CORPUS-BASED ANALYSIS OF ENGLISH SPOKEN BY JAPANESE STUDENTS IN VIEW OF THE ENTIRE PHONEMIC SYSTEM OF ENGLISH

Nobuaki MINEMATSU† *Gakuto KURATA*† *Keikichi HIROSE*‡

† Graduate School of Information Science and Technology, University of Tokyo
‡ Graduate School of Frontier Sciences, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, JAPAN
{mine, gakuto, hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

English and Japanese are quite different languages both phonetically and linguistically and it is often very difficult for Japanese students to master English pronunciation. To help students improve their pronunciation proficiency, a Japanese national project of "Advanced Utilization of Multimedia for Education" has started in 2000 and under this project, a large database of English words and sentences *read* by 200 Japanese students was built mainly for CALL system development. This paper describes a corpus-based analysis and comparison of American English (AE) and Japanese English (JE) by using WSJ database and the new JE database. Here, Hidden Markov Models (HMMs), which are widely-used acoustic modeling techniques of speech recognition, were firstly made for individual phonemes in the two kinds of English, and then, a tree diagram was drawn for the entire phonemes of each HMM set. The analysis and comparison of the two trees showed many interesting characteristics of JE, some of which are well-known habits observed in JE pronunciation. The authors consider that this study showed *statistical* differences between AE and JE in view of the entire phonemic system of English for the first time.

1. INTRODUCTION

Development of CALL systems requires large databases of learners' speaking/writing the target language because the current speech and language technologies are mainly based upon statistical methods. However, when the national project began, databases of Japanese students' speaking English were quite rare except that a few spontaneous speech databases existed[2, 3]. But the current speech recognition technologies are not mature enough to deal with even native spontaneous speech adequately due to its large variations[4]. Considering that acoustic/linguistic variations in non-native speech are much larger than in native speech, the databases of non-native spontaneous speech may not have so good advantage for the CALL system development and the databases of English *read* by Japanese are required instead, which still have good advantage to help students learning a new language. For these technical and educational reasons, a database of English sentences/words read by Japanese students was built in the national project. The first author of this paper was in charge of the database design and development and a brief description of the database will be found later.

In this paper, the new JE database was utilized not for CALL system development but for analysis of JE pronunciation. An AE read database can be easily found, a major one of which is WSJ. Using these two databases, corpus-based analysis and comparison of AE and JE pronunciation was done. Firstly, using PRONLEX lexicon[5], phonemic transcriptions were automatically generated for the two databases because they only included prompted sentences. Using the phonemic transcriptions, HMMs of the AE phonemes and those of the JE phonemes were built by embedded training. After that, a tree diagram was drawn for the entire phonemes of each of the two kinds of English. Analysis and comparison of the two trees gave us many interesting characteristics of JE and AE, which may be used for English teachers to know about JE as well as to train students. The authors consider that this work showed statistical differences between AE and JE especially in view of the entire phonemic system of English for the first time.

2. DEVELOPMENT OF AN ENGLISH DATABASE READ BY JAPANESE STUDENTS

Prior to the development, a lot of preliminary discussions were done to design the database[6]. In this paper, only the specification of the database and the adopted recording strategy are summarized.

2.1. Specifications of the database

Syllabi to teach English pronunciation were considered and two sets of material, sentences/words to learn segmental aspect of the pronunciation and those to do its prosodic aspect, were prepared.

2.1.1. Words/sentences for the segmental aspect of pronunciation

Table. 1 shows the final sets of words and sentences prepared for the segmental aspect. Since many of the material were selected based upon phonemic balance, rare words can be found easily. In the recording, words whose phonemic representations the speakers don't know at all were not adequate as reading material. Therefore, phonemic symbols were assigned to each word on a reading sheet. Even with the phonemic symbols, various pronunciation errors were highly expected for lack of the speakers' knowledge on correct articulation of English pronunciation. A set of minimal pair words included unknown words, for which, the speakers were asked to pronounce sequences of the assigned phonemic symbols.

2.1.2. Words/sentences for the prosodic aspect of pronunciation

Table. 2 lists the final sets of words and sentences for the prosodic aspect. The word set included words and phrases which can have their stressed syllables at different positions. In the sentence set

Table 1. Word and sentence sets prepared for the segmental aspect of English pronunciation

set	size
Phonetically-balanced words	300
Minimal pair words	600
MOCHA-TIMIT phonetically-balanced sentences	460
Sentences including phoneme sequences difficult for Japanese to pronounce correctly	32
Sentences designed as test set	100

Table 2. Word and sentence sets prepared for the prosodic aspect of English pronunciation

set	size
Words with various accent patters	109
Sentences with various intonation patterns	94
Sentences with various rhythm patterns	121

with various intonation patterns, the following types of sentences were included; 1) sentence pairs each of which are the same except that one has a comma at a certain position and the other does not at the position. This causes different intonation patterns, 2) sentence pairs each of which are identical except that their focused words differ, 3) sentences with different intonation patterns according to their syntactic structure and/or their meaning, and so forth. In the other sentence set, sentences with rhythm patterns of various difficulties were prepared. Prosodic symbols for intonation and rhythm were also assigned adequately to the material on a reading sheet.

2.2. Adopted strategy of the recording

Selection of speakers should be done carefully because the selected speakers should cover as wide a range of English pronunciation proficiency as possible. If only voluntary speakers are adopted for the recording, the database shall contain only speech samples of good speakers of English. To realize the adequate selection, we requested each of the recording sites to select randomly Japanese students and have them participate in the recording as speakers. Twenty organizations such as universities and colleges cooperated in the recording and English speech samples spoken by 100 male and 100 female students were collected. All the sentences in Tables **1** and **2** were divided into 8 groups and all the words were into 5 groups. The amount of the recording per speaker was a sentence group ($\sim$120 sentences) and a word group ($\sim$220 words). Therefore, each sentence was read by about 12 speakers and each word was read by about 20 speakers for each gender.

Speakers were asked to have practice in pronouncing sentences and words on the given sheets before the recording. During the recoding, they were requested to read the material repeatedly until they could do what they thought was the correct pronunciation. Even with this strategy of recording, many pronunciation errors were easily expected because of Japanese students' lack of knowledge on the correct articulation of English pronunciation.

3. CORPUS-BASED ANALYSIS AND COMPARISON OF AE AND JE PRONUNCIATION

3.1. Training of HMMs with the AE and JE databases

Monophones with diagonal matrices were adopted as HMMs because visualization of the results of the analysis required HMMs of the simple structure. To build the HMMs with embedded training, phonemic transcriptions of individual speech samples were required, which were generated by looking up PRONLEX pronunciation lexicon. In the lexicon, each word has only one pronunciation form basically, called citation form. In the transcriptions, a short pause was allowed between two consecutive words. In JE speech samples, these pauses were frequently observed due to low fluency. Speech samples were digitized at 16bit/16kHz sampling and 12 MFCCs, 12 ΔMFCCs, and Δpower were extracted from the signals with 25 ms frame length and 10 ms frame shift. The initial HMMs were trained using TIMIT database and they were used in the subsequent embedded training with WSJ database for AE models and with the newly built database for JE ones. The number of sentences from the former database was 25,652 spoken by 245 male speakers and that from the latter was 8,282 by 68 speakers. The other 32 speakers were testing speakers and not used in the training. **Table. 3** shows a phoneme set used here.

3.2. Tree diagrams of the entire phonemes of AE and JE

An HMM is composed of a number of states and several transitions between two states. Distance between a state and another can be calculated using adequate distance measure. Here, with Bhattacharyya distance measure, distance matrix was made for each of AE and JE HMM set. This matrix shows distances between any two of all the states in the HMM set, which include distance between a state of a phoneme and a state of another. This distance matrix enables us to draw a tree diagram of the entire phonemes based upon Ward's method, which is one method of hierarchical clustering. **Figure. 1** shows two tree diagrams of AE and JE. Leaf nodes correspond to states of the HMMs. Comparison of the two trees are expected to give us many characteristics of AE and JE.

3.3. Comparison between the two tree diagrams

3.3.1. Magnitude of variances of the AE and JE HMMs

Firstly, the broadness of parameter distributions was compared between AE and JE. **Figure. 2** shows ratios of averaged variances of MFCCs in JE to those in AE. Here, the averaged variances were calculated for each state over cepstrum dimensions. The figure shows that the variances in JE are larger than those in AE although the JE training data size is much smaller. Considering that the JE database contains carefully read speech only, the above fact implies that the large broadness of parameter distributions in JE is due to inter-speaker variations of pronunciation proficiency. This finding led us to devise a novel method of adapting HMMs for non-native speech recognition based upon speakers' proficiency. This work is described in detail in another paper of this conference[7].

3.3.2. Difficult phoneme pairs for Japanese to discriminate

Positions of difficult phoneme pairs for Japanese students to distinguish are investigated in each of the two trees. **Table. 4** shows distance between each pair and positions of some pairs in the table are shown in **Figure. 1**. For example, /s/ and /th/ are found

Table 3. Phoneme set used in the analysis

b, d, g, p, t, k, jh, ch, s, sh, z, zh,
f, th, v, dh, m, n, ng, l, r, w, wh, y,
hh, iy, ih, eh, ey, ae, aa, aw, ay, ah,
ao, oy, ow, uh, uw, er, ax

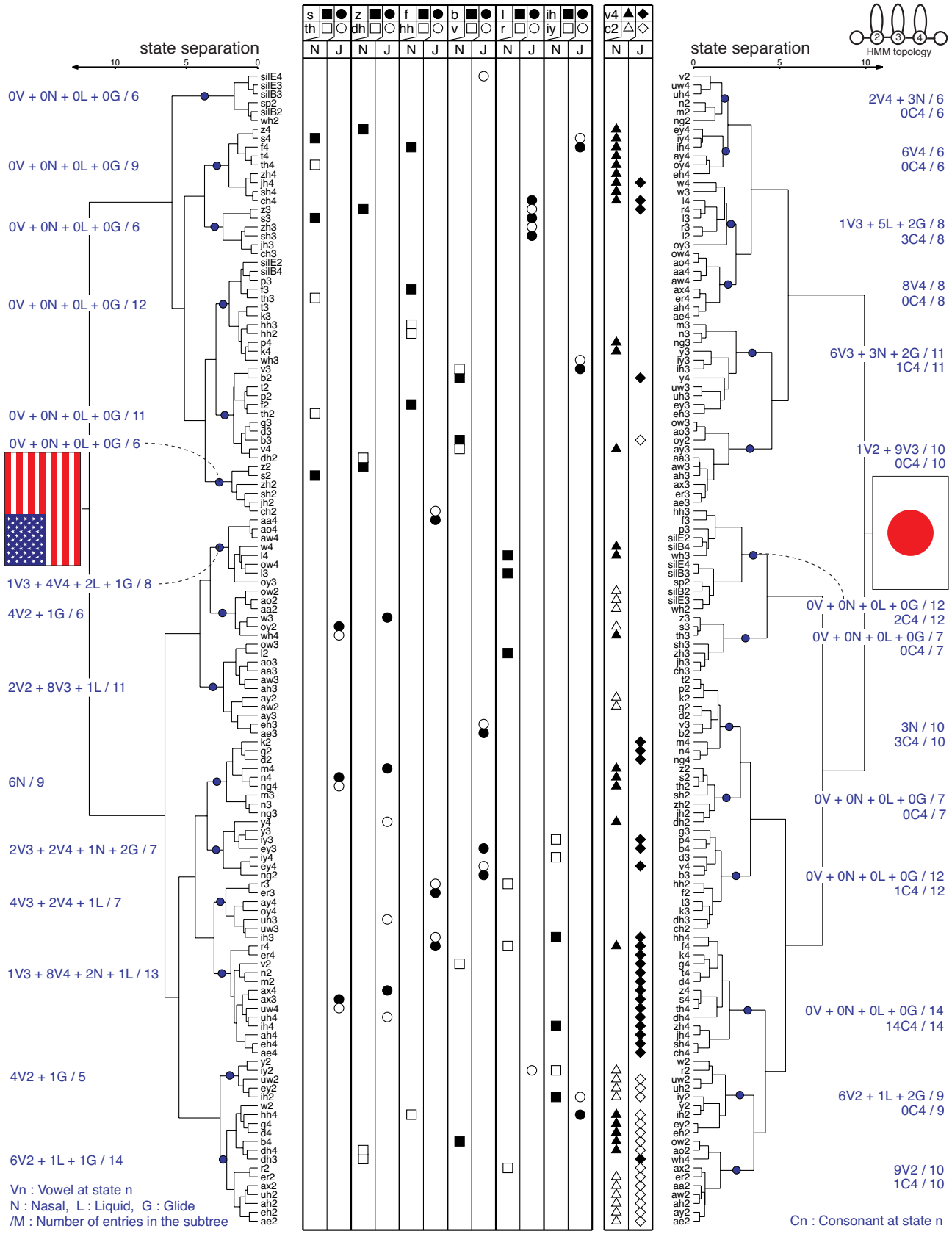

Fig. 1. Two tree diagrams for the entire phonemes of American English and Japanese English

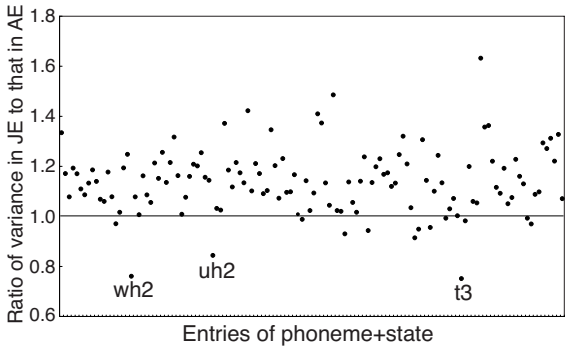

Fig. 2. Ratio of variance in JE to that in AE

Table 4. Ratios of Bhattacharyya distance in JE to that in AE

pair	s2	s3	s4	pair	s2	s3	s4
/r/&/l/	0.43	0.43	0.32	/hh/&/f/	0.56	0.52	0.76
/s/&/th/	0.30	0.18	0.31	/b/&/v/	0.97	0.89	0.63
/s/&/sh/	0.53	0.58	0.74	/ih/&/iy/	0.47	0.45	0.39
/th/&/sh/	0.48	0.57	0.70	/ih/&/y/	0.41	0.54	0.79
/z/&/zh/	0.60	0.70	0.87	/uh/&/uw/	0.51	0.48	0.55
/z/&/dh/	0.45	0.49	0.59	/ae/&/aa/	0.49	0.53	0.79
/z/&/jh/	0.46	0.61	0.79	/ae/&/ah/	0.51	0.41	0.35
/zh/&/jh/	0.56	0.57	0.75	/aa/&/ah/	0.51	0.36	0.68
/zh/&/dh/	0.52	0.56	0.62	/er/&/ah/	0.28	0.30	0.40
/dh/&/jh/	0.44	0.45	0.66	/er/&/aa/	0.41	0.34	0.51
/n/&/ng/	0.86	0.77	0.71	/er/&/ae/	0.39	0.30	0.47

quite close to each other in JE although they are located far away in AE. The distance of the table is represented in the form of ratio of the distance between the phoneme pair in JE to that in AE and the ratios are always less than 1.0. It can be definitely said that Japanese tend to confuse a phoneme of each pair with the other. Especially, mid and low vowels such as /ah/, /ae/, /aa/ are much confusing with each other. This is because the Japanese language has only one mid and low vowel of /a/ and students tend to replace all the English mid and low vowels with a Japanese vowel of /a/.

3.3.3. Vowel insertion between consecutive consonants

Figure. 1 also shows positions of state-4s of consonants and state-2s of vowels. It is found that most of the states of the both types are located under a single subtree in JE. This is because of the well-known JE habit of vowel insertion. In Japanese, every consonant is followed by a vowel. Therefore, Japanese tend to insert an additional vowel between two consecutive consonants when speaking English. Since these pronunciation errors were not represented in the transcription used to train HMMs, state-4s of consonants are expected to have similar spectrums to those in state-2s of vowels.

3.3.4. Schwa and the other vowels in AE and JE pronunciation

Schwa is often considered as the most difficult vowel for Japanese to pronounce correctly. Here, the five nearest phonemes to schwa are investigated for each state in AE and JE, which is shown in **Table. 5**. Phonemes near to schwa in AE are various vowels and this accords with a fact that unstressed vowels of any kind approach schwa sounds. On the other hand, most of the phonemes near to schwa in JE are mid and low vowels. This is because Japanese perceive a schwa sound as a Japanese mid and low vowel of /a/ and they produce a Japanese /a/ sound for a schwa sound.

Table 5. The five nearest phonemes to schwa in AE and JE

state	1st	2nd	3rd	4th	5th
ax2/AE	ih2(0.68)	uh2(0.73)	d4(0.75)	ah2(0.76)	eh2(0.86)
ax3/AE	ih3(0.87)	uh3(0.88)	eh4(0.93)	ae4(0.94)	uw4(0.96)
ax4/AE	uw4(0.69)	ih4(0.72)	uh4(0.76)	ah4(0.80)	eh4(0.84)
ax2/JE	ae2(0.46)	ah2(0.51)	aa2(0.51)	ay2(0.65)	aw2(0.69)
ax3/JE	ah3(0.57)	ae3(0.61)	aa3(0.72)	aw3(0.80)	uh3(0.87)
ax4/JE	ah4(0.54)	ae4(0.61)	aa4(0.73)	aw4(0.78)	uh4(0.86)

3.3.5. Structural differences between the two trees

Here, characteristics of subtrees are examined. In **Figure. 1**, constituent states of some subtrees are indicated with respect to vowels, nasals, liquids, and glides. In AE, it is found that all the states of the four classes are under the left-hand side of the entire tree and that 86 % of the states in the left-hand side belong to the four classes. Nasals, liquids, and glides have common characteristics that there is only a partial closure or an unimpeded oral or nasal escape of air. They are said to share many phonetic characteristics with vowels[8]. The AE tree clearly shows this property. On the other hand, this property cannot be seen in the JE tree. While all of the state-3 vowels and all of the state-4 vowels are under the left-hand side of the tree, all of the state-2 vowels but /oy2/ are under the right-hand side. Only about 70 % of the states of the above three consonants are found under the left-hand side. As mentioned in section **3.3.3**, more than half state-4s of consonants and state-2s of all the vowels are found under a single subtree, which clearly represents Japanese habit of inserting an additional vowel between consecutive consonants. Further, it is very interesting that state-3 vowels and state-4 vowels of JE tend to be separately located in subtrees under the left-hand side of the entire tree.

4. CONCLUSIONS

Corpus-based comparison of AE and JE were carried out. Here, HMMs were firstly built with AE and JE databases and tree diagrams were drawn for the two kinds of English. Comparison between the trees gave us some characteristics of AE and JE, which are related to 1) broadness of parameter distributions, 2) difficult phoneme pairs for Japanese to discriminate, 3) vowel insertion between two consecutive consonants, 4) schwa and the other vowels, and 5) structural differences between the two trees. This paper showed statistical differences between AE and JE in view of the entire phonemic system of English for the first time.

5. REFERENCES

[1] http://www.nime.ac.jp/tokutei120/index.html

[2] H. Isahara, T. Saiga, and E. Izumi, "The TAO speech corpus of Japanese learners of English," Proc. ICAME'2001 (2001)

[3] http://cslu.cse.ogi.edu/corpora/fae/index.html

[4] S. Nakagawa, "A survey on automatic speech recognition," Trans. IEICE, vol.J83-D-II, no.2, pp.433–457 (2000, in Japanese).

[5] http://www.ldc.upenn.edu/Catalog/LDC97L20.html

[6] N. Minematsu et al., "English speech database read by Japanese learners for CALL system development," Proc. LREC'2002, pp.896–903 (2002)

[7] N. Minematsu et al., "Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition," Proc. ICSLP'2002 (2002, accepted)

[8] A. C. Gimson, "An Introduction to the Pronunciation of English," Edward Arnold Ltd. (1980)