社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

日本人英語音声に対する米国語聴取者による誤認識の予測

郭　長深[†]　峯松　信明[†]　広瀬　啓吉[††]

† 東京大学大学院 情報理工学系研究科
〒 113-0033 東京都文京区本郷 7–3–1
†† 東京大学大学院 新領域創成科学研究科
〒 113-0033 東京都文京区本郷 7–3–1
E-mail: †{kaku,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　幅広い英語習熟度をカバーしている日本人読み上げ音声 DB 中から、習熟度、文長、言語的複雑さに関するバランスを考慮した上で音声試料を選定し、米国人に聴取、書き取らせるという実験を行なった。得られた書き取り結果を用いて、どのような発音エラー（の組み合わせ）が母語話者の誤認識、誤理解を誘発するのかについて分析を行なった。即ち、音声・言語処理技術を用いて各種の分節的、韻律的、言語的パラメータを抽出し、これらのパラメータを説明変数として単語単位の書取り率を CART 法を用いて予測した。と同時に、日本人英語教師による書取り率予測も行ない、本タスクの困難さについても検討した。自動予測実験の結果、予測に最も寄与するパラメータとして（上位三位まで）、強勢シラブルの規則性、言語的容易さ、音素生成の適切さ、が得られた。

キーワード　日本人英語、聴取実験、誤聴取、CART 法、リズム、親密度、音素尤度

Prediction of American listeners' misrecognition of English words spoken by Japanese

Changchen GUO[†], Nobuaki MINEMATSU[†], and Keikichi HIROSE[††]

† Graduate School of Information Science and Technology, University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
†† Graduate School of Frontier Sciences University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
E-mail: †{kaku,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　This study tries to automatically estimate the probability of individual spoken words of Japanese English (JE) being perceived correctly by American listeners and to clarify what kind of combinations of segmental, prosodic, and/or linguistic errors are more fatal to the correct recognition. Firstly, from a large speech database of JE, a balanced set of 360 utterances of 90 male speakers were selected. Secondly, a listening experiment was done where 6 Americans were asked to transcribe these 360 JE utterances. Next, using speech and language technology, values of many segmental, prosodic, and linguistic attributes of words, which are related to the pronunciation errors, were automatically calculated from the JE utterances. Finally, relation between the misrecognized words and the attribute values were analyzed with Classification And Regression Tree (CART) method to automatically predict the probability of each of the JE words being correctly transcribed. The prediction performance was compared with the human prediction performance which was obtained by another experiment, where a Japanese teacher of English was requested to estimate the probability by hearing the utterances while looking at the intended sentences.

Key words　Japanese English, misrecognition, listening test, CART, rhythm, familiarity, phoneme likelihood

1. Introduction

Recently in the foreign language education, students' ability to engage in meaningful conversational interaction in the target language has been more and more focused. Previously, the pronunciation learning often put the emphasis on acoustic similarity between students' pronunciation and native speakers' one. But most of the cases, the students are anxious of whether their English can be understood well, not of how acoustically close their pronunciation is to native speakers' one. Based on these considerations, we believe that a method to automatically estimate the intelligibility of the pronunciation should be devised with speech and language technologies. There are many factors including segmental, prosodic and linguistic ones that influence the intelligibility and the misrecognition will be caused by a combination of various factors. What kind of combination of the factors are more related to the intelligibility of the pronunciation ? If this question is solved, it will be very useful in the pronunciation learning. Computer-Aided Language Learning (CALL) has played a great role in helping Japanese students learning English. But most of the current CALL systems are built based on the analysis of one or two independent factors and can never tell the learners whether their utterances can be understood correctly or not. We think this kind of CALL systems are insufficient and need the improvement.

In this paper, a CART-based method of predicting how probably individual JE words are correctly transcribed by Americans is described, where values of segmental, prosodic, and linguistic attributes of the utterances were used. Experiments showed that the prediction performance is much higher than the human prediction performance, which was obtained from a Japanese teacher of English.

2. Some issues on pronunciation training

2. 1 Foreign accented vs. intelligibility

What kind of pronunciation should be pursued in foreign language learning ? In English education in Japan, the criterion seems to have been changed from reducing the Japanese accented pronunciation to gaining the intelligibility of pronunciation. The first criterion is a sufficient condition to the second one, which is a requisite condition to the first one. Many speech applications were developed to automatically rate the pronunciation proficiency. But every one is based upon acoustic matching between (quasi-)native acoustic models and learner's utterances and this strategy is for the first criterion because only the acoustic matching cannot separate the two criteria. The authors tentatively developed a method of estimating the intelligibility of pronunciation without any acoustic matching [1]. This method uses knowledge of the entire structure of English vocabulary and evaluates which phonemes should be clearly separated in the learner's acoustic space of pronunciation to reduce the confusedness. This method, however, only focuses on segmental features and cannot judge whether individual JE words are perceived correctly by native speakers of English.

2. 2 Differences in speech perception

Reading, writing, speaking, and listening abilities are said to be the four elemental abilities when learning a language. Communication ability is often added to them. To achieve the four abilities, learners take lessons of vocabulary, grammar, pronunciation, and listening. But especially when learning a foreign language which is phonetically and linguistically different on a large scale from the native one, the authors believe that there is another element which should be acquired even before the four abilities. The missing element is the perception. Several studies of language learning focused on the perceptual differences between learners and native speakers of the target language [2], [3]. They tried to induce a paradigm shift of capturing input speech from learners' ways to native speakers' ones. Knowledge on the native speakers' perception shall be effective to improve the intelligibility because the knowledge will instruct what kind of pronunciation errors are more fatal to them.

"Listen to me." This is a phrase repeated in class by teachers. But Japanese students don't know how to listen because their manner of perception is not adequate. "Repeat after me." This is another phrase repeated thousand times. But they don't know how to repeat because they don't know the perception of native speakers. The current work tries to provide the learners with the knowledge of what kind of pronunciation errors are hated by native speakers.

2. 3 Segmental, prosodic, or linguistic ?

Spoken language has various aspects, which can be divided into two ones; acoustic and linguistic. The former can be further divided into segmental aspect and prosodic one. "Segmental or prosodic ?" This is a well-known and still controversial issue in pronunciation learning. The intelligibility of the pronunciation can be interpreted as how easily the mental lexicon is accessed correctly with a given utterance. Factors of facilitating the lexical access is well discussed in studies of speech perception [5], [6]. These studies led the authors to the belief that what has to be discussed is not "segmental or prosodic" but "what kind combinations of segmental errors, prosodic errors, and linguistic errors are more fatal to native speakers' correct perception." Based on this belief, we defined pronunciation of the individual words as a set of values of the segmental, prosodic, and linguistic attributes. Using this definition, in this work, research focus was put on estimation of the probability of each spoken word of JE being correctly recognized by native speakers.

3. JE read speech database

JE speech database [7] was used in the trascribing experiment. All the utterances of the DB were made by Japanese learners' reading given sentence sheets. In this meaning, there are no grammatical or linguistic errors at all in the DB. However, the sentence set used in the experiment was a phonemically-rich set and, to achieve the richness, the set included rather rare words and phrases. These can be used as somewhat unnatural wording examples. The DB only contains speech samples which were judged by the speakers (learners) to be correctly pronounced but it still has a large number of pronunciation errors [8].

4. Transcription of JE speech samples by native speakers

4. 1 Selection of sentences and speakers

The DB contains about 24,000 sentence utterances of 100 male and 100 female speakers. Since it is impossible to type every utterance, a part of them should be adequately selected for the experiment. Out of several sentence sets in the DB, a phonemically-rich sentence set was selected, which has 460 different sentences. Out of the set, 360 sentences were unbiasedly selected according to how many words are in the sentence and perplexity of the sentence (how unpredictable the words are). For the sentence length, considering capacity of human STM (7 chunks), the sentences were divided into three groups, 1) less than 6 words, 2) 6 or 7 words, and 3) more than 7 words. As for the perplexity, we also prepared three groups, 1) less, 2) rather, and 3) more predictable. In other words, we prepared 9 subsets of about 40 sentences each, which varied in their linguistic complexity.

The DB contains the pronunciation proficiency of the individual speakers rated by five native English teachers. Referring to the rating results, the unbiased selection of speakers is also possible for each of the 9 sentence subsets. We selected 90 male speakers by excluding 10 with extremely high scores. Finally, we got 360 (90×4) speech samples.

4. 2 Measurement of typing ability of the subjects

In the transcribing (typing) experiment, the subjects were asked to write down what they just heard without any guessing. But no guessing during listening is strictly impossible. In order to prevent the subjects from deep guessing, we designed the experiment so that the minimum duration of typing should be given to the subjects according to length of the sentence and typing ability of the individual subjects. To realize this design, the ability of quick typing was measured for each subject in the following manner.

For each of given speech samples, length of the pause (T_p) was measured by a simple power threshold method. Using the length of the sentence (T_s) and T_p, the presentation interval from the end of the sentence to the beginning of the following one was set to

$$T = \alpha(T_s - T_p) - T_s,$$

where α is determined in advance according to the typing ability of the subject. The subject was allowed to start typing just after hearing the initial word of the sentence, and therefore, the duration allowed for typing the sentence was $\alpha(T_s - T_p)$. Using speech samples of *native* speakers, α was determined for each subject, ranging from 3.0 to 4.0.

4. 3 Transcription of JE speech

4. 3. 1 Subjects

6 Americans participated in the experiment. It is very interesting to analyze the typing results of English native speakers without any exposure to the Japanese language or JE speech. Since it is very hard to look for these people in Japan, however, we adopted subjects on a condition that their native language was American English and their stay in Japan was less than a year. 1 Canadian, who has never talked with Japanese, also took part in the experiment.

4. 3. 2 Procedures of the experiment

The control of the interval between the two consecutive sentence speech samples was described in section 4. 2. If this control is done for every sentence, it may result in increasing simple typing errors. To avoid this, we gave correction time to the subjects every three presentations of the sentences. Here, the time was provided as long as they wanted but they were strongly requested not to guess any additional words.

120 sets of 3 sentences were presented sequentially to the subjects through headphones, who were asked to write down with a PC what they heard. The transcriptions made by the subject will show us whether he/she recognized the individual words correctly. But it is still uncertain whether he/she received some meaningful linguistic content by hearing the utterance. So, we prepared another simple task, where the subjects were asked to indicate whether they had some questions on the utterance or not. The indication was done after each transcription by writing "X" when they had some and "O" when they had none.

Matching between the transcriptions and the sentence sheets prompted in the recording will give us the data of the misrecognized words. But we ignored the mismatches only by their word forms, walk and walked for example, although the number of the mismatches of this type was quite small. Finally, we got the data of the probability of the individual words being correctly recognized by Americans, ranging from 0/6 to 6/6 by a step of 1/6.

4. 4 Transcription of noisy Japanese utterances by Japanese

Results of the above transcription experiments will show how well Americans with some exposure to JE can recognize JE words correctly and a Canadian with no exposure can. However, if the rates of their performance are given, it is rather difficult for Japanese students to perceive the reduced intelligibility in the JE words. Here, we tried to obtain noisy Japanese speech samples which showed the equivalent performance to be recognized correctly. In this experiments, Japanese speech samples of various Signal-to-Noise (SN) ratios were presented to 18 Japanese students and they were asked just to transcribe them without deep guessing. 30 sentences were extracted from an ATR 503 phonetically-rich sentence set by considering sentence length and linguistic complexity. 6 male speakers were used to generate the speech stimuli, where those of 5 speakers were used in the transcription and those of the other one were used just as examples of noisy speech. The 30 sentences were divided into 3 sets, set-A to set-C, 10 sentences each. Every set included 2 sentences of each speaker. 3 SN ratios (0.0, –2.5, –5.0 dB) were used and they were assigned to the 3 sets. These procedures gave 6 different stimulus set, each including the same 30 sentences, spoken by the same 5 speakers and degraded by different SN ratios. Each of the 6 stimulus set was transcribed by 3 Japanese subjects. Results of this experiment will roughly show to what SN ratio "Japanese being" corresponds in speaking English.

5. Acoustic and linguistic analysis of the JE utterances

5. 1 Phoneme errors

All the JE utterances were time-aligned with a phoneme sequence arranged by referring to the prompted sentence and PRONLEX pronunciation lexicon. After that, the phoneme sequence was converted into a phoneme network to predict phoneme errors (replacement, deletion and insertion) of the pronunciation. The conversion rules needed deep knowledge of JE and were written by carefully considering characteristics of pronunciation errors found in JE. Recognizing the utterances with the phoneme network gives us the phoneme errors. The acoustic models used here were multi-mixture monophones trained with TIMIT database, where speakers with strong local accents or strong linking between phones were excluded although they were native.

5. 2 Stress errors

The resulting phoneme sequence obtained after the recognition was segmented into syllables by using a syllabification software named `tsylb`, which can syllabify an arbitrary sequence of phonemes. Then, each syllable was automatically judged whether it was stressed or not by using acoustic models of stressed syllables and unstressed ones [9], which were trained for each syllable group by using database of carefully spoken sentences in view of sentence stress. Coarse spectrum envelope, power, pitch, duration, and voicing degree

```
------- +1.000 +1.000  - silB [       0- 3600000]<-63.33> == silB [       0- 3600000]<-63.33>  silB match -
iris    -1.645 -1.645  S    Y [ 3600000- 5800000]<-60.60> ==    Y [ 3600000- 5700000]<-60.33> Y_cor match S
iris    -1.645 -1.645  -    r [ 5800000- 6100000]<-90.74> ==    y [ 5700000- 6200000]<-73.09> y_rep match -
iris    -1.645 -1.645  W    I [ 6100000- 7200000]<-69.31> ==    i [ 6200000- 7200000]<-58.44> i_rep match S
iris    -1.645 -1.645  -    s [ 7200000- 8000000]<-68.13> ==    T [ 7200000- 9300000]<-63.58> T_rep match -
------  +1.000 +1.000  - null [ 8000000- 8000000]< +0.00> == null [ 9300000- 9300000]< +0.00>  null match -
thinks  -4.292 -3.731  -    T [ 8000000- 9600000]<-64.39> ==    D [ 9300000- 9600000]<-72.58> D_rep match -
thinks  -4.292 -3.731  S    I [ 9600000-10000000]<-71.58> ==    i [ 9600000-10600000]<-58.34> i_rep match S
thinks  -4.292 -3.731  -    G [10000000-11300000]<-68.55> ==    G [10600000-11300000]<-76.30> G_cor match -
thinks  -4.292 -3.731  -    k [11300000-12400000]<-79.76> ==    k [11300000-12400000]<-79.76> k_cor match -
thinks  -4.292 -3.731  -    s [12400000-14300000]<-63.36> ==    s [12400000-14300000]<-63.36> s_cor match -
------- +1.000 +1.000  -   sp [14300000-23700000]<-56.24> ==   sp [14300000-23700000]<-56.24>    sp match -
```

Figure. 1 An example of the segmenal, prosodic, and linguistic analysis of a JE utterance

were used as the acoustic parameters for the modeling with different HMM topologies for different syllable groups. The syllable groups were designed based upon the structure of syllables, V, CV, VC, and CVC for example. The stress detection performance of the models was measured in the speaker-closed experiment and it was 96%.

5. 3 Linguistic unpredictability

Unpredictability of the individual words were estimated by using 1-gram and 2-gram language models trained with WSJ newspaper text corpus. 1-gram values can be used as rough estimates of familiarity of the words. As shown in [6], familiarity of a word is one of the main factors which influence the mental lexical access.

Figure 1 shows an example of the analysis. Values of 1-gram and 2-gram, lexical stress of the word, results of the time-alignment, results of the recognition with the phoneme network, classification of phoneme errors (replacement, deletion, or insertion), and results of the stress detection and so on are shown. In this analysis, no detection or judgment was done in terms of intonation. This is because most of the sentences were declarative ones and in this case, there is little difference in intonation between Japanese and English. As for speech rhythm, intervals between two consecutive stressed syllables, which were automatically detected, were used as a predicting factor.

6. Prediction of the misrecognized words with CART

6. 1 Preparation of predicting factors

Probability of the JE words being correctly recognized was estimated with CART method, where a decision tree was built with training data. A question is properly assigned to each node of the tree and answering the questions leads to a leaf node which indicates how probably the word is recognized correctly. In this experiment, the predicted factors are the probability of the JE words being correctly recognized by the 6 Americans which ranges from 0/6 to 6/6, and the predicting factors have to be prepared by using parameter values obtained in the acoustic/linguistic analysis.

Using the parameter values, various predicting factors were prepared, which were divided into three groups; segmental, prosodic, and linguistic factors. These factors can be categorized into four types from a different viewpoint; frame, phoneme, syllable, word, and sentence level. A sentence level factor was calculated for each sentence, and in this case, the unique value was assigned to every word in the sentence. Table 1 lists a set of the predicting factors used. The table also shows to which level each factor belongs.

Table. 1 Predicting factors prepared for CART

segmental factors	level
#phonemes	P
#vowels	P
#consonants	P
#vowel replacements	P
distance vector of vowel rep.	P
#vowel insertions	P
#vowel deletions	P
#cons. rep.	P
distance vector of cons. rep.	P
#cons. insertions	P
#cons. deletions	P
#mismatches	P
word-level likelihood	W
phoneme-level likelihood	P
averaged likelihood	F
prosodic factors	level
#stressed syllables	Sy
stressed syl. %correct	Sy
stressed syl. accuracy	Sy
#stressed syllables correctly produced	Sy
#rep. of stress with unstress	Sy
#rep. of unstress with stress	Sy
#inserted stressed syllables	Sy
#inserted unstressed syllables	Sy
word duration	W
averaged syllable duration	Sy
pause length before the word	W
pause length after the word	W
averaged stress-to-stress interval	S
variance of stress-to-stress intervals	S
linguistic factors	level
part of speech	W
position in the sentence	W
1-gram score	W
2-gram score	W

6. 2 Training of the decision trees

Transcriptions of 360 utterances (about 2,600 words) by 6 subjects gave us a large number of data of words recognized or misrecognized. Using the data, cross-validation was carried out to test the decision tree, where data of 89 speakers were used for training and those of the remaining 1 speaker

```
((interval_bunsan < 7.81918e+12)
((pause_nxt < 510000)
(((N 0) (0 0.0126582) (1 0.0202532) (2 0.0126582) (3 0.0481013)
 (4 0.0683544) (5 0.15443) (6 0.683544)  6))))
 (((N 0) (0 0) (1 0.0159574) (2 0.00531915) (3 0.0319149) (4 0.0265957)
 (5 0.12234) (6 0.797872) (7 0)(8 0) 6))))
 ((unigram < -1.6613)
 ((duration/syl < 1.18e+06)
 ((phoneme_udo < -348.952)
 ((duration/syl < 750000)
 (((N 0) (0 0.235294) (1 0.294118) (2 0) (3 0) (4 0) (5 0.176471)(6 0.29)
 (7 0)(8 0) 1))
 (((N 0) (0 0) (1 0.125) (2 0) (3 0) (4 0) (5 0.25) (6 0.625) (7 0) (8 0) 6)))
 (((N 0) (0 0.0714286) (1 0) (2 0.142857) (3 0) (4 0.0714286) (5 0.142857)
 (6 0.571429)(7 0)(8 0) 6)))
```

Figure. 2 An example of a part of the decision tree

were used for testing. By rotating the testing speaker, every speaker was used in testing. Each of the training words has its probability ranging from 0/6 to 6/6 and the distribution of the probability over the words is very biased, where words of 6/6 occupy 55 % of all the words. This bias sometimes causes an unexpected tree, with which all the words are judged to be recognized correctly without any question. To avoid this, besides the normal training method, we tentatively examined another tricky method of counting $n/6$ ($n < 6$) data more than once so that the distribution becomes unbiased. The problem of the biased distribution is attributed to the definition of the target function which should be maximized during training. Since we cannot change the definition adopted in the CART package [10], we tested the tricky method experimentally. An example of a part of the trained tree is shown in Figure 2. Estimation of the probability was done with different sets of the predicting factors. The experimental conditions are shown in Table 2. As for the performance measurement, recall and precision factors were calculated by ignoring estimation errors by ±1/6.

6. 3 Prediction by a human English teacher

In order to compare the prediction performance of CART method with that of human teachers, a listening test was carried out. 1 Japanese teacher of English joined this experiment so far. Each of the 360 JE utterances was presented to the teacher and she was asked to listen to it without looking at the intended sequence of words. After that, she read the intended words and rated each of the words in terms of how probably the individual words are correctly transcribed by Americans. Rating was done with a 7-level scale, ranging from 0 to 6. The teacher was allowed to listen to the JE utterances as many times as she wanted. But the first listening was done without looking at the intended sentence.

Table. 2 Experimental conditions

CASE-1	only with segmental factors
CASE-2	only with prosodic factors
CASE-3	only with linguistic factors
CASE-4	only with acoustic factors
CASE-5	with all the factors

6. 4 Discussions

Performance of the 6 Americans' correct transcription is shown in Table 3. The table shows the performance separately for proficiency levels of the speakers and also shows the rate of "X", indicating that the listeners have something uncertain on the utterances. It is interesting that speakers of ∼3 and ∼3.5 have almost the same performance of their *words*' being correctly recognized but there is a significant difference between their rates of "X". It implies that speakers of the higher levels should have better skills for meaningful speech communication. The average performance of word-level transcription is 79.3% for the case of the 6 Americans and 68.7 % for the case of the one Canadian. Figure 3 shows the performance of word-level transcription of noisy Japanese utterances. From this figure, we can roughly estimate the SN ratios indicating the performance of 79.3 % and 68.7 %, which are -1.2 dB and -3.3 dB. These imply that "Japanese being" corresponds to adding -1.2 dB noise when talking with native speakers with some exposure to JE and -3.3 dB noise when talking with those without it.

Table 4 shows recall and precision rates in various conditions. C-1 to C-5 show the results of five conditions of Table 2. The baseline (B·L) is chance-level performance, which was calculated by assuming that the estimation was done randomly. In this calculation, the ignorance of the mismatch by ±1/6 was considered. The table shows that the CART performance naturally and strongly depends upon the biased distribution of the probabilities over the training data and the falling tendency from 6/6 to 0/6 is clearly found. Therefore, although the highest performance is achieved in CASE-5 with all of the segmental, prosodic, and linguistic factors, recall rates of 0/6 to 4/6 are smaller than the chance-level performance. C-5' shows the results of the tricky tree

Table. 3 Performance of the transcription

prof. level	#spk.	#uttr.	%correct	rate of X
∼2	2	16	64.1%	83.3%
∼2.5	27	216	75.4%	56.7%
∼3	38	304	82.3%	44.7%
∼3.5	21	168	83.4%	33.7%
∼4	2	16	91.3%	20.8%

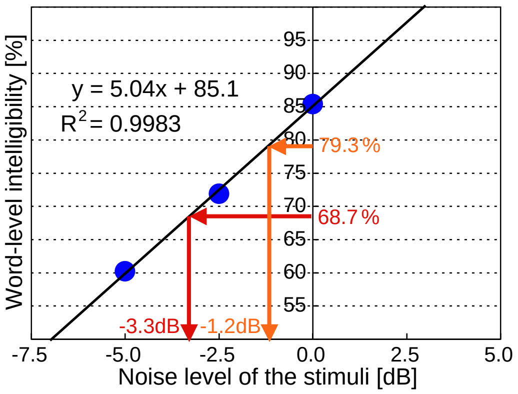

Figure. 3 Word-level intelligibility for noisy Japanese utterances

Table. 4 Prediction performance [%]

		0/6	1/6	2/6	3/6	4/6	5/6	6/6
C-1	recall	9.6	10.6	11.1	9.3	25.8	95.4	97.4
	prec.	34.2	42.9	41.7	51.7	59.4	76.9	75.4
C-2	recall	15.9	4.7	3.7	18.6	33.5	96.3	95.6
	prec.	37.8	100	61.5	51.7	60.0	70.3	76.6
C-3	recall	15.9	21.2	12.0	17.0	28.6	96.0	96.7
	prec.	43.9	48.2	58.1	50.0	57.2	64.7	78.9
C-4	recall	15.9	11.7	18.5	17.8	27.4	95.2	96.4
	prec.	42.8	53.6	44.7	54.7	53.8	81.0	76.0
C-5	recall	25.5	27.0	20.4	17.8	32.9	95.1	96.1
	prec.	42.3	53.8	51.9	62.5	56.6	79.7	79.3
C-5'	recall	67.8	85.7	84.2	75.0	71.7	75.9	59.7
	prec.	38.2	46.1	28.3	44.2	50.0	95.8	93.6
H	recall	2.5	5.8	6.8	13.3	10.1	95.8	93.6
	prec.	12.5	40.9	24.1	41.8	24.7	74.2	73.8
B·L	recall	28.5	35.4	43.3	43.8	42.9	43.5	29.8
	prec.	7.1	11.0	15.6	22.6	33.0	79.4	73.4

training, where the unbiased problem of training data is artificially solved. The performance of this tricky training is significantly higher than that of the baseline both in terms of recall and precision. Data preparation for training the tree has to be carefully designed.

The table also shows the human performance as 'H' and it can be said that it is very difficult for a Japanese teacher of English to rate the intelligibility of English spoken by Japanese. Especially, for the cases of recall rates of 0/6 to 4/6, the performance is lower than the chance-level. It implies that she did binary judgment, good or bad, and not quantitative judgment, *how* good or bad. However, only one Japanese teacher of English participated in the experiment so far and the reliability of the data is still low. We are collecting data of the human performance separately for the cases of native teachers and non-native teachers.

The CART package showed some dominant questions for the decision. Here, the most dominant one was obtained as "variance of stress-to-stress intervals" even though it is sentence-level attribute, the second was on "1-gram score", and the third was on "phoneme-level likelihood". These results imply the following. Rhythmical pronunciation is the most important key for high intelligibility. Next, easy and plain wording should be learned. Lastly, correct pronunciation of the individual phones should be acquired.

7. Conclusions

In this work, the intelligibility of pronunciation, not the acoustic similarity to native pronunciation, was strongly focused and acoustic and linguistic factors reducing the intelligibility were examined through listening and transcribing experiments. Using the obtained transcription, CART analysis was done to automatically predict how probably each word of the JE utterances can be transcribed correctly. The prediction was also done by a Japanese teacher of English. Results showed that the CART method trained with unbiased data is much better than the human teacher in terms of the prediction performance. Further, we roughly estimated the SN ratios reducing the transcription performance to the performance observed in native listeners' transcribing JE utterances. Results imply that "Japanese being" corresponds to adding -1.2 dB noise when talking with native speakers rather familiar with JE and -3.3 dB noise when talking with those unfamiliar. As future works, we are planning to collect more human prediction data separately for two cases of native and non-native teachers and to pursue some other effective parameters to improve the prediction performance.

References

[1] N.Minematsu, *et al.*, "Corpus-based analysis on pronunciation and perception of Japanese English in view of the entire structure of English phonemes and vocabulary." Proc. Autumn Meeting of Phonetic Society of Japan, pp.97–102 (2002, in Japanese).

[2] T.Otake, *et al.*, "Phonological units in speech segmentation and phonological awareness," Proc. ICSLP'98, pp.2179–2182 (1998).

[3] K.Tajima, *et al.*, "Perceptual learning of second-language syllable rhythm by elderly listeners," Proc. ICSLP'02, pp.249–252 (2002).

[4] S. Amano, *et al.*, "Estimation of mental lexicon size with word familiarity database," Proc. ICSLP'1998, pp.2119–2122 (1998).

[5] N.Minematsu, *et al.*,"The influence of semantic and syntactic information on spoken sentence recognition,"Proc. ICSLP'92, pp.153–156 (1992).

[6] N.Minematsu, *et al.*,"Role of accent nuclei and word familiarity in accelerating Japanese word recognition,"Proc. ICSP'99, pp.601–606 (1999).

[7] N.Minematsu, *et al.*,"English speech database read by Japanese learners for CALL system development,"Proc. LREC'02, pp.896–903 (2002).

[8] N.Minematsu, *et al.*,"Corpus-based analysis of English spoken by Japanese students in view of the entire phonemic system of English,"Proc. ICSLP'02, pp.1213–1216 (2002).

[9] N.Minematsu, *et al.*,"Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development,"Proc. ICSLP'02, pp.745–748 (2002).

[10] http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/x3475.htm