Predicting Word Intelligibility of Japanese Accented English*

○ T. Pongkittiphan[1], N. Minematsu[1], T. Makino[2], H.-P. Shen[3], K. Hirose[1]
(The Univ. of Tokyo[1], Chuo Univ.[2], National Cheng Kung Univ.[3])

1 Introduction

English is the only one common language for international communication. Statistics show that there are about 15,000 millions of users of English but only a quarter of them are native speakers, while the rest of them are speaking English with foreign accent [2]. This clearly indicates that foreign accented English is more globally spoken and heard than native English. Although foreign accent often cause miscommunication, native English can becomes unintelligible to non-native listeners because speech intelligibility depends on various factors including the nature of listeners [3].

However, it has been a controversial issue which of native sounding pronunciation and intelligible enough pronunciation should be the target of English pronunciation learning. Recently, the concept of World Englishes [4] is more and more widely accepted by teachers, where it is claimed that, instead of mastering native-like pronunciation, foreign accented pronunciation is acceptable if it is intelligible enough. However, the pronunciation intelligibility is difficult to define because it depends on various factors e.g. the language background of listeners, the speaking context and the speaking proficiency of a speaker [5] [6].

It is known that Japanese learners tend to have poorer speaking skill of English than learners in other Asian countries. One possible reason is there are big differences in the phonological and phonotactic systems between Japanese and English. Therefore, when Japanese learners have to repeat after the English teacher, many of them don't know well how to repeat. In other words, it is difficult for learners to know what kinds of mispronunciations are more fatal to the perception of listeners.

Saz et al. [7] proposed a Basic Identification of Confusable Contexts (BICC) technique to detect the minimal-pairs-based confusable context in a sentence, which might lead to a miscommunication. The subjective evaluation was done by letting subjects read the sentences modified by altering minimal pairs and rate how confusable each sentence is. However, this reflects a lexical and textual confusion perceived by reading sentences not by hearing spoken utterances.

To end this, in this study, by using the results of intelligibility listening tests [1], for given English sentences, we propose a method of automatically predicting the words that will be intelligible or unintelligible to American listeners if those words are spoken with Japanese accent.

2 ERJ Intelligibility Database

Minematsu et al. [1] conducted a large listening test, where 800 English utterances spoken by Japanese (JE) were presented to 173 American listeners. Those utterances were carefully selected from the ERJ (English Read by Japanese) speech database [8]. The American listeners were those who had no experience talking with Japanese and asked to listen to the selected utterances and immediately repeat what they just heard. Then, their responses were transcribed word by word manually by experimenters. Each utterance was heard by 21 listeners on average and a total of 17,416 transcriptions were obtained. In addition to JE utterances, 100 English utterances spoken by speakers of general American English (AE) were used and their repetitions were transcribed in the same way.

---

* "日本人英語を対象とした単語了解度の検出", ポンキッティパン　ティーラポン[1]、峯松信明[1]、牧野武彦[2]、沈涵平[3]、広瀬啓吉[1]　(東京大学[1]、中央大学[2]、成功大学[3])

Following that work, in this study, an expert phonetician, the third author, annotated all the JE and AE utterances with IPA symbols. The IPA transcription shows what is phonetically happening in each of the JE and AE utterances. It would be very interesting to observe the phonetic differences between a JE utterance and an AE one of the same sentence and analyze the word-by-word transcriptions of the JE utterance. The results of which will show what kind of phonetic differences between JE and AE tend to cause misperception. However, it is a pity that the sentences in the JE 800 utterances and those in the AE 100 ones are not overlapped well. So, the above analysis is currently difficult to realize, but the IPA transcriptions of the 900 utterances and the 17,416 word-by-word transcriptions, i.e. misperceptions, will be included in the next release of the ERJ.

Then in this paper, by using the results of the listening test, we firstly define the words in the read sentences that became *very unintelligible* or *rather unintelligible* due to Japanese accent.

Next, we investigate automatic detection of those words by using their lexical and linguistic features that can be extracted directly from textual information. Moreover, referring to actual JE utterances, we also use phonemic information of word defined in CMU pronunciation dictionary, which can be used as one reference of the correct English pronunciations. Actually, CMU pronunciation, itself, contains only the phonemic pronunciation of words based on American English phoneme. We converted its phonemic pronunciation to corresponding phonetic one by following the mapping from its American English phoneme to IPA phone officially defined by its researcher team.

3 Pronunciation Distance

To calculate the pronunciation distance between two IPA symbols, a phonetic-level pronunciation distance matrix is prepared by two following steps.

At first, we calculate the occupancy of each IPA phone with diacritic marks found in 800 JE utterances, and selected only 176 phones which can cover 95% of all existing phones. The phonetician, the third author, was asked to pronounce each of these phones three times, and has to be careful of diacritical difference within the same IPA phone.

Then, we construct a three-state HMM for each phone in which each state has a Gaussian distribution. The Bhattacharyya distance between two corresponding states of each phone pair was calculated, and the 176×176 phonetic-level pronunciation distance matrix was constructed.

The remaining 5% of IPA phones that are not included in 176×176 distance matrix are later replaced by their closest IPA phone by removing diacritic mark or altering to nearest phone considering articulation manner of pronunciation.

Using dynamic time wrapping (DTW) technique, the accumulated pronunciation distance of two IPA sequences of a word pair can be calculated. The larger the distance is, the more the word pair is considered to be phonetically different. This pronunciation difference might affect the perception of native listeners and make the word become unintelligible.

Shen et al. [9] also used this pronunciation distance matrix and the same simplification method in speakers clustering task, and its experimental results showed that this pronunciation matrix is reliable and effective.

4 Prediction of Word Intelligibility

4.1 Definition of “will-be-unintelligible” words

The ERJ contains the pronunciation proficiency score (1.0 to 5.0) for each speaker, which was rated by five American teachers of English. To focus on the listening test results of only typical Japanese speakers, we removed the data of too poor speakers (<2.5) and those of too good speakers (>4.0). The resulting data had 756 utterances and 5,754 words in total.

As described in Section 2, each spoken word was heard by 21 American listeners on average and the correct perception rate was obtained for each. In this study, to describe the word perception qualitatively, the words whose perception rate is less than 0.1 are defined as *very unintelligible* due to Japanese accent and the

words whose rate is from 0.2 to 0.3 are defined as *rather unintelligible*. The occupancies of *very unintelligible* and *rather unintelligible* words were 18.9% and 34.2%, respectively.

4.2 Preparation of features for automatic prediction

From preliminary experiments, we found two things. 1) Since we wanted a binary (intelligible/unintelligible) classifier of input data, we firstly trained CART as binary classifier but results were not good. Then, we trained CART as predictor of perception rate of each word and, comparing the output to a threshold, binary classification was made possible. We found this strategy to be effective. 2) Since we wanted to train CART distinctively between intelligible words and unintelligible words, we intentionally removed words of intermediate level (0.4 to 0.6) of perception rate only from training data. This removal was effective although those data were actually included in testing data.

The features used for CART-based detection were prepared by using the CMU pronunciation dictionary and the n-gram language models trained with 15 millions words from the OANC text corpus [10]. Table 1 shows these features that are categorized into 4 groups; lexical, linguistic and other features.

The feature [C], which is the maximum number of consecutive consonants in the word, is derived by considering Japanese pronunciation habits of English that is caused by phonological and phonotactic differences between the two languages. The smallest unit of speech production in Japanese is called mora, which has the form of either CV or V. However, consecutive consonants, with the form of CCV or CCCV, are very common in English. Japanese speakers sometimes insert an additional vowel after a consonant, which increases the number of syllables in that word and is expected to decrease the intelligibility of that word easily, e.g. the word 'sky' (S-K-AY) is often pronounced as (S-UH-K-AY), where additional UH vowel is added.

Table 1 The features prepared for CART

[A] lexical features for a word
▪ #phonemes in the word ▪ #consonants in the word ▪ #vowels (=#syllables) in the word ▪ forward position of 1st stress in the word ▪ backward position of 1st stress in the word ▪ forward position of 2nd stress in the word ▪ backward position of 2nd stress in the word ▪ word itself (word ID)
[B] linguistic features for a word in a sentence
▪ part of speech ▪ forward position of the word in the sentence ▪ backward position of the word in the sentence ▪ the total number of words in the sentence ▪ 1-gram score of the word ▪ 2-gram score of the word ▪ 3-gram score of the word
[C] phonological and phonotactic feature for a word
▪ the maximum number of consecutive consonants
[D] pronunciation distance
▪ phonetic-level DTW distance of the word

Table 2 Precisions, recalls, and F1-scores [%]

		[A]	[B]	[AB]	[AB] +C	[AB] +CD
very unintelligible	P	44.19	42.42	60.67	74.01	77.34
	R	3.71	22.70	47.68	58.64	60.11
	F1	6.85	29.58	53.39	65.44	67.88
rather unintelligible	P	57.04	57.08	70.12	73.72	79.92
	R	11.02	45.12	58.66	67.46	71.17
	F1	18.48	50.49	63.92	70.45	75.29

The last feature [D] is the DTW-based phonetic-level pronunciation distance of the word. This is the only feature that is extracted from IPA transcriptions of JE utterances, while [A], [B] and [C] are features that can be extracted only from text. As described in section 3, if the pronunciation of word in JE utterance is phonetically different from that of CMU pronunciation dictionary if the pronunciation of word in JE utterance is phonetically different to some degrees from that of CMU dictionary, the word will be misrecognized by native listeners.

4.3 Experimental results

We have four kinds of features; [A], [B], [C] and [D], and have two levels of "will-be-unintelligible" words; *very unintelligible*

and *rather unintelligible*. Table 2 shows the results of precisions, recalls, and F1-scores of 10 cross-validation experiments.

By using only either lexical [A] or linguistic [B] features, each method has low F1-scores, while combination of [A] and [B] can increase the F1-score significantly to 53.39% and 63.92% for very and rather unintelligible words, respectively.

An interesting finding is that, when adding the feature [C], the maximum number of consecutive consonants, the F1-score is improved significantly again from 53.39% to 65.44% and from 63.92% to 70.45% for each case.

Furthermore, after including the last feature [D], the F1-score is further increased to 67.88% and 75.29%, which is quite obvious because we use the actual phonetic pronunciation of JE utterances.

The precisions in the table claim that almost 75% of the words that were identified as very or rather unintelligible are correctly detected. As described in Section 4.1, the occupancies of very and rather unintelligible words were 18.9% and 34.2%, which correspond to the precisions when detecting unintelligible words randomly.

When omitting the last feature D, although no acoustic observation is used, it can detect "will-be-unintelligible" words very effectively. Considering these facts, the proposed method is able to show which words of a presentation manuscript Japanese learners should be very careful of to make their English oral presentation more intelligible.

Use of phonetic information did improve the prediction performance. However, the CMU pronunciation dictionary defines a pronunciation of word when saying it isolatedly, which cannot explain the actual phenomenon of continuous speech articulation in which the change of phones can be found. For that, we now continue annotating additional utterances to get IPA transcriptions of the sentence utterances of AE to get complete overlap between JE and AE. With these transcriptions, we can add another IPA-based phonetic feature to improve the detection performance. We're also interested in replacing manual IPA-based features with features obtained automatically by ASR.

5 Conclusions

This study examines the prediction of word intelligibility of Japanese accented English. Defining the words that are *very unintelligible* and *rather unintelligible* to native listeners, the proposed method can effectively predict unintelligible words even using only the information extracted from text. Moreover, adding of phonetic-level pronunciation distance later improves the prediction performance. In the future, acoustic and phonetic information extracted from annotated AE utterances will be used for performance improvement.

References

[1] N. Minematsu et al., "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) Database", Proc. Interspeech, pp. 1481-1484, 2011.

[2] Y. Yasukata., "English as an International Language: Its past, present, and future", Tokyo: Hitsujishobo, pp. 205-227, 2008.

[3] J. Flege., "Factors affecting the pronunciation of a second language", Keynote of PLMA, 2002.

[4] J. Jenkins., "The phonology of English as an international language", Oxford University Press, 2000.

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

[6] J. Bernstein., "Objective measurement of intelligibility", Proc.ICPhS, 2003.

[7] O. Saz and M. Eskenazi., "Identifying confusable contexts for automatic generation of activities in second language pronunciation training", Proc. SLaTE, 2011.

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

[9] H.-P. Shen et al., "Speaker-based pronunciation clustering using world Englishes and pronunciation structure", Proc.ASJ Spring, 2013

[10] The Open American Nation Corpus (OANC), http://www.anc.org/data/oanc/.