Generation of CG Animations Based on Articulatory Features for Pronunciation Training

Yurie Iribe*, Takuro Mori*, Kouichi Katsurada* and Tsuneo Nitta*,**

(Received 13 April 2012 and accepted in revised form 30 October 2012)

Abstract We describe a system for pronunciation training that dynamically generates CG animations to express pronunciation visually from speech based on articulatory features. The system specifically displays the results of phoneme recognition and CG animations of articulatory movements of both learners and a teacher that are estimated from their speech. Learners can thus notice their mispronunciation movements and find the correct method of pronunciation by comparing their incorrect pronunciation movements with the correct ones on the animations. We conducted an experiment to evaluate the effectiveness of the animated pronunciations and we acquired a correctness of 93% for articulatory features with our proposed system. As a result, we clarified that CG animations could adequately visualize the teacher’s articulatory movements and those of learners. Further, the improvement to the pronunciation score with the proposed system was double that with the existing system. These results verified that the new system was an effective training system.

Keywords: interactive pronunciation training, articulatory feature (AF), articulatory movement, CG animation, speech recognition

1. Introduction

Face-to-face lectures are ideal in pronunciation education in foreign language education. A teacher teaches accurate pronunciations to students by explaining the movements and relative locations of the tongue, palate, and lip rounding when they are making utterances. The teacher also points out mistakes learners are making in pronunciation and how to correct them. However, it is difficult for the teacher and learners to arrange opportunities to study together one-on-one because they are often busy. There is not enough time to study pronunciation during regular school hours. It is therefore necessary to provide an environment and tools that enable learners to study pronunciation at any time and from anywhere. Computer-based pronunciation training systems are very useful as they allow learners to study pronunciation at any time.

Computer-assisted pronunciation training (CAPT) systems have been introduced into English language education in recent years[1, 2]. CAPT systems typically analyze learners’ speech by using speech recognition technology, and they point out problems with pronunciation of specific phonemes in words and automatically score the quality of pronunciation[3–5]. They also often indicate the differences between incorrect and correct pronunciation by displaying learners’ speech waves and correct speech waves[6, 7]. However, although learners can recognize that their speech is different from the teacher’s, they cannot understand how to correctly move the appropriate organ for articulation.

However, one CAPT system can display the resonance frequency of the vocal tract of vowels uttered by a teacher and learners on an F1-F2-F3 plane[8–10]. Here, Fi is the i-th resonance frequency of the vocal tract. Nevertheless, it is difficult for students to understand how they can correct their articulation from the F1-F2-F3 plane without knowledge of phonetic science. Although these studies only targeted vowels, our proposed system includes articulation of consonants as well as vowels.

Other studies have examined the creation of animations and video to correct pronunciation in advance[11–13], but they have not dynamically produced animations of learners’ erroneous pronunciations. A system should guide learners in how to correct mispronunciations as is done in face-to-face lectures. There has also been a study that provided visual feedback on phonetic corrections through acoustic to articulatory inversion[14]. This research gave visual articulatory feedback for one specific speaker; however, it is important for
pronunciation training systems to provide feedback to various speakers. We therefore provide articulatory movements for unspecified speakers by converting speech to articulatory features based on multi-layer neural networks trained for the speech of unspecified speakers.

The system we propose provides several interactive methods for learners that not only point out problems with their individual pronunciations, but that also visually represent the teacher’s and learners’ articulatory movements (movements of the tongue, palate, and lips) by using CG animations. The system appropriately provides learners with guidance on their inaccurate articulation in pronunciation as well as correct articulation in real time. As a result, learners can study how to move the articulatory organs while visually comparing animations of their mispronunciations and the correct pronunciations. Such animations make it easy to understand articulatory movements. The proposed system directly estimates articulatory features (the place and manner of articulation, which express features of articulatory movements) from the teacher’s and the learners’ speech automatically to represent their articulatory movements.

In this paper, Section 2 outlines the proposed system and Section 3 describes the method of estimating articulatory features (AFs). Section 4 explains the system for generating CG animations in detail. Section 5 discusses our experimental evaluation to confirm the accuracy of the animated pronunciations that were generated. The last section summarizes the paper.

2. Outline of Interactive Pronunciation Training System

Our main aim was to develop a system that could represent how to correct movements of articulatory organs while comparing mispronunciations with correct pronunciations through animations. Therefore, we had to accurately and automatically acquire detailed information on the movements involved in pronunciation from speech. That is, it was important to estimate movements from learners’ imprecise pronunciations. Therefore, the proposed system automatically expressed the movements of pronunciation by estimating AFs and generated CG animations based on these features.

Figure 1 outlines the proposed system, which consists mainly of an AF estimator, speech recognition by a hidden Markov model (HMM), and a CG animation generator. First, the system estimates AFs from a learner’s and a teacher’s speech to analyze movements in their pronunciation.

We used existing technologies that we had already developed for speech recognition(15). The CG animations were generated by using AFs, the phoneme sequence, and phoneme boundary extracted by using these technologies.

Moreover, our system was designed for American English because the proposed system was based on the Texas Instruments and Massachusetts Institute of Technology (TIMIT)(16) data set.

3. Estimation of Articulatory Features

3.1 Articulatory features

We developed an automatic speech recognition (ASR) system that reduced the effect of high-level additive noise based on calculations of a distinctive phonetic feature (DPF)(17) that systematized common features to some phonemes(15,18). The system recognized speech by estimating distances between phonemes based on DPF.

However, humans change the shape of their vocal tracts and move articulatory organs such as the lips, teeth, alveolar arch, palate, tongue, and pharynx to vocalize speech. These motions are called articulatory movements. All attributes of the place of articulation (e.g., back vowels, front vowels, and palate) and manner of articulation (e.g., fricative, plosive, and nasal) in articulatory movements are called articulatory features (AFs).

We revised the DPFs for phonemes to features for articulation (AFs) to apply our ASR based on DPFs to multiple languages. We defined the AFs for English
phones in international phonetic symbols (the International Phonetic Alphabet: IPA) based on phonetics with phoneticians, to assign AFs to phones in common use throughout the world\textsuperscript{(19–22)}. Concretely, AFs were expressed by assigning \( /H11001/H11002\) as the feature of each articulation in a phoneme in this paper (Figure 2). The three pairs of features \((++, +, -)\) for diphthongs in Figure 2 show the variations of the place of articulation. For instance, the vowel height “open-close” of [eI] changes to close from close mid during the utterance.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{artofset}
\caption{Articulatory Feature Set for IPA Symbols.}
\end{figure}
We generated an AF table of 28 dimensions corresponding to 42 English phonemes.

3.2 Estimation of articulatory features

We also used our previously developed AF extraction technology\(^{(23)}\). Figure 3 outlines the flow for the AF extractor. Input speech is sampled at 16 kHz and a 512-point fast Fourier transform (FFT) of the 25-ms Hamming-windowed speech segment is applied every 10 ms. The resulting FFT power spectrum is then integrated into 24-channel band pass filtering (BPF) outputs with Mel-scaled center frequencies. The BPF outputs at the extraction stage of acoustic features are first converted into local features (LFs) by applying three-point linear regression (LR) along the time and frequency axes. LFs represent variations in the spectrum pattern along two axes. After these two LFs with 24 dimensions are compressed into LFs with 12 dimensions using a discrete cosine transform (DCT), 25-dimensional (12 \(\Delta t\), 12 \(\Delta f\), and \(\Delta P\), where \(P\) stands for the log power of a raw speech signal) feature vectors called LFs are extracted. Our previous work applied LFs as the input for multilayer neural networks (MLNs) to estimate AFs. We assigned +/- to each articulation in a phoneme as AFs in MLNs as a teacher signal. The [+] is actually awarded a value of “1”. The [-] is actually given a value of “0”. Although the teacher signal is given 0 or 1, the actual AFs of MLN output are continuous values in a range from 1 to 0. An example of AF sequences of “are more latent” is shown in Figure 4. Because phoneme /l/ is a voiced sound, the “voiced” in Figure 4 is awarded a [+] . Because phoneme /t/ is a voiceless sound, the “voiced” in Figure 4 is awarded a [-] . The “unvoiced” in Figure 4 is awarded a [+].

3.3 Evaluation of articulatory features

We utilized phoneme recognition technology using the AFs that we developed\(^{(15)}\). The phoneme boundary and phoneme sequence were obtained by inputting AFs that were estimated as described in Section 3.2 to the HMM. The score for each phoneme was calculated by comparing the correct AF with the estimated AF based on the phoneme boundary. The phoneme sequences...
4. Pronunciation Instruction Based on CG Animations

A CG animation of the inside of the mouth of the teacher and the learner was generated based on AFs and phoneme sequences, and a phoneme boundary was obtained from their speech. Because the CG animations were automatically generated as their speech was input, each learner could immediately confirm the motions involved in pronunciation.

4.1 Analysis of articulatory features

The phoneme sequences, phoneme boundaries, and values [0 or 1] of AFs that indicate the movements and locations of articulatory organs were obtained as described in Section 3.2. Figure 5 shows part of the results obtained from AFs for the word “read”. The system recognizes the phoneme /r/ from frames 22 to 37 of speech. One frame corresponds to 10 ms. The phoneme boundary that was designated between the start and end points of each phoneme was used for the reproduction timing of animation. The AF sequence in this paper was defined based on the AF values of 28 dimensions (the place and manner of articulation) in each frame (Figure 5①). This section describes how the articulatory movements to generate animations based on AFs were determined. The procedure involved four steps.

1. The value of each AF of each phoneme was estimated from speech. In the example in Figure 5, the AFs from frames 22 to 37, which represented the phoneme boundary of /r/, were estimated (Figure 5①).

2. The time average of each AF was calculated for a phoneme boundary (e.g., the /r/ phoneme boundary is frames 22 to 37). The mean value of the “alveolar” in phoneme /r/ was 0.58 (Figure 5②).

3. AFs were classified according to the place and manner of articulation based on the articulatory phonetics summarized in Table 1. For instance, the nasal, fricative, approximant, and lateral approximant were classified as the manner of articulation (consonants) (Figure 5③). The vowel height was determined from the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw.

4. For each AF category, the most effective AF was chosen from the AFs of that AF category for all AF categories. Here, the most effective AF was defined as the AF that had the highest value. Because the mean value of “alveolar” was the maximum in the “place of articulation (consonant)” category in the /r/ section, the alveolar movement was registered as one AF of /r/.

4.2 Generation of animations

It is necessary to present the movement of the mouth cavity viewed from the side and the lips viewed from the front to clearly demonstrate to the learner how to correctly pronounce /r/.

It is also necessary to accurately express the movements of all articulatory organs. We generated some animations in advance independently of the place and manner of articulation included in AFs, and then overlapped these animations. The order in which animations were generated involved four steps.

1. The patterns of movement for each articulatory organ were prepared with Adobe Illustrator CS3 before-
hand. Various animations of movements such as the alveolar (fricatives), alveolar (plosives), postalveolar (approximants) of the tongue (consonants), and closed, open, and narrowing of the lips (bilabial) were prepared. The animation movements were prepared based on phonetics.

2. We used the “shape tween” ActionScript for the animated motion. The shape tween is a function that changes the shape of an object gradually whenever one frame is advanced by matching the top of the object in the first frame with that in the last frame for each articulatory organ. The animation for one articulatory organ is about 15 frames (15 ms), and is assigned a unique ID.

3. The animation for the mouth cavity and lips was generated by overlapping respective animations corresponding to the AF of a phoneme as determined in Section 4.1 (Figure 6). The animations were generated for all phonemes using the same method.

4. Next, the animations of respective phonemes were connected (Figure 7). The last shape of the previous phoneme’s animation might not have been the same as the default shape for the following phoneme. The system therefore connected two animations by com-

| Category of Articulatory Feature | Articulatory Feature |
|----------------------------------|----------------------|
| Vowel/consonant                  | Vowel, consonant     |
| Voice band                       | Voiced, unvoiced     |
| Manner of articulation (consonant)| Plosive, nasal, fricative, flap or tap, approximant, lateral approximant|
| Place of articulation (consonant)| Bilabial, labiodental, dental, alveolar, postalveolar, palatal, velar, glottal |
| Vowel height (vowel)             | Close, close-mid, open-mid, open |
| Vowel backness (place of tongue) | Front, back, central |
| Other (vowel)                    | Round, tense, rhoticity |

**Figure 6.** Generation of Animation for Each Phoneme.

**Figure 7.** Connecting Animations of Respective Phonemes.
implementing the starting frame of the next phoneme’s animation and five frames earlier than the last frame of the previous phoneme’s animation by applying key frames to the animations. Key frames have become popular in the field of computer graphics and they define the start and end points of smooth transitions. The connection between phoneme animations becomes more natural by using key frames.

4.3 Pronunciation instruction functions

Four effective functions for pronunciation training were included to appropriately and specifically display the learner’s correct and incorrect articulatory movements. Figure 8 has an example of a pronunciation training window. We provide clear details on these functions below.

(1) Comparing correct with incorrect animations

When the play button on the screen is clicked, an animation of the mouth cavity and lips is played with the learner’s speech. When the play button on the correct pronunciation side is clicked, the animation is reproduced with the teacher’s speech which had been recorded beforehand. When the simultaneous button is clicked, both the learner’s animation and the correct animation are simultaneously reproduced. The learner can thus visually compare his/her own pronunciation with the correct pronunciation.

(2) Highlighting of wrong articulatory organs

The system highlights the wrong articulation organs with a red circle to teach the learner how and where he/she should make corrections by comparing the AFs of the learner and the teacher if the learner’s pronunciation is wrong (red circle in Figure 8). As a result, the learner can see which articulatory movements are wrong, and how the articulatory organs should be moved to pronounce phonemes correctly. Because the movements of the lips and the tongue for the /æ/ of “map” are different (the learner pronounces /æ/ in Figure 8, a marker is displayed on the lips and tongue of both the learner and teacher.

(3) Articulatory movements of each phoneme (Figure 8)

It is also important to confirm the correct method of pronunciation for each phoneme. Therefore, the pronunciation animation for the phoneme can be played by clicking on the phonemic symbols (/m/, /æ/, and /p/) on the screen.

Figure 8. Example of Pronunciation Instruction.
Conversion of speech rate (Figure 8\textcopyright)

Slow-motion replay of speech and animations at three speeds of $\times 1$, $\times 0.5$, and $\times 0.25$ is possible in the system, allowing the learner to see the movements of pronunciation in slow motion by adjusting the play speed. We built a function to convert the speech rate to enable the speed of speech to be adjusted.

5. Preliminary Evaluation

5.1 Data set

We calculated the correct rate of AFs to generate CG animations that confirmed the accuracy of the animations. Moreover, we evaluated the rate of progress for subjects’ pronunciation scores and conducted a preliminary subjective evaluation to demonstrate the effectiveness of using the CG animations when the system we developed was used.

We used English speech data of English native speakers for AF training and HMM training in this experiment. There were three training data sets for the detailed speech data.

- **D1**: Training data set for AF estimator training: 2600 TIMIT\textsuperscript{(16)} sentences of English speech (325 male native English speakers)
- **D2-1**: Testing data set of AFs: Eight words of English speech (one male native English speaker)
- **D2-2**: Testing data set of AFs: Eight words of English speech (11 male Japanese speakers who were the subjects in this experiment)

D2-1 data included speech data that an English (American English) native speaker who was an English school teacher had pronounced. They consisted of eight words: “box (\textipa{bɒks})”, “dish (\textipa{dɪʃ})”, “good (\textipa{ɡʊd})”, “lead (\textipa{liːd})”, “map (\textipa{meɪp})”, “room (\textipa{rʊm})”, “sing (\textipa{sɪŋ})”, and “think (\textipa{θɪŋk})”. D2-2 data included speech data that Japanese subjects in Section 5.3 pronounced. They consisted of eight words: “read (\textipa{riːd})”, “bird (\textipa{bɜːd})”, “good (\textipa{ɡʊd})”, “think (\textipa{θɪŋk})”, “map (\textipa{meɪp})”, “sea (\textipa{siː})”, “sing (\textipa{sɪŋ})”, and “bought (\textipa{bɔːt})”.

The MLN for AF training consisted of three layers. There were 25\times 3 (LF) input layers, 150 hidden layers, and 28\times 3 (AF) output layers in the MLN to estimate AFs.

5.2 Correct rate of articulatory features

First, we evaluated the correct rate (%) of AFs (AFCR) because the accuracy of CG animations depended on their correct rate.

$$AFCR = \frac{N_c}{N \times 28} \times 100[\%]$$

$N_c$ is the total number of correctly recognized AFs in each frame. The test data of D2-1 and D2-2 were directly labeled onto every speech by forced alignment of HMM. The correct AFs in each frame were converted from the labeling information (phoneme sequences and phoneme boundary) based on the AF table in Figure 2. $N_c$ was counted by comparing the recognized AFs with

![Figure 9. Correct Rate [%] of Articulatory Features for Each Phoneme.](image-url)
the correct AFs. N is the total number of frames. The number 28 indicates the dimensions of AFs.

Figure 9 shows the correct rate (expressed as a percentage) for all phonemes included in test data D2-1 and D2-2. The mean value of AFCR (“ALL” in Figure 9) was around 90% for subject speech and 96% for the teacher’s speech. The AFCR of subject speech was the mean of all subjects’ AFCR’s. Here, because D2-2 data did not contain /d/ and /l/, their AFCRs for teacher’s speech were each 0%. Here, the phoneme sequence for “box” was counted as /b o k s/.

As a result, we clearly saw that the designed MLN system could convert speech to AFs based on accurately labeled information. Although /u/, /u/, /j/, and /z/ were phonemes not included in the eight words given to subjects, their AFs were counted. This indicated that Japanese speakers pronounced them by replacing them with other phonemes. In contrast, the AF of /l/ was only estimated from the teacher’s speech regardless of whether it was contained in the word “read”. This demonstrated that the Japanese subjects pronounced the “r” in “read” by replacing it with /l/. Our system estimated this pronunciation of /l/ through the MLN.

Figure 10 presents the AFCR results for all AFs. Almost all AFs for the teacher’s speech were adequately estimated at over 90%. However, the results for subject speech were dispersed overall. As a result, D2-1 speech data were labeled almost exactly and AFs were adequately estimated by the MLN. However, because D2-2 speech data were not labeled exactly by forced alignment, we assumed the MLN could not sufficiently estimate AFs. Because these speech data were labeled based on the TIMIT through forced alignment, we assumed that particular Japanese phonemes would not be adequately labeled. Because English speech by Japanese speakers may contain unique Japanese phonemes, it is necessary to label phonemes for Japanese speech in English. We plan to train an AF estimator not only for English speech but also for Japanese speech in the future.

The D1 data set included the speech data of native English speakers who came from different hometowns. Therefore, even if they pronounced the same phoneme, their articulation might differ. The proposed system in this experiment generated animations of articulatory movements for standard American English. For example, two articulatory variants of /r/ were generally found. For bunched /r/, the dorsum of the tongue is bunched in the region of the palate. For retroflex /r/, the tip of the tongue is curled up. We employed bunched /r/ that is commonly used in American English (20). However, we must estimate the articulatory variants for different speakers to visualize their articulatory movements more accurately from speech. Therefore, we plan to map speech to AFs by observing the movements of the tongue, palate, and pharynx in detail while a person is uttering words by using data obtained from magnetic resonance imaging (MRI).

Additionally, the articulatory matrix (Figure 2) we designed still did not take into account the co-articulation that changes articulatory movements according to the phonetic environment. For instance, palatalization and assimilation are phenomena such that the place of articulation and/or the manner of articulation changes.
according to succeeding phonemes. The rule of co-articulation should be applied to the articulatory matrix. The MLN should be trained to accurately represent the relations between speech and AFs based on co-articulation rules and MRI images.

Moreover, our system should utilize speech data defined based on phones to accurately estimate articulatory movements because the AFs were assigned according to the IPA (phones). However, the system was built based on phonemes in this paper because most current speech data are labeled for phonemes. We intend to label speech data based on phonemes and apply these data to our system.

Last, although this experiment was used for a preliminary evaluation of the robustness of the designed MLN through the correct rate of AFs, the accuracy of AFs estimated from speech and animations generated based on them should be evaluated by comparing articulatory movements that are actually observed. Therefore, we intend to verify these accuracies by using phoneticians and MRI data.

5.3 Evaluation of subjects

We compared the proposed system with an existing system in an experiment to demonstrate its effectiveness. The existing system was used speech recognition technology to display correct and incorrect phonemes in text and correct articulatory movements in pictures as seen in Figure 11. The subjects were 11 Japanese native speakers, who were in a beginner’s class for English pronunciation. Figure 12 shows the experimental scenario.

The experiment was carried out in five steps.
1. The subjects were divided into groups 1 and 2 to avoid subject bias.
2. Eight English words including a phoneme that Japanese people were not good at were prepared. The eight words were divided into two groups of four words each. The words were divided into word groups A (“read”, “bird”, “good”, and “think”) and B (“map”, “sea”, “sing”, and “bought”) to avoid word bias.
3. We explained how the system could be used before the experiment.
4. We recorded the subjects’ pronunciation of speech before the experiment, and recorded the score evaluated by the system. The pronunciation quality of each speech was automatically scored using an existing method (24).
5. Group 1 pronounced word group A after they had studied these pronunciations with the existing system for one week. Group 2 pronounced word group B after they had studied pronunciation with the proposed system for one week. Next, both groups changed systems and pronounced the other word group.
6. The system calculated the score for all speech.

Table 2 lists the results for the experiments. The pronunciation score with the proposed system was double that with the existing system. The score increased by about 30% especially when group 2 used the new system, even though the score for word group A in both groups was low before the study. Learners with the existing system could not understand how to correct mistakes in pronunciation even though they noticed incorrect phonemes. Nevertheless, learners could clearly and
efficiently understand how to correct their mispronunciations with the proposed system, which visually showed the correct articulatory movements and the learners’ articulatory movements. However, for the rate of increase in word group B, the difference between the new and existing systems was only 1.2%, which did not indicate a significant difference. Nevertheless, the rate of increase for “sing” in word group B was 17.1% for the existing system and 2.2% for the new system. There were cases where some subjects pronounced /u/ behind /ŋ/. However, the differences between the generated animation for the subject’s mispronunciation of /ŋu/ and the animation for the correct pronunciation of /ŋ/ were difficult for subjects to understand. That is, the animations could not indicate how to correct the mispronunciation. Subjects in the existing system checked /sɪŋu/ displayed as a recognition result for their pronunciation and their mispronunciation was corrected by their being careful not to insert /u/. It was sufficient to only display the recognition result depending on the mispronunciation. We should also improve the user interface to display the place and method to clearly correct mispronunciations by combining them with the recognition results. However, the rate of increase for the existing system was 1.7% for “bird” in word group A, but that for the new system was 27.2%. Although most subjects were able to pronounce /b/ correctly, they could correct this by comparing them with the generated animations. The animations would be effective for broad movements such as plosives.

Although the eight words used in this experiment contained phonemes that were difficult for Japanese to pronounce, it was clear that only the content and numbers of these words were insufficient. The effects in this evaluation were confined to these phonemes. We therefore intend to conduct a long-term experiment with many words that contain various phonemes to verify the effectiveness of animations for pronunciation training.

We administered the five-stage evaluation listed in Table 3 with questions that focused on the ease of use and usefulness of the system. Questions 2, 7, 8, and 9 were evaluated positively, which indicated that the sys-
tem is a useful method of studying pronunciation. However, question 1 did not get a good evaluation item. Subjects answered “It is difficult to notice the difference between correct and incorrect animations” and “Although the new system highlighted the wrong articulation organs with red circles, I sometimes lost sight of the circles when they were displayed at the same time”. When the movements and places of some articulatory organs were both wrong, these displays became complicated. Therefore, if learners mistake some articulation movements, we need to improve the user interface so that the red circles are not indicated together, but are displayed step by step. Additionally, the system should indicate mispronunciations with circles as well as text to give simple advice on pronunciation. For example, it needs to display messages like “The tip of the tongue does not touch the back of the front teeth when pronouncing the letter L” in text with the tip of the tongue surrounded by a red circle. Moreover, questions 5 and 6 were also poorly evaluated, perhaps due to the lack of smoothness in connecting some animations between phonemes. We intend to construct animations that use a physical model and MRI to generate more natural animations in the future.

6. Conclusion

We developed a system of dynamically generating CG animations to express pronunciation from speech based on AFs, and we conducted experiments that confirmed the effectiveness of the animations of pronunciations. Learners could understand mispronunciations and correct pronunciations through specific animations of pronunciations. We intend to improve the system to make the motions in animations more natural, and study more effective ways of teaching pronunciation in the future.

Acknowledgements

This research was supported by a Grant-in-Aid for Young Scientists (B) (Subject No. 21700812) and the Kayamori Foundation of Information Science Advancement.

References

(1) Delmonte, R.: “SLIM prosodic automatic tools for self-learning instruction”, Speech Communication, Vol. 30, No. 2-3, pp.145–166 (2000).
(2) Gamper, J. and Knapp, J.: “A review of intelligent CALL systems”, Computer Assisted Language Learning, Vol.15, No. 4, pp.329–342 (2002).
(3) Neumeyer, L., Franco, H., Digalakis, V. et al.: “Automatic scoring of pronunciation quality”, Speech Communication, Vol. 30, No. 2–3, 83–93 (2000).
(4) Witt, S.M. and Young, S.J.: “Phone-level pronunciation scoring and assessment for interactive language learning”, Speech Communication, Vol. 30, No. 2–3, pp. 95–108 (1995).
(5) Deroo, O., Ris, C., Gielen, S. et al.: “Automatic detection of mispronounced phonemes for language learning tools”, Proc. of ICSLP-2000, Vol. 1, pp. 681–684 (2000).
(6) AmiVoice CALL Lite -pronunciation- http://www.advanced-media.co.jp/products/amivoicecalllite.html
(7) ATR CALL BRIX: http://www.atr-ht.jp/products/brix/index.html
(8) Jo, C.-H., Kawahara, T., Doshita, S. et al.: “Japanese Pronunciation Instruction System Using Speech Recognition Methods”, IEICE Trans. Comm. & Inf. Syst., Vol. E83-D, No 11, pp. 1960–1968 (2000).
(9) Sonic Print: http://www.ardicad.co.jp/sp/index.html
(10) Wik, P. and Escrivanb, D. L.: “Say ‘Aaaaa’ interactive vowel practice for second language learning”, Proc. SLaTE (2009).
(11) Wong, K. H., Lo, W. K. and Meng, H.: “Allophonic variations in visual speech synthesis for corrective feedback inCAPT”, Proc. ICASSP 2011, pp. 5708–5711 (2011).
(12) Wang, S., Higgins, M. and Shima, Y.: “Training English pronunciation for Japanese learners of English online”, The JALT Call Journal, Vol. 1, No. 1, pp. 39–47 (2005).
(13) Phonetics Flash Animation Project: http://www.uiowa.edu/~acadtech/phonetics/
(14) Badin, P., Youssef, A. B., Bailly, G. et al.: “Visual articulatory feedback for phonetic correction in second language learning”, Proc. SLaTE (2010).
(15) Huda, M. N., Kawashima H. and Nitta T.: “Distinctive phonetic feature (DPF) extraction based on MLNs and inhibition/enhancement by network”, IEICE Transactions on Information and Systems, Vol. 92, No. 4, pp. 671–680 (2009).
(16) TIMIT Acoustic-Phonetic Continuous Speech Corpus: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1
(17) Hiki, S.: “Speech Information Processing”, University of Tokyo Press, Tokyo (1973) (in Japanese).
(18) Fukuda, T. and Nitta, T.: “Noise-robust ASR by using dis-
tinctive phonetic features approximated with logarithmic normal distribution of HMM”, Proc. of Eurospeech 2003, Vol. 3, pp. 2185–2188 (2003).

(19) Hiki, S.: “A panphonic version of “The North Wind and the Sun” for the illustration of the IPA of Japanese consonants”, J. of Acoustical Society of Japan, Vol. 66, No. 10, pp. 485–486 (2010) (in Japanese).

(20) Takebayashi, S. and Saito, H.: “Eigo: English Phonetics”, Taishukan Shoten, Tokyo (2008) (in Japanese).

(21) Fujimura, Y.: “Speech Science”, University of Tokyo Press, Tokyo (1972) (in Japanese).

(22) Morris, H.: “On distinctive features and their articulatory implementation”, Natural Language & Linguistic Theory, Vol. 1, No. 1, pp. 91–105 (1983).

(23) Huda, M. N., Katsurada, K. and Nitta, T.: “Phoneme recognition based on hybrid neural networks with inhibition/enhancement of Distinctive Phonetic Feature (DPF) trajectories”, Proc. Interspeech’08, pp. 1529–1532 (2008).

(24) Nagaoka, H., Iribe, Y., Katsurada, K. et al.: “Articulatory feature extraction and its evaluation for English pronunciation learning”, Information Processing Society of Japan, SLP80, No. 13 (2009).

**Yurie Iribe** received her M.S. and Ph.D. degrees from the Graduate School of Human Informatics of Nagoya University, Japan. She is currently an Assistant Professor at the Information and Media Center, Toyohashi University of Technology, Japan. Her recent research interests include education support and speech recognition. She is also a member of JSiSE, IEICE, IPSJ, JSAI, and ASJ.

**Takuro Mori** received his M.E degree from Toyohashi University of Technology in 2012. He is currently with KYB Corporation. His research interests include education systems for pronunciation training.

**Kouichi Katsurada** received his Ph.D. degree from Osaka University in 2000. He has been with Toyohashi University of Technology since 2000. He is currently an Associate Professor at the Center for International Relations and Department of Computer Science and Engineering, Toyohashi University of Technology. His research interests include multimodal interaction, facial image processing, and spoken term detection. He is a member of IEEE, ISCA, IEICE, IPSJ, JSAI, and ASJ.

**Tsuneo Nitta** received his Ph.D. from Tohoku University, Japan. He worked at the R&D Center and Multimedia Eng. Lab. of Toshiba Corp. from 1970 to 1998. He subsequently joined Toyohashi University of Technology as a Professor in the Graduate School of Eng. In 2012, he became a TUT Emeritus Professor and a Visiting Professor at both TUT and Waseda University. His current research interests include speech recognition, speech synthesis, and multi-modal interaction. He is an IPSJ Fellow and a member of IEICE, ASJ, JSAI, and IEEE.