory movements in AGSP with linear function by using Jacobian matrix. However, the results did not show clear relations between those two observations. In the present study, therefore, we investigated the relations between acoustic difference and articulatory increments using a neural network, and evaluated the reliability of the experimental data.

4.1 Construction of a neural network

We used a three-layer neural network to investigate the relations between acoustic difference and articulatory modification, which is shown in Fig. 3. This network consists of the input layer, hidden layer and output layer. The variables of these layers are represented by $\boldsymbol{i}=(i_1,\cdots,i_m)$, $\boldsymbol{h}=(h_1,\cdots,h_n)$ and $o=(o_1,\cdots,o_j)$, respectively. Each neuron in the hidden layer and output layer acts according to

$$h_n = \mathrm{f}\left(\sum_{m=1} \mathbf{V}_{nm} \cdot i_m + a_n\right) \qquad (2)$$

$$o_j = \mathrm{f}\left(\sum_{n=1} \mathbf{W}_{jn} \cdot h_n + b_j\right) \qquad (3)$$

where $\mathbf{V}_{nm}$ shows the weight matrices from the m-th neuron in the input layer and to n-th neuron in the hidden layer, and $\mathbf{W}_{jn}$ also is the weight matrices between hidden layer and output layer. $\mathbf{a}=(a_1,\cdots,a_n)$ and $\mathbf{b}=(b_1,\cdots,b_j)$ are bias in the hidden layer and output layer, respectively. Additionally, $\mathrm{f}(x)$ is a log-sigmoid function given by

$$\mathrm{f}(x)=\frac{1}{1+e^{-x}} \qquad (4)$$

Learning of the weights in each layer was performed using Levenberg-Marquardt algorithm.

4.2 Evaluation

In the learning, we used the observation of acoustic data as input and the observation of articulation as output to investigate the relationship between acoustic differences and articulatory increments. To do so, the first three formant frequencies ($\mathbf{F}^S=(F1,F2,F3)$) were first extracted using LPC analysis with a frame length of 25 ms and a frame shift of 10 ms, and we calculated the average of the formant frequencies over six frames in stable speech segments.

Next, the articulatory data (T1–T4, UL, LL, and LI) were extracted during the same period as that used in the acoustic analysis, and then an average was taken on time. Here, these articulatory data comprise two-dimensional midsagittal information; the horizontal and vertical positions of articulators, as seen in Fig. 2. We applied principal component analysis (PCA) to the articulatory data to reduce the dimensions. The first three components ($\mathbf{A}^S=(A^{PC1},A^{PC2},A^{PC3})$) are used to describe the articulatory features. The first three components for Subjects-A, B, and C can explain 93%, 86%, and 90% each of the variance, respectively.

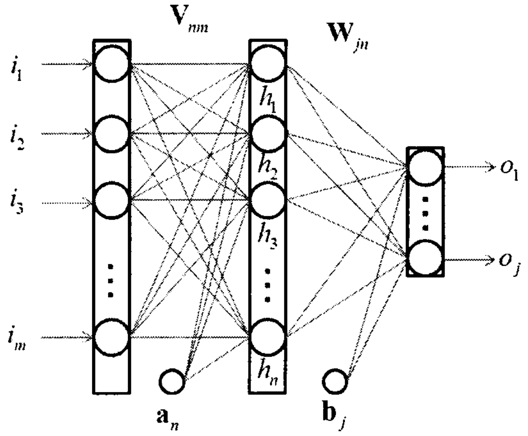

Figure 3: Three-layer feedforward neural network for mapping from acoustic parameters to parameters of articulatory modification

Subsequently, we obtained the input and output data for the AGSP framework. The input data were given by

$$\boldsymbol{i}=\Delta\mathbf{F}_{k-1}=(\Delta F1_{k-1},\Delta F2_{k-1},\Delta F3_{k-1}) \qquad (5)$$
$$\text{where} \quad \Delta\mathbf{F}_{k-1}=\mathbf{F}^S_{k-1}-\mathbf{F}^T$$
$$\mathbf{F}^T:\text{Target formant}$$

The output data were defined as

$$o=\Delta\mathbf{A}_{k-1}=(\Delta A^{PC1}_{k-1},\Delta A^{PC2}_{k-1},\Delta A^{PC3}_{k-1}) \qquad (6)$$
$$\text{where} \quad \Delta\mathbf{A}_{k-1}=\mathbf{A}^S_k-\mathbf{A}^S_{k-1}$$

Here, these input and output data were normalized in the range between 0 and 1 for learning. 95% of the data was randomly selected as the training data set, and the remaining 5% of the data was used for testing.

We conducted computer learning on the three-layer neural network, with 3 neurons in the input layer, 6 neurons in the hidden layer, and 3 neurons in the output layer. The data for each vowel were divided into training set and test set in 10 different ways, so that we have 10 sets of data. In order to avoid the local minimum problem, we used 10 different initial weight matrices for each training set, and chose the best one. For each set, the training was iterated 10,000 times, and then each remaining test data was used for testing. The correlation coefficient between the observed articulatory data and

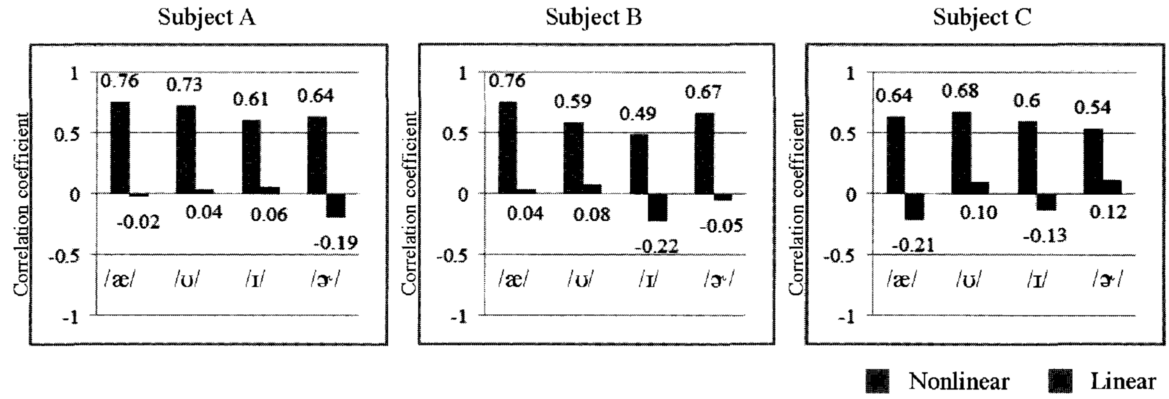

Figure 4: Comparison of the average correlation coefficients between observed and predicted values on articulatory modifications using a neural network (Nonlinear) and Jacobian matrix (Linear) for mapping

predicted data was averaged over 10 data sets for each vowel and for each subject. The correlation coefficients are shown in Fig. 4, with a comparison between the results from nonlinear and linear mappings. The results show that after applying nonlinear learning approach, the articulatory increments and acoustical differences show higher correlation from 0.5 to 0.76. This implies that the experimental data are reliable to some extent, and available for the modeling.

5. Discussion and Conclusions

In this study, we proposed a simple functional model of Auditory-Guided Speech Production (AGSP) involved in the speech learning process. Based on the behavior of human imitation, we attempted to use formulae to describe the process for learning unfamiliar speech sounds and constructing a new acoustic category during imitation.

We conducted an experiment to monitor the learning process through imitation, by means of acoustic and articulatory data obtained from the EMA. In this study, the relations between articulations and acoustics were investigated using a neural network. The result showed that they had higher correlation, while it could not appear the relations in a linear analysis. The experimental data in the present study suggested that there are nonlinear relations between acoustic differences and articulatory modifications.

Although we proposed a model for the learning process, the model has not been refined by the observed data yet. To formulate the AGSP function, we need to develop a method that can be used with the nonlinear system such as the extended Kalman filter. These tasks remain for future studies.

Acknowledgements

We would like to thank Dr. Atsuo Suemitsu for his continuing assistance. This study was supported by a Grant-in-Aid for Scientific Research (C) (22500150) from the Japan Society for the Promotion of Science (JSPS), and is also supported partly by the National Thousand Talents Program of China.

References

[1] J. Dang, K. Akagi and K. Honda: Communication between speech production and perception within the brain-observation and simulation, J. Computer Science and Technology, Vol.21, No.1, pp.95-105, 2006.

[2] T. Uchiyamada, X. Lu, J. Dang and M. Akagi: Investigation of compensation of speech production for the transformed auditory feedback based on articulatory measurement (in Japanese), ASJ Spring Meeting, pp. 455–456, 2007.

[3] J.A. Jones and K.G. Munhall: Learning to produce speech with an altered vocal tract: The role of auditory feedback, J. Acoust. Soc. Am., Vol.113, No.1, pp.532–543, 2003.

[4] F.H. Guenther, S.S. Ghosh and J.A. Tourville: Neural modeling and imaging of the cortical interactions underlying syllable production, Brain & Language, Vol.96, pp.280–301, 2006.

[5] B.J. Kröger, J. Kannampuzha and C. Neuschaefer-Rube: Towards a neurocomputational model of speech production and perception, Speech Communication, Vol. 51, pp.793–809, 2009.

[6] K. Fujii, J. Wei, A. Suemitsu and J. Dang: The relationship between speech production and perception in the process of learning vowels, ICCS2010, pp.291–292, 2010.

[7] K. Fujii, A. Suemitsu and J. Dang: Investigation of relationship between speech perception and articulatory movement during learning process of vowels (in Japanese), ASJ Auditory Research Meeting, Vol.40, No.1, pp.665–670, 2010.

[8] J.R. Westbury, G. Turner and J. Dembowski: X-RAY MICROBEAM SPEECH PRODUCTION DATABASE USER'S HANDBOOK, Version 1.0., University of Wisconsin, Madison, 1994.

[This paper was published in Journal of Signal Processing, Vol. 15, No. 4.pp287-290, july 2011]