Generating Animated Pronunciation from Speech through Articulatory Feature Extraction

Yurie Iribe[1], Silasak Manosavanh[2], Kouichi Katsurada[2], Ryoko Hayashi[3], Chunyue Zhu[4] and Tsuneo Nitta[2]

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

iribe@imc.tut.ac.jp, mano@vox.tutkie.tut.ac.jp, katsurada@cs.tut.ac.jp
rhayashi@kobe-u.ac.jp, shu_s_y@koala.kobe-u.ac.jp, nitta@cs.tut.ac.jp

Abstract

We automatically generate CG animations to express the pronunciation movement of speech through articulatory feature (AF) extraction to help learn a pronunciation. The proposed system uses MRI data to map AFs to coordinate values that are needed to generate the animations. By using magnetic resonance imaging (MRI) data, we can observe the movements of the tongue, palate, and pharynx in detail while a person utters words. AFs and coordinate values are extracted by multi-layer neural networks (MLN). Specifically, the system displays animations of the pronunciation movements of both the learner and teacher from their speech in order to show in what way the learner's pronunciation is wrong. Learners can thus understand their wrong pronunciation and the correct pronunciation method through specific animated pronunciations. Experiments to compare MRI data with the generated animations confirmed the accuracy of articulatory features. Additionally, we verified the effectiveness of using AF to generate animation.

Index Terms: animated pronunciation, pronunciation learning, articulatory feature

1. Introduction

Computer Assisted Language Learning (CALL) systems have been introduced for language education in recent years [1][2]. CALL systems typically analyze a learner's speech by using speech recognition technology, and point out pronunciation problems with specific phonemes in words and automatically score the pronunciation quality [3][4][5]. However, although the learner can thus realize that his/her speech is different from the teacher's, the learner cannot understand how to correctly move the appropriate articulation organ. The system should show how to do this when the learner makes a wrong pronunciation, in the same way that teachers teach. On the other hand, although other studies have examined making correct pronunciation animations and video in advance [6][7][8], they do not automatically produce animations of the learner's wrong pronunciation. The proposed system visually represents the teacher's and the learner's articulatory movements (movement of the tongue, palate, and lips) by using CG animations. As a result, the learner can study how to move an articulatory organ while visually comparing their mispronunciation animation with the correct pronunciation animations. To represent the teacher's and the learner's articulatory movements, the proposed system extracts the articulatory features (AFs) from the learner and teacher speeches automatically. Next, the system converts speech from articulatory features into coordinate distances based on MRI data by two MLN. This paper describes the method of automatically generating animated pronunciations from speech. In section 2 we describe the method of articulatory feature extraction, coordinate distance extraction and CG animation generation. Section 3 discusses the experimental evaluation to confirm the accuracy of the generated animated pronunciation. The last section summarizes this paper.

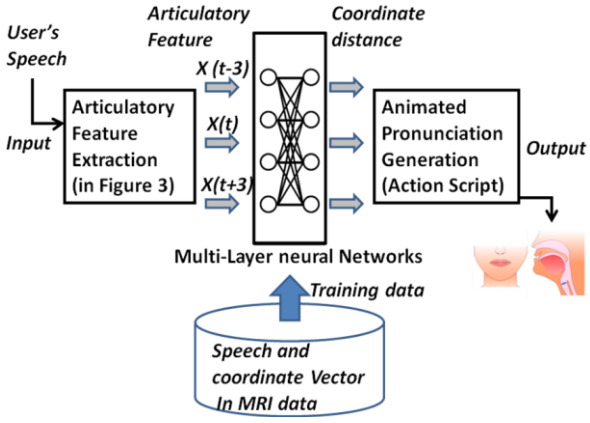

Figure 1: *System outline*.

2. CG Animation Generation System

2.1. System outline

Figure 1 shows an outline of the system. The system consists mainly of articulatory feature extraction by first multi-layer neural networks (MLN), coordinate distance extraction by second MLN, and CG animation generation programs.

We use the articulatory features composed of place of articulation and manner of articulation extracted from the speech, and use them to generate highly accurate CG animations. Concretely, the articulatory features are extracted from the speech input to first MLN, and the articulatory features and the coordinate distances of the MRI data are trained by second MLN. As for articulatory extraction, we use existing developed technologies as described in the next paragraph. The CG animation is generated based on the y-coordinate distances (Δy) extracted from 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.

2.2. Articulatory feature extraction

In order to vocalize, human beings change the shape of the vocal tract and move articulatory organs such as the lips, alveolar arch, palate, tongue and pharynx. This is called articulatory movement. Each attribute of the place of articulation (back vowel, front vowel, palate, etc.) and manner of articulation (fricative, plosive, nasal, etc.) in the articulatory movement is called an articulatory feature. In short, articulatory features are information (for instance, closing the lips to pronounce "m") about the movement of the articulatory organ that contributes to the articulatory movement. In this paper, articulatory features are expressed by assigning +/- as the feature of each articulation in a phoneme. For example, the articulatory feature sequence of "/jiNkoese/ (space satellite)" in Japanese is shown in Figure 2. Because phoneme N is a voiced sound, "voiced" in Figure 2 is given [+] (Actually, [+] is given a value of "1" (right side of Figure 2)) as the teacher signal. Because phoneme k is a voiceless sound, "voiced" in Figure 2 is given [-]. Actually, [-] is given a value of "0" (right side of Figure 2) as the teacher signal and "unvoiced" in Figure 2 is given [+]. We generated an articulatory feature table of 15 dimensions corresponding to 25 Japanese phonemes. We defined the articulatory features based on distinctive phonetic features (DPF) involved in Japanese phonemes in international phonetic symbols (International Phonetic Alphabet; IPA) [9].

We also used our previously developed articulatory feature (AF) extraction technology [10]. The extraction accuracy is about 95 %. Figure 3 shows the AF extractor. An input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segment is applied every 10 ms. The resultant FFT power spectrum is then integrated into a 24-ch BPFs output with mel-scaled center frequencies. At the acoustic feature extraction stage, the BPF outputs are first converted to local features (LFs) by applying three-point linear regression (LR) along the time and frequency axes. LFs represent variation in a spectrum pattern along two axes. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using a discrete cosine transform (DCT), a 25-dimensional (12 Δt, 12 Δf, and ΔP, where P stands for the log power of a raw speech signal) feature vector called LF is extracted. Our previous work showed that LF is superior to MFCC as the input to MLNs for the extraction of AFs. LFs then enter a three-stage AF extractor. The first stage extracts

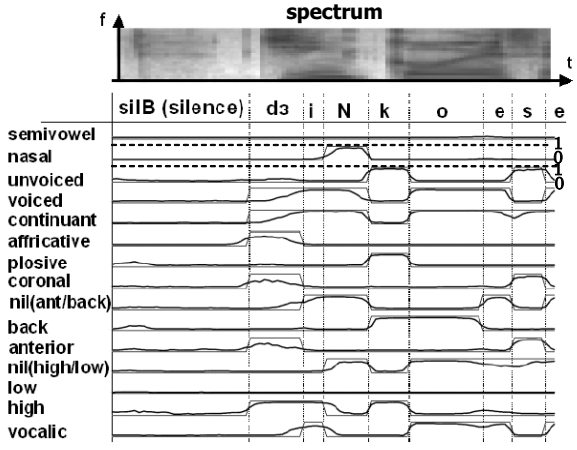

Figure 2: *Articulatory feature sequence: /jiNkoese (artificial satellite)/.*

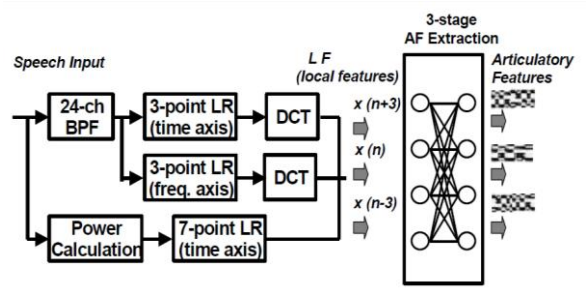

Figure 3: *Articulatory feature extraction*.

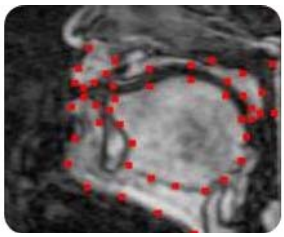

Figure 4: *Feature points on MRI data.*

45-dimensional AF vectors from the LFs of input speech using two MLNs, where the first MLN maps acoustic features, or LFs, onto discrete AFs and the second MLN reduces misclassification at phoneme boundaries by constraining the AF context. The second stage incorporates inhibition/enhancement (In/En) functionalities to obtain modified AF patterns. The third stage decorrelates three context vectors of AFs.

2.3. Coordinate distance extraction

We use the magnetic resonance imaging (MRI) data to map AFs to coordinate values that are necessary to generate CG animations. MRI captures images within the body by using magnetic fields and electric waves. We used MRI data captured in three dimensions, which shows in detail the movements of the person's tongue, larynx, and palate while making an utterance. CG animations are generated based on coordinate distances. Concretely, MLN inputs AFs extracted from speeches included in the MRI data and outputs coordinate distances. As a result, after the user's voice is input, the coordinate vectors adjusted to the speech are extracted, and a CG animation is generated based on them. This section describes the extraction of the feature points on the MRI data and the method of calculating the y-coordinate distance from them.

We assigned feature points to the mouth shape on the MRI data (tongue, palate, lips, and lower jaw) beforehand. To generate CG animations automatically, the proposed system uses the distance of the y coordinate of each feature point. We assigned 15 tongue points, 2 lip points, and 18 palate points as the initial feature points in view of the frequency of movement of the articulatory organs. Figure 4 shows these feature points.

The relative, not absolute, coordinate distance is used for CG animations because the feature points of each articulatory organ in the MRI data vary among individuals.

The coordinate distances are extracted as follows. Firstly, we imported 10-ms speech and image segment in the MRI data because speech segment is 10 ms. The coordinate value of each feature point is extracted by the optical flow calculation program for each frame. The input data for the program is the MRI images and coordinate vectors of the initial feature points. Next Acquisition of many MRI data costs time and money, so we decreased the number of dimensions of MLN training data

in order to train MLN effectively by even a small amount of MRI data. Therefore, only eight feature points having large changes in movement are trained in MLN (Figure 5). Moreover, the proposed system calculates only the y-coordinate distance of each feature point used as MLN training data to decrease the number of dimensions. The y-coordinate distance is calculated by subtracting y from y'. The x-coordinate value is the same as x-coordinate of the initial feature point (Figure 5). That is, the distance is calculated only the y-axis.

Specifically, to fix the palate with a little movement and to acquire the change of the uvula, we set feature point ⑦. Moreover, to express the movement of the tongue, the system determines the distance of the y coordinate (Δy) from feature point ① to feature point ⑤. To acquire the movement of the lips, the change in y-coordinate distance (Δy) between feature point ⑥ and feature point ⑧ was calculated. The spline curve mainly supplements the y-coordinate distance (Δy) of other initial feature points based on the above-mentioned eight points. On the other hand, to consider co-articulation, the system calculates the y-coordinate distance of the preceding and subsequent frames (t-3, t+3) in each frame (t), and trains these data in MLN. That is, the output of MLN is 8×3 dimensions.

Next, we explain the training method of MLN. AF is obtained by converting the speeches that accompany the MRI data. MLN projects the extracted AF to the y-coordinate distance. The number of dimensions of MLN is articulatory features (15×3 dimensions) as inputs and y-coordinate distances (8×3 dimensions) as outputs.

2.4. CG animation generation programs

We used the moving average method, spline curve, and median filter to construct smooth CG animations by using the y-coordinate distance extracted from MLN.

Firstly, the system smoothes the movement of the tongue, palate, upper lip, and lower jaw by the moving average method to average the coordinate vectors of each frame. Moreover, the spline curve is used to complement between 8 feature points (training by MLN) and other feature points. This generates a CG animation having a smooth curve and movement. The movement is drawn based on the y-coordinate distance, but it moved twitchily, so we used a median filter to smooth the movement. The median value means the intermediate value when it is arranged finite data in descending order. The present study outputs as the median value the intermediate value of five data: the y-coordinate value of the third frame is used as the median value when the coordinate values of five frames are sorted in ascending order.

The pronunciation learning system is designed to play CG animations on a web browser so that various users can use it. The CG animation program was implemented with Actionscript3.0 to operate on 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. The animation can be played slowly at half speed. Users can see the pronunciation in slow-motion by adjusting the play speed

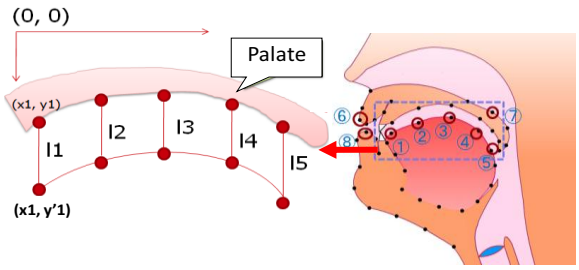

Figure 5: *Feature point used in MLN training.*

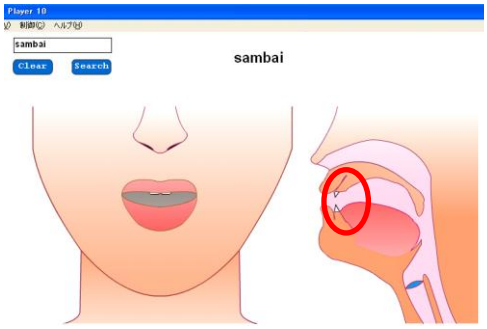

Figure 6: *CG animation of pronunciation movement.*

Table 1. *Words and syllables included in MRI data*

Japanese vowels and consonants	/a/ /i/ /u/ /e/ /o/ /ka/ /ki/ /ku/ /ke/ /ko/ /sa/ /si/ /su/ /se/ /so/ /ta/ /ti/ /tu/ /te/ /to/ /na/ /ni/ /nu/ /ne/ /no/ /ha/ /hi/ /hu/ /he/ /ho/ /ma/ /mi/ /mu/ /me/ /mo/ /ya/ /yi/ /yu/ /ye/ /yo/ /ra/ /ri/ /ru/ /re/ /ro/ /wa/ /ga/ /gi/ /gu/ /ge/ /go/ /za/ /zi/ /zu/ /ze/ /zo/ /da/ /di/ /du/ /de/ /do/ /ba/ /bi/ /bu/ /be/ /bo/ /pa/ /pi/ /pu/ /pe/ /po/
Contracted sounds	/kya/ /kyu/ /kyo/ /sya/ /syu/ /syo/ /cya/ /cyu/ /cyo/ /nya/ /nyu/ /nyo/ /hya/ /hyu/ /hyo/ /mya/ /my/ /myo/ /rya/ /ryu/ /ryo/ /gya/ /gyu/ /gyo/ /zya/ /zyu/ /zyo/ /bya/ /byu/ /byo/ /pya/ /pyu/ /pyo/
Sound of the kana /N/	/saNbai/, /saNdai/, /saNnin/, /saNko/, /saNen/, /saNwari/, /saNsai/
Double consonant /Q/	/iQpai/, /iQtai/, /iQko/, /iQsai/, /iQsyo/, /iQtu/, /iQcho/

3. Evaluation

We calculated the correlation coefficient between the coordinate values of generated CG animations and 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 of the case of AF with the case of LF as MLN inputs.

3.1. Experimental setup and method

We used MRI data pronounced by a 39-year-old Japanese male who specializes in Japanese-language education and who has received phonology training. The data is consisted of pictures and Japanese speeches when the subject pronounced in an MRI machine.

We used 5 vowels and 99 syllables, 11 words as MLN training data and 3 words ("sandai," "sanbai," "sanko") as test data among 41 Japanese words included in the MRI data. Table 1 shows the Japanese MRI data used by MLN.

Each MLN has three layers. The number of input layer is 75, hidden layer is 150, and output layer is 45 in the first MLN to extract AF. The number of input layer is 45, hidden layer is 90, and output layer is 24 in the second MLN to extract coordinate distances.

3.2. Experimental Results

Here, we discuss mainly the results of three words with the kana /N/ because the pronunciation movement of this sound differs according to the back phoneme. As a typical example, /N/ in "sanbai" is the same as the nasal sound of the English /m/ with both lips shut. As for /N/ of "sandai," it is the nasal sound when uttering with the tongue tip touching the alveolar

ridge behind the anterior teeth as in the English /n/. The /N/ of "sanko" is created without the tongue tip touching the alveolar ridge behind the anterior teeth unlike /N/ of "sandai". It is the nasal sound that is made by stopping the flow of air to the mouth and breathing out from the nose. When uttered, the back of the tongue rises just a little. We evaluated whether the animated pronunciation including /N/ was accurately generated according to the different back phoneme. Thus, the experimental method compared the correlation coefficient of the coordinate value of CG animation automatically generated from the speech and the coordinate value of the MRI data for each frame for three words. The key point is that these three words are not trained in MLN.

Firstly we calculated the correct rate of AF that is important to generate CG animation (Figure 7). Although the overall average was about 82%, it is necessary to improve AF extraction because the correct rate of /d/ was low. Next we also compared LF with AF as the input of the second MLN to show the effectiveness of using AF extracted from speech. Figure 8 also shows the results of them. The all in Figure 8 is average correlation coefficient of target phonemes. AF shows a higher correlation coefficient than LF overall. The results showed the pronunciation movement was expressed more accurately by mapping the speech to the articulatory feature.

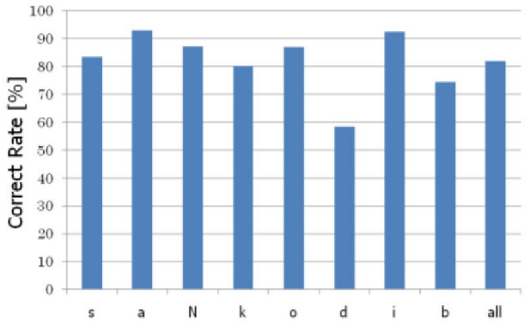

Figure 7: *AF correct rate for each phoneme.*

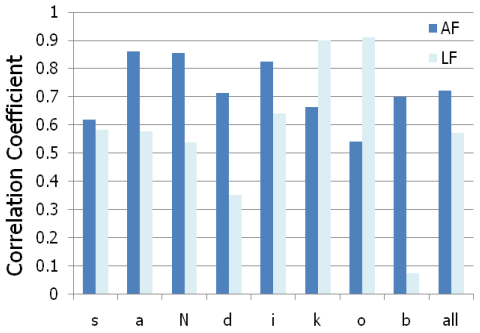

Figure 8: *Correlation coefficient for each phoneme.*

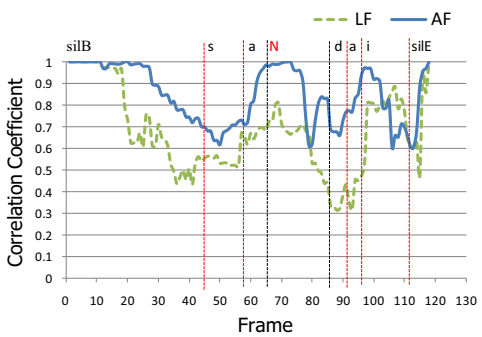

Figure 9: *Correlation coefficient of "sandai".*

Although the pronunciation movement of /N/ differs according to the back phoneme (that is coarticulation), the result of /N/ is about 0.85 which is high. The results showed that the proposed system can accurately generate CG animations while considering coarticulation. Although the AF correct rate of /o/ in Figure 7 was high, its correlation coefficient in Figure 8 was not good. Therefore, it is important to improve second MLN. Figure 9 shows the correlation coefficient per frame. The change rate of the correlation coefficient in a phoneme boundary is large depending on phoneme (Figure 9). As for /N/, the correlation coefficient decreases rapidly from around 80ms. The small amount of MRI data was used in this experiment. To generate more accurate animation, we intend to use more MRI data in future. Moreover, we will generate not only Japanese animation but also English animation by using English MRI data.

4. Conclusions

We developed a system to automatically generate 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 which confirmed the accuracy of the generated CG animations. The correlation coefficient was more than about 0.7, and we confirmed that smooth animations were generated from speech automatically. We will also improve the system to make the animation motions more natural, and build a pronunciation instructor system including the CG animation program. In the future, we will conduct experiments to compare AF and MFCC as the inputs of MLN.

5. Acknowledgements

This research was supported by a Grant-in-Aid for Young Scientists (B) (Subject No. 21700812).

6. References

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

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

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

[4] Witt, S. M. and Young, S. J. (1995). "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, *30*(2-3), 95–108.

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

[6] Wang, S., Higgins, M. and Shima, Y. (2005). "Training English pronunciation for Japanese learners of English online," *The JALT Call Journal, 1*(1), 39–47.

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

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

[9] Halle, M. (1983). "On distinctive features and their articulatory implementation," *Natural Language & Linguistic Theory*, *1*(1), 91–105.

[10] Huda, M. N., Katsurada, K. and Nitta, T. (2008). "Phoneme recognition based on hybrid neural networks with inhibition/enhancement of Distinctive Phonetic Feature (DPF) trajectories," *Proc. Interspeech'08,* pp. 1529–1532.