Development of English Speech Database Read by Japanese to Support CALL Research

Nobuaki MINEMATSU[1], *Yoshihiro TOMIYAMA*[2], *Kei YOSHIMOTO*[3], *Katsumasa SHIMIZU*[4], *Seiichi NAKAGAWA*[5], *Masatake DANTSUJI*[6], *and Shozo MAKINO*[2]

1: University of Tokyo, 2: Telecom Technology Europe, 3: Tohoku University,
4: Nagoya Gakuin University, 5: Toyohashi University of Technology, and 6: Kyoto University
eng-db@gavo.t.u-tokyo.ac.jp

Abstract

This paper describes development of ERJ (English Read by Japanese) database, which is designed to support CALL (Computer Assisted Language Learning) research. The DB is divided into two parts, English read by Japanese and that by Americans with the same reading sheets in both. The reading material is composed of four sections, prosodic/phonetic $\times$ sentence/word sections. 202 Japanese students were randomly selected from various universities all over Japan and they were requested to repeat reading the sheets until they judged that they did the correct pronunciation. The reading sheets contained helpful guidelines to the correct pronunciation, such as phonemic symbols, prosodic marks, and so on. Before the recording, the students were allowed to do some practices. The resulting DB can be viewed as a volume of English pronunciations which are correct at least for Japanese students. After collecting speech samples, five American teachers of English were asked to rate utterances of a part of the DB based on three criteria, phonetic (segmental), rhythmic, and intonational aspects of pronunciation. This paper also describes several on-going projects with this DB.

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 ability for Japanese students to communicate in English is quite poor in comparison with that for students in other Asian countries to do. To save this situation, some 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 into language learning and education.

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 students. However, we have one big problem in the development. Since most of the current speech technologies are based upon statistical methods, they naturally require large DBs. To develop recognizers of *native* speech, a large number of DBs were already built and distributed worldwide. As for DBs of non-native speech, we could find only several ones partly because these DBs should be built dependently on both the native language and the target language of students and the development cost is quite high. Moreover, the non-native speech DBs usually contain spontaneous speech only, e.g. Q & A style conversations [1, 2] and free conversations on telephone line [3]. 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 DB of not spontaneous but *read* speech of 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 [4]. It is easily assumed that acoustic variations and distortions found in non-native speech are much larger than those in native speech. It implies that a DB of non-native spontaneous speech will have only limited advantage. These educational and technical requirements led us to build an English speech DB *read* by Japanese students.

2. Preparation for developing the database

As described above, 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 knowledge acquired so far on the target language, and so forth. Since it is very difficult to design the DB so that it contains all the kinds of the distortions, the following guidelines were made for the DB development.

- The target language is General American (GA).
- Speakers are university or college students and graduate school students of Japanese.
- 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 only on the acoustic distortions which are found rather commonly and frequently in Japanese speaking of English. They are mainly caused for lack of knowledge on correct articulation of English phones.

Preliminary discussions were also done on recording conditions and reading material according to syllabuses of teaching English pronunciation. Even if we followed the above guidelines, the acoustic distortions were expected to be still very large if the recording was done in inadequate way. To determine the recording condition, we categorized situations of students' speaking English for pronunciation learning into several types according to hints given to the students in advance.

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.

Condition of type-3 was selected because we judged that type-1 and type-2 should often generate too many student-specific and/or temporary pronunciation errors and that the model utterances for students as in type-4 could not be always prepared in self-learning situation. Even in type-3, we expected that various acoustic distortions could be observed in students' pronunciation for their lack of knowledge on the correct articulation.

Syllabuses of English education show that various issues should be treated in pronunciation learning. However, they can be divided into two aspects; segmental (phonetic) aspect and prosodic aspect. In the database development, we determined to prepare sentence sets and word sets for each aspect. For the former aspect, a phonemically-balanced sentence set, a sentence set including sequences of phonemes difficult for Japanese students to pronounce fluently, 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, 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 prepared for the database. On reading sheets, phonemic and/or prosodic symbols were assigned if required. Before the recording, we gave instructions to students so that they could understand correctly what these symbols meant.

3. Specification of the database

3.1. Phonemic and prosodic symbols

Phonemic symbols of TIMIT database and those of CMU pronunciation dictionary were used as reference sets. After modifying these sets, the phonemic symbols were determined, which are listed in **Table. 1**. Most of the English-Japanese dictionaries represent schwa sounds by different symbols, 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 symbols prepared, it was difficult to determine how to pronounce words including /AX/. In this case, we asked them to look up their own English dictionaries before recording.

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 several kinds of arrows. 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 adequately by an American teacher of English. Some examples of the reading material with these symbols are shown in section **4**.

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

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

Table. 2 shows the sets of words and sentences finally prepared 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. 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. Preparation of two types of sheets was because reading sentences with referring to phonemic symbols was expected to induce unnatural pronunciation. With phonemic symbols, some speakers may read not a sentence but a sequence of isolated words. As for word sets, since some words were 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 was 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, it included words and phrases which can have their stressed syllable at different positions. In the sentence set with various intonation patterns, the following sentences were included; 1) sentence pairs each pair 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 intona-

tion patterns according to their syntactic structure and/or their meaning, and so forth. In the sentence set with various rhythm patterns, stress marks were assigned by an American teacher 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 **4** shows some examples of the word sets and the sentence sets.

The DB contains speech samples of these sets of Japanese students and Americans, and results of American teachers' proficiency rating of the Japanese students. The recording and the rating procedures are described below.

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 balanced selection, we requested each of the recording sites to select randomly Japanese students in 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 102 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 (∼120 sentences) and a word group (∼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

The following unique recording procedures were used.

1. Before the recording, speakers were asked to practice pronouncing sentences and words on the given sheets. In the practice, they were allowed 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 strategy, many pronunciation errors were still easily expected for lack of knowledge on the correct articulation. 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.

The resulting DB is a volume of English pronunciations judged as correct by students. This implies that the DB indicates the performance of English pronunciation teaching in Japan.

4.3. Recording of American English samples

The same material was read by 20 (8 male and 12 female) Americans to be recorded. Here, only General American speakers were adopted because GA was treated as the target language. One speaker read a half of all the sentence sets (∼480 sentences) and a half of all the word sets (∼550 words).

Some examples of the reading material are shown in **Tables 4** to **6**. 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 the pronunciation have prosodic symbols such as intonation patterns (arrows) or rhythm patterns.

5. Pronunciation proficiency rating

Speech samples of non-native speakers are not of great use in CALL research if they are provided without any additional information. The DB can be beneficial with pronunciation proficiency scores of individual speakers or utterances. 5 American teachers of English, who had good experience of teaching English pronunciation to Japanese and good knowledge of phonetics, joined the rating experiment. The rating was done from three viewpoints, segmental, rhythmic, and intonational aspects of the pronunciation. In recording speech samples of Japanese students, phonemic, rhythmic, and intonational symbols were referred to by the students. Then, it was quantitatively rated by the teachers whether these symbols were adequately realized. 5-scale rating was adopted for each of the three aspects. For word utterances, 20 words and 10 words were rated for each student in terms of phonetic aspect and lexical stress respectively. For sentence utterances, 10 sentences, 5 sentences, and 5 sentences were rated for each student for segmental, rhythmic, and intonational aspects of the pronunciation. A teacher rated approximately 3,800 sentences and 5,700 words.

6. Use of the database in CALL research

The DB was already utilized in various CALL researches. In this section, several examples are shown. The authors believed that the DB is the largest English speech database *read* by Japanese ever made and that the DB contains a very wide range of the pronunciation proficiency. Considering this uniqueness of the DB, the first author did two interesting researches[5, 6].

In [5], a large listening experiment was carried out, where six Americans listened to a part of the DB once per utterance and typed what they heard. Using the typing results, it was clarified what kinds of (combinations of) segmental, prosodic, and linguistic errors are more fatal to speech communication. Further, an automatic method was proposed to predict how likely each word in connected speech in the DB is perceived correctly by the six Americans. Human teachers, 3 Japanese and 4 Americans, were also requested to predict the typing accuracy by listening to and looking at the intended sentences. The machine prediction was very comparable to the best performance of the four American teachers. It was commonly found in machine prediction and in Americans' prediction that very intelligible and very unintelligible pronunciations are rather easy to predict. In Japanese prediction, very different tendency was observed. For Japanese teachers, it was the most difficult to identify the pronunciations completely unintelligible to the six Americans as unintelligible. This result implies that Japanese teachers and students cannot recognize or perceive fully why students' pronunciations are not understood correctly. As is mentioned in [7], people of different languages have different ears. English education for oral communication should focus much more on perceptual differences between Japanese and English ears.

In [6], a novel method was proposed to describe individual students differently. Most of the CALL researches are based

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]
```

Table 5: 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_0097
         She knows you, doesn't she ?
         [SH IY1] [N OW1 Z] [Y UW1] [D AH1 Z AX0 N T] [SH IY1]
```

Table 6: Examples of sentences of various rhythm patterns with phonemic symbols and prosodic symbols

```
S1_0106  Come to tea with John.
        / +    -  +   -    @  /
         [K AH1 M] [T UW1] [T IY1] [W IH1 DH] [JH AA1 N]
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]
```

upon phonetics and phonology. But phonetics was born to describe phones and phonology was born to describe languages. Strictly speaking, they are not good sciences to describe individual students. From this viewpoint, a new method, which can be viewed as yet another speech science, was proposed to solve this problem. The method has no dimensions to represent static distortions inevitably involved in speech production / recording process. In other words, differences in age, shape, size, gender, microphone, room, and line cannot be seen in the new representation of speech. What can be seen there is only dependency of English pronunciation on speakers' mother tongues. 202 students in the DB were classified according to *types* of Japanese English which were defined based on the proposed method. This is completely impossible only with phonetics and phonology because they are not sciences for students originally.

7. Conclusions

In this paper, development of ERJ (English Read by Japanese) database was described, where main focus was placed upon pronunciation errors caused for lack of knowledge on correct articulation of English phones. 202 students were randomly selected and they were asked to read sentence and word sets repeatedly until they could do what they thought was the correct pronunciation. After collecting American speech samples with the same reading sets, a part of the utterances were rated by American teachers of English. The DB was already used in many CALL researches and two of them were briefly introduced. The authors hope that the DB could be of great help to researchers and teachers and provide students with better environment of learning English pronunciation. If readers have any interest in this DB, please email to eng-db@gavo.t.u-tokyo.ac.jp.

8. References

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

[2] Y. Tono, "The standard speaking corpus: a 1 million-word spoken corpus of Japanese," Proc. ASIALEX'2001 (2001)

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

[4] S. Nakagawa, "A survey on automatic speech recognition," Trans. Institute of Electronics, Information and Communication Engineers, vol.J83-D-II, 2, pp.433–457 (2000, in Japanese)

[5] N. Minematsu, C. Guo, and K. Hirose, "CART-based factor analysis of intelligibility reduction in Japanese English," Proc. EUROSPEECH, pp.2069-2072 (2003)

[6] N. Minematsu, "Yet another acoustic representation of speech sounds," submitted to ICASSP'2004

[7] A. Cutler, "Listening to a second language through the ears of a first," Interpreting, vol.5, no.1, pp.1–23 (2001)