

Development of English Speech Database Spoken by Japanese Learners

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

Due to recent advances in speech processing techniques, we can see a 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. While we can find lots of speech corpora available from distribution sites such as Linguistic Data Consortium[1] and European Language Resources Association[2], the number of speech corpora built especially for CALL system development is strongly restricted. In this paper, we introduce a national project of the CALL system development. One of the objectives of this project is to construct English speech database spoken by Japanese students. Here, we describe the purpose of the database development and the size, specifications, and recording strategies adopted for the database in addition to preliminary discussions and investigations done before the development.

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 non-native speakers of English in other Asian countries to do. To save this situation, 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[3][4],” financially supported by the Ministry of Education, Culture, Sports, Science and Technology. This project progressively promotes speech and natural language techniques into language education and learning.

Recent advances in speech technologies have made it possible to develop CALL systems especially

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 require large databases. For the development of speech recognizers of *native* languages, a large number of databases of various languages were already built and distributed worldwide. As for speech databases of non-native languages, 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 cost of the development is quite high. Moreover, most of the non-native speech databases found contain spontaneous speech only, e.g. Q & A style conversations[5][6] and free conversations on telephone line[7]. 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 not of spontaneous speech but of *read* speech by non-native speakers. It should be noted that speech recognition technologies are not mature enough to deal with even native speech adequately in the case that it is generated spontaneously[8]. Furthermore, it is easily assumed that acoustic variations and distortions found in speech are much larger in non-native speech than in native speech. These educational and technical necessities led us to build an English *read* speech database spoken by Japanese students.

In Section 2, preliminary discussions are shown on what kind of sentences or words should be read and so forth. After that, the specifications of the speech database are given in Section 3. In the recording, we adopted a unique strategy in order to select 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 you some examples of reading material. Finally, this paper is summarized in Section 6.

2. Preparations done for the database development

2.1. Databases required by the CALL system development

As told in the previous section, non-native speech contains larger acoustic and linguistic distortions than native speech does. The magnitude of these distortions is supposed to be determined by various factors such as the target language and the native one 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 variety of the distortions among students. 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 the non-native speech. In the case of the native language of students being Japanese and their target language being English, it is also quite 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 American English (GA).
- Learners are students of universities or colleges including their graduate schools.
- Focus is placed only on acoustic distortions. Linguistic distortions such as grammatical errors are not considered.
- 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 due to lack of knowledge on articulation for English pronunciation.

2.2. Outline of the database specifications

Preliminary discussions were done on recording conditions and reading material according to the syllabus of mastering English pronunciation.

Even if we follow the guidelines in Section 2.1, the acoustic distortions are still large compared with those in native speech. We can categorize the conditions of students' speaking English for pronunciation learning into several manners in terms of information or hints given to students. Since the acoustic distortions should partly be dependent on the manners, we selected one out 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 without any hint.
2. Students read words or sentences. In this case, text or spelling information is given.
3. Students read words or sentences with phonetic/prosodic symbols. In this case, phonetic/prosodic information is given *as text* in addition to spelling.
4. Students read words or sentences after hearing a model utterance from a teacher of English. In this case, acoustic information, both segmental and prosodic, is additionally given, which is so-called a "repeat-after-me" manner.

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

As for reading material, we considered the syllabus of mastering English pronunciation. Although we have a large number of matters which 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 phonetically-balanced sentence set, a sentence set including sequences of phonemes difficult for Japanese students to pronounce correctly, a phonetically-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 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 can have their stressed syllable at different positions in the words, a set of compound words, and so on were generated.

On the sheets of the reading material, as told above, phonetic symbols and/or prosodic ones were assigned if required. Before recording, we gave instructions to speakers so that they could understand these symbols correctly. For each of the phonetic symbols used here, we arranged so that speakers could hear a word example through the Internet which included the phonetic symbol and did not appear in the word sets. As for symbols of prosodic phrase break, which were a part of the prosodic symbols, speakers were allowed to make sure of how to realize the phrase break by hearing several sentence examples also through the Internet.

3. Specifications of the database

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

Phonetic symbols of TIMIT database[9] and those of CMU pronunciation dictionary[10] were used as reference sets. After modifying these sets, the phonetic symbols for the assignment were determined, which are listed in Table 1. Most of the English-Japanese dictionaries owned by Japanese students represent schwa sounds by more than one phonetic symbol, which are selectively used mainly according to the orthography. In the phonetic symbols 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 the word. In this case, we asked them to look up their own English dictionary. When assigning the phonetic symbols to words, their pronunciation of citation form was considered for each word.

As for the prosodic symbols, primary/secondary stress symbols, intonation symbols, and/or rhythm symbols were assigned if necessary. A number, one of 0, 1, and 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 one of a rising arrow, a falling one, a rising-falling one, and 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. Positions and levels of these symbols were determined by a native English teacher or by looking up textbooks of English. Readers can find examples of reading material with these symbols in Section 5.

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

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 these words, speakers were requested to pronounce a sequence of phonetic symbols which were assigned to each word (See Section 5). For sentence sets, we prepared two types of reading sheets for each of the sets. One was with phonetic symbols for every word, which was used

Table 1: Phonetic symbols assigned to sentences and words

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
phonetically-balanced words[12]	300
minimal pair words	600
MOCHA-TIMIT[11] phonetically-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

only in pronunciation practice before recording, and the other was without them, which was referred to during recording. The preparation of two types of sheets is because reading sentences by referring to phonetic symbols is expected to induce unnatural pronunciation. In other words, with the phonetic symbols for every 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 phonetic symbols was not expected here. This is because most of the words in word sets were short and plain except for the unknown words, while sentence sets sometimes had rare words.

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

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 which can have their stressed syllable at different positions in the words, compound words, and so on. 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 focused words differ between the two, 3) sentences with intonation patterns specific to their syntactic structure, 4) those with intonation patterns specific to 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 assigned to each syllable by an American teacher of English

based upon the principle that the phrase-final syllable tends to be with 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 structure. Section 5 shows some examples of the word sets and the sentence sets with phonetic/prosodic symbols.

4. Recording of speech samples

4.1. Selection of speakers

Selection of 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 speakers with rather good pronunciation ability. It should contain English speech by poor speakers as well as by good speakers. To realize the adequate selection, we requested each of the recording sites to select randomly Japanese students of the site and make them participate in the recording as speakers. Twenty organizations such as universities and colleges cooperated on the recording and English speech samples spoken by 200 Japanese students, 100 male and 100 female students, were collected. All the sentences in Tables 2 and 3 were divided into 8 groups and all the words were into 5 groups. The required amount of reading material per speaker was a sentence group (~120 sentences) and a word group (~220 words). Therefore, each sentence was read by about 12 different speakers and each word was read by about 20 different speakers for each gender.

4.2. Procedures of the recording

As told in Section 2.1, in the development of the database, neither acoustic distortions observed only in a particular student's utterances nor those observed only temporarily were considered. In other words, main focus was placed on the distortions found rather commonly in Japanese speaking of English. Besides, during the recording, there should be no unknown words for speakers because pronunciation errors due to lack of English vocabulary are another problem than errors due to lack of knowledge on articulation for English pronunciation.

To avoid the unwanted events shown above, the following procedures were adopted for the recording.

1. Before the recording, speakers were requested to have practice in pronouncing sentences and words on the given sheets. In the practice, they were allowed to refer to the reading sheets with phonetic 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 with this strategy of recording, many pronunciation errors were easily expected because of lack of knowledge on English pronunciation. If speakers failed in correct pronunciation three times, they were allowed to skip the material and go to the next one.

3. After the recording, each of speech samples was checked by members of the recording site. If they found any technical errors in some sentences or words, the recording was done again for them.

Through the recording procedures above, the database shall contain English speech samples which the speakers judged themselves that were correctly pronounced. Therefore, the remaining pronunciation errors are supposed to be purely because of lack of speakers' knowledge on English pronunciation.

5. Examples of reading material

Some examples of reading material are shown in the following pages. All the words in the examples are with phonetic 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 or rhythm patterns.

6. Conclusions

In this paper, a national project for developing an English read speech database spoken by Japanese students was introduced. In the development, main focus was placed upon pronunciation errors caused by lack of knowledge on articulation for English pronunciation. Two types of reading material were prepared. One is related to segmental aspect of English pronunciation and the other is to prosodic aspect. Two hundred speakers were randomly selected and they were requested to do practice in pronouncing sentences and words in given sheets before recording. During the recording, they were asked to read sentences and words repeatedly until they could do what they thought was the correct pronunciation. We believe that the database will progressively promote developing English CALL systems for Japanese learners.

7. References

- [1] <http://www ldc.upenn.edu/>
- [2] <http://www.icp.inpg.fr/ELRA/home.html>

Table 4: Examples of phonetically-balanced sentences with phonetic symbols and word stress symbols

S1_0051	Ambidextrous pickpockets accomplish more. [AE2 M B AXO D EH1 K S T R AXO S] [P IH1 K P AA2 K AXO T S] [AXO K AA1 M P L AXO SH] [M AO1 R]
S1_0052	Her classical repertoire gained critical acclaim. [HH ER1] [K L AE1 S AXO K AXO L] [R EH1 P AXRO T W AA2 R] [G EY1 N D] [K R IH1 T AXO K AXO L] [AXO K L EY1 M]
S1_0053	Even a simple vocabulary contains symbols. [IY1 V AXO N] [AXO] [S IH1 M P AXO L] [V OWO K AE1 B Y AXO L EH2 R IYO] [K AXO N T EY1 N Z] [S IH1 M B AXO L Z]
S1_0054	The eastern coast is a place for pure pleasure and excitement. [DH IH1] [IY1 S T AXRO N] [K OW1 S T] [IH1 Z] [AXO] [P L EY1 S] [F AO1 R] [P Y UH1 R] [P L EH1 ZH AXRO] [AE1 N D] [AXO K S AY1 T M AXO N T]
S1_0055	The lack of heat compounded the tenant's grievances. [DH AXO] [L AE1 K] [AH1 V] [HH IY1 T] [K AXO M P AW1 N D AXO D] [DH AXO] [T EH1 N AXO N T S] [G R IY1 V AXO N S AXO 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 AEO N S IH1 S K OWO] [IH1 Z] [W AH1 N EY1 TH] [AE1 Z] [P AA1 P Y AXO L AXO S] [AE1 Z] [N Y UW1] [Y AO1 R K]
S1_0062	Its extreme width was eighteen inches. [IH1 T S] [AXO K S T R IY1 M] [W IH1 D TH] [W AA1 Z] [EYO T IY1 N] [IH1 N CH AXO Z]
S1_0063	Who ever saw his old clothes ? [HH UW1] [EH1 V AXRO] [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 AXO NG] [Y UW1] [DH AXO] [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 phonetically-balanced words with phonetic 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 phonetic 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 AXO NG]	sailing	[S EY1 L AXO 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]

- [3] <http://www.nime.ac.jp/tokutei120/index.html>
- [4] S. Nakagawa, "The CALL project in Japan," Proc. EuroCALL, pp.23 (2000)
- [5] H. Isahara, T. Saiga, and E. Izumi, "The TAO speech corpus of Japanese learners of English," Proc. ICAME'2001 (2001)
- [6] Y. Tono et al., "The standard speaking corpus: a 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography," Proc. ASIALEX'2001 (2001)
- [7] <http://cslu.cse.ogi.edu/corpora/fae/index.html>
- [8] S. Nakagawa, "A survey on automatic speech recognition," Trans. Institute of Electronics, Information and Communication Engineers, vol.J83-D-II, no.2, pp.433-457 (2000, in Japanese).
- [9] <http://www ldc.upenn.edu/Catalog/LDC93S1.html>
- [10] <ftp://ftp.cs.cmu.edu/project/speech/dict>
- [11] <http://www.cstr.ed.ac.uk/artic/mocha.html>
- [12] J. P. Egan, "Articulation testing methods," Laryngoscope, vol.58, no.0, pp.955-991 (1948).

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


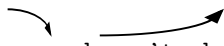
S1_0086	That's from my brother who lives in London. [DH AE1 T S] [F R AH1 M] [M AY1] [B R AH1 DH AXRO] [HH UW1] [L IH1 V Z] [AXO N] [L AH1 N D AXO N]
S1_0087	That's from my brother, who lives in London. [DH AE1 T S] [F R AH1 M] [M AY1] [B R AH1 DH AXRO] [HH UW1] [L IH1 V Z] [AXO N] [L AH1 N D AXO N]
S1_0091	 Cauliflower, broccoli, cabbage, sprouts, and onions. [K AA1 L AXO F L AW2 AXRO] [B R AA1 K AXO L IYO] [K AE1 B AXO JH] [S P R AW1 T S] [AE1 N D] [AH1 N Y AXO N Z]
S1_0094	Is this elevator going up or down ? [IH1 Z] [DH IH1 S] [EH1 L AXO V EY2 T AXRO] [G OW1 AXO 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 AXO N T] [SH IY1]

Table 9: Examples of sentences of various rhythm patterns with phonetic 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 IYO]
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 IYO] [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 [AXO] [D AA1 R K] [R UW1 M]	W1_0207	almond-eyed [AA2 M AXO N D AY1 D]
W1_0202	a darkroom [AXO] [D AA1 R K R UW2 M]	W1_0208	broad-minded [B R AO1 D M AY1 N D AXO D]
W1_0203	a light housekeeper [AXO] [L AY1 T] [HH AW1 S K IY2 P AXRO]	W1_0209	free-range [F R IY1 R EY1 N JH]
W1_0204	a lighthouse keeper [AXO] [L AY1 T HH AW2 S] [K IY1 P AXRO]	W1_0210	blue-black [B L UW1 B L AE1 K]
W1_0205	the brief case [DH AXO] [B R IY1 F] [K EY1 S]	W1_0211	forward-looking [F AO1 R W AXRO D L UH2 K AXO NG]
W1_0206	the briefcase [DH AXO] [B R IY1 F K EY2 S]	W1_0212	built-in [B IH1 L T IH1 N]