Measurement of Objective Intelligibility of Japanese Accented English Using ERJ (English Read by Japanese) Database

Nobuaki Minematsu[1], *Koji Okabe*[2], *Keisuke Ogaki*[1], *and Keikichi Hirose*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo
[2]NEC Corporation

{mine,ogaki,hirose}@gavo.t.u-tokyo.ac.jp, k-okabe@bx.jp.nec.com

10.21437/Interspeech.2011-310

Abstract

In many schools, English is taught as international communication tool and the goal of English pronunciation training is generally to acquire intelligible enough pronunciation, which is not always native-sounding pronunciation. However, the definition of the intelligible pronunciation is not easy because it depends on the speaking skill of a speaker, the predictability of a content, and the language background of a listener. One kind of accented pronunciation, which is intelligible enough for some listeners, is often less intelligible for others. This paper focuses on objective intelligibility of Japanese English through the ears of American English speakers with little exposure to Japanese English. A large listening test was conducted using ERJ (English Read by Japanese) database. A balanced subset of this database were presented over a telephone line to the American listeners who were asked to repeat what they heard. Totally, 17,416 repetitive responses were collected and they were transcribed manually. This paper describes the design of this experiment and some results of analyzing the results of transcription.

Index Terms: English pronunciation training, foreign accent, intelligibility, listening test, ERJ database

1. Introduction

We're living in the era of internationalization. For international language communication, we have no other choice than to use English. Last year, several large Japanese companies adopted English as official language even used between two Japanese in domestic offices. Since April this year, English lessons have been introduced to every public primary school in Japan.

Native-sounding vs. intelligible, this has been a controversial issue in discussing the goal of pronunciation training. The former is a sufficient condition to the latter and the latter is a requisite condition to the former. Today, in many schools, English is taught as international communication tool and the goal of English pronunciation training is generally to acquire intelligible enough pronunciation [1]. In this case, however, we can point out two critical problems. The first one is the difficulty of defining the intelligible pronunciation and the second one is, in a classroom situation, the difficulty for learners to know how intelligible their utterances are and what kind of mispronunciations can be accepted or cannot be accepted by listeners.

Generally speaking, the intelligibility of a given non-native English utterance depends on the speaker's pronunciation skill, the predictability of the content, and the language background of listeners [2]. Here, we ignore the second factor because it is independent of pronunciation training. Dependency of the intelligibility on listeners indicates that one kind of accented pronunciation, which is intelligible enough for some listeners, can be unintelligible for others. The most intelligible pronunciation of a specific listener is the listener's own pronunciation. For many Japanese, the most intelligible English will not be British/American English but Japanese accented English [3]. Logically speaking, if a native speaker of English wants to give an intelligible lecture to Japanese, he/she may have to master Japanese English before the lecture. We can say that it is difficult to discuss the intelligibility of utterances without specifying the language background of listeners. In this paper, we focus on the intelligibility of English sentences spoken by Japanese to Americans with no experience of talking with Japanese.

In a classroom, learners often repeat what English teachers say or some spoken materials recorded by native speakers. In many cases, however, learners don't know well whether their utterances can be understood correctly by listeners. This is partly because what learners are given is always *authorized* pronunciation, i.e. teachers' good pronunciation. They are given what to do but not given what not to do. Good pronunciation is needed for learners to try to imitate but we consider that inadequate pronunciations can be a good information source for learners to improve their pronunciation only if the following two conditions are satisfied. One is that explicit description on why that pronunciation is inadequate or which words are misunderstood has to be shown to learners and the other is that the inadequate pronunciations have to have a high coverage of pronunciation errors that can be observed in the same group of learners, i.e. Japanese learners. Recently, many research efforts were given to develop machines to detect pronunciation errors using automatic speech recognition (ASR) techniques [4]. Using these machines, a learner can check which words of his/her utterances were inadequate but it will be difficult to know about some other errors which may be made afterward by the learner. Here, as told above, we can say that inadequate pronunciations made by others can be a good information source. If collection of these pronunciations satisfying the above two conditions is possible, these data will be helpful both for learners and teachers.

To this end in this paper, a large listening test was conducted, where 173 Americans with no experience of talking with Japanese listened to 800 English utterances made by 100 male and 100 female Japanese. The Americans were asked to repeat what they just heard and their repetitive responses were transcribed manually. The speech materials were selected from ERJ database [5, 6], which is a large database of Japanese English. After the listening test, we obtained 17,416 transcriptions (repetitions) and an utterance gave us 21 transcriptions on average, indicating that an utterance was heard by 21 Americans on average. The following sections describe the design of this experiment in detail and some results of analyzing the transcriptions, which will be included in the next release of ERJ.

Table 1: Word and sentence sets for the segmental aspect

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 2: Word and sentence sets for the prosodic aspect

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

2. ERJ database

In the listening test, we had to present speech stimuli that contained a wide enough variety of pronunciation errors. To satisfy this condition, we defined a subset of utterances from those in ERJ database. In this section, we describe ERJ database briefly.

2.1. Selection of reading material

Syllabuses of English pronunciation training is mainly divided into two aspects; segmental and prosodic. As for reading material, sentence sets and word sets are used for both aspects, shown in Table 1 and Table 2. On reading sheets, phonemic/prosodic symbols are assigned to each word to facilitate recording procedures, namely, learners did not have to look up an English dictionary for recording. Exactly speaking, this database does not reflect the learners' *true* pronunciation or reading proficiency directly because the reading sheets include many hints for pronunciation, i.e. phonemic/prosodic symbols.

2.2. Selection of speakers

To realize a balanced selection of speakers, 100 male and 100 female university students are randomly selected at twenty institutes all over Japan. All the sentences in Table 1 and Table 2 are divided into eight groups and all the words in the tables are into five groups. The required amount of recording per speaker is a sentence group ($\sim$120 sentences) and a word group ($\sim$220 words). Each sentence is read by twelve speakers and each word is read by twenty speakers for both genders. The total number of sentence utterances is about 24.7K and that of word utterances is about 45.5K. Besides the Japanese learners, eight male and twelve female General American speakers read the material. One speaker reads a half of sentences of all the sets ($\sim$480 sentences) and a half of words of all the sets ($\sim$550 words).

2.3. Recording Japanese utterances and American ones

Before the recording, Japanese learners are allowed to practice reading sentences and words on their reading sheets. In the recording, they are asked to read the sentences and words repeatedly until they can do what *they think* is correct pronunciation. Then, the resulting database is a collection of English utterances judged as correct by Japanese learners. As for recording American utterances, no special instruction is given and they read the sentences and the words in a normal speaking rate.

2.4. Rating Japanese learners' pronunciation

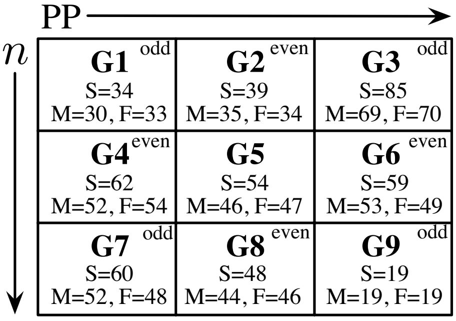
To a part of sentence utterances ($\sim$3.8K utterances) and word utterances ($\sim$5.7K utterances), five-scale goodness scores of pronunciation are assigned by five American teachers of English, who have good experience of teaching English to Japanese and good knowledge of phonetics. Here, 1 and 5 mean very poor and very good, respectively. Since the reading material is divided into several groups (see Table 1 and Table 2), the utterances are rated using different strategies depending on which group each utterance belongs to. For example, for the phonemically-balanced sentences, the teachers rate them based on whether indented phonemes can be perceived well.

3. Selection of Japanese utterances from ERJ for the listening test

Since it is practically impossible to present all the utterances in ERJ database, we had to select a subset of the utterances in the database. To measure the intelligibility of Japanese accented English utterances objectively, these utterances should include a large enough variety of errors including linguistic ones such as grammatical errors. Since ERJ database contains only read speech, however, it comes to have no linguistic error. Further, the sentences designed for prosodic variation tend to be those with simple syntactic structure. If they are used for the listening test, due to syntactic simplicity, the obtained intelligibility will be biased. On the other hand, although the sentences of the TIMIT-based phonemically-balanced sentence set also have no linguistic error, to achieve a high balance, they include rather rare words and phrases including proper names. These are considered as somewhat unnatural wording examples. For the listening test, we decided to define a subset of sentences by selecting sentences from the phonemically-balanced set.

First, we selected some sentences from the 460 sentences and, to each of the selected sentences, we adequately assigned a Japanese learner. For sentence selection, the following two linguistic parameters were used: 1) the number of words in a sentence and 2) perplexity of that sentence calculated using bigram language models trained with three years' news articles included in WSJ database. Here, the vocabulary was defined as the most frequent 65K words in the articles plus 117 words to have no unknown word in the 460 sentences. According to the number of words of a sentence (n), the 460 sentences were divided into three groups of a ($n \leq 6$), b ($7 \leq n \leq 8$), and c ($9 \leq$n). Their sizes are 158, 175, and 127 for a, b, and c. Similarly, according to perplexity (PP), the 460 sentences were divided into another three groups of A (PP$\leq$2000), B (2000$<$ PP $\leq$7000), and C (7000 $<$PP). Very high values of PP are because 1) we have no unknown word and 2) the domain of sentences in the TIMIT set is different from that of news articles. The sizes of A, B, and C are 156, 141, and 163. By combining these two parameters of n and PP, we had 9 groups of A-a, B-a, C-a, ..., A-c, B-c, and C-c, which are referred to as groups 1 to 9 hereafter. The average number of sentences in a group was 51.1.

For speaker assignment, the following strategy was taken. By referring to the goodness scores of phoneme pronunciation included in ERJ database, the 100 male speakers and the 100 female speakers were separately sorted and we obtained even-numbered 50 speakers and odd-numbered 50 speakers for each gender. A speaker in the even-numbered male speakers was assigned to a sentence of group m (m=2, 4, 6, and 8). Namely, he was assigned to four sentences. A speaker in the odd-numbered male speakers was assigned to a sentence of group l (l=1, 3, 7,



G1 odd
S=34
M=30, F=33
	G2 even
	S=39
	M=35, F=34
		G3 odd
		S=85
		M=69, F=70

G4 even
S=62
M=52, F=54
	G5
	S=54
	M=46, F=47
		G6 even
		S=59
		M=53, F=49

G7 odd
S=60
M=52, F=48
	G8 even
	S=48
	M=44, F=46
		G9 odd
		S=19
		M=19, F=19

S = #sentences in the group
M = #selected sentences from the group for male
F = #selected sentences from the group for female

Figure 1: Speaker assignment to each sentence group

and 9). Figure 1 shows this strategy schematically. Similar assignment was done for female speakers. This assignement strategy was not always able to be carried out because the sizes of some groups were less than 50. In this case, neighboring groups or group 5 were used instead. It should be noted that, for each gender, a particular sentence was assigned only once. Finally, each of the 100 male and the 100 female speakers was assigned to four sentences[1]. Totally, we obtained 400 utterances from the 100 male speakers and another 400 utterances from the 100 female speakers. The sentence overlap between the two sets of 400 utterances is 381. As shown in Figure 1, we designed this balanced subset very carefully.

For American speaker assignment, the following procedure was carried out. From the overlapped 381 sentences, 100 sentences were manually selected so carefully that the selected sentences were distributed reasonably evenly for the 9 sentence groups. During sentence selection, speaker assignment was also done in such a way that gender ratio (M:F) in a sentence group was 4:6. This is because ERJ database has eight male and twelve female American speakers. The number of sentences per American speaker was five ($5 \times (8 + 12) = 100$).

4. The listening test

4.1. Selection of subjects

This listening test was conducted at Indiana University. The subjects satisfying the following conditions were collected.

- The subject's mother tongue is American English.
- He/she has no hearing problem.
- He/she has had no experience of talking with Japanese.

173 subjects were collected and their average age was 20.5. Eighty percent of the subjects were from the State of Indiana.

4.2. Power normalization and white noise addition

Power normalization was done for all the stimuli because ERJ's utterances were recorded in different sites, which resulted in providing utterances of different average power.

For American utterances, in addition to clear utterances, we prepared noisy ones. Here we prepared utterances with different signal-to-noise ratios (SNRs). SN=-5, -2.5, 0.0, 2.5, 5.0, and ∞. ∞ means clear (noise-free) utterances. Totally, we prepared $6\times100=600$ utterances of American English. The reason of preparing noisy utterances is that we were interested in to which value of SNR the Japanese accent is equal in terms of intelligibility, i.e. word identification rate.

Table 3: #speakers for each group of pronunciation goodness

score	≤2.0	≤2.5	≤3.0	≤3.5	≤4.0	≤4.5	≤5.0
male	2	27	43	16	5	0	2
female	0	8	36	25	19	7	0

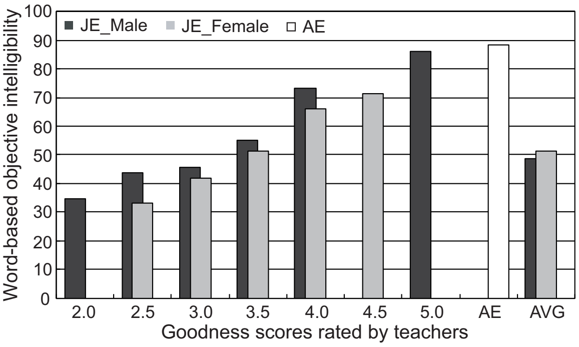


Figure 2: Word-based intelligibility for different learner groups

4.3. Procedure of the listening test

The total number of utterances to be presented is 1,400 (400 male Japanese utterances, 400 female Japanese utterances, and 600 American utterances with/without noise). Out of these, randomly selected 175 utterances were presented to a subject, where more than one utterance of the same sentence were not presented to that subject. To facilitate presentation, recording, and transcription, this listening test was done through a telephone line. A subject makes a call to the designated site and this starts the listening test. After the subject declared his/her ID and answered several questions, 175 utterances were presented. Each one was presented only once. The task was to repeat what they just heard. The repetitive response was monitored automatically on the other side and when the end of the response was detected, the next utterance was presented. All the responses were transcribed by experienced transcribers. Here, involuntarily spoken utterances such as "I don't know" or filled pauses were also transcribed. The total amount of required time for a subject was approximately 30 minutes. 173 subjects joined this listening test and we obtained 17,416 transcriptions for Japanese accented utterances and 12,859 transcriptions for American utterances with/without white noise.

4.4. Discussion on the obtained trascriptions

For each of the Japanese utterances and the American utterances, 21 American subjects repeated it, resulting in 21 transcriptions on average. For each transcription, we counted the number of correctly transcribed words semi-automatically, where errors only by conjugational suffix or by singular/plural form were treated as correct. A word-based intelligibility score (word identification rate) was calculated for each utterance.

By referring to the goodness scores in ERJ database, the 100 male Japanese and the 100 female Japanese were clustered into seven groups, shown in Table 3. For each group, the average intelligibility score was calculated, which is shown in

[1] A few speakers were assigned to three sentences and another few speakers were to five sentences. The reason for that is not specified here because it is trivial.

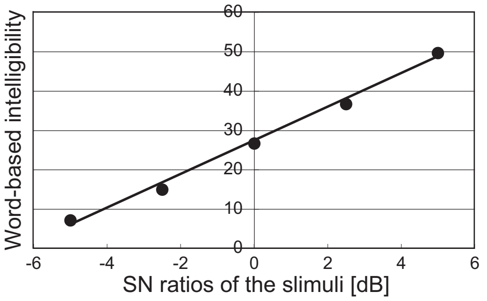


Figure 3: Word-level intelligibility for noisy speech

Figure 2. In this figure, the intelligibility of American utterances with no noise is also plotted. Their intelligibility is not perfect but it is 89.2%. This is considered to be because, as described in Section 3, the phonemically-balanced set includes some rare words and phrases and, in the listening test, two consecutive utterances are contextually independent.

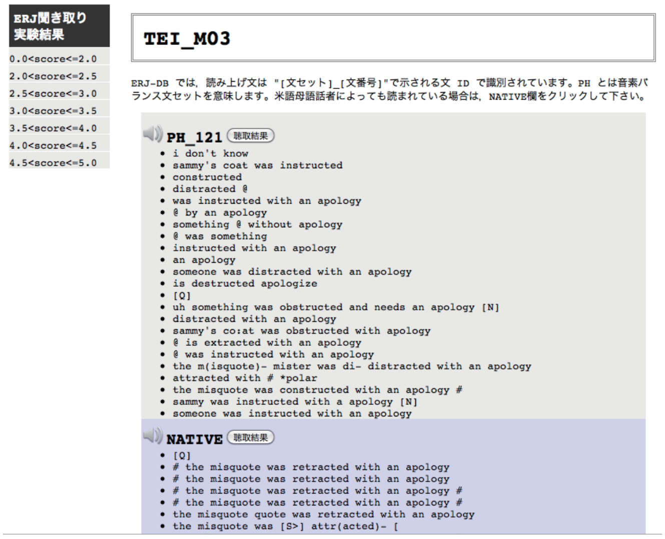
As for Japanese learners, clear linear relation is observed between the goodness scores in ERJ and the word-based objective intelligibility scores. While the intelligibility of learner group of 5.0 is almost the same as that of native speakers, learners' intelligibility of group of 2.0 is 34.6%. The average intelligibility is 48.8% and 51.2% for male and female learners.

All the utterances contained in ERJ database are those judged as correct pronunciation by learners themselves. Contrary to their expectations, these results indicate that half of the words are not understood correctly on average by native speakers of American English with little exposure to Japanese English. It is true that the experimental condition of this listening test is artificial in that no sentence-level contextual information is available and some rare words and phrases exist to achieve a good phonemic balance. But we can claim that the fact of very low intelligibility of Japanese English pronunciations, which are correct at least to learners, should be reflected seriously on how English pronunciation should be taught to Japanese learners. Not only by hearing and repeating teachers' pronunciations, but also by accessing to unintelligible pronunciations and understanding why they are unintelligible, learners' pronunciation may be improved. Based on this consideration, we developed a web-based system for learners and teachers to listen to every utterance used in the listening test and read every transcription obtained in the test. This system will be described shortly.

Figure 3 shows word-based intelligibility for noisy American utterances. Here, linear relation is observed again and the SNRs corresponding to 49.8% and 52.0% are 5.3 [dB] and 5.8 [dB]. These results imply that the Japanese accent is equal to about 5.5 [dB] white noise addition in terms of intelligibility.

5. Development of a transcription browser

For easy access to the many raw facts of unintelligible pronunciations and their transcriptions made by Americans, we developed a web system for browsing. In the system, following Table 3, the learners are clustered into seven groups and a user can select which learner group to browse. In the web page of the specified group, a list of the learners of that group is shown and by selecting a learner, a user can hear the utterances of that learner and read the transcriptions obtained from these utterances. Figure 4 shows a part of the page of learner TEI_M03. The intended sentence is also shown in the page.


ERJ聞き取り
実験結果

- 0.0<score<=2.0
- 2.0<score<=2.5
- 2.5<score<=3.0
- 3.0<score<=3.5
- 3.5<score<=4.0
- 4.0<score<=4.5
- 4.5<score<=5.0

TEI_M03

ERJ-DB では，読み上げ文は "[文セット]_[文番号]"で示される文 ID で識別されています。PH とは音素バランス文セットを意味します。米語母語話者によっても読まれている場合は，NATIVE欄をクリックして下さい。

PH_121 聴取結果

- i don't know
- sammy's coat was instructed
- constructed
- distracted @
- was instructed with an apology
- @ by an apology
- something @ without apology
- @ was something
- instructed with an apology
- an apology
- someone was distracted with an apology
- is destructed apologize
- [Q]
- uh something was obstructed and needs an apology [N]
- distracted with an apology
- sammy's co:at was obstructed with apology
- @ is extracted with an apology
- @ was instructed with an apology
- the m(isquote)- mister was di- distracted with an apology
- attracted with # *polar
- the misquote was constructed with an apology #
- sammy was instructed with a apology [N]
- someone was instructed with an apology

NATIVE 聴取結果

- [Q]
- # the misquote was retracted with an apology
- # the misquote was retracted with an apology
- # the misquote was retracted with an apology #
- # the misquote was retracted with an apology #
- the misquote quote was retracted with an apology
- the misquote was [S>] attr(acted)- [

Transcriptions of TEI_M03's utterance of sentence PH_121 and a native speaker's utterance of the same sentence.

Figure 4: The web page for TEI_M03

6. Conclusions

This paper described the aim and the design of a very large listening experiment, where Americans repeated Japanese accented English utterances immediately after hearing them. Their responses were transcribed. The speech stimuli were selected very carefully from ERJ database, paying much attention to balanced coverage of the pronunciation errors included in the database. The obtained transcriptions show a very low intelligibility of Japanese English to Americans with little exposure to Japanese English. In the transcriptions, we can find a huge number of facts of miscommunications, which are expected to show what not to do for Japanese to speak to Americans. We hope that these facts will be effectively used in a classroom or a self-learning situation to improve communication ability of Japanese learners. All the transcriptions and the browsing system will be included in the next release of ERJ database. We are also planning to include phonetic transcriptions of the 800 Japanese English utterances in the new release.

7. Acknowledgements

We are very grateful to Jared Bernstein and Masaki Suzuki for their support in collecting data at Indiana University.

8. References

[1] D. Crystal, "English as a global language," Cambridge University Press, New York, 1995.

[2] J. Bernstein, "Objective measurement of intelligibility," Proc. ICPhS, pp.1581–1584, 2003.

[3] M. Pinet *et al.*, "Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction," Proc. Int. Workshop on Second Language Studies (L2WS), 2010.

[4] M. Eskenazi, "An overview of spoken language technology for education", Speech Communication, 51, 10, 832–844, 2009.

[5] N. Minematsu *et al.*, "Development of English speech database read by Japanese to support CALL research," Proc. Int. Conf. Acoustics, 557–560, 2004.

[6] N. Minematsu *et al.*, "Development of English speech database read by Japanese and Americans for CALL system development," Journal of Japan Society for Educational Technology, 27, 3, 259–272, 2004 (in Japanese).