

Pronunciation Understood — How intelligible do you think you are? —

N. Minematsu, S. Asakawa, K. Okabe, and K. Hirose
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
{mine, asakawa, okabe, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This study aims at automatically estimating probability of individual words of Japanese English (JE) being perceived correctly by American listeners and clarifying what kinds of (combinations of) segmental, prosodic, and linguistic errors in the words are more fatal to their correct perception. From a JE speech database, a balanced set of 360 utterances by 90 male speakers are firstly selected. Then, a listening experiment is done where 6 Americans are asked to transcribe all the utterances. Next, using speech and language technology, values of many segmental, prosodic, and linguistic attributes of the words are extracted. Finally, the relation between transcription rate of each word and its attribute values is analyzed by the Classification And Regression Tree (CART) method to predict probability of each of the JE words being transcribed correctly. Performance of the machine prediction is compared with that of the human prediction by four American teachers and three Japanese ones. This method is shown to be comparable to the best American teacher of the four. This paper also describes differences in perceiving intelligibility of the pronunciation between American teachers and Japanese ones.

Keywords: *Objective intelligibility, listening test, speech recognition, CART, speech rhythm*

1 Introduction

What kind of pronunciation should be pursued in language learning? In English education in Japan, the criterion seems to have been changed from acquiring the *native-sounding* pronunciation to achieving the *intelligible* pronunciation. Foreign accented pronunciations do not always reduce the intelligibility (Flege 2002) and, if the latter criterion is adopted, it is important to clarify what kind of (combinations of) acoustic and linguistic errors in the utterances are more relevant to miscommunication.

What is the intelligibility of the pronunciation? In the current paper, it is defined as easiness of accessing to a listener's mental lexicon with given utterances. Then, why some foreign accented pronunciations are accepted and the others are not? Factors affecting the mental lexical access have been discussed by many researchers (Amano 1998) and it is commonly assumed that different factors have different influences. Also, it is easily supposed that it strongly depends upon a listener's language background which factors are influential and how much they are. The authors wonder whether non-native teachers can judge the intelligibility of students' pronunciations adequately only by their ears. This is because their language background is different from that of native speakers. Some previous studies of language learning discussed the perceptual differences between learners and native speakers of the target language (Otake 1998; Tajima 2002; Cutler 2000; Cutler 1994). They tried to induce a strategic change in learner's way of capturing input speech. "Listen to me." This is a phrase repeated thousand times in class by teachers. But Japanese students may not know how to listen because their manner of perception may not be adequate. "Repeat after me." This is another phrase repeated thousand times. But they may not know how to repeat because they may not know the perception of native listeners.

2 English speech database read by Japanese

English speech database Read by Japanese (ERJ database)(Minematsu 2004) was used in the transcription experiment. All the utterances were made by Japanese learners' carefully reading a list of prepared sentences. This means that 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(Minematsu 2002a).

3 Transcription experiment

3.1 Selection of sentences and speakers

The DB contains about 24,000 sentence utterances by 100 male and 100 female speakers. Since it is impossible to use every utterance in the experiment, a part of them had to be adequately selected. 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 selected without bias according to the number of words in the sentence and its perplexity. For the sentence length, considering capacity of human STM (7 chunks), the sentences were divided into 3 groups, 1) less than 6 words, 2) 6 or 7 words, and 3) more than 7 words. As for the perplexity, we also prepared 3 groups, 1) less, 2) rather, and 3) more predictable. Perplexity was calculated using bigram language models trained with the news articles included in WSJ database. Finally, we prepared 9 subsets of about 40 sentences each, which varied in their linguistic complexity.

The DB contains pronunciation proficiency labels of every speaker rated by five native teachers. With the labels, unbiased selection of speakers was also possible for each subset. We selected 90 male speakers by excluding 10 with extremely high or low scores. Finally, 360 (90×4) speech samples were prepared.

3.2 Measurement of quick typing ability of the subjects

In the transcription (typing) experiment, the subjects were asked to type what they just heard on a PC without any guessing. But no guessing during listening is strictly impossible. In order to prevent lengthy consideration before guessing, we designed the experiment so that a minimum duration of typing time should be provided for the subject according to length of the sentence and his/her typing ability. To realize this design, ability of quick typing was measured for each subject in the following manner.

For a given speech sample, length of the pause (T_p) was measured by a simple power threshold method. Using 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 was determined in advance for each subject. The subject was allowed to start typing just after hearing the initial word, and therefore, the actual duration allowed for typing the sentence was $\alpha(T_s - T_p)$. Using native speech samples, α was determined for each subject, which ranged from 3.0 to 4.0.

3.3 Transcription of Japanese English speech

6 adult Americans participated in the experiment. The authors originally tried to find American subjects without any exposure to JE speech. Since it was very hard to find these people in Japan, however, we adopted subjects on a condition that their native language was American English (AE) and period of their stay in Japan was less than a year. 1 Canadian, who has never talked to a Japanese, also took part in the experiment.

The interval between stimuli was carefully controlled to minimize guessing as described previously. To ensure against typing errors, we gave correction time to the subjects every three presentations of the

stimuli. Here, the time was provided as long as they wanted but they were strongly requested not to type any additional words.

120 sets of 3 sentences were presented sequentially to the subjects through headphones, who were required to type on a PC what they heard. The obtained transcriptions would show us whether they recognized the individual words correctly. But it was still uncertain whether they received some meaningful content or information from the utterance. Then, we prepared another task, where the subjects were asked to indicate whether they had some questions on the utterance. The indication was done after each transcription by typing "X" when they had some and "O" when they had none.

Matching between the transcriptions and the reading sheets used in the recording would give us the words that could not be transcribed correctly. We ignored mismatches only by their word forms, walk and walked for example, although the number of mismatches of this type was quite small. Finally, we got data of probability of the individual words being correctly recognized by the six American listeners, ranging from 0/6 to 6/6.

4 Acoustic and linguistic analysis

4.1 Phoneme error detection

Every JE utterance was time-aligned with a phoneme sequence obtained by referring to its prompted sentence and PRONLEX pronunciation lexicon. Next, the phoneme sequence was converted into a phoneme network to predict phoneme errors of the pronunciation (replacement, deletion, and insertion). The conversion rules were written by carefully considering characteristics of JE. Automatic speech recognition of the utterances with the network gave us the phoneme errors. 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.

4.2 Stress error detection

The resulting phoneme sequence was segmented into syllables by tsylb software, which can syllabify an arbitrary sequence of phonemes. After that, each syllable was automatically judged whether it was stressed or not with acoustic models of stressed syllables and unstressed ones. The acoustic models were trained separately for each syllable group by using a database of sentences spoken carefully regarding sentence stress. Coarse spectrum envelope, power, pitch, duration, and voicing degree were utilized as acoustic parameters (Minematsu 2002b). The syllable groups were determined by syllable structures, V, CV, VC, and CVC for example (V has three variants; a short monophthong, a long monophthong, or a diphthong. C is a sequence of consonants whose length is more than or equal to 1). Stress detection performance with the acoustic models was tested in a speaker-closed experiment and it was 96%.

4.3 Linguistic unpredictability

Unpredictability of the individual words (perplexity) in the 360 utterances was 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 a word, which is one of the main factors affecting the mental lexical access.

In the analysis done in this paper, 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.

5 Prediction of the probability

5.1 Preparation of predicting factors

Probability of each of the JE words being correctly recognized was predicted with the CART method, where a decision tree was built with training data. A question on a predicting factor was properly assigned

Table 1: *Predicting factors prepared for CART*

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

to each node of the tree and answering the questions led to a leaf node which indicated how likely the word was to be recognized. The predicting factors had to be prepared by using the parameter values obtained in the acoustic/linguistic analysis and Table 1 lists a set of the factors used. They are divided into three groups; segmental, prosodic, and linguistic factors. These factors can be categorized into five levels; frame, phoneme, syllable, word, and sentence level. A sentence level factor was calculated for each sentence and the unique value was assigned to every word in it.

5.2 Training of the decision tree

Transcriptions of the 360 utterances (about 2,600 words) by the 6 subjects gave us data of the correct recognition probability. 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 were used for testing. By changing the testing speaker, every speaker was used in the testing. It was found that distribution of the probability over the words was biased, where words of 6/6 occupied 55 % of all the words. Then, the normal training generated the tree which treated the 6/6 words with too high priority. To avoid this, besides the normal training method, we tentatively examined a modified method of counting $n/6$ ($n < 6$) data more than once so that the distribution became unbiased. By this method, each category, 0/6 to 6/6, could be treated equally. In the experiments, estimation of the probability was done with different conditions of the predicting factors, which are shown in Table 2. As for performance measurement, recall and precision factors were calculated by ignoring estimation errors by $\pm 1/6$.

5.3 Prediction by American/Japanese teachers of English

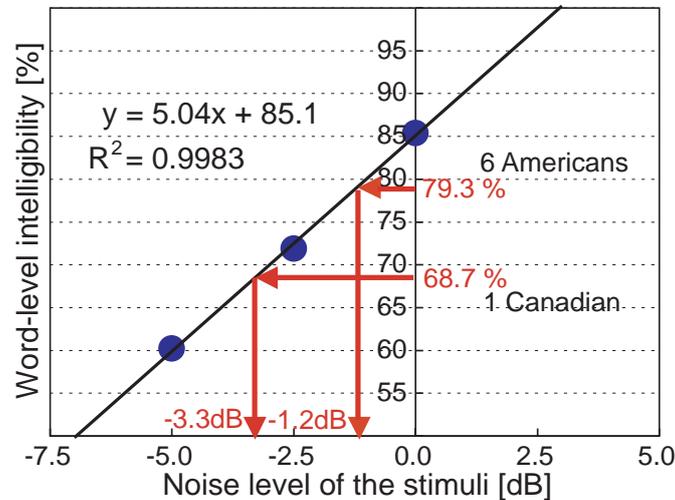
In order to compare CART prediction performance with human performance, a listening test was carried out. 4 American and 3 Japanese teachers of English participated in this experiment. Firstly, detailed descriptions of the transcription experiments were given to the teachers. Then, each of the 360 JE utterances was presented and they were asked to listen to it without looking at the intended sentence. After that, they read the sentence and rated each word in terms of how probable it was supposed to be transcribed correctly by Americans. The rating was done with a 7-level scale, ranging from 0 to 6. The

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

Table 3: *Performance of the transcription*

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 1: *Word-level intelligibility for noisy utterances*

teachers were allowed to listen to the JE utterances as many times as they wanted. But the first listening had to be done without looking.

5.4 Results and discussions

Table 3 shows performance of the 6 Americans' transcription separately for proficiency levels of the speakers and rate of "X", indicating that the listeners had something uncertain on the utterances. It is interesting that speakers of ~3 and ~3.5 levels have almost the same probability of their *words*' being correctly recognized but there is a significant difference between their rates of "X". This implies that speakers of higher levels should have better skills for meaningful speech communication. Average performance of the word-level transcription is 79.3% for the 6 Americans and 68.7% for the 1 Canadian.

The transcription experiment showed that JE are 79.3% intelligible on average. Students can understand mathematically the 79.3% intelligibility but may hardly perceive it by a real experience. In order for the students to experience the 79.3% intelligibility, we tried to generate the 79.3% intelligible speech samples by adding white noise on Japanese speech samples. A small experiment of transcribing Japanese noisy utterances was done and its results are shown in Figure 1. It shows that 79.3% and 68.7% correspond to signal-to-noise ratios of -1.2 dB and -3.3 dB respectively. It implies that "Japanese being" corresponds to -1.2 dB white noise addition when talking to native speakers with some exposure to JE and -3.3 dB white noise without it.

Table 4 shows recall (R) and precision (P) in various conditions. C-1 to C-5 show results of the five conditions of Table 2. BL means baseline and it is chance-level performance, which was calculated by assuming random estimation. In this calculation, the ignorance of $\pm 1/6$ mismatch was also considered. The table shows that CART performance naturally and strongly depends upon the biased distribution of the probability and falling tendency from 6/6 to 0/6 is clearly found. Although the highest performance

Table 4: Prediction performance in various conditions[%]

		0/6	1/6	2/6	3/6	4/6	5/6	6/6	avg.
C-1	R	9.6	10.6	11.1	9.3	25.8	95.4	97.4	37.0
	P	34.2	42.9	41.7	51.7	59.4	76.9	75.4	54.6
C-2	R	15.9	4.7	3.7	18.6	33.5	96.3	95.6	38.3
	P	37.8	100	61.5	51.7	60.0	70.3	76.6	65.4
C-3	R	15.9	21.2	12.0	17.0	28.6	96.0	96.7	41.1
	P	43.9	48.2	58.1	50.0	57.2	64.7	78.9	57.3
C-4	R	15.9	11.7	18.5	17.8	27.4	95.2	96.4	40.4
	P	42.8	53.6	44.7	54.7	53.8	81.0	76.0	58.1
C-5	R	25.5	27.0	20.4	17.8	32.9	95.1	96.1	44.9
	P	42.3	53.8	51.9	62.5	56.6	79.7	79.3	60.9
C-5'	R	67.8	85.7	84.2	75.0	71.7	75.9	59.7	74.3
	P	38.2	46.1	28.3	44.2	50.0	95.8	93.6	56.6
A	R	55.4	47.6	19.7	24.7	24.3	89.2	91.4	50.3
	P	45.0	53.1	43.5	42.4	45.1	68.0	83.6	54.4
A'	R	44.5	49.5	47.7	52.5	62.7	87.5	83.2	61.1
	P	53.8	62.0	40.0	41.5	44.1	89.1	91.1	60.2
J	R	15.1	21.7	25.7	34.4	44.4	82.0	76.1	42.8
	P	20.6	40.1	25.2	31.7	32.6	81.2	80.5	44.5
BL	R	28.5	35.4	43.3	43.8	42.9	43.5	29.8	38.2
	P	7.1	11.0	15.6	22.6	33.0	79.4	73.4	34.6

is achieved in C-5 out of the five cases, their recall rates of 0/6 to 4/6 are lower than those of chance-level. It is clear that this low performance is due to the biased distribution. C-5' shows results of the modified training, where the bias problem was artificially solved so that every category could be treated equally. Performance of this modified training is significantly higher than the chance-level performance both in terms of recall and precision. Data preparation for training the tree should be carefully done according to desired characteristics of the tree. The CART package had a function to show the most effective factor for the prediction by assuming that all the factors were independent. Although this assumption was not always valid, this function gave us interesting results. It showed that the most effective factor was "variance of stress-to-stress intervals" even though it was a sentence-level attribute, the second was "1-gram score", and the third was "phoneme-level likelihood". These results may imply the following. Rhythmical pronunciation is the most important key for high intelligibility. Next, plain wording should be learned. Lastly, correct pronunciation of individual phones should be acquired.

Performance comparison between human and machine was done using F-measure, which is often used to integrate two measures, recall and precision, into one to facilitate the comparison. F-measure is calculated as $2PR/(P+R)$. Before describing the human-machine comparison, however, several findings are shown here on differences in American and Japanese teachers' perceiving the intelligibility. Figure 2 shows F-measures of all the kinds of the probabilities obtained in the various conditions, where "American best" indicates the best prediction out of the 4 American teachers. A, A' and J in Table 4 show recall and precision of American avg., American best, and Japanese avg. respectively. In the case of American avg., F-measures of 0, 1, 5, and 6 are much higher than those of 2, 3, and 4. This denotes that it is much easier to assess the pronunciations with very low or very high intelligibility and much harder to label the pronunciations with middle intelligibility. In the case of Japanese avg., similar tendencies are found with regard to the 2, 3, 4, 5, and 6 cases. But F-measures of 0 and 1 are very low and that of 0 is the lowest among the seven cases. This surprising result indicates that it is the most difficult for Japanese teachers of English to judge the completely *unintelligible* pronunciations as unintelligible. The CART analysis implied that the unintelligible pronunciations were likely to be broken with regard to

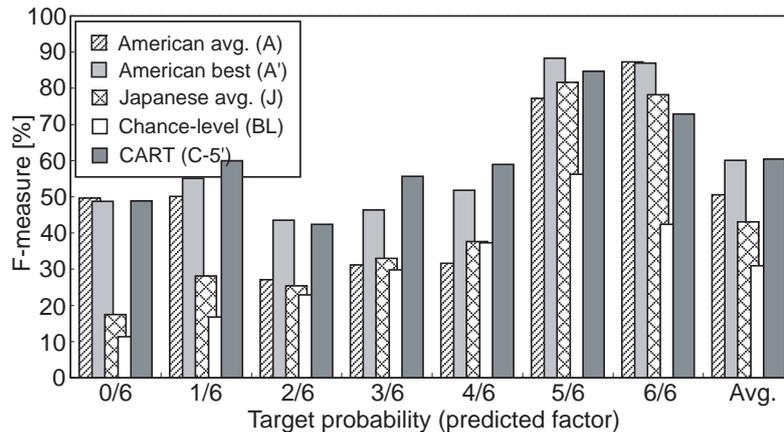


Figure 2: *F-measures in various conditions*

English rhythm. Previous studies indicated that Japanese and English have completely different rhythmic structures (Ramus 2002). Japanese teachers may hardly be able to perceive the broken rhythm. What about Japanese students? It is obvious that they also can hardly do that. As noted in Section 1, if this perceptual difference is not a focus in English education in Japan, the students may never be able to do *listen and repeat*.

In the case of American best, it can be seen that the pronunciations with middle-level intelligibility are predicted rather well. Further, the figure definitely indicates that performance of the proposed CART-based method is completely comparable to the best human prediction performance. This result verifies high validity of the proposed method.

6 Conclusions and future works

Intelligibility of the pronunciations of language learners, not their acoustic similarity to the native pronunciation, was focused upon and acoustic and linguistic factors reducing the intelligibility were examined through CART. Although the transcription experiment was rather a small one, the evaluation experiments showed that the proposed method could predict how likely individual words in JE utterances were perceived correctly by Americans, as well as by the best human teacher of the seven who joined the experiments. Further, this paper clarified a very critical problem of English education to Japanese people, that is perception. The authors hope that this work might be a trigger for a perception-based language learning regime.

Currently, the authors are doing another data collection by presenting Japanese English utterances to a larger number of American listeners with very little exposure to Japanese English. The data collection is being done in Indiana University. Similar analysis are planned with the new data as future work. The authors are also interested in the relationship between perception ability and production ability, especially whether improvement of the pronunciation intelligibility may help the perception ability.

References

- J. Flege, "Factors affecting the pronunciation of a second language", Keynote of ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language (2002)
- S. Amano and T. Kondo, "Estimation of mental lexicon size with word familiarity database," Proc. Int. Conf. Spoken Language Processing (ICSLP), pp.2119–2122 (1998)
- T. Otake and K. Yoneyama, "Phonological units in speech segmentation and phonological awareness," Proc. Int. Conf. Spoken Language Processing (ICSLP), pp.2179–2182 (1998)

- K. Tajima, R. Akabane-Yamada, and T. Yamada, "Perceptual learning of second-language syllable rhythm by elderly listeners," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.249–252 (2002)
- A. Cutler, "Listening to a second language through the ears of a first," *Interpreting*, vol.5, no.1, pp.1–23 (2000)
- A. Cutler, "Segmentation problems, rhythmic solutions," *Lingua*, 92, pp.81–104 (1994)
- N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. Int. Conf. Acoustics(ICA), pp.557–560 (2004)
- N. Minematsu, G. Kurata, and K. Hirose, "Corpus-based analysis of English spoken by Japanese students in view of the entire phonemic system of English," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.1213–1216 (2002a)
- N. Minematsu, S. Kobashikawa, K. Hirose, and D. Erickson, "Acoustic modeling of sentence stress using differential features between syllables for English rhythm learning system development," Proc. Int. Conf. Spoken Language Processing(ICSLP), pp.745–748 (2002b)
- F. Ramus, "Acoustic correlates of linguistic rhythm: perspectives," Proc. Int. Conf. Speech Prosody, pp.115–11 (2002)