S. L. Wong et al. (Eds.) (2010). Proceedings of the 18th International Conference on Computers in Education. Putrajaya, Malaysia: Asia-Pacific Society for Computers in Education.

Pronunciation Instruction using CG Animation based on Articulatory Features

Yurie IRIBE[a]*, Takuro MORI[b], Kouichi KATSURADA[b] & Tsuneo NITTA[b]
[a]*Information and Media Center, Toyohashi University of Technology, Japan*
[b]*Graduate School of Engineering, Toyohashi University of Technology, Japan*
*iribe@imc.tut.ac.jp

Abstract: We describe a pronunciation instruction system that dynamically generates CG animations to express the pronunciation from speech based on articulatory features. Specifically, the system displays the results of phoneme recognition and animation of the pronunciation movements of both the learner and teacher from their speech in order to show in what way the learner's pronunciation is wrong. Learners can thus understand their wrong pronunciation and the correct pronunciation method through specific pronunciation animations. Experiments confirmed the effectiveness of the animations.

Keywords: Pronunciation instruction, animation, articulatory feature, speech recognition

Introduction

In pronunciation education, face-to-face lectures are ideal. The teacher teaches accurate pronunciations to students by explaining the movement and relative location of the tongue and palate when making utterances. The teacher also points out each learner's pronunciation mistakes and how to correct them. However, it takes time to acquire correct pronunciation and learners typically do not have enough time to attend face-to-face lectures. In Japan in particular, there is not enough time to study pronunciation at school. Therefore, computer-based pronunciation training systems are very useful as they allow learners to study pronunciation at any time.

With this background, Computer Assisted Language Learning (CALL) systems have been introduced for English language education in recent years [1][2]. CALL systems typically analyze a learner's speech by using speech recognition technology, and point out pronunciation problems with specific phonemes in words and automatically score the pronunciation quality [3][4][5]. However, although the learner can thus realize that his/her speech is different from the teacher's, the learner cannot understand how to correctly move the appropriate articulation organ. The system should show how to do this when the learner makes a wrong pronunciation, in the same way that teachers do. The proposed system provides several interactive methods for learners, not only pointing out their individual pronunciation problems, but also visually representing the teacher's and learner's articulatory movements (movement of the tongue, palate, and lips) by using CG animation. The system provides guidance on the learner's wrong pronunciation movement as well as the correct pronunciation movement, appropriately and in real-time. As a result, the learner can study how to move an articulatory organ while visually comparing their mispronunciation animation and the correct pronunciation animation. Such animations are easy to understand. Although other studies have examined making correct pronunciation animation and video in advance [6][7], they do not dynamically produce animations of the learner's wrong pronunciation. To represent the teacher's and learner's articulatory movements, the proposed system directly extracts articulatory features (place of articulation and manner of articulation) from their speeches.

1. Outline of Pronunciation Training System

Figure 1 shows an outline of the system. The system consists mainly of articulatory extraction, phoneme recognition, and pronunciation instruction functions. As for articulatory extraction and phoneme recognition, we use existing technologies we have already developed. The CG animation is generated by using the phoneme sequence, phoneme boundary and articulatory features extracted using these technologies.

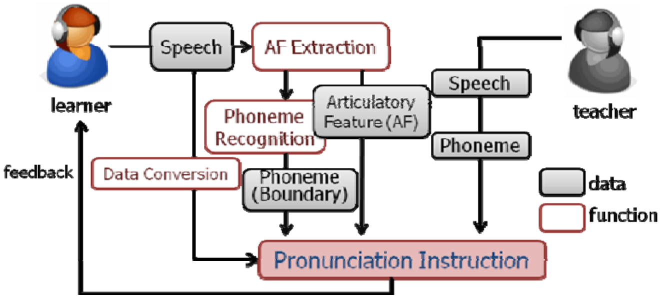


Fig. 1: System outline

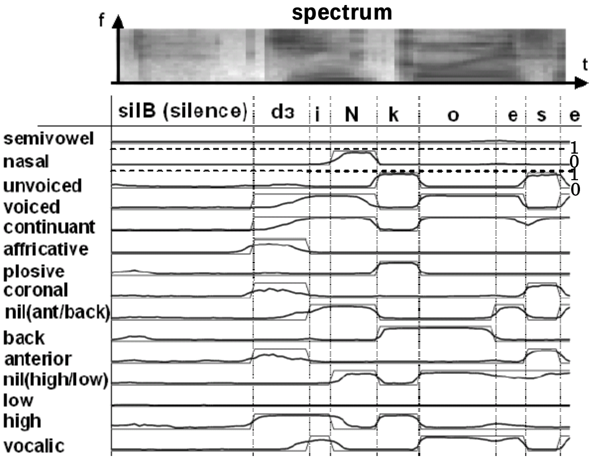


Fig. 2: Articulatory feature sequence: /jiNkoese (artificial satellite)/

2. Articulatory Feature Extraction

2.1 Articulatory Features

In order to vocalize, human beings change the shape of the vocal tract and move articulatory organs such as the lips, teeth, alveolar arch, palate, tongue and pharynges. 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 (AF). In short, articulatory features mean information (for instance, closing the lips to pronounce "m") on the movement of the articulatory organ that contributes to the articulatory movement. In this paper, articulatory features are expressed by assigning +/- as the feature of each articulation in a phoneme. For example, the articulatory feature sequence of "/jiNkoese (Space satellite)" in Japanese is shown in Figure 2. Because phoneme N is a voiced sound, "voiced" in Figure 2 is given [+]. (Actually, [+] is given a value of "1" (right side of Figure 2).) Because phoneme k is a voiceless sound, “voiced” in Figure 2 is given [-]. Actually, [-] is given a value of “0” (right side of Figure 2)) and “unvoiced” in Figure 2 is given [+]. We generated an articulatory feature table of 28 dimensions corresponding to 42 English phonemes. We defined the articulatory features based on distinctive phonetic features (DPF) involved in English phonemes in international phonetic symbols (International Phonetic Alphabet; IPA)[8].

2.2 Articulatory Feature Extraction

We also used our previously developed articulatory feature (AF) extraction technology [9]. The extraction accuracy is about 95 %. Figure 3 shows the AF extractor. An input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segment is applied every 10 ms. The resultant FFT power spectrum is then integrated into a 24-ch BPFs output with mel-scaled center frequencies. At the acoustic feature extraction stage, the BPF outputs are first converted to local features (LFs) by applying three-point linear regression (LR) along the time and frequency axes. LFs represent variation in a spectrum pattern along two axes. After compressing these two LFs with 24 dimensions into LFs with 12 dimensions using a discrete cosine transform (DCT), a 25-dimensional (12 Δt, 12 Δf, and ΔP, where P stands for the log power of a raw speech signal) feature vector called LF is extracted. Our previous work shows that LF is superior to MFCC as the input to MLNs for the extraction of AFs, or distinctive phonetic features (DPFs). LFs are then entered into 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 using the Gram-Schmidt (GS) orthogonalization procedure before connecting with the HMM-based classifier.

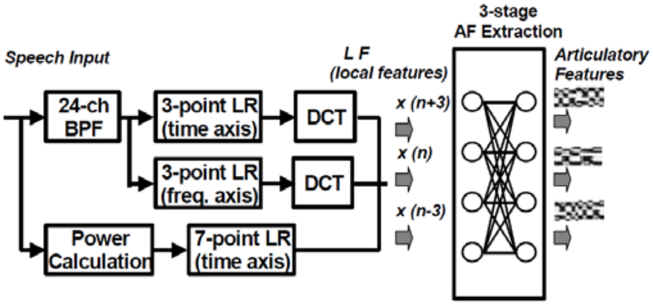


Fig. 3 Articulatory feature extraction

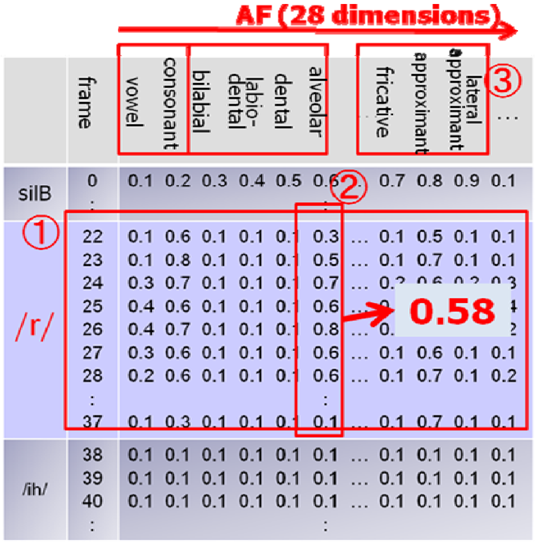
2.3 Phoneme Recognition

For recognizing phonemes, we utilize the phoneme recognition technology using AFs that we have developed [10]. The phoneme boundary and phoneme sequence are obtained by inputting AFs extracted as described in Section 2.2 to the Hidden Markov Model (HMM). The score of each phoneme is calculated by comparing the correct AF with the extracted AF based on the phoneme boundary. Additionally, correct phonemes and wrong phonemes are decided by comparing each AF. Finally, the results of evaluating the learner’s speech such as the score and whether each phoneme and phoneme sequence are correct or wrong are fed back to the learner.

3. Pronunciation Instruction based on CG Animation

A CG animation of the inside of the mouth of the teacher and learner is generated based on the AFs, phoneme sequence, and phoneme boundary obtained from their speech. Because the CG animation is generated immediately as their speech is input, the learner can confirm the motion of pronunciation in real time.

3.1 Articulatory Feature Analysis


AF (28 dimensions)
①
②
③
0.58

	frame	vowel	consonant	bilabial	labio-dental	dental	alveolar		fricative	approximant	lateral approximant	…
silB	0	0.1	0.2	0.3	0.4	0.5	0.6		0.7	0.8	0.9	0.1
	:						:					
/r/	22	0.1	0.6	0.1	0.1	0.1	0.3	…	0.1	0.5	0.1	0.1
	23	0.1	0.8	0.1	0.1	0.1	0.5	…	0.1	0.7	0.1	0.1
	24	0.3	0.7	0.1	0.1	0.1	0.7	…	[illegible]	[illegible]	[illegible]	0.3
	25	0.4	0.6	0.1	0.1	0.1	0.6	…	[illegible]	[illegible]	[illegible]	[illegible]
	26	0.4	0.7	0.1	0.1	0.1	0.8	…	[illegible]	[illegible]	[illegible]	[illegible]
	27	0.3	0.6	0.1	0.1	0.1	0.6	…	0.1	0.6	0.1	0.1
	28	0.2	0.6	0.1	0.1	0.1	0.6	…	0.1	0.7	0.1	0.2
	:						:					
	37	0.1	0.3	0.1	0.1	0.1	0.1	…	0.1	0.7	0.1	0.1
/ih/	38	0.1	0.1	0.1	0.1	0.1	0.1	…	0.1	0.1	0.1	0.1
	39	0.1	0.1	0.1	0.1	0.1	0.1	…	0.1	0.1	0.1	0.1
	40	0.1	0.1	0.1	0.1	0.1	0.1	…	0.1	0.1	0.1	0.1
	:						:					

Fig. 4: Example of AF table ("read")

The phoneme sequences, phoneme boundaries and value [0 or 1] of each articulatory organ (AF) are obtained as described in Section 2.2. Figure 4 shows part of the results of AFs for the word "read". The system recognizes the phoneme /r/ from frame 22 to frame 37 of the speech. One frame corresponds to 10 ms. The phoneme boundaries (frame to frame of each phoneme) are used for the reproduction timing of the animation. The AF sequence in this paper is defined based on the AF values of 28 dimensions (place of articulation and manner of articulation) in each frame (Figure 4①). In this section, we describe how to decide the articulation motion to generate the animation based on AFs. The order is as follows:

1. The AFs of each phoneme are extracted. In the example of Figure 4, the AFs from frame 22 to frame 37, which is the phoneme boundary of /r/, are extracted (Figure 4①).
2. The mean value of each AF is calculated in the section. In the example, the mean value of "alveolar" in the section of /r/ is 0.58 (Figure 4②).
3. AFs are classified according to the point of articulation or manner of articulation based on the articulatory phonetics shown in Table 1. For instance, nasal, fricative, approximant**,** lateral approximant, etc. are classified as manners of articulation (consonant) (Figure 4③). Next, one effective AF is chosen from each category, which is the AF whose mean value becomes the maximum in the category. That is, the articulatory organ that moved when the learner pronounces is specified. Because the mean value of "alveolar" is the maximum in the "point of articulation (consonant))" category in the /r/ section, the alveolar registers as one AF of /r/ in Figure 4. That is, it is judged that the speech was an alveolar movement (the tip of the tongue pressed against the alveolar ridge behind the teeth) when the learner pronounced /r/.

Table 1: Examples of AF categories

Category of AF	AF
vowel / consonant	vowel, consonant
voice band	voiced, unvoiced
manner of articulation (consonant)	plosive, nasal, fricative, flap, approximant**,** lateral approximant
point of articulation (consonant)	labial, labiodental, dental, alveolar, postalveolar, glottal, palatal
lip (vowel)	raised, half-closed, half-open, lowered
place of tongue (vowel)	front vowel, back vowel, central vowel
other (vowel)	round, tense, R-colored vowel

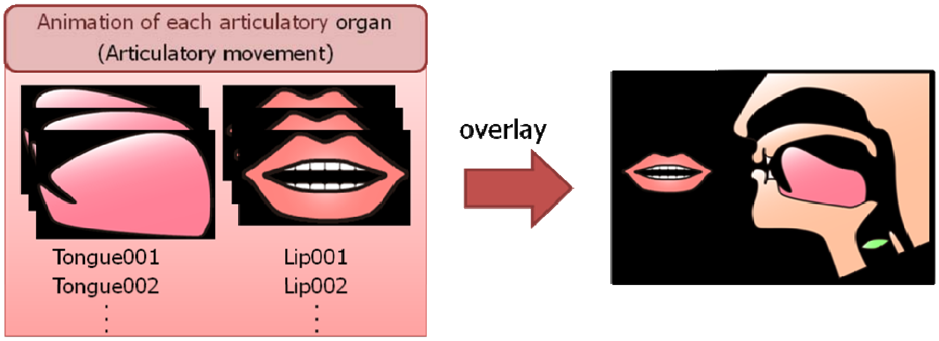


Fig. 5: Animation generation for each phoneme

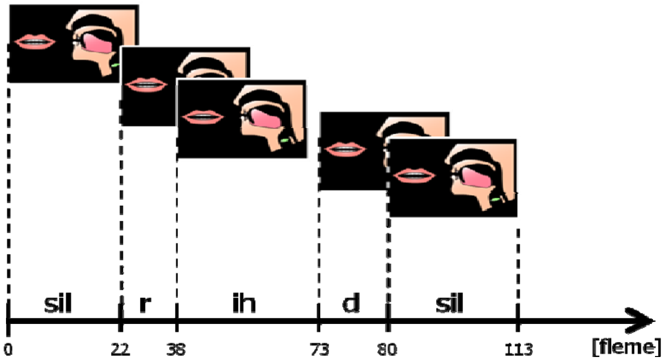


Fig. 6: Animation joint between each phoneme

3.2 Animation Generation

It is necessary to present the movement of the mouth cavity viewed from the side and the lips viewed from the front in order to clearly show the pronunciation method to the learner. It is also necessary to express the movement of each articulatory organ accurately. We generate beforehand some animations independently for each articulatory organ, and then overlap these animations. The order of animation generation is as follows:

1. The movement pattern of each articulatory organ is made with Adobe Illustrator CS3 beforehand. Various animations of movements such as alveolar (fricative), alveolar (plosive), postalveolar (approximant) of the tongue (consonant), closed, open, narrowness of lips, bilabial, etc. are made. The animation movements are made based on phonology.
2. We use the “shape tween” ActionScript for the animated motion. The shape tween is a function which changes the shape of an object gradually whenever one frame advances 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 of the mouth cavity and lips is generated by overlapping respective animations corresponding to the AF of a phoneme decided in Section 3.1 (Figure 5). The animations are generated for all phonemes using the same method.
4. Next, the animations of respective phonemes are connected. The shape of an articulatory organ would change too rapidly if the previous phoneme’s animation that has finished changing is simply connected to the default animation of the following phoneme. The system therefore connects the vector (feature point) of the next phoneme’s animation and the vector (feature point) of the previous phoneme’s animation five frames earlier, respectively. This method makes the connection between phoneme animations more natural.

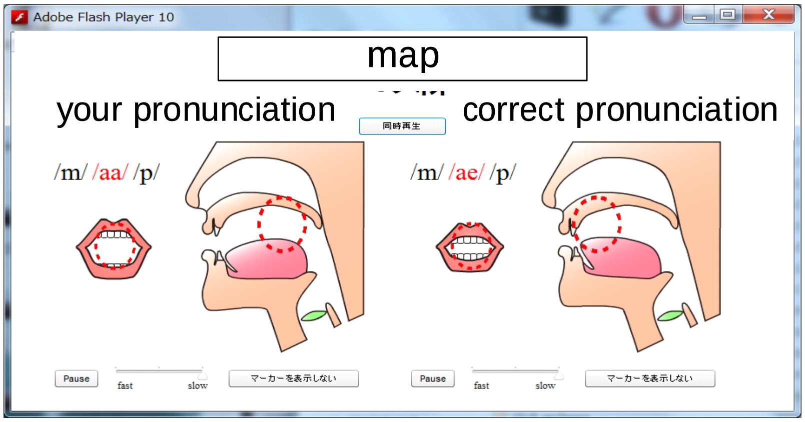


Fig. 7: Example of pronunciation instruction

3.3 *Pronunciation Instruction Functions*

To show the learner's mispronunciation movement and the correct pronunciation movement appropriately and specifically, the following functions are included.

- When the play button on the screen is clicked, the animation of the mouth cavity and lips is reproduced with the learner's voice. When the play button on the correct pronunciation side is clicked, the animation is reproduced with the voice of the teacher who was recorded beforehand. When the simultaneous button is clicked, both the learner's animation and the correct animation are reproduced at the same time. The learner can thus visually compare their own pronunciation with the correct pronunciation.
- To teach how and where the learner should make corrections, the system highlights the wrong articulation organs with a red circle by comparing the AFs of the learner and the teacher if the learner's pronunciation is wrong (red circle in Figure 7). As a result, the learner can see which articulatory movement is wrong, and how the articulatory organs should be moved to pronounce correctly. In Figure 7, because the movement of the lips and the tongue for /æ/ of "map" is different (the learner pronounces /aa/), a marker is displayed on the lips and tongue of both the learner and teacher.
- It is also important to confirm the correct pronunciation method in each phoneme. Therefore, the pronunciation animation of the phoneme can be played by clicking on the phonemic symbol (/m/,/æ/,/p/) on the screen.
- Slow-motion replay of the voice and animation at three speeds, ×1, ×0.5, and ×0.25, is possible, allowing the learner to see the pronunciation in slow-motion by adjusting the play speed. We built a speech rate conversion function for adjusting the speed of the voice.

4. Experimental Evaluation

4.1 *Experimental Setup*

To show the effectiveness of the proposed system, we compared it with an existing system by experiment. The existing system was a mounted speech recognition technology which shows correct/wrong phonemes by text. The subjects were 11 Japanese native speakers, who were in a beginner's class for English pronunciation. The order of the experiment was as follows.

1. Eight English words including a phoneme that Japanese people are not good at were prepared. The eight words were divided into two groups of four words each.
2. We recorded the subject's pronunciation voice before the experiment, and recorded the score evaluated by the system.
3. To avoid subject bias, the subjects were divided into group 1 and group 2.
4. To avoid word bias, the words were divided into word group A and word group B.
5. After group 1 studied using the existing system, they pronounced word group A. Meanwhile, after group 2 studied using the proposed system, they pronounced word group B. Next, each group changed systems and pronounced the other word group. We explained how to use the system before the experiment. The system calculated the score for all the voices.

   We also used TIMIT [11] as training data set for MLN and HMM.

Table 2: Change of pronunciation scores

	Group 1	**Group 2**
Existing system	Word group A 51.4 points → 55.8 points (↑ 8.40%)	Word group B 65.2 points →76.0 points (↑ 16.7%)
Proposed system	Word group B 62.5 points → 73.8 points **(↑ 17.9%)**	Word group A 50.1 points → 65.3 points **(↑ 30.5%)**

Word group A: "read", "bird", "good", "think", Word group B: "map", "sea", "sing", "bought"

Table 3: Evaluation of English pronunciation instruction system
(5: Very good; 4: Good; 3: Fair; 2: Poor; 1: Very poor)

Explanation	Average
Q1. Was the system easy to use?	2.7/5.0
Q2. Was the system interesting?	3.7/5.0
Q3. Did you realize your weakness?	3.2/5.0
Q4. Did you understand the difference between your pronunciation and the teacher's pronunciation?	3.5/5.0
Q5. Did you think that your pronunciation animation described your wrong pronunciation accurately?	2.5/5.0
Q6. Did you think that the correct pronunciation animation described the correct pronunciation accurately?	2.8/5.0
Q7. Was the slow-motion replay function useful?	3.7/5.0
Q8. Was the phoneme animation function useful?	3.7/5.0
Q9. Were the comparison display and highlighting of pronunciation movement useful?	3.8/5.0
Q10. Would you like to use the system again?	3.3/5.0

4.2 Experimental Results

Table 2 shows the results of the experiments. The improvement in pronunciation score with the proposed system was double that with the existing system. Especially, the score increased by about 30% when group 2 used the system, even though the score for word group A in both groups was low before the study. With the existing system, the learners could not understand how to correct mistakes in pronunciation even though they noticed wrong phonemes. On the other hand, with the proposed system which visually showed the correct pronunciation movement and the learner's pronunciation movement, the learners could clearly and efficiently understand how to correct their wrong pronunciations. However, the score was low even when using the system because the duration of the experiment was short. We therefore intend to conduct a long-term experiment with many words because too few words were used in this experiment. We obtained a five-stage evaluation as shown in Table 3, for questions focusing on ease of use and the usefulness of the system. Q2, Q7, Q8 and Q9 received good evaluations, showing that the system is a useful method of studying pronunciation. However, Q5 and Q6 received low evaluations, perhaps due to lack of smoothness of connecting the animation between phonemes. In the future, we will construct animations that use a physical model. Moreover, we plan to generate more natural animations by using magnetic resonance imaging (MRI) data for Japanese speakers and native speakers. Longer-term testing in actual lectures is also necessary.

Conclusions

We developed a system to dynamically generate CG animations to express pronunciation from speech based on articulatory features, and conducted experiments which confirmed the effectiveness of the pronunciation animations. Learners can understand their wrong pronunciations and the correct pronunciation method through specific pronunciation animations. We will improve the system to make the animation motions 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).

References

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

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

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

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

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

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

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

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

[9] Huda, M.N., Kawashima, H. and Nitta, T. (2009). Distinctive Phonetic Feature (DPF) Extraction Based on MLNs and Inhibition/EnhancementbNetwork, I*EICE Transactions on Information and Systems, 92*(4), pp. 671–680

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

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