INTRODUCING ARTICULATORY ANCHOR-POINT TO ANN TRAINING FOR CORRECTIVE LEARNING OF PRONUNCIATION

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

[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
[4]Green Computing System R&D Center, Waseda University, Japan

ABSTRACT

We describe computer-assisted pronunciation training (CAPT) through the visualization of the articulatory gestures from learner's speech in this paper. Typical CAPT systems cannot indicate how the learner can correct his/her articulation. The proposed system enables the learner to study how to correct their pronunciation by comparing the wrongly pronounced gesture with a correctly pronounced gesture. In this system, a multi-layer neural network (MLN) is used to convert the learner's speech into the coordinates for a vocal tract using Magnetic Resonance Imaging data. Then, an animation is generated using the values of the vocal tract coordinates. Moreover, we improved the animations by introducing an anchor-point for a phoneme to MLN training. The new system could even generate accurate CG animations from the English speech by Japanese people in the experiment.

Index Terms— Computer aided instruction, Interactive pronunciation training, Articulatory feature extraction, Articulatory gesture CG-generation

1. INTRODUCTION

Computer-assisted pronunciation training (CAPT) has recently been introduced for language education [1-3]. Typical CAPT systems evaluate the pronunciation of learners and point out their articulation errors by using speech recognition technology [4-6]. Moreover, some of them can indicate the differences between incorrect and correct pronunciation by displaying a speech waveform or playing the 1st and 2nd formant frequencies [7]. The learners will then be aware of the differences; however, they cannot correct their pronunciation using only such information unless they have sufficient knowledge in phonetics. On the other hand, some studies have introduced sagittal articulatory information by using animations for correct gestures [8], [9], but because these approaches do not simultaneously feedback the learner's incorrect gesture, these types of articulatory feedback do not in fact help learners. The CAPT systems should guide learners on how to adjust their articulatory organs when correcting pronunciation errors. There have also been studies that provide visual feedback on phonetic corrections through acoustic to articulatory inversion [10]. This research provides visual articulatory feedback for one specific speaker. We have studied pronunciation training based on articulatory feature extraction from unspecified speakers' speech [11], [12] that helps visualize the learner's pronunciation errors. We believe that the step-by-step learning process using CG animation enables a learner to study how to correct his/her pronunciation by comparing it with a correct one. In the proposed system, a multi-layer neural network (MLN) is used to convert the learner's speech into the coordinates of a vocal tract using Magnetic Resonance Imaging (MRI) data [13]. The articulatory features (AFs; place and manner of articulation) extracted from speech is used as the input for the MLN [14]. Then, a CG generation process outputs the articulatory gesture using the values of the coordinates. However, the accuracies of the CG animations we have generated were insufficient. Therefore, the new system modifies the coordinate vectors corresponding to some anchor points and trains them as teacher signals in the MLN. The anchor points are configured according to phoneme in order to anchor the coordinate vectors of an important position of a articulatory organ when a phoneme is pronounced. The position of an anchor point depends on each phoneme. The system improves the MLNs by focusing on important articulatory positions. Lastly, a comparison of the generated CG animation from the given speech and the actual MRI data was conducted. In particular, we conducted a comparison of an MLN trained with the modification of the anchor points and one without them.

2. GENERATION OF ANIMATED PRONUNCIATION

2.1. System outline

Figure 1 presents a system outline. We introduce articulatory features (AFs) composed of the place and manner of an articulation extracted from given speech as described in the next paragraph, and use them to generate highly accurate CG animations. The MLN is trained on the basis of the articulatory features being used as the input data and the coordinate vectors as the output data. Here, the coordinate vectors corresponding to the anchor points for each phoneme are modified. The revised coordinate vectors are used for the teacher signal of the MLN. The proposed system improves MLNs by focusing on the training of the important articulatory positions (anchor points). The CG animation is generated on the basis of the coordinate values extracted from the trained MLN. As a result, a user's speech is input into our system, and then a CG animation is automatically generated to visualize the articulatory gestures.

2.2. Articulatory feature extraction

Articulatory features are the attributes of the movement of the articulatory organs that contribute to the articulatory movements (for instance, closing the lips to pronounce "m"). We defined the articulatory feature table for 28 dimensions corresponding to 43 English phonemes based on phonetics [15-17]. We also used our previously developed articulatory feature (AF) extraction technology [14]. Figure 2 shows the AF extractor. At the acoustic feature extraction stage, the BPF outputs are first converted to local features (LFs) by using a three-point linear regression (LR) along the time and frequency axes. The LFs represent the variations in the spectrum pattern along two axes. The 45-dimensional AF vectors are extracted from the LFs of the input speech using two MLNs, where the first MLN maps the acoustic features, or LFs, onto discrete AFs and the second MLN reduces the misclassification at the phoneme boundaries by constraining the AF context.

2.3. Coordinate vector extraction

We use magnetic resonance imaging (MRI) data to obtain the coordinate values of the shape of the articulatory organs. The two-dimensional MRI data detail the movements of the person's tongue, larynx, and palate during an utterance using phonation-synchronized imaging [13]. The data-set for the MRI and speech used in this research contains 176 vocabulary-words uttered 192 times by two native English speakers (one female and one male) and two Japanese speakers (one female and one male). The CG animations are generated on the basis of the coordinate vectors. The MLN trains the articulatory features as the input, which are extracted from the speech recorded in the MRI data collections, and the coordinate vectors of the articulatory organs acquired from the MRI images as the output (Fig. 3).

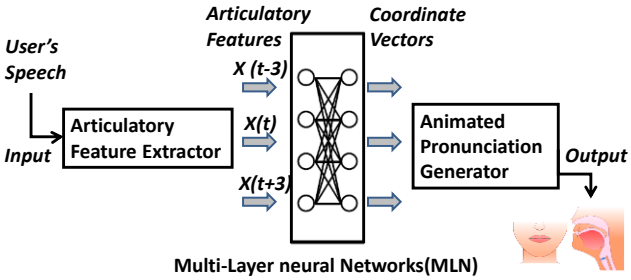

Figure 1: System Outline.

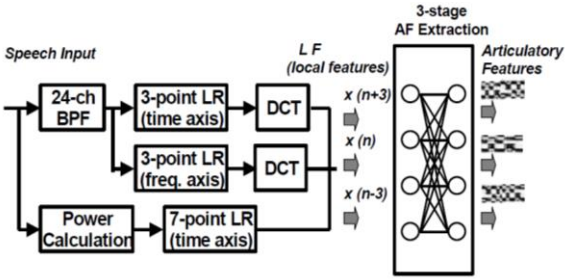

Figure 2: Articulatory feature extraction.

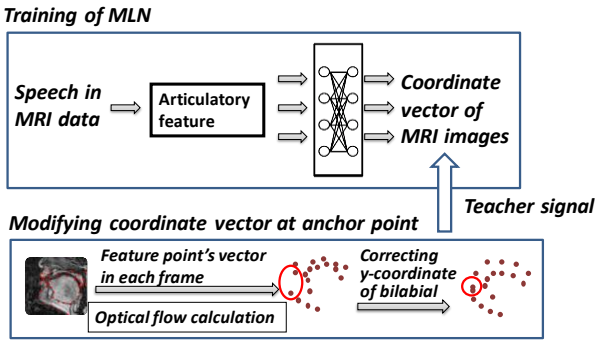

Figure 3: Training of MLN to extract coordinate vector.

As a result, after the user's speech is input, the coordinate vectors adjusted to the speech are extracted and then a CG animation is generated. In this section, the extraction of the feature points from 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 organs' (tongue, palate, lips, and lower jaw) shapes on the MRI data beforehand. The number of initial feature points was 43 (black-colored points in Fig. 4). Then, we selected six feature-points that are important in pronunciation training (Fig. 5). The coordinate value of each feature point is extracted by calculating the optical flow in each frame (Fig. 3). The input data for the optical flow program is the coordinate vector set at the initial feature points.

The coordinate vectors of each feature point are calculated. The dimensions in each MLN-unit: (a) input unit 84 (28 × 3) articulatory features and (b) output unit 36 (6 × 2 × 3) x-y coordinate vectors.

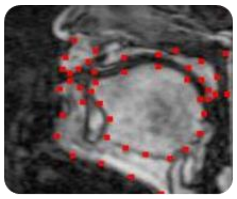

Figure 4: Feature points on MRI data.

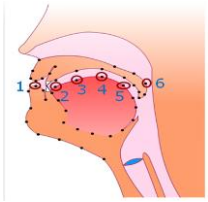

Figure 5: Feature points used in MLN training.

Table 1: Anchor and feature points for each phoneme

Phoneme	Anchor points	Feature points in Fig. 5
p, b, m	Bilabial	1
f, v	Labiodental	1
θ, ð, t_J, d_J	Dental	2
t, d, z, ɾ, l, ɹ, dz_J, ts_J, s	Alveolar	2
ʃ, ʃ_J, tʃ, dʒ, ʒ, ʒ_J	Post alveolar	3
k^j, g^j	Palatal	3
k, g	Velar	4
ŋ	Velar (Backward)	5
m, n, ŋ	Uvula	6

2.4. Modification of coordinate vectors at anchor point

We use the MRI data to obtain the detailed movements of human articulation. However, some shapes of the articulatory organs in the MRI data are indistinct because the MRI data is generated by superposing the images capturing the articulatory movements uttered 192 times per word. Therefore, these images are blurred around the edges of some of the articulatory organs. As a result, some of the coordinate values are incorrect because they are not accurately trackable by the optical flow. The proposed system, therefore, sets some anchor points based on the places and manner of the articulation. The anchor points are the important places on an articulatory gesture that change with the utterance. For instance, an anchor point of /θ/ is on the upper teeth because it is a phoneme articulated with the tongue against the upper teeth. To anchor the coordinate vector of a position of an important articulatory organ when a phoneme is pronounced, the system trains the MRI by using the modified coordinate vector for the anchor point. The approach can efficiently clarify the motions of the CG animations and teach learners the important articulatory movement. Table 1 lists the anchor points of each phoneme and a feature point (in Fig. 5) corresponding to them. The "_J" phoneme is a unique Japanese phoneme not included in the English phonemes. We defined the anchor point for the unique Japanese phoneme because Japanese speakers may pronounce some unique Japanese phonemes while uttering English.

First, the coordinate vectors of the feature point corresponding to a phoneme in the coordinate vectors calculated in Section 2.3 are modified on the basis of Table 1. For example, when humans pronounce /p/, /b/, or /m/, both lips are closed at first. Therefore, we defined bilabial (closing both lips) as the anchor point for /p/, /b/, and /m/. The coordinate value of feature point ① indicates the lower lip is modified to touch the upper lip for /p/, /b/, and /m/ in the teacher signal for the MLN, as shown in Fig. 3. In the case of /t/, /d/, and /z/, the coordinate value of feature point ② indicates the tip of the tongue is modified to touch the superior alveolar ridge between the upper teeth and the hard palate. Other coordinate vectors without anchor points are not modified. The MLN trains the revised coordinate vectors and coordinate vectors without anchor points together as a teacher signal.

Figure 6: Animated Pronunciation of "read" for learner/teacher.

2.5. CG animation generation programs

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 image at first. Then, the position relations between six important feature-points used for training the MLN and the 37 remaining feature-points are calculated. The movement is drawn on the basis of the coordinate vectors, but since this movement is often unstable, we also introduce a median filter to smooth it out. A spline curve is used to complement six specific feature-points and the other feature-points by maintaining the position relation.

A pronunciation training system is built as a web application so that various users can access it on the web. The CG animation program is implemented with Actionscript3.0 so that it can be used on a web browser with a Flash Player plug-in. Figure 6 shows a screenshot of a CG animation developed in the present study. Our system teaches the learner how and where he/she should make corrections by comparing a correct pronunciation animation and an incorrect pronunciation animation.

3. EVALUATION

The correlation coefficients between the coordinate values in CG animations extracted by the MLN and their corresponding values in the articulatory gestures investigated

for the MRI data were evaluated. In particular, the system automatically generated CG animations from the English speech by Japanese speakers. We then verified the effectiveness of anchor points.

3.1. Experimental data and setup

The MRI data used in the evaluation was taken in a single shot, in which two native English speakers and two Japanese speakers uttered 176 English words. The data set used in the experiment is as follows.

D1: *Training data set for an AF-coordinate vector converter: 176 short words of English speech and images included in the MRI data (one male and one female native English speaker, one male and one female Japanese speaker).*

D2: *Testing data set for an AF-coordinate vector converter: 75 words of English speech included in the MRI data (Japanese speaker). The experiments were conducted by using a leave-one-out cross-validation method.*

The MLN for the AF extractor was designed using not only the Texas Instruments and Massachusetts Institute of Technology (TIMIT) database [18] as the English speech, but also the Corpus of Spontaneous Japanese (CSJ) database [19], because English speech by Japanese speakers may contain unique Japanese phonemes. Each MLN has three layers. The number of input layer is 75, hidden layer is 150, and output layer is 84 in the first MLN to extract AF. The number of input layer is 84, hidden layer is 168, and output layer is 36 in the second MLN to extract coordinate distances.

3.2. Experimental results

Figure 7 shows the correlation coefficients for each feature point to compare the animations modifying a feature point corresponding to an anchor point with the animations not modifying it. The positions of each feature point are shown in Figure 5. As a result, the correlation coefficients of the animations for the English speech from Japanese speakers averaged out at 0.79. Moreover, the correlation coefficient of the animations with feature points modified at each anchor point increased by about 14% more than the animations without anchor points. In particular, it was extremely valuable to improve the coordinate vectors of feature points ① and ②, because the motions of the lip and the tip of the tongue are important for instructing articulatory movements. The results demonstrate that the animations generated from the English speech from Japanese speakers could also be smooth motions.

Figure 8 shows the correlation coefficient for each phoneme. Since the Japanese speakers in this experiment uttered unique Japanese phonemes in their English speech, the Japanese phonemes of t_J, d_J, ts_J, and ʒ_J are included in Figure 8. That is, the animations of the Japanese phonemes

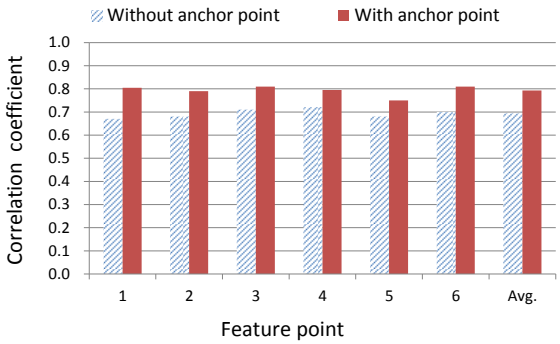

Figure 7: Correlation coefficient of each feature point (Japanese speaker: D2).

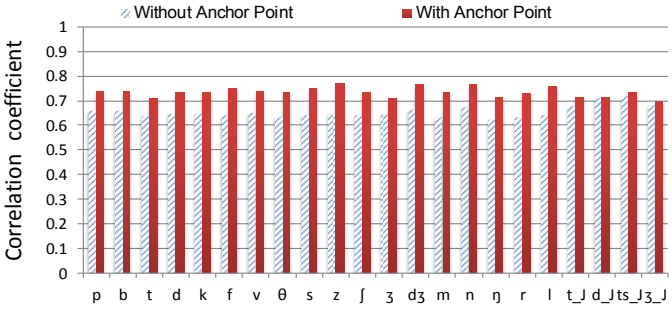

Figure 8: Correlation coefficient of each phoneme (Japanese speaker: D2).

could be generated by accurately estimating them from the English speech. As a result, our proposed system can point out these unique Japanese mispronunciations through the generated animations. For the averaged correlation coefficient of all the phonemes, the animations with feature points modified for the anchor points came out as 0.74 and the animations without anchor point were 0.67. Although the system could express important articulatory manners and places by setting some anchors, other feature points in the animations should be further corrected.

4. CONCLUSION

We developed a system that can generate the CG animation of articulatory movements by extracting the articulatory features from speech. Pronunciation errors made by learners can be seen by displaying the articulatory movements of his/her tongue, palate, lip, and lower jaw on a screen as a comparative animation with a teacher. We improved the animations by modifying some of the coordinates corresponding to the anchor points and training them in the MLN. As a result, the correlation coefficient of the animations modified at each anchor point increased by about 14% more than the animations without an anchor point and averaged out at 0.79. We confirmed that smooth animations can also be generated from the English spoken by a Japanese speaker. Future works include educational evaluations of our proposed system by conducting language classes.

6. REFERENCES

[1] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, 51(10), pp.832-844, 2009.

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

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

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

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

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

[7] P. Wik, D. L. Escribano,"Say 'Aaaaa' Interactive Vowel Practice for Second Language Learning," *Proc. SLaTE*, 2009.

[8] 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.

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

[10] P. Badin, A. B. Youssef, G. Bailly , F. Elisei, T. Hueber:"Visual Articulatory Feedback for Phonetic Correction in Second Language Learning," *Proc. SLaTE*, pp. 1-10, 2010.

[11] Tsuneo Nitta, Silasak Manosavan, Yurie Iribe, Kouichi Katsurada, Ryoko Hayashi and Chunyue Zhu, "Pronunciation Training by Extracting Articulatory Movement from Speech," *Proc. of IS ADEPT (International Symposium on Automatic Detection of Errors in Pronunciation Training)*, pp.211-216, 2012.

[12] 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.

[13] H. Takemoto, K. Honda, S. Masaki,Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *J. Acoust. Soc. Am*, 119 (2), pp.1037-1049, 2006.

[14] T. Nitta, T. Onoda, M. Kimura, Y. Iribe, K. Katsurada, "One-model speech recognition and synthesis based on articulatory movement HMMs," *Proc. Interspeech 2010*, pp.2970-2973, 2010.

[15] S. Hiki,:"A Panphonic version of "The North Wind and the Sun" for the illustration of the IPA of Japanese consonants," *Journal* of *Acoustical Society of Japan*, 66(10), pp.485-486 in Japanese, 2010.

[16] S. Takebayashi, H.Saito. "English phonetics," *Taishukan Shoten*, Tokyo, 2008.

[17] Morris, H.: "On distinctive features and their articulatory implementation," *Natural Language & Linguistic Theory*, 1(1), pp. 91–105, 1983.

[18] TIMIT Acoustic-Phonetic Continuous Speech Corpus http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

[19] The Corpus of Spontaneous Japanese http://www.ninjal.ac.jp/english/products/csj/