English Speech Database Read by Japanese Learners for CALL System Development

N. Minematsu[†], Y. Tomiyama[‡], K. Yoshimoto[*], K. Shimizu[⋆], S. Nakagawa[◇], M. Dantsuji[‡], and S. Makino[*]

[†]Univ. of Tokyo, [‡]Kyoto Univ., [*]Tohoku Univ.,
[⋆]Nagoya Gakuin Univ., and [◇]Toyohashi Univ. of Tech.
eng-db@gavo.t.u-tokyo.ac.jp

Abstract

With the help of recent advances in speech processing techniques, we can see various kinds of practical speech applications in both laboratories and the real world. One of the major applications in Japan is CALL (Computer Assisted Language Learning) systems. It is well-known that most of the recent speech technologies are based upon statistical methods, which require a large amount of speech data. Although we can find many speech corpora available from distribution sites such as Linguistic Data Consortium, European Language Resources Association, and so on, the number of speech corpora built especially for CALL system development is very small. In this paper, we firstly introduce a Japanese national project of "Advanced Utilization of Multimedia to Promote Higher Educational Reform," under which some research groups are currently developing CALL systems. One of the main objectives of the project is to construct an English speech database read by Japanese students for CALL system development. This paper describes specification of the database and strategies adopted to select speakers and record their sentence/word utterances in addition to preliminary discussions and investigations done before the database development. Further, by using the new database and WSJ database, corpus-based analysis and comparison between Japanese English and American English is done in view of the entire phonemic system of English. Here, tree diagrams of the two kinds of English are drawn through their HMM sets. Results show many interesting characteristics of Japanese English.

1. Introduction

It is widely known that Japanese and English are very different languages linguistically and phonetically and this difference makes it quite difficult for Japanese students to master English. It is reported that the ability for Japanese students to speak, listen to and/or write English is quite poor in comparison with that for students in other Asian countries to do. To save this situation of Japanese students, various national projects have been formed so far. One of them is Scientific Research on Priority Area (A), "Advanced Utilization of Multimedia to Promote Higher Educational Reform," which has started in 2000 under financial support of the Ministry of Education, Culture, Sports, Science and Technology. This project progressively promotes speech and language technologies as well as multimedia technologies into language education and learning.

Recent advances in speech technologies have made it possible to develop CALL systems for pronunciation learning. In Japan, many speech researchers and language teachers are aiming at developing tools and systems helpful for Japanese students of English. However, we have one big problem for the development. Since most of the current speech technologies are based upon statistical methods, they naturally require large databases. To develop recognizers of *native* speech, a large number of databases of various languages were already built and distributed worldwide. As for databases of non-native speech, we can find only several ones partly because these databases should be built dependently on both the native language and the target language of students, and therefore the development cost is quite high. Moreover, most of the non-native speech databases contain spontaneous speech only, e.g. Q & A style conversations (Isahara et al., 2001; Tono, 2001) and free conversations on telephone line (CSLU@OGI, 2000). When learning a new language, as the first step, students are often required to pronounce sentences/words written on a textbook repeatedly. To introduce speech technologies into this situation, what is required and desired is a database of not spontaneous speech but *read* speech by non-native speakers. It should be noted that speech recognition technologies are not mature enough to deal with even native spontaneous speech adequately due to its large variations (Nakagawa, 2000). It is easily assumed that acoustic variations and distortions found in non-native speech are much larger than in native speech. It implies that a database of non-native spontaneous speech will have limited advantage. These educational and technical necessities led us to build an English speech database *read* by Japanese students.

In section **2.**, preliminary discussions are shown on what kind of sentences or words should be prepared as reading material. Specification of the speech database is given in section **3.**. We adopted a unique recording strategy to collect only the speech samples which Japanese students themselves judged that were correct in terms of pronunciation, which is shown in section **4.**. Section **5.** gives some examples of the reading material, some of which are with phonemic/prosodic symbols. Using the new database and WSJ database, corpus-based analysis and comparison of Japanese English and American English is done. Some interesting characteristics of Japanese English are described in section **6.**. Finally, this paper is concluded in section **7.**.

2. Preparations done before the database development

2.1. Databases required for the CALL system development

As told in the previous section, non-native utterances have larger acoustic and linguistic distortions than native

ones. The magnitude of these distortions is supposed to depend on various factors such as the target language and the native language of students, their dialect, their age, the amount of acquired knowledge on the target language, and so on. This fact often causes a very large inter-student variety of the distortions. In the development of CALL systems, it is desirable to use a database which contains all the acoustic and/or linguistic distortions possible observed in non-native speech. In the case that the native language of students is Japanese and their target language is English, phonetic/linguistic differences between the two languages make it very difficult to describe all the distortions systematically, which is currently one of the main issues in the error analysis of Japanese students' speaking and writing English. Based upon these considerations, we made guidelines below to follow for the database development.

- The target language is General American (GA).
- Speakers are Japanese students of universities or colleges and their graduate schools.
- Main focus is placed only upon acoustic distortions. Linguistic distortions such as grammatical errors are not considered in the development.
- Neither acoustic distortions observed only in a particular student's utterances nor those observed only temporarily are considered. In the current work, main focus is put on the acoustic distortions which are found rather commonly and frequently in Japanese speaking of English. They are mainly for lack of knowledge on correct articulation for English pronunciation.

2.2. Outline of the database specification

Before the database development, preliminary discussions were done on recording conditions and reading material according to the syllabus of teaching English pronunciation. Even if we follow the guidelines in section **2.1.**, the acoustic distortions are expected to be still very large compared to those in native speech. Then, we categorized situations of students' speaking English for pronunciation learning into several types based upon information or hints given to the students. Since the acoustic distortions should partly depend on the type, we selected one of them so that the distortions might be reduced to be treated adequately and correctly by the current speech technologies.

1. Students speak English fully spontaneously and freely without any hint or help.
2. Students read given words or sentences. In this case, text or orthographical information is given.
3. Students read given words or sentences with phonemic/prosodic symbols. In addition to orthographical information, phonemic/prosodic one is given *as text*.
4. Students read given words or sentences after hearing model utterances spoken by an English teacher. Here, *acoustic* information, both segmental and prosodic, is additionally given to students.

Out of the above types, we selected the third one. This is because we judged that the first and second types should often generate too many student-specific and/or temporary pronunciation errors and that the model utterances for students as in the forth type were not always prepared in self-learning of English. Even in the third manner, we expected that various acoustic distortions could be observed in students' pronunciation for their lack of knowledge on correct articulation of English pronunciation.

As for the reading material, we considered the syllabus of teaching English pronunciation. Although various matters should be taught to students, we divided them into two aspects; segmental (phonetic) aspect and prosodic aspect. In the database development, we determined to prepare sentence sets and word sets for each of the two aspects. For the former aspect, a phonemically-balanced sentence set, a sentence set including sequences of phonemes difficult for Japanese students to pronounce correctly, a phonemically-balanced word set, a set of minimal pair words, and so forth were prepared. As for the latter aspect, a set of sentences with various intonation patterns, some of which depend upon syntactic structure of the sentence and others are related to meaning of the sentence, a set of sentences with various rhythm patterns, which are defined as a sequence of stressed syllables and unstressed ones, a set of words which are allowed to have their stressed syllables at different positions in the words, a set of compound words, and so on were designed for the database.

On the sheets of the reading material, phonemic symbols and/or prosodic ones were assigned if required. Before the recording, we gave instructions to speakers (Japanese students) so that they could understand these symbols correctly. For each of the phonemic symbols used here, we arranged a web page so that the speakers could hear a word example which did not appear in the reading sheets but included the phonemic symbol. As for phrase break symbols, which are part of the prosodic symbols, the speakers were allowed to make sure of how to realize the phrase break by hearing several sentence examples on the web.

3. Detailed specification of the database

3.1. Phonemic symbols and prosodic ones assigned to words and sentences

Phonemic symbols of TIMIT database and those of CMU pronunciation dictionary were used as reference sets. After modifying these sets, the phonemic symbols for the assignment were determined, which are listed in **Table. 1**. Most of the English-Japanese dictionaries of Japanese students represent schwa sounds by more than one phonetic symbol, which seem to be selectively used mainly according to the orthography. In the phonemic symbol set adopted here, we have only one symbol /AX/ for schwa sounds. Some speakers claimed that, only with the assigned symbols, it was difficult to determine how to pronounce words

Table 1: Phonemic symbols assigned to reading material

B, D, G, P, T, K, JH, CH, S, SH, Z, ZH, F, TH, V, DH, M, N, NG, L, R, W, Y, HH, IY, IH, EH, EY, AE, AA, AW, AY, AH, AO, OY, OW, UH, UW, ER, AXR, AX

Table 2: Word and sentence sets prepared in terms of the segmental aspect of English pronunciation

set	size
phonemically-balanced words	300
minimal pair words	600
TIMIT-based phonemically-balanced sentences	460
sentences including phoneme sequences difficult for Japanese to pronounce correctly	32
sentences designed for test set	100

Table 3: Word and sentence sets prepared in terms of 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

including /AX/. In this case, we asked them to look up their own English dictionary before recording. When assigning the phonemic symbols to words, citation form of their pronunciations was used for each word.

As for the prosodic symbols, primary/secondary stress symbols, intonation symbols, and/or rhythm symbols were assigned if necessary. A number, 0, 1, or 2, was given to each vowel, which represented three levels of word stress; primary stress (1), secondary stress (2), and no stress (0). Intonation was indicated by a rising arrow, a falling one, a rising-falling one, or a falling-rising one. Rhythm pattern was represented by a sequence of sentence stress, which also had three levels; stress nucleus (@), normal stress (+), and no stress (-). One of them was assigned to each syllable in a sentence. Levels of these symbols were adequately determined by a native teacher of English or by looking up textbooks of English. Some examples of the reading material with these symbols are shown in section **5.**.

3.2. Word sets and sentence sets prepared in terms of the segmental aspect of English pronunciation

Table. 2 shows the final sets of words and sentences prepared for the database development in terms of the segmental aspect. A set of minimal pair words included unknown words, for which, speakers were requested to pronounce a sequence of phonemic symbols assigned to them (See section **5.**). For sentence sets, we prepared two types of reading sheets for each of the sets. One was with phonemic symbols for every word, which was used only for pronunciation practice before the recording, and the other was without them, which was referred to during the recording. The preparation of two types of sheets is because reading sentences with referring to phonemic symbols is expected to induce unnatural pronunciation. With the phonemic symbols for each word, some speakers may not read a sentence but a sequence of isolated words. As for word sets, since some words are unknown, reading sheets of the first type were only prepared. Unlike sentence sets, unnatural pronunciation due to the phonemic symbols was not expected here. This is because most of the words in the word sets were short and plain except for the unknown words, while the sentence sets had rare words especially in the case of the phonemically-balanced set.

3.3. Word sets and sentence sets prepared in terms of the prosodic aspect of English pronunciation

Table. 3 lists the final sets of words and sentences prepared in terms of the prosodic aspect. In the word set, as told in section **2.2.**, it included words and phrases which can have their stressed syllable at different positions. Different location of the stressed syllable gives different meanings such “white house” and “White House”. In the sentence set with various intonation patterns, the following sentences were included; 1) sentence pairs each of which are the same except that one has a comma at a certain position in it and the other does not at the position. This causes different intonation patterns between the two, 2) sentence pairs each of which are identical except that they have different focused words, 3) sentences with various intonation patterns according to their syntactic structure and/or their meaning, and so forth. In the sentence set with various rhythm patterns, sentence stress was grouped into three levels and one of the levels was adequately assigned to each syllable by an American teacher of English based upon a principle that the stressed syllable in the last content word in a phrase has stress nucleus (the strongest stress) in the phrase. In this sentence set, several sentences composed a subset, where subsequent sentences were arranged to be more difficult in terms of their syntactic/rhythmic structure. Section **5.** shows some examples of the word sets and the sentence sets with phonemic/prosodic symbols.

4. Recording of speech samples

4.1. Selection of the speakers

Selection of the speakers should be done carefully because it is desired that the speakers should cover as wide a range of English pronunciation ability as possible. If only voluntary speakers are collected for the recording, the database shall contain only English speech samples of rather good speakers of English. It should contain English speech of poor speakers as well as good speakers. To realize the adequate selection, we requested each of the recording sites to select randomly Japanese students of the site 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 Japanese students were collected. All the sentences in Tables **2** and **3** were divided into 8 groups and all the words in the tables were into 5 groups. The required amount of the recording per speaker was a sentence group ($\sim$120 sentences) and a word group ($\sim$220 words). Therefore, each sentence and each word were read by about 12 speakers and 20 speakers respectively for each gender.

4.2. Procedures of the recording

In the database development, neither acoustic distortions observed only in a particular student’s utterances nor those observed only temporarily were considered. In other words, main focus of the database development was placed on the distortions or deviations found rather commonly in

Japanese speaking of English. Besides, during the recording, there should be no unknown words for the speakers because pronunciation errors due to lack of English vocabulary are another problem than errors due to lack of knowledge on correct articulation of English pronunciation. To avoid these unwanted events, the following recording procedures were adopted in the development.

1. Before the recording, speakers were asked to practice pronouncing sentences and words on the given sheets. In the practice, they were permitted to refer to the reading sheets with phonemic and prosodic symbols.
2. In the recording, speakers were asked to read sentences and words on the given sheets repeatedly until they could do what *they thought* was the correct pronunciation. Even in this recording strategy, many pronunciation errors were still easily expected for lack of knowledge on correct articulation of English pronunciation. If speakers made the same pronunciation error three times repeatedly, they were allowed to skip the material and go to the next one.
3. After the recording, each of speech samples was checked by technical staff of the recording site. If they found any technical errors in some sentences or words, the recording was done again for them.

Through the above recording procedures, the database shall contain English speech samples which the speakers judged themselves that were correctly pronounced. Therefore, the pronunciation errors in the database are supposed to be purely for lack of the speakers' knowledge on English articulation. The authors consider that the database focusing purely on the lack of that knowledge is quite unique.

5. Examples of reading material

Some examples of the reading material are shown in the following pages (**Tables 4** to **10**). All the words in the examples are with phonemic symbols and every vowel has its stress mark, 0, 1, or 2. Some examples for the prosodic aspect of English pronunciation have prosodic symbols such as intonation patterns (arrows) or rhythm patterns.

6. Corpus-based analysis and comparison of American English and Japanese English

The database was built mainly to be utilized in the CALL system development. In this section, however, the database was used to statistically analyze the acoustic/phonetic differences between American English (AE) and Japanese English (JE). As mentioned above, this database is quite unique in that only the speech samples which Japanese speakers themselves judged that were correctly pronounced. This means that the pronunciation errors in the database can be viewed as inevitable pronunciation distortions or deviations generated in the current methodologies of English pronunciation education. Considering this unique property of the database, the first author of this paper carried out a corpus-based analysis and comparison of AE and JE. He considers that this analysis showed statistical differences between AE and JE *in view of the entire phonemic system of English* for the first time.

6.1. Training of AE HMMs and JE HMMs

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 new JE 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. Although all the analysis was done only with male speech, the first author thinks of no significant differences between the two genders. **Table. 11** shows a phoneme set used here, which is slightly different from the phoneme set assigned to the reading sheets of the recording. This is because the automatic generation of the transcriptions required us to use PRONLEX lexicon and it adopts a phoneme set of **Table. 11**.

6.2. Tree diagrams of the entire phonemic system 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, represented by its phoneme identity and its state number. In the figure, HMM topology is also shown and parameter distributions are defined only in states 2 to 4.

6.3. Comparison between the two tree diagrams

Comparisons between the AE tree and the JE tree were done based upon the following five viewpoints, some of which are closely related to well-known habits of JE.

Table 11: 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

Table 4: Examples of phonemically-balanced sentences with phonemic symbols and word stress symbols

```
S1_0051  Ambidextrous pickpockets accomplish more.
         [AE2 M B AX0 D EH1 K S T R AX0 S] [P IH1 K P AA2 K AX0 T S]
         [AX0 K AA1 M P L AX0 SH] [M AO1 R]
S1_0052  Her classical repertoire gained critical acclaim.
         [HH ER1] [K L AE1 S AX0 K AX0 L] [R EH1 P AXR0 T W AA2 R] [G EY1 N D]
         [K R IH1 T AX0 K AX0 L] [AX0 K L EY1 M]
S1_0053  Even a simple vocabulary contains symbols.
         [IY1 V AX0 N] [AX0] [S IH1 M P AX0 L] [V OW0 K AE1 B Y AX0 L EH2 R IY0]
         [K AX0 N T EY1 N Z] [S IH1 M B AX0 L Z]
S1_0054  The eastern coast is a place for pure pleasure and excitement.
         [DH IH1] [IY1 S T AXR0 N] [K OW1 S T] [IH1 Z] [AX0] [P L EY1 S] [F AO1 R]
         [P Y UH1 R] [P L EH1 ZH AXR0] [AE1 N D] [AX0 K S AY1 T M AX0 N T]
S1_0055  The lack of heat compounded the tenant's grievances.
         [DH AX0] [L AE1 K] [AH1 V] [HH IY1 T] [K AX0 M P AW1 N D AX0 D] [DH AX0]
         [T EH1 N AX0 N T S] [G R IY1 V AX0 N S AX0 Z]
```

Table 5: Examples of sentences including phoneme sequences which are difficult for Japanese to pronounce fluently

```
S1_0061  San Francisco is one-eighth as populous as New York.
         [S AE1 N] [F R AE0 N S IH1 S K OW0] [IH1 Z] [W AH1 N EY1 TH] [AE1 Z]
         [P AA1 P Y AX0 L AX0 S] [AE1 Z] [N Y UW1] [Y AO1 R K]
S1_0062  Its extreme width was eighteen inches.
         [IH1 T S] [AX0 K S T R IY1 M] [W IH1 D TH] [W AA1 Z] [EY0 T IY1 N]
         [IH1 N CH AX0 Z]
S1_0063  Who ever saw his old clothes ?
         [HH UW1] [EH1 V AXR0] [S AO1] [HH IH1 Z] [OW1 L D] [K L OW1 DH Z]
S1_0064  I could be telling you the five-fifths of it in two-three words.
         [AY1] [K UH1 D] [B IY1] [T EH1 L AX0 NG] [Y UW1] [DH AX0]
         [F AY1 V F IH1 F TH S] [AH1 V] [IH1 T] [IH1 N] [T UW1 TH R IY1]
         [W ER1 D Z]
```

Table 6: Examples of phonemically-balanced words with phonemic symbols and word stress symbols

```
W1_0041 rub   [R AH1 B]     W1_0044 strife [S T R AY1 F]   W1_0047 there [DH EH1 R]
W1_0042 slip  [S L IH1 P]   W1_0045 such   [S AH1 CH]      W1_0048 toe   [T OW1]
W1_0043 smile [S M AY1 L]   W1_0046 then   [DH EH1 N]      W1_0049 use   [Y UW1 S]
```

Table 7: Examples of minimal pair words with phonemic symbols and word stress symbols

```
luck    [L AH1 K]          lack    [L AE1 K]
robe    [R OW1 B]          rope    [R OW1 P]
sink    [S IH1 NG K]       sing    [S IH1 NG]
burn    [B ER1 N]          barn    [B AA1 R N]
selling [S EH1 L AX0 NG]   sailing [S EY1 L AX0 NG]
stuck   [S T AH1 K]        stock   [S T AA1 K]
meat    [M IY1 T]          mitt    [M IH1 T]
pitch   [P IH1 CH]         bitch   [B IH1 CH]
```

6.3.1. Magnitude of variances of AE and JE HMMs

Firstly, the broadness of parameter distributions were compared between AE and JE HMMs. **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 (Minematsu et al., 2002).

6.3.2. Phoneme pairs difficult for Japanese to discriminate perceptually

Positions of difficult phoneme pairs for Japanese students to distinguish perceptually are investigated in each of the two trees. **Table. 12** 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 quite close to each other in JE although they are located far away in AE.

Table 8: Examples of sentences of various intonation patterns with phonemic symbols and prosodic symbols

```
S1_0086  That's from my brother who lives in London.
         [DH AE1 T S] [F R AH1 M] [M AY1] [B R AH1 DH AXR0] [HH UW1] [L IH1 V Z]
         [AX0 N] [L AH1 N D AX0 N]
S1_0087  That's from my brother, who lives in London.
         [DH AE1 T S] [F R AH1 M] [M AY1] [B R AH1 DH AXR0] [HH UW1] [L IH1 V Z]
         [AX0 N] [L AH1 N D AX0 N]
S1_0091
         Cauliflower, broccoli, cabbage, sprouts, and onions.
         [K AA1 L AX0 F L AW2 AXR0] [B R AA1 K AX0 L IY0] [K AE1 B AX0 JH]
         [S P R AW1 T S] [AE1 N D] [AH1 N Y AX0 N Z]
S1_0094
         Is this elevator going up or down ?
         [IH1 Z] [DH IH1 S] [EH1 L AX0 V EY2 T AXR0] [G OW1 AX0 NG] [AH1 P]
         [AO1 R] [D AW1 N]
S1_0097
         She knows you, doesn't she ?
         [SH IY1] [N OW1 Z] [Y UW1] [D AH1 Z AX0 N T] [SH IY1]
```

Table 9: Examples of sentences of various rhythm patterns with phonemic symbols and prosodic symbols

```
S1_0105  Come to tea.
        / +    -  @ /
         [K AH1 M] [T UW1] [T IY1]
S1_0106  Come to tea with John.
        / +    -  +   -    @  /
         [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N]
S1_0107  Come to tea with John and Mary.
        / +    -  @ / -    +   -    @ -/
         [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N] [AE1 N D] [M EH1 R IY0]
S1_0108  Come to tea with John and Mary at ten.
        / +    -  @ / -    +   -    +   -   @ /
         [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N] [AE1 N D] [M EH1 R IY0]
         [AE1 T] [T EH1 N]
```

Table 10: Examples of words of various accent patterns with phonetic symbols and word stress symbols

```
W1_0201 a dark room
        [AX0][D AA1 R K][R UW1 M]
W1_0202 a darkroom
        [AX0][D AA1 R K R UW2 M]
W1_0203 a light housekeeper
        [AX0][L AY1 T][HH AW1 S K IY2 P AXR0]
W1_0204 a lighthouse keeper
        [AX0][L AY1 T HH AW2 S][K IY1 P AXR0]
W1_0205 the brief case
        [DH AX0][B R IY1 F][K EY1 S]
W1_0206 the briefcase
        [DH AX0][B R IY1 F K EY2 S]
W1_0207 almond-eyed
        [AA2 M AX0 N D AY1 D]
W1_0208 broad-minded
        [B R AO1 D M AY1 N D AX0 D]
W1_0209 free-range
        [F R IY1 R EY1 N JH]
W1_0210 blue-black
        [B L UW1 B L AE1 K]
W1_0211 forward-looking
        [F AO1 R W AXR0 D L UH2 K AX0 NG]
W1_0212 built-in
        [B IH1 L T IH1 N]
```

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/.

6.3.3. Vowel insertion between consecutive consonants

It is found that most of the state-4s of consonants and the state-2s of vowels 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. Then, Japanese tend to insert an additional vowel between two consecutive consonants when speaking English. Since these errors were not represented in the transcription used to train HMMs, state-4s of consonants are expected to have similar spectrums to state-2s of vowels.

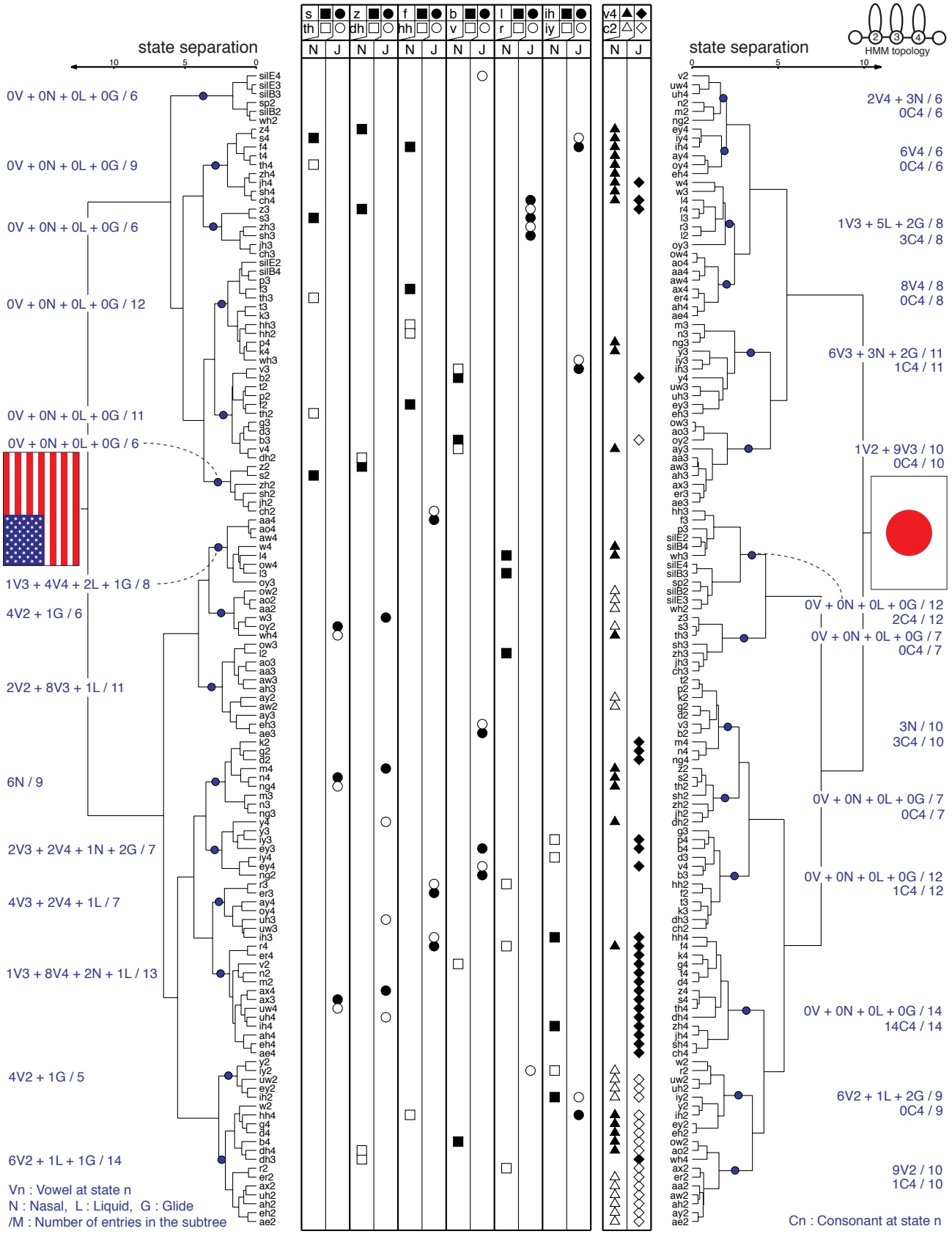

Figure 1: Two tree diagrams for the entire phonemes of American English and Japanese English

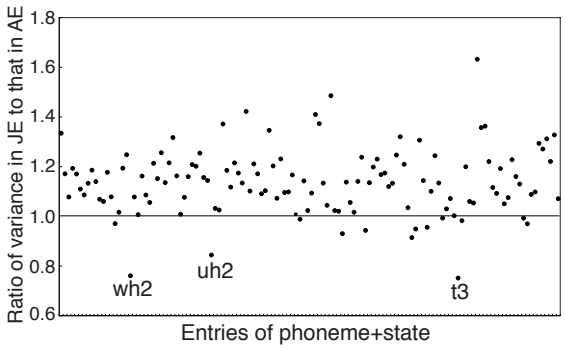

Figure 2: Ratio of variance in JE to that in AE

Table 12: Ratios of state 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

Table 13: The five nearest phonemes to schwa

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)

6.3.4. Schwa and the other vowels in AE and JE

Schwa is often viewed as the most difficult vowel for Japanese to pronounce correctly. Here, the five nearest phonemes to schwa are investigated for each of its states in AE and JE, which is shown in **Table. 13**. 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.

6.3.5. 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(Gimson, 1980). 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 **6.3.3.**, 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 vowel insertion. 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.

7. Conclusions

In this paper, the development of an English speech database read by Japanese students was described. Main focus was placed upon pronunciation errors for lack of knowledge on correct articulation of English pronunciation. Two types of reading material were prepared. One is related to segmental aspect of English pronunciation and the other is to its prosodic aspect. Two hundred speakers were randomly selected and they were asked to read sentences and words repeatedly until they could do what they thought was the correct pronunciation. Therefore, the pronunciation errors in the database are supposed to be purely for lack of the speakers' knowledge on English articulation. Using the new database, corpus-based comparison of American English and Japanese English was carried out. Here, HMMs were firstly built with AE and JE databases and a tree diagram was drawn for each of 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.

8. References

CSLU@OGI. 2000. http://cslu.cse.ogi.edu/corpora/fae.

A. C. Gimson. 1980. An introduction to the pronunciation of english.

H. Isahara, T. Saiga, and E. Izumi. 2001. The tao speech corpus of japanese learners of english. *Proc. ICAME'2001*.

N. Minematsu, G. Kurata, and K. Hirose. 2002. Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition. *Proc. ICSLP'2002 (submitted)*.

S. Nakagawa. 2000. A survey on automatic speech recognition. *Trans. Institute of Electronics, Information and Communication Engineers*, vol.J83-D-II(2):433–457.

Y. Tono. 2001. The standard speaking corpus: a 1 million-word spoken corpus of japanese.