# IMPROVEMENT OF ANIMATED ARTICULATORY GESTURE EXTRACTED FROM SPEECH FOR PRONUNCIATION TRAINING

*Yurie Iribe[1], Silasak Manosavan[1], Kouichi Katsurada[1], Ryoko Hayashi[2], Chunyue Zhu[3] and Tsuneo Nitta[1]*

[1]Graduate School of Engineering, Toyohashi University of Technology, Japan
[2]Graduate School of Intercultural Studies, Kobe University, Japan
[3]School of Language and Communication, Kobe University, Japan

## ABSTRACT

Computer-assisted pronunciation training (CAPT) was introduced for language education in recent years. CAPT scores the learner's pronunciation quality and points out wrong phonemes by using speech recognition technology. However, although the learner can thus realize that his/her speech is different from the teacher's, the learner still cannot control the articulation organs to pronounce correctly. The learner cannot understand how to correct the wrong articulatory gestures precisely. We indicate these differences by visualizing a learner's wrong pronunciation movements and the correct pronunciation movements with CG animation. We propose a system for generating animated pronunciation by estimating a learner's pronunciation movements from his/her speech automatically. The proposed system maps speech to coordinate values that are needed to generate the animations by using multilayer perceptron neural networks (MLP). We use MRI data to generate smooth animated pronunciations. Additionally, we verify whether the vocal tract area and articulatory features are suitable as characteristics of pronunciation movement through experimental evaluation.

***Index Terms***— Vocal Tract Area, Articulatory Gesture, Pronunciation Animation, Pronunciation Training

## 1. INTRODUCTION

Computer-assisted pronunciation training (CAPT) was introduced for language education in recent years [1][2]. CAPT typically scores pronunciation quality and points out a learner's wrong phonemes by using speech recognition technology [3][4]. Moreover, it often indicates the differences between incorrect and correct pronunciation by showing the learner's speech wave and the correct speech wave. Although the learner can thus realize that his/her speech is different from the teacher's, the learner cannot understand how to make the correct pronunciation movement. In particular, as for the speech wave, only a phonetician can realize the reasons for the differences. The system should show how the wrong articulatory organs move and how to correct this movement when the learner makes a wrong pronunciation. Although other studies have examined making correct pronunciation animations and video in advance [5][6], the studies do not automatically produce animations of the learner's wrong pronunciation from speech. We indicate these differences by visualizing the learner's wrong pronunciation movements (movement of the tongue, palate, and lips) and the correct pronunciation movements by using CG animation (Figure 1). As a result, the learner can study how to move the articulatory organs while visually comparing their mispronunciation animation with the correct pronunciation animation. We confirmed the educational effectiveness of the animations for pronunciation training in our previous research [7]. Therefore we propose generating animated pronunciations by automatically estimating the pronunciation movement from speech. Concretely, the proposed system maps speech to coordinate values that are needed to generate the animations by using multi-layer neural networks (MLN). It applied multilayer perceptron neural networks in MLN. We use MRI data to represent smooth human articulatory movements accurately. MRI data is applied as MLN training data. Additionally, we compare whether the vocal tract area and articulatory features are appropriate as MLN input through experimental evaluation. In this paper, the method for automatically generating animated pronunciations from speech is described. In section 2, we describe the method for the vocal tract area calculator, coordinate vector extraction, and CG animation generation. In section 3, an experimental evaluation to confirm the accuracy of the generated animated pronunciation is discussed. In the last section, the paper is summarized.

## 2. ANIMATED PRONUNCIATION GENERATION

### 2.1. System outline

Figure 2 shows an outline of the system. The system consists mainly of the vocal tract area extractor, coordinate vector extraction by MLN, and CG animation generation based on the coordinate vectors. Coordinate vectors are acquired by transforming the vocal tract area extracted from speech. Our previous research applied the articulatory features (place of

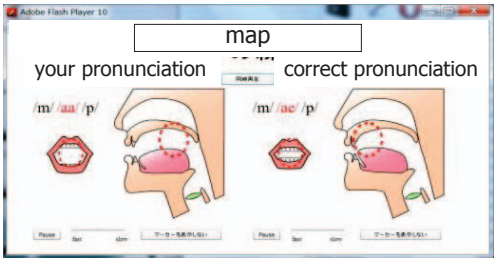


Figure 1: Animations of learner's and correct pronunciation

articulation and manner of articulation) to extract articulatory movement from speech [8]. Concretely, articulatory features were extracted from the speech input to the MLN. In this paper, to generate more accurate animation, we use the vocal tract area in place of articulatory features because we regarded that it may be suitable for area mapping with the coordinate values of the articulation organs. We verify this possibility with an evaluation.

The CG animation is generated on the basis of coordinate values extracted from a trained MLN. As a result, the user's speech is input in our system, and a CG animation that visualizes the pronunciation movement is automatically generated. Moreover, this paper describes the animation generation of English using more phonemes than in Japanese.

### 2.2. Vocal tract area extraction

The vocal tract area is determined from the following vocal tract area function with the following formula.

$$A_{m-1} / A_m = (1+k_m)/(1-k_m), (m = M, \ldots, 1) \quad (1)$$

$A_m$ : m dimension of the vocal tract area, $k_m$ : PARCOR coefficient

PARCOR coefficients are equivalent to the reflection coefficients in a lossless acoustic tube model of the vocal tract. A vocal tract area function expresses the vocal tract area from the glottis to the lips as a function of the distance from the glottis, and it is related to the distance between the palate and the tongue. The vocal tract area is acquired by calculating PARCOR parameters converted from speech signals. The vocal tract area (13 dimensions) is combined with the other two frames, which are three points prior to and following the current frame (VT(t, t-3), VT(t, t+3)) to form articulatory movement. MLN input is the vocal tract area (13 × 3 dimensions).

### 2.3. Coordinate vector extraction

We apply magnetic resonance imaging (MRI) images to obtain the coordinate values of the shape of an articulatory organ. MRI machines capture images within the body by using magnetic fields and electric waves. MRI data captured in two dimensions details the movements of the person's tongue, larynx, and palate while making an utterance. CG animations are generated on the basis of coordinate vectors. The MLN trains the vocal tract area extracted from the speech included in the MRI data as input and the coordinate vectors of the articulatory organs acquired from the MRI

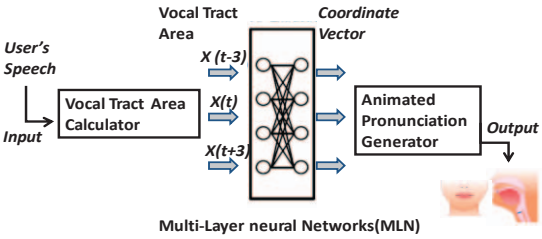


Figure 2: System outline

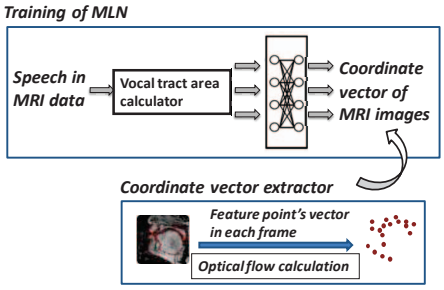


Figure 3: Coordinate vector extractor

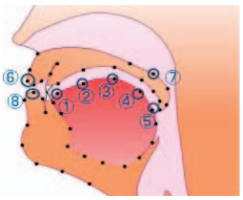

Figure 4: Feature points used in MLP training

images as output (Figure 3). As a result, after the user's speech is input, the coordinate vectors adjusted to the speech are extracted, and a CG animation is generated. In this section, the extraction of the feature points on the MRI data and the method for calculating the coordinate vectors of each feature point are described.

We assigned initial feature points to the articulatory organ's shapes (tongue, palate, lips, and lower jaw) on the MRI data beforehand. The number of initial feature points was 43 (black color points in Figure 4). We decreased the number of the dimensions of the MLN training data in order to train the MLN effectively by using even a small amount of MRI data. Therefore, we selected only eight feature points that vary infinitely and are important for the pronunciation teaching method (Figure 5). Many feature points should be assigned if a lot of MRI data can be used. These feature points were obtained in the following order.

1. We imported 10-ms speech and image segments to the MRI data
2. The coordinate value of each feature point was extracted by calculating the optical flow for each frame. The input data for the optical flow program is the coordinate vectors of the initial feature points.
3. Only the y-coordinate distance of each feature point was calculated to decrease the number of dimensions.

The x-coordinate value was the same as the x-coordinate of the initial feature point.

The number of the dimensions of the MLN was the vocal tract area (15 × 3 dimensions) used as input and y-coordinate vectors (8 × 3 dimensions) used as output.

### 2.5. CG animation generation programs

We correct the y-coordinate vectors by using a spline curve and a median filter to form the CG animations. We assigned 43 points (15 tongue points, 2 lip points, 16 palate points, and 10 lower jaw points) as the initial feature points of the MRI images. The position relation between the 8 feature points (trained by the MLN) and the remaining 35 feature points are calculated. The spline curve is used to complement between the eight feature points and other feature points by keeping the position relation. The movement is drawn on the basis of the y-coordinate distance, but since this movement is twitchy, we use a median filter to smooth it out.

The system is built as a web application so that various users can use it on the web, and the system can be incorporated in various web dictionaries. The CG animation program was implemented with Actionscript3.0 to operate in a web browser with a Flash Player plug-in installed. Figure 6 shows a screen shot of a CG animation developed in the present study.

## 3. EVALUATIONS

We calculated the correlation coefficient between the coordinate values of the generated CG animations and the MRI data to confirm the accuracy of the animations. Moreover, to show the effectiveness of using articulatory features to extract coordinate distances, we compared the correlation coefficients for the case of AF with the case of LF as MLN inputs.

### 3.1. Experimental data and setup

To evaluate the animation generated from speech, the correlation coefficient between the animations and MRI images is calculated. Moreover, the correlation coefficient of the articulation features and a vocal tract area as MLN input is also compared. The MRI data used in the evaluation was

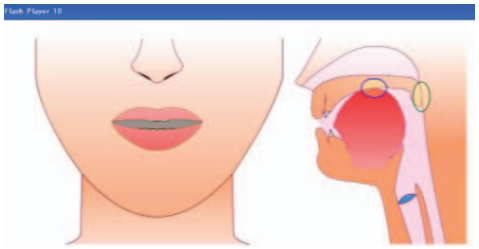

Figure 5: CG animation of "basket"

taken in a single shot, in which one female English native speaker uttered 37 English words. These were recorded at ATR-Brain Activity Imaging Center [9]. We filtered out noise from speech included MRI data. The data set used for the experimental evaluation is as follows.

***D1:*** *Training data set for AF-coordinate vector or VT-coordinate vector converter training: 36 words of English speech and images included in the MRI data (one female English native speaker)*

***D2:*** *Testing data set for AF- coordinate vector or VT-coordinate vector converter adaptation: One word of English speech included in the MRI data (one female English native speaker)*

Experimental results are acquired by using the leave-one-out cross-validation method. The 37 animations were generated in this experiment. Each MLN has three layers. The number of input layer is 75, hidden layer is 150, and output layer is 45 in the MLN to extract AF. The number of input layer is 45, hidden layer is 90, and output layer is 24 in the MLN to extract coordinate vector.

### 3.2. Experimental results

Figure 6 shows the correlation coefficient for each phoneme. As for the average correlation coefficient ("all" in Figure 6) of all phonemes, the vocal tract area was 0.83 and articulatory feature was 0.78. A comparatively high correlation coefficient was acquired in spite of the small amount of training data. On the whole, the correlation coefficient of the vocal tract area was higher than that of the articulatory feature.

It is clear that the translation to coordinate vectors has

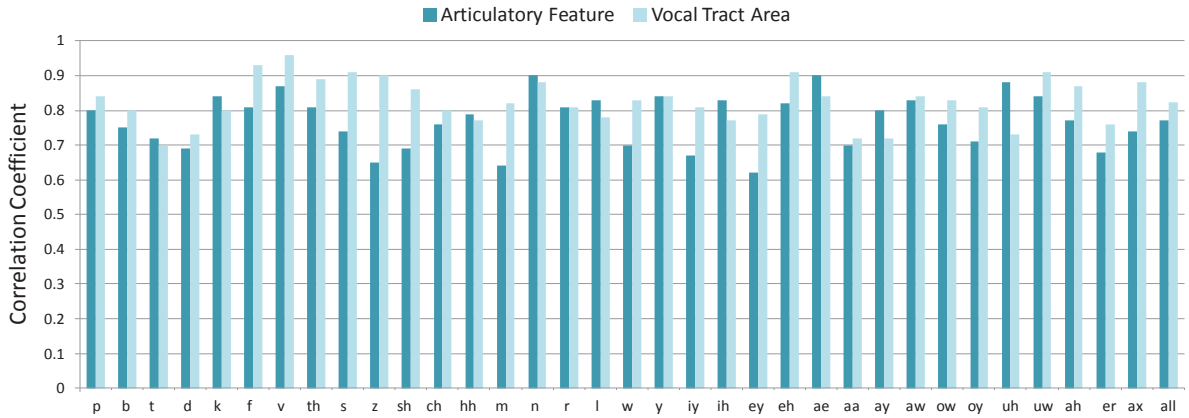


Figure 6: Correlation coefficient of each phoneme

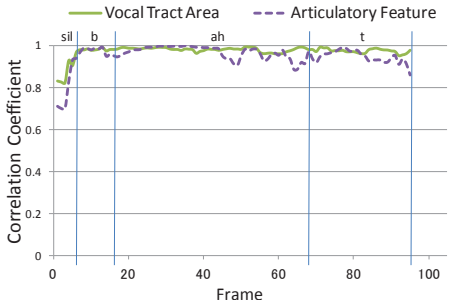

Figure 7: Correlation coefficient of word "but"

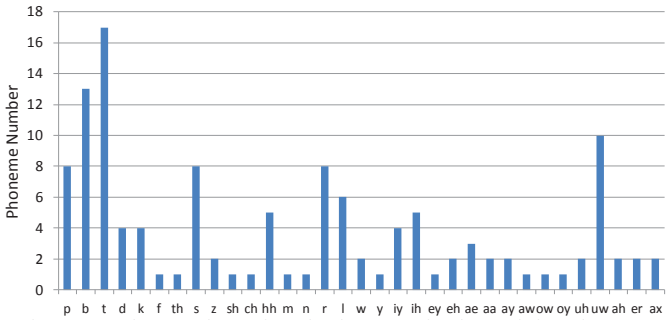

Figure 8: The number of each phoneme in the training data

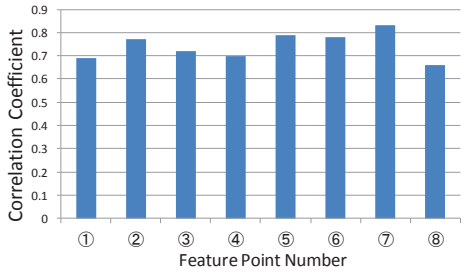

Figure 9: Correlation coefficient of each feature point.

higher adaptability with the vocal tract area than do the articulatory features. However, although we evaluated with a speaker dependent data set in this experiment, the vocal tract area typically changes in response to a speaker's age and sex. On the other hand, it is well expected that the correlation coefficient of the articulatory features would be higher than that of the vocal tract area in the case of a speaker-independent experiment because it would not be dependent on a speaker. We intend to evaluate with speaker-independent data. Figure 7 shows the result of the word "but." Generally, before humans utter speech, an articulation organ is already beginning to move. It is clear that the articulatory movement ["sil" (from frames 0 to 7)] before the speech of phoneme "b" is expressed accurately from Figure 7. It is effective to train by combining the preceding and subsequent frames (t-3, t+3) in current frame t in the MLN. Figure 8 shows the number of phonemes contained in the training data. Despite the fact that phoneme "t" has the heaviest number of all the phonemes, the correlation coefficient is not very high. Since training is insufficient depending on the phoneme, further improvement is required.

Figure 9 shows the correlation coefficient for each articulatory organ. The horizontal axis shows feature points in Figure 4, and more specifically, from feature point ① to feature point ⑤ refers to the tongue, feature point ⑥ refers to the upper lip, feature point ⑦ refers to the soft palate, and feature point ⑧ refers to the lower lip. Although the soft palate shows high correlation, the correlation of the lower lip is not very good. Moreover, although the tongue is an average 0.7, since it is a very important organ for various pronunciations, it is necessary to improve the articulatory gesture of the tongue and lower lip. We plan to intensively train important articulatory manners and articulatory positions in the MLN by forming some anchor points.

## 4. SUMMARY

We developed a system for automatically generating CG animations to express pronunciation movement through articulatory features extracted from speech. The pronunciation mistakes of the user can be pointed out by expressing the pronunciation movements of the user's tongue, palate, lips, and lower jaw as animated pronunciations. We conducted experiments that confirmed the accuracy of the generated CG animations. The correlation coefficient was more than about 0.83, and we confirmed that smooth animations were generated from speech automatically. We will build a pronunciation instructor system that includes the CG animation program.

## 6. REFERENCES


[1] R. Delmonte, "SLIM prosodic automatic tools for self-learning instruction," *Speech Communication*, 30(2-3):145–166, 2000.

[2] J. Gamper and J. Knapp, "A Review of Intelligent CALL Systems," *Computer Assisted Language Learning*, 15(4): 329–342, 2002.

[3] L. Neumeyer, H.Franco, V.Digalakis and M. Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, *30*(2-3), 83–93, 2000.

[4] O. Deroo, C. Ris, S.Gielen and J. Vanparys, "Automatic detection of mispronounced phonemes for language learning tools," *Proceedings of ICSLP-2000*, vol. 1, 681–684, 2000.

[5] Phonetics Flash Animation Project: http://www.uiowa.edu/~acadtech/phonetics/

[6] K.H. Wong, W.K.Lo and H. Meng, "Allophonic variations in visual speech synthesis for corrective feedback in capt," *Proc. ICASSP 2011*, pp. 5708-5711, 2011.

[7] Y.Iribe, T.Mori, K.Katsurada and T.Nitta, "Pronunciation Instruction using CG Animation based on Articulatory Feature," *Proc of. ICCE2010 (International Conference on Computers in Education)*, pp. 501-508, 2010.

[8] Y.Iribe, S.Manosavanh, K .Katsurada, R.Hayashi, C.Zhu and T.Nitta, "Generation Animated Pronunciation from Speech through Articulatory Feature Extraction," *Proc of. Interspeecch' 11*, pp. 1617-1621, 2011.

[9] K.Honda,"Evolution of vowel production studies and observation techniques," *Acoustical Science and Technology,* Vol. 23, No. 4, pp. 189-194, 2002.