ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Real-time Visualization of English Pronunciation on an IPA Chart Based on Articulatory Feature Extraction

Yurie Iribe[1], Takurou Mori[1], Kouichi Katsurada[1], Goh Kawai[2] and Tsuneo Nitta[1]

[1] Graduate School of Engineering, Toyohashi University of Technology, Aichi, JAPAN
[2] Research Faculty of Media and Communication, Hokkaido University, Hokkaido, JAPAN

iribe@imc.tut.ac.jp, mori@vox.cs.tut.ac.jp, katsurada@cs.tut.ac.jp, goh@kawai.com, nitta@cs.tut.ac.jp

10.21437/Interspeech.2012-253

Abstract

In recent years, Computer Assisted Pronunciation Technology (CAPT) systems have been developed that can help Japanese learners to study foreign languages. We have been developing a pronunciation training system to evaluate and correct learner's pronunciation by extracting articulatory-features (AFs). In this paper, we propose a novel pronunciation training system that can plot the place and manner of articulation of learner's pronunciation on an International Phonetic Alphabet (IPA) chart in real time. First, the proposed system converts input speech into AF-sequences by using multi-layer neural networks (MLNs). Then, the AF-sequences are converted into x-y coordinates and plotted on an IPA chart to show his/her articulation in real time. Lastly, we investigate plotting accuracies on the IPA chart through experimental evaluation.

Index Terms: pronunciation training, articulatory feature, IPA chart

1. Introduction

Computer Assisted Pronunciation Technology (CAPT) systems have been developed in recent years that can support learners to study foreign language pronunciation [1][2]. The purpose of such systems is to enable learners to have the correct-articulation training that a good instructor gives. One of these systems can estimate the pronunciations of a teacher and a learner and display them on an F1-F2 plane [3][4]. Here, Fi is the i-th resonance frequency of a vocal tract. However, it is difficult for the learners to understand how to correct his/her articulation from the F1-F2 plane without the knowledge of phonetic science, because the learners usually do not know the relationship between F1-F2 and articulation gestures. Additionally, although some systems display speech waves to show some differences between pronunciation of a teacher and a learner, most learners cannot understand the essences of these differences. The system should accurately show the learners how to correct wrong articulation, in the same way that teachers do.

We propose a novel English pronunciation training system that has functions by which learners can intuitively understand the difference in articulation between the teacher and the learner and quickly correct his/her articulation. The system can plot the place and manner of articulation of his/her speech on an International Phonetic Alphabet (IPA) chart [5] in real time. In the case of vowel learning, the learner's pronunciation is plotted on a trapezoid map of the IPA chart in which the vertical axis shows the vowel height "open-close" and the horizontal axis expresses the vowel backness "front-back" in accordance with the tongue position. The vowel height is named for the vertical position of the tongue relative to either the roof of the mouth or the aperture of the jaw. We call the charts "pronunciation maps". Figure 1 represents the vowel pronunciation map. Each learner can asymptotically approach to correct his/her articulation by checking the X- and Y-axes on the map that indicate the place and manner of articulation.

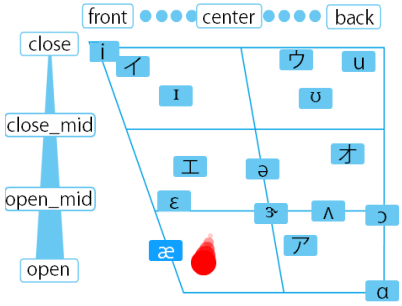


Figure 1: *Window of vowel pronunciation maps.*

2. Pronunciation Training System "Pronunciation Map"

Figure 2 outlines the proposed system. It extracts the articulatory features (AFs) from a learner's speech to acquire his/her articulation and plots the place and manner of articulation on the pronunciation maps by translating the extracted AFs into x-y coordinates in real time (each 10 ms). For the plotted time of each phoneme, the proposed system applies the phoneme boundary extracted from HMM through AF sequences. The next section describes AF extraction.

2.1. Articulatory feature extraction

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, AFs 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, AFs are expressed by assigning +/- as the feature of each articulation in a phoneme. For example, the AF sequence of "/jiNkoese/ (space satellite)" in Japanese is shown in Figure 3.

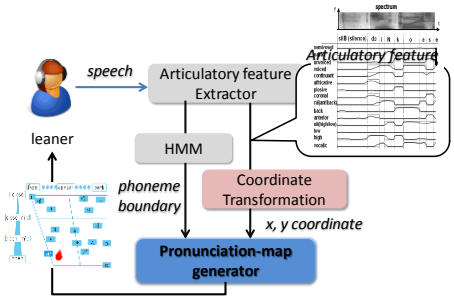


Figure 2: *Outline of pronunciation maps.*

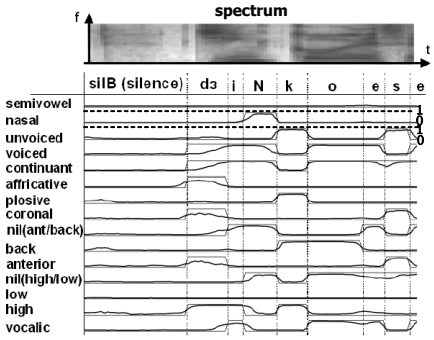


Figure 3: *Articulatory feature sequence: /jiNkoese (artificial satellite)/.*

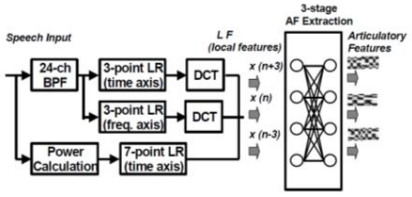


Figure 4: *Articulatory feature extractor*.

Because the phoneme N is a voiced sound, "voiced" in Figure 3 is given [+]. (In fact, [+] is given a value of "1" as the teacher signal.) Because phoneme k is a voiceless sound, "voiced" in Figure 3 is given [-]. Actually, [-] is given a value of "0" as the teacher signal and "unvoiced" in Figure 3 is given [+]. We generated an AF table of 28 dimensions corresponding to 62 English and Japanese phonemes. We defined the AFs on the basis of distinctive phonetic features (DPF) involved in Japanese phonemes in international phonetic symbols (International Phonetic Alphabet; IPA). We also used our previously developed AF extraction technology [6]. Figure 4 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 band pass filters (BPFs) output with mel-scaled center frequencies. At the acoustic feature extraction stage, the BPF outputs are first converted into 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 Mel-frequency cepstral coefficient (MFCC) as the input to multi-layer neural networks (MLNs) for

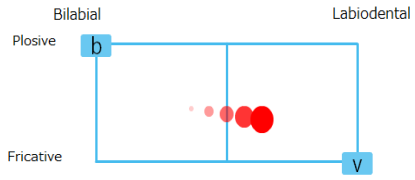


Figure 5: *Window of consonant pronunciation map*.

the extraction of AFs [6]. 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.

2.2. Pronunciation maps based on IPA Chart

The vowel pronunciation map (in Figure 1) and consonant pronunciation map (in Figure 5) indicate a learner's articulation with a red dot on IPA chart showed phonetic symbol in real time (every 10 ms). These maps are displayed independently. The learner can check his/her articulation while pronouncing in real time. Additionally, because the system will plot the red dot around the target phonemic symbol if the learner pronounces correctly, the learner can correct his/her articulation movement by checking the position and distance among the red dot, x-axis of "front-back", and y-axis of "open- close". The map converts AF of multi- dimensions to 2 dimensions so that the learners are able to check them easily, t. The proposed system extracts AF of 28×3 dimensions each 10 ms from a learner's speech. The vowel pronunciation map removes AF of 10 dimensions (front, center, back, open, open-mid, close, close-mid, tense, round, rhoticity) corresponding to vowels in extracted AF, and translates them into x-y coordinates by coordinate transformation. The consonant pronunciation map translates AF of 14 dimensions corresponding to consonants into x-y coordinates. The AFs of consonants include bilabial, labiodentals, dental, alveolar, post alveolar, palatal, velar, glottal, plosive, nasal, tap or flap, fricative, approximant, lateral approximant. The values of each AF are averaged in each phoneme boundary calculated from HMM. The system plots the averaged values on the pronunciation maps.

2.2.1. Vowel pronunciation-map

The window of the vowel oipronunciation map is displayed in Figure 1. It plots learner's place of articulation on the basis of the scale of the vowel backness (x-axis) and the vowel height (y-axis). Additionally, the trajectory of learner's articulation is expressed with some pale red dots (in Figures 1 and 5).
Next, we explain coordinate transformation through AF in details. The transformations into x-coordinate and y-coordinate are calculated by (1) and (2), respectively.

$$x = D_{width} \times (\alpha + \Delta_x)/4 \quad (1)$$

D_{width} is horizontal length of pronunciation map, α is the value given by dividing horizontal axis length by number of the vowel backness (front, central, back), Δ_x is the value given by

subtracting second largest value of AF from maximum value of AF in the vowel backness. The range of AF is 0 to 1. Lastly, x is normalized by dividing by 4, which is the number partitioned by front, front-center (center of front and center), center, center-back (center of center and back), and back. For example, if the maximum value is 1.0 (front in this example) and the value of the next AF (center in this example) is 0.7, the red dot is plotted in "somewhat more forward than center of front tongue and center tongue" as Figure 1. That is, by taking into account not only maximum AF but also next AF, the system can express the learner's detailed the place of articulation on the map.

$$y = D_{height} \times (\beta + \Delta_y) / 6 \quad (2)$$

D_{height} is vertical length of the pronunciation maps, β is the value given by dividing the length of the vertical axis by the number of the vowel height (central, central-mid, open, open-mid), and Δ_y is the value given by subtracting the value of the next AF from the maximum value of AF in the vowel height. The vowel pronunciation map translates x-y coordinates in accordance with the shape of a trapezoid because it is a trapezoid.

2.2.2. Consonant Pronunciation map

The window of the consonant pronunciation map is displayed in Figure 5. The consonant pronunciation map plots learner's articulation with the red dot. The phonemes on the map are pairs of the phonemes with which Japanese speakers easily make substitution errors. For example, they are [b] and [v], [ɹ] and [l] and [ɾ], etc. Articulatory features on horizontal axis and vertical axis indicate AFs that are different among phonemes shown on the map. For instance, different AFs between 'b' and 'v' such as Figure 5 are plosive and fricative for articulatory manner and bilabial and labiodental for the place of articulation. Moreover, if different AFs such as [ɹ], [l], [ɾ] are only articulatory manner, only horizontal axis is shown on the map. The map emphasizes his/her wrong articulation by clearly expressing the differences in articulation among these similar phonemes.

The transformation from AF to x-coordinate and y-coordinate applies formula (3) and (4) if the number of phoneme pairs is 2, formula (5) and (6) if the number of phoneme pairs is 3, respectively.

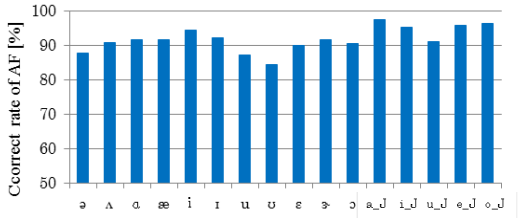


Figure 6: *Correct rates of articulatory features for vowel [%].*

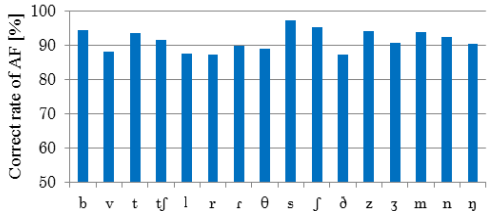


Figure 7: *Correct rates of articulatory features for some consonant [%].*

$$x = D_{width} \times (1 + AF_{x1} - AF_{x0}) / 2 \quad (3)$$

$$y = D_{height} \times (1 + AF_{y1} - AF_{y0}) / 2 \quad (4)$$

$$x = D_{width} \times (\alpha + \Delta_x) / 4 \quad (5)$$

$$y = D_{height} \times (\beta + \Delta_y) / 4 \quad (6)$$

AF_{x0} and AF_{x1} indicate the value of AF expressed on the horizontal axis. AF_{y0} and AF_{y1} indicate the value of AF expressed on the vertical axis. D_{width}, D_{height}, Δ_x, and Δ_y are calculated with the same method with vowel pronunciation map. The leaners can notice the detailed errors of articulation such as "my tongue's position is correct, but fricative cannot be produced" by reading the position of red dot on horizontal axis and vertical axis, because the AF extractor can recognize the manner and the place of articulation separately.

3. Evaluation

We evaluated the correct rate of articulation feature and plotting accuracy on the map to confirm accuracy of the pronunciation maps.

3.1. Data set

We used Japanese speech data for AF training in this experiment because English speech by Japanese speakers may include phonemes unique to Japanese. Moreover the system trained TIMIT and CSJ to recognize accurately English phonemes and the unique Japanese phonemes from English speech by Japanese speakers.

D1-1: *Training data set for AF extractor training: 2600 TIMIT [7] sentences of English speech (325 male English native speakers)*

D1-2: *Testing data set for adaptation of AF and plot to the maps: 896 TIMIT sentences of English speech (112 male English native speakers)*

D2-1: *Training data set for AF extractor training: CSJ 22 hours of Japanese speech (92 male Japanese native speakers)*

D2-2: *Testing data set for adaptation of AF and plot to the maps: CSJ 2.5 hours of Japanese speech (10 male Japanese native speakers)*

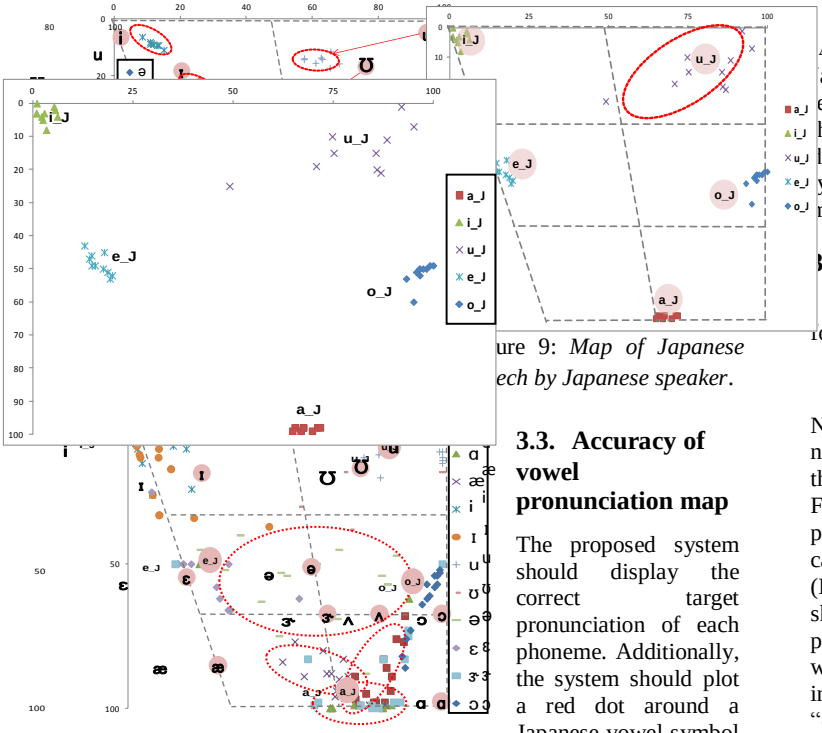
D3: *Testing data set for adaptation of plot to the maps: 15 "The North Wind and The Sun [8]" sentences of English speech (15 male Japanese speakers)*

3.2. Correct rates of articulatory features

First, we computed the correct rate of the articulation feature (AFCR) that has a high possibility of affecting plotting accuracy.

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

N_c is the total number of recognized AFs correctly, and N is total frame number. The number 28 indicates dimensions of AF. We regard native speaker's speech as correct pronounced speech. Figures 6 and 7 show the correct rates of vowel AFs and consonant AFs of each phoneme, respectively. The correct rates of AFs to every English phoneme (including phonemes shared with Japanese) averages 93%, and that of unique Japanese phonemes ([a_J], [i_J], [u_J], [e_J], [o_J]) averages about 96%. The results clearly show that the proposed system can extract AFs accurately.



ure 9: *Map of Japanese ech by Japanese speaker*.

3.3. Accuracy of vowel pronunciation map

The proposed system should display the correct target pronunciation of each phoneme. Additionally, the system should plot a red dot around a Japanese vowel symbol if a Japanese speaker pronounces Japanese vowels in order to understand the differences between English and Japanese vowels. Therefore, we compared correct coordinate (each phoneme's coordinates are shown on the IPA chart) with coordinates translated with English native speaker's speech (D1-2) through an AF extractor trained by D1-1. Figure 8 shows the results. The circle in Figure 8 indicates the correct phoneme position. As for D1-2, it plots 8 dots according to each speaker's hometown (8 areas). Similarly, we also evaluated Japanese speech with D2-1 and D2-2. Figure 9 shows the results. The phonemes [i], [ɪ], [i_J], for which the AFs correct rates are comparatively high in Figure 7, are plotted in coordinates about the same as correct coordinates. However, the phonemes [u], [ʊ], [u_J], for which AFs correct rates are comparatively low in Figure 7, are plotted in coordinates separate from correct coordinates. We consider that the accuracy of pronunciation map correlate strongly with correct rates of AFs from the results. Therefore, AF extractor needs to be improved. Next, we evaluated the plotted accuracy for English speech by a Japanese speaker (D3) that may include Japanese vowels. In this experiment, the AF extractor is trained by D1-1 and D2-1. The results are shown in Figure 10. The phonemes [æ], [ʌ], [ɑ] are plotted around Japanese vowel [a_J]. This means that Japanese speakers pronounce by replacing [æ], [ʌ], [ɑ], which are not included in Japanese phonemes, with the apanese vowel [a_J], because our system achieves a certain evel of accuracy from results of Figures 8 and 9. Additionally, he phoneme [ə], which Japanese are not good at pronouncing, liverges extensively. The proposed system can display some ypical English error pronunciation by a Japanese speaker on the nap.

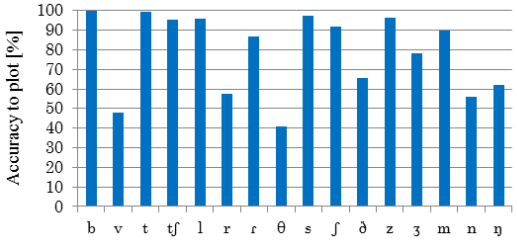


Figure 11: *Accuracy to plot consonant speech*.

3.4. Accuracy of consonant pronunciation map

Accuracy of consonant pronunciation map is calculated by the ollowing formula.

$$\text{Accuracy} = \frac{N_p}{n} \times 100\,[\%]$$

N_p is the number plotted in the correct area of a phoneme p, and n is the phoneme's number included in test data. For instance, the correct area of phoneme /v/ shows right side of the map in Figure 5. The proposed system must plot each correct phoneme's position for a English native speaker's speech. Therefore, we calculated the accuracy by an English native speaker's speech (D1-2) through an AF extractor trained by D1-1. Figure 11 shows the results. Accuracy over 50% was achieved for every phoneme except [v] and [θ]. On the other hand, low accuracy was achieved for the AF correct rates of [θ], [ð], and [v], just as in Figure 7. Although AFs of [θ], [ð] and [v] are "fricative", "dental" and "labiodentals", accuracies of [s] and [ʃ] are very high for same "fricative". Therefore, the AF extractor must be improved so that is can recognize "dental" and "labiodental" correctly.

4. Conclusions

We developed a pronunciation maps to plot his/her articulation on an IPA chart in real-time. We conducted some experiments to confirm the accuracy of the pronunciation maps for vowels and consonants. The results confirmed that the average accuracy is over 70%. In the future, we will improve the system to extract "back", "dental", and "labiodentals" accurately. Moreover, we intend to develop a function to indicate AFs of "round" and "tense", which are very important for vowels. Lastly we will conduct some experiments for educational effect so that our system has already implemented in CAPT system.

5. References

[1] Maxine Eskenaze., "An overview of spoken language technology for education", Speech Communication., vol.51: 832-844, 2009.

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

[3] Chul-Ho Jo, Tatsuya Kawahara, Shuji Doshita, Masataka Dantsuji., "Japanese Pronunciation Instruction System Using Speech Recognition Methods", IEICE TRANS. INF. & SYST., VOL.E83-D, No.11: 1960-1968, 2000.

[4] Sonic Print - http://www.arcadia.co.jp/SP/index.html

[5] IPA vowel chart: http://www.langsci.ucl.ac.uk/ipa/

[6] 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, 1529–1532, 2008.

[7] Garofolo , J.S. et al.: TIMIT Acoustic Phonetic Continuous Speech Corpus, Linguistic Data Consortium,1993.

[8] Kondo, Y. The development of automatic speech evaluationsystem for learners of English. Unpublished doctoral dissertation,Waseda University, Tokyo, 2010.