Articulatory Movements from Speech for Pronunciation Training

Silasak Manosavanh[a], Yurie Iribe[a*], Kouichi Katsurada[a], Ryoko Hayashi[b], Chunyue Zhu[c] & Tsuneo Nitta[a]

[a]*Graduate School of Engineering, Toyohashi University of Technology, Japan*
[b]*Graduate School of Intercultural Studies, Kobe University, Japan*
[c]*School of Language and Communication, Kobe University, Japan*
*iribe@imc.tut.ac.jp

Abstract: In this paper, we describe computer-assisted pronunciation training (CAPT) through the visualization of learner's articulatory gesture. Typical CAPT systems evaluate pronunciation by using speech recognition technology, however, they cannot indicate how the learner can correct his/her articulation. The proposed system enables the learner to study how to correct pronunciation by adjusting the articulatory organs highlighted on a screen and comparing with the correctly pronounced gesture. In the system, a multi-layer neural network (MLN) is used to convert learner's speech into the coordinate of a vocal tract using MRI data. Then, a CG generation process outputs articulatory gesture using the values of the vocal tract coordinate. Moreover we improved the animations by modifying vocal tract coordinate of important articulatory organ and training them in MLN. Lastly the comparison of the extracted CG animation from speech and the actual MRI data is investigated. The new system could generate accurately CG animations from English speech by Japanese as well as English native speech in this experiment.

Keywords: Interactive pronunciation training, Articulatory feature extraction, Articulatory gesture CG-generation

Introduction

Computer-assisted pronunciation training (CAPT) has been introduced for language education in recent years [1], [2]. Typical CAPT systems evaluate the pronunciation of learners and point out the articulation error by using speech recognition technology [3], [4], and [5]. Moreover, some of them can indicate the differences between incorrect and correct pronunciation by displaying speech waveform or 1st and 2nd formant frequencies. The learners can aware of the differences; however, they cannot correct the pronunciation only by such information unless they have sufficient knowledge in phonetics. On the other hand, some studies have introduced sagittal articulatory information by animations or video of correct gesture [6], [7], however, because these approaches do not feedback the learner's incorrect gesture at the same time, these types of articulatory feedback do not in fact help learners. The CAPT systems should guide learners how to adjust articulatory organs when correcting pronunciation error. We have studied pronunciation training based on articulatory feature extraction from speech [8], [9] that realizes visualization of learner's pronunciation error. We expect that the step-by-step learning process using CG animation enables a learner to study how to correct his/her pronunciation by adjusting articulatory movement highlighted on a screen and comparing with the correct one.

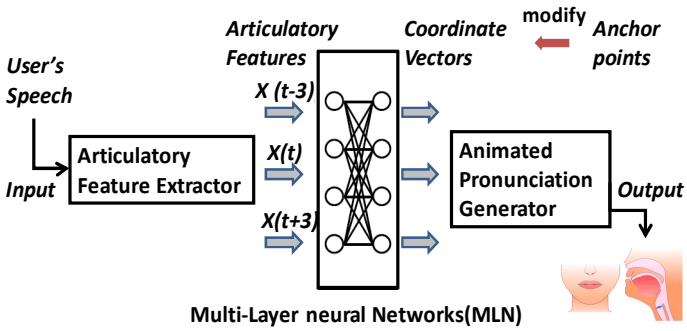

Figure 1: System Outline.

In the proposed system, a multi-layer neural network (MLN) is used to convert learner's speech into the coordinate of a vocal tract using MRI data [10]. The articulatory features (AFs; place of articulation and manner of articulation) extracted from speech is applied for the input of MLN [11]. Then, a CG generation process outputs articulatory gesture using the values of the coordinate. However, the accuracies of the generated CG animations were not sufficient. Therefore the new system modifies coordinate vectors corresponding to anchor points and gives them as teacher signals of the MLN. Here, anchor points indicate important articulatory position. They are configured according to phoneme in order to anchor the coordinate vector of an important position of a articulatory organ when a phoneme is pronounced. Namely anchor point changes depending to phoneme. The system improves MLNs by focusing to the training of important articulatory positions. Lastly, the comparison of the generated CG animation from speech and the actual MRI data was investigated. In particular we conducted the comparison of MLN trained with anchor point modification and trained without them because the goal of this paper is to improve the accuracies of the CG animations with anchor point modification.

In section 1, the estimation of articulatory feature, the coordinate vector extraction, the modification of coordinate vectors in anchor points, and the CG animation generation are described. In section 2, experimental results are discussed by comparing the extracted animation with MRI data. In the last section, the paper is summarized.

1. Generation of Animated Pronunciation

1.1 System outline

Figure 1 shows a system outline. The system consists mainly of an AF extractor [8], a coordinate vector extractor using MLN, and a CG animation generator. We introduce the articulatory features (AFs) composed of place of articulation and manner of articulation extracted from the speech, and use them to generate highly accurate CG animations. As for articulatory features extraction, we use existing developed technologies as described in the next paragraph. MLN is trained on the basis of the articulatory features as input data and the coordinate vectors as output data. Here the coordinate vectors corresponding to anchor points are modified for each phoneme. The revised coordinate vectors are applied in teacher signal of MLN. The coordinate vectors are acquired by transforming the AFs. This means that the new system improves MLNs by focusing to the training of important articulatory manners and positions. The CG animation is generated on the basis of coordinate values extracted from the trained MLN. As a result, user's speech is input in

our system, and then a CG animation is automatically generated to visualize the articulatory movements.

Table 1: A part of articulatory features set

	IPA symbol	Bilabial	Labiodental	Alveolar	Velar	Nasal	Fricative	Plosive	Voice	Voiceless
Consonant	p	+	-	-	-	-	-	+	-	+
	b	+	-	-	-	-	-	+	+	-
	d	-	-	+	-	-	-	+	+	-
	k	-	-	-	+	-	-	+	-	+
	g	-	-	-	+	-	-	+	+	-
	f	-	+	-	-	-	+	-	-	+
	v	-	+	-	-	-	+	-	+	-
	m	+	-	-	-	+	-	-	+	-

1.2 Articulatory feature extraction

In order to vocalize, humans 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. Table 1 shows articulatory features set. We generated an articulatory feature table of 28 dimensions corresponding to 43 English phonemes. We defined the articulatory features based on distinctive phonetic features (DPF) 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 2 shows the AF extractor. An input speech is sampled at 16 kHz and a 512-point FFT (Fast Fourier Transform) 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 (Band-pass Filters) 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 (Mel-frequency Cepstrum Coefficients) as the input to MLNs for the extraction of AFs. LFs then enter a three-stage AF extractor. The first stage extracts 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. Figure 3 shows actual articulatory feature sequence outputted from MLN. Articulatory features are continuous values.

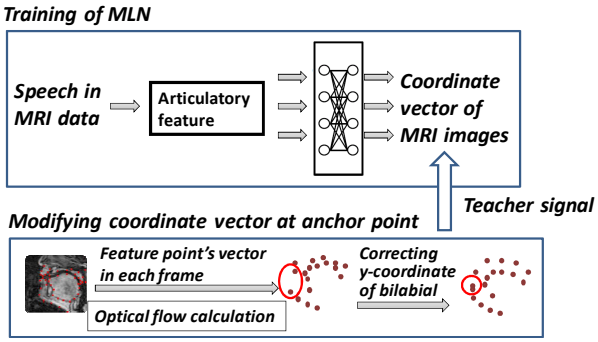

Figure 4: Training of MLN through extracting coordinate vector.

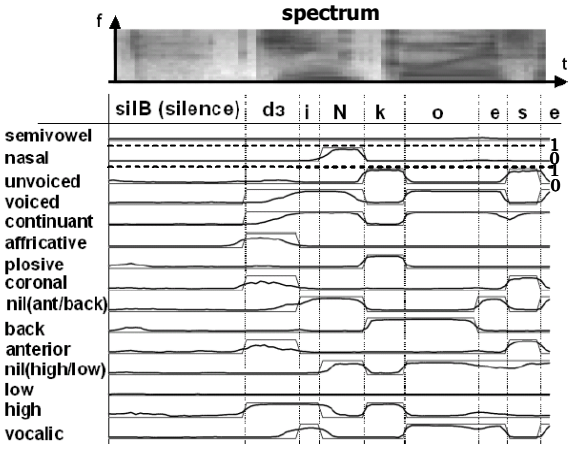

Figure 3: Articulatory feature sequence: /jiNkoese (artificial satellite)/.

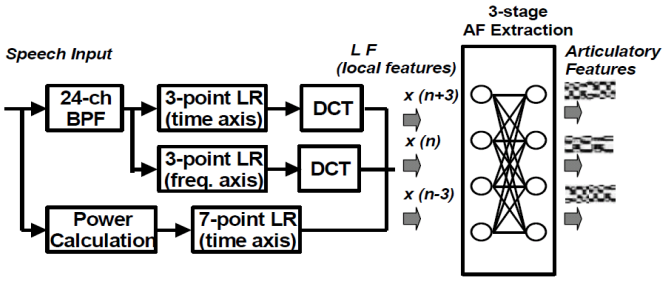

Figure 2: Articulatory feature extraction.

1.3 Coordinate vector extraction

We apply magnetic resonance imaging (MRI) data 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. The MRI data captured in two dimensions detail the movements of the person's tongue, larynx, and palate during utterance using a phonation-synchronized imaging [10]. The data-set of MRI and speech used here is 176 vocabulary-words uttered 192 times by two English native speakers (one female and one male) and

two Japanese speakers (one female and one male). CG animations are generated on the basis of coordinate vectors. The MLN trains articulatory features as input, that are extracted from speech recorded at the MRI data collections, and the coordinate vectors of the articulatory organs acquired from the MRI images as output (Figure 4). As a result, after the input of user's speech, 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 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-colored points in Figure 5). Then, we decreased the number of dimensions in the MLN in order to train the MLN effectively with a small amount of MRI data. We selected six feature-points that are important at pronunciation training (Figure 6). The feature points used here are obtained by the following steps.

1. 10-ms speech and image segment from the MRI data are imported.
2. The coordinate value of each feature-point is extracted by calculating the optical flow in each frame (Figure 4). The input data of the optical flow program is a coordinate vector set at the initial feature points.
3. The coordinate vectors of each feature point are calculated. The numbers of dimensions in each MLN-unit are; (a) input unit 84 (28 × 3) articulatory features, (b) output unit 36 (6 × 2 × 3) x-y coordinate vectors.

1.4 Modification of coordinate vectors for anchor Point

We apply MRI data to obtain the detailed movements of human's articulation. However some shapes of articulatory organs in the MRI data are indistinct because the MRI data is generated by superposing the images captured the articulatory movements uttered 192 times per a word. Some coordinate values are wrong not to be able to track them by optical flow accurately. The proposed system therefore sets some anchor points based on the places of articulation. Anchor points means an important place of an articulatory organ that change with uttering. To anchor the coordinate vector of a position of an important articulatory organ when a phoneme is pronounced, the system trains MRI by using the modified coordinate vector for anchor point. The approach could be efficient to clarify the motions of CG animations and teach the learners the important articulatory movement. Table 2 shows an anchor point of each phoneme and a feature point (in Fig.6) corresponding to the anchor point. The phoneme added "_J" means unique Japanese phoneme not included English phoneme.

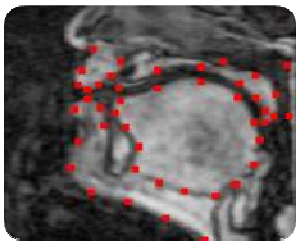

Figure 5: Feature points on MRI data.

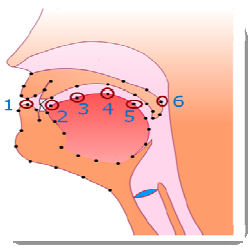

Figure 6: Feature points used in MLN training.

Table 2: Anchor point and the feature point for each phoneme

Phoneme	Anchor point	Feature point in Fig.6
p, b, m	Bilabial	1
f, v	Labiodental	1
θ, ð, t_J, d_J	Dental	2
t, d, z, ⍰, l, ⍰, dz_J, ts_J, s	Alveolar	2
⍰, ⍰_J, t⍰, d⍰, ⍰, ⍰_J	Post alveolar	3
k^j, g^j	Palatal	3
k, g	Velar	4
ŋ	Velar (Backward)	5
all phonemes except m, n, ŋ	Uvula	6

First the coordinate vector of the feature point corresponding to a phoneme in coordinate vectors calculated in section 1.3 is modified on the base of Table 2. For example when humans pronounce /p/, /b/, /m/, both lips are closed firstly. Therefore we difined bilabial (closing both lips) as the anchor point of /p/, /b/, /m/. Figure 6 explains the example. The coordinate value of feature point ① indicates the lower lip for /p/, /b/, /m/ in teacher signal is modified to coincide with it of the upper lip. In the case of /t/, /d/, /z/, the coordinate value of feature point ② indicates the tip of the tongue is modified to touch it to the superior alveolar ridge between the upper teeth and the hard palate. Other coordinate vectors without anchor points are not modified. MLN trains the revised coordinate vectors and coordinate vectors without anchor points together as teacher signal.

1.5 CG animation generation programs

We, firstly, 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. Then, the position relations between six important feature-points used for training at MLN and remaining 37 feature-points are calculated. The spline curve is used to complement six feature-points and other feature-points by keeping the position relation. The movement is drawn on the basis of coordinate vectors, however, since this movement is often unstable, we introduce a median filter to smooth it out. A pronunciation training system is built as a web application so that various users can access on the web. The CG animation program is implemented with Actionscript3.0 to operate on web browser with Flash Player plug-in. Figure 7 shows a screen shot of a CG animation developed in the present study. The system highlights the wrong articulation organs with a red dot by comparing positions of each feature point between the learner's animation and the teacher's animation. For instance, Figure 7 draws a red dot on the tip of the tongue, because the learner touches the tip of the tongue to palate for utterance of /r/. Our system teaches the learner how and where he/she should make corrections with the red dot.

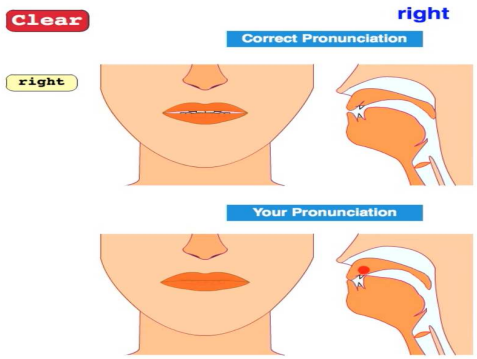

Figure 7: Animated Pronunciation of "right" for learner/teacher.

2. Evaluation

The correlation coefficients between the coordinate values in CG animations extracted by MLN and their corresponding values in articulatory gestures investigated on the MRI data were evaluated. Especially the new system generated automatically CG animations from English speech by Japanese speakers as well as English native speech in this experiment. We verified the accuracy of their animations. Additionally, the effectiveness of anchor point was investigated.

2.1 Experimental data and setup

The MRI data used in the evaluation was taken in a single shot, in which two English native 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 English native speakers, one male and one female Japanese speakers).
D2-1: Testing data set for an AF-coordinate vector converter: One word of English speech included in the MRI data (English native speaker).
D2-2: Testing data set for an AF-coordinate vector converter: One word of English speech included in the MRI data (Japanese speaker).

The MLN for the AF extractor [11] was designed using not only TIMIT (the Texas Instruments and Massachusetts Institute of Technology) database [12] as English speech but CSJ (The Corpus of Spontaneous Japanese) database [13], because English speech by Japanese speakers may contain unique Japanese phone. Experiments are conducted by using a leave-one-out cross-validation method, that is, 175 trials are investigated.

2.2 Experimental results

To compare the animations modified a feature point corresponding to an anchor point with the animations not modified it, Figure 8 and Figure 9 show the correlation coefficient for each feature point. The articulatory organs of each feature point are shown in Figure 6. The horizontal axis shows feature points, they are; from feature point ① refers to the lower lip, feature point ② to ⑤ refers to the tongue, and feature point ⑥ the soft palate. As a result, the correlation coefficient of animations modified feature points of an anchor point increased by about 0.11 point than the animations without anchor point in the case of English speaker, and it improved 0.10 point in the case of Japanese speaker. Especially it was extremely valuable to be improved coordinate vectors of feature point ① and feature point ②, because the motions of the lip and the tip of the tongue are important to instruct articulatory movements. Moreover the correlation coefficient of the animations for English speech of Japanese speakers was average 0.79. The results demonstrate that the animations generated from English speech of Japanese speakers could be also smooth motions.

Figure 10 and Figure 11 show the correlation coefficient for each phoneme. Because Japanese speaker in this experiment uttered unique Japanese phonemes in English speech, Japanese phonemes of t_J, d_J, ts_J and ʒ_J are included in Figure 11. That is, the animations of Japanese phonemes could be generated by estimating accurately them from English speech. As a result, our proposed system can point out such a unique Japanese mispronunciation through the generated animations. For the averaged correlation coefficient of all the phonemes, the animations with the modification for anchor point were 0.74 and the animations without anchor point were 0.67. Though the system could

express important articulatory manners and places by setting some anchors, other feature points in animations should be further corrected.

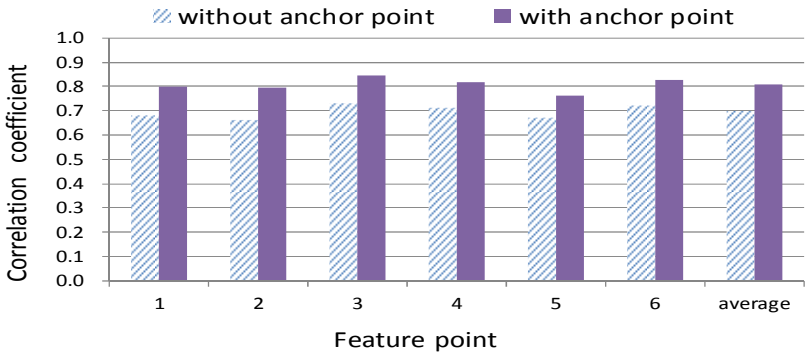

Figure 8: Correlation coefficient of each feature point (English native speaker: D2-1).

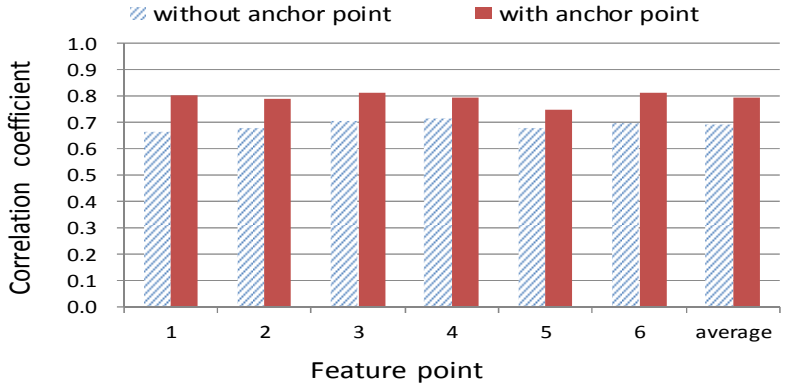

Figure 9: Correlation coefficient of each feature point (Japanese speaker: D2-2).

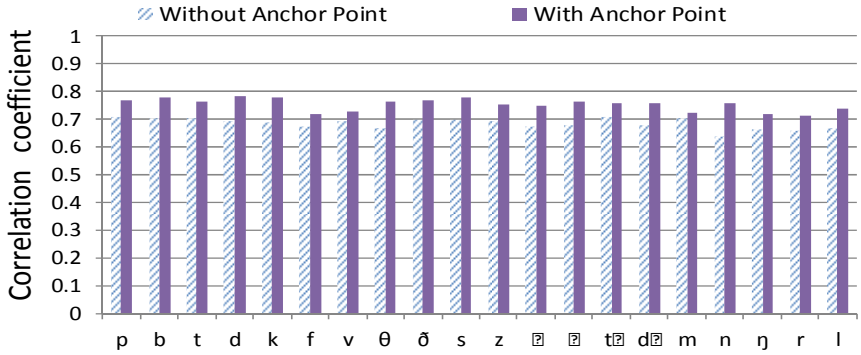

Figure 10: Correlation coefficient of each phoneme (English native speaker: D2-

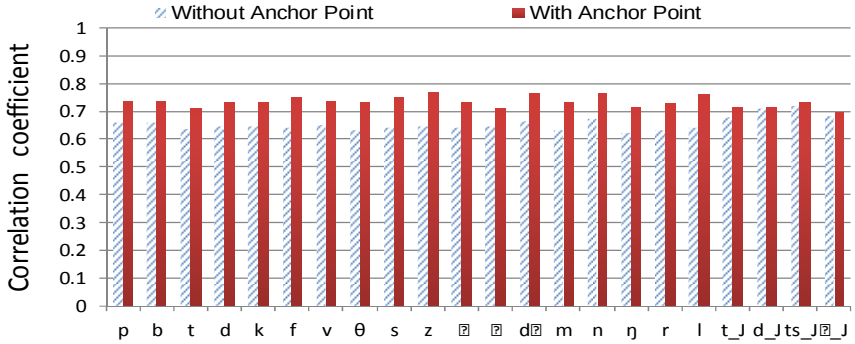

Figure 11: Correlation coefficient of each phoneme (Japanese speaker: D2-2).

3. Conclusion

We developed a system that can generate CG animation of articulatory movements by extracting articulatory features from speech. Pronunciation errors of a learner 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. In this paper, we improved the animations by modifying some coordinates corresponding to anchor points. Experimental evaluation showed the correlation coefficient of the CG animations with articulatory gestures investigated in the MRI data. As a result, the correlation coefficient of animations modified in anchor point increased by about 0.11 point than the animations without anchor point. The correlation coefficient of the animations for English speech of Japanese speaker was average 0.79. We confirmed that the smooth animations can be also generated from English speech of Japanese speaker as well as English native speaker. Future works include educational evaluations of our proposed system by conducting in language classes.

Acknowledgements

This research was supported by a Grant-in-Aid for Young Scientists (B) (Subject No. 24720254) and the Kayamori Foundation of Information Science Advancement.

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

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

[7] 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.

[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] Y. Iribe, S. Manosavanh, K . Katsurada, R. Hayashi, C. Zhu and T. Nitta, "Improvement of animated articulatory gesture extracted from speech for pronunciation training," *Proc. of ICASSP'1*2, 2012

[10] 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.

[11] 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.

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

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