一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

# 日本人英語音声を対象とした単語了解度の自動予測

ポンキッティパン ティーラポン† 峯松 信明† 牧野 武彦‡ 沈 涵平†† 広瀬 啓吉†

†東京大学工学部 〒113-8656 東京都文京区本郷 7-3-1
‡中央大学経済学部 〒192-0393 東京都八王子市東中野 742-1
††国立成功大学 〒701 台湾台南市大学路 1 号

E-mail: †{teeraphon, mine, hirose}@gavo.t.u-tokyo.ac.jp, ‡ mackinaw@tamacc.chuo-u.ac.jp,
†† happy@ gavo.t.u-tokyo.ac.jp

**あらまし** 本研究では，日本人が英文を読み上げた場合に日本語訛りによって聞き取り難くなってしまう単語を自動的に予測する手法を検討する。我々の先行研究[1]では，日本人による 800 の読み上げ文音声を 173 名の母語話者に呈示して書き起こさせ，発声中の単語毎に聞き取り率を求めている。本研究ではこの実験結果を用いて「日本語訛りによって聞き取り難くなる単語発声」を定義し，その単語発声を自動的に予測することを考える。意図された文とその読み上げ音声から，言語的素性，語彙的素性のみを使って CART (Classification And Regression Tree)による予測を試みた。次に，英語と日本語の音韻体系の違い，音素配列の違いを考慮して新たな素性を導入し，更には，入力音声と当該文の母語話者発声に対する IPA 書き起こしに基づく素性も導入した。言語的素性及び語彙素性のみを用いた手法に対し，新しく導入した二素性は予測率の向上に大きく貢献することが分った。最終的に，提案手法は「非常に聞き取り難くなる単語」と「やや聞き取り難くなる単語」を，F1 スコア 69.59%及び，78.36%で予測可能であることが分った。

**キーワード** 了解度，第二言語学習，外国なまり，日本人英語データベース，CART，国際音声記号

# Automatic Prediction of Intelligibility of Spoken Words in Japanese Accented English

Teeraphon PONGKITTIPHAN† Nobuaki MINEMATSU† Takehiko MAKINO‡ Han-Ping SHEN††
and Keikichi HIROSE†

†Faculty of Engineering, The University of Tokyo 7-3-1 Hongo Bunkyo, Tokyo, 113-8654 Japan
‡Faculty of Economics, Chuo University 742-1 Higashinakano Hachioji, Tokyo, 192-0351, Japan
††National Cheng Kung University No.1, University Road, Tainan City, 701, Taiwan

E-mail: †{teeraphon, mine, hirose}@gavo.t.u-tokyo.ac.jp, ‡ mackinaw@tamacc.chuo-u.ac.jp,
†† happy@ gavo.t.u-tokyo.ac.jp

**Abstract** This study examines automatic prediction of the words that will be unintelligible if they are spoken by Japanese speakers of English. In our previous study [1], 800 English utterances spoken by Japanese speakers, which contained 6,063 words, were presented to 173 American listeners and correct perception rate was obtained for each spoken word. By using the results, in this study, we define the words that are very unintelligible through Japanese accented English pronunciation and also define the words that are rather unintelligible. Then, by using Classification And Regression Tree (CART) with linguistic features and lexical features only, we examine automatic detection of these words. After that, we introduce an additional feature derived by considering phonological and phonotactic differences between Japanese and English, and another feature derived by calculating the phonetic pronunciation distance observed from manually-annotated IPA transcriptions of Japanese English and American English. This additional features are found to be very effective and our proposed method can detect very unintelligible words and rather unintelligible words automatically with F1-scores of 69.56 and 78.36 [%], respectively, if phonetic transcriptions are given.

**Keyword** Speech Intelligibility, Second Language Learning, Foreign Accent, ERJ Database, CART, IPA

## 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 causes 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 their English teacher, many of them don't know well how to repeat adequately. 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 with their IPA transcriptions, 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-800) 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 via a telephone call and immediately repeat what they just heard. Then, their responses were transcribed word by word manually by expert transcribers. 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-100) were used and their repetitions were transcribed in the same way.

Following that work, in this study, an expert phonetician, the third author, annotated all the JE-800 and AE-100 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, the sentences in the JE-800 utterances and those in the AE-100 ones are not overlapped well. So, the same phonetician also annotated another 419 utterances spoken by one female speaker. This corpus is called "AE-F-419", which completely covers all utterances of JE-800 and AE-100, and the analysis of JE-800 comparing to AE-F-419 can be done at phonetic level.

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-800 utterances, we also use phonetic information of IPA transcriptions of AE-F-419 utterances, which can be used as one reference of the correct American English pronunciations.

## 3. Pronunciation Distance and Intelligibility

### 3.1. Construction of pronunciation distance matrix

Comparison of a JE utterance in JE-800 and its corresponding AE utterance in AE-F-419 is done by comparing their IPA transcriptions. Pronunciation distance is the distance calculated by comparing the two IPA transcriptions and it requires the phone-based pronunciation distance matrix, which is prepared by the following two steps.

At first, we calculate the occupancy of each IPA phone with diacritic marks found in JE-800 utterances, and selected only 153 phones which can cover 95% of all existing phones. The phonetician, the third author, was asked to pronounce each of these phones twenty times by paying good attention to 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 153×153 phonetic-level pronunciation distance matrix was constructed.

The remaining 5% of IPA phones that are not included in the 153×153 distance matrix are later replaced by their closest IPA phone by removing diacritic mark or altering to nearest phone considering the articulation manner of pronunciation.

Using word-based 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 more unintelligible if it is larger. Note that, in this study, when calculating the DTW pronunciation distance, we use the IPA transcriptions of AE-F-419 utterances as the correct pronunciation references of American English.

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

### 3.2. Preliminary analysis of pronunciation distance

In this section, we show a result to support our assumption, saying that if the pronunciation of word in JE-800 utterances is phonetically different to some degrees from the correct pronunciation of American English, the word will be misrecognized by native listeners.

According to previous study [1], the ERJ contains the pronunciation proficiency score (1.0 to 5.0) for each speaker, which was rated by five American teachers of English shown in Table 1. Figure 1 shows the word-based correct perception rates for different learner groups, and words spoken by speakers with higher pronunciation proficiency score tend to be more intelligible.

Using this subjective evaluation result, we first investigate the correlation between the pronunciation proficiency score and pronunciation distance of words in JE-800. As described in Section 3.1, we use DTW technique to calculate the pronunciation distance of words in JE-800 utterances comparing to the correct pronunciation of AE-F-419's ones, and the obtained distance is normalized by the number of DTW phone comparisons. As a result, the average of word-based pronunciation distance is calculated and grouped by the level of proficiency shown in Table 2. The visualized version in Figure 2 shows that the pronunciation proficiency score and the average of word-based pronunciation distance have a considerably strong correlation. The utterances of high-level speakers have lower phonetic pronunciation difference than those of low-level speakers.

Table 1 #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 |

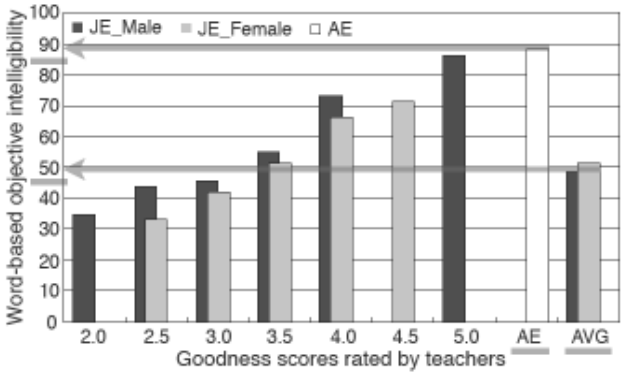


Figure 1: Word-based correct perception rates for different learner groups

Table 2 The average of word-based pronunciation distance classified by pronunciation proficiency score

| Proficiency | JE-800 (ALL) | JE-F-400 | JE-M-400 |
|---|---|---|---|
| ≤ 2.0 | 2.09 | | 2.09 |
| ≤ 2.5 | 1.90 | 1.87 | 1.93 |
| ≤ 3.0 | 1.87 | 1.89 | 1.87 |
| ≤ 3.5 | 1.76 | 1.70 | 1.89 |
| ≤ 4.0 | 1.63 | 1.60 | 1.81 |
| ≤ 4.5 | 1.61 | 1.61 | |
| ≤ 5.0 | 1.42 | | 1.42 |

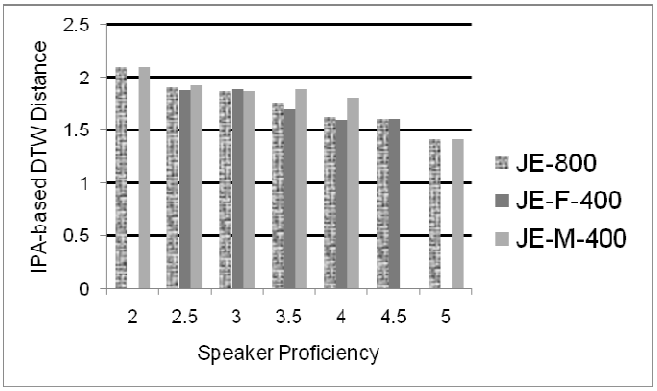


Figure 2: Visualization of correlation between speaker proficiency and word-base pronunciation distance

The same analysis is done on the common sentences found in AE-100, AE-F-419, JE-F-400, and JE-M-400. The number of sentences is 100. Here, DTW-based distances are calculated from AE-100, JE-F-400, and JE-M-400 comparing to AE-F-419. The result shows AE-100 has the smallest pronunciation distance which is 1.083, while JE-female-100 and JE-male-100 have 1.497 and 1.582, respectively. These again confirm that the intelligible utterances have smaller phonetic pronunciation distance and less phonetically different from the correct pronunciation of American English.

## 4. Prediction of Word Intelligibility

### 4.1. Definition of "will-be-unintelligible" words

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 3 shows these features that are categorized into 4 groups; lexical, linguistic and other features.

Table 3 The features prepared for CART

| |
|---|
| [A] lexical features for a word |
| ▪ #phonemes in the word<br>▪ #consonants in the word<br>▪ #vowels (=#syllables) in the word<br>▪ forward position of 1st stress in the word<br>▪ backward position of 1st stress in the word<br>▪ forward position of 2nd stress in the word<br>▪ backward position of 2nd stress in the word<br>▪ word itself (word ID) |
| [B] linguistic features for a word in a sentence |
| ▪ part of speech<br>▪ forward position of the word in the sentence<br>▪ backward position of the word in the sentence<br>▪ the total number of words in the sentence<br>▪ 1-gram score of the word<br>▪ 2-gram score of the word<br>▪ 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 4 Precisions, recalls, and F1-scores [%]

| | | [A] | [B] | [AB] | [AB] +C | [AB] +CD |
|---|---|---|---|---|---|---|
| very unintel ligible | P | 44.19 | 42.42 | 60.67 | 74.01 | 78.97 |
| | R | 3.71 | 22.70 | 47.68 | 58.64 | 62.15 |
| | F1 | 6.85 | 29.58 | 53.39 | 65.44 | 69.56 |
| rather unintel ligible | P | 57.04 | 57.08 | 70.12 | 73.72 | 81.51 |
| | R | 11.02 | 45.12 | 58.66 | 67.46 | 75.44 |
| | F1 | 18.48 | 50.49 | 63.92 | 70.45 | 78.36 |

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.

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 automatically. As described in section 3, if the pronunciation of word in JE-800 utterances is phonetically different to some degrees from that of AE-F-419's ones, the word will be misrecognized by native listeners.

### 4.3. Experimental results and discussion

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 4 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 69.56% and 78.36%, 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. This phonetic information extracted by manually-annotated IPA transcription is considered to be very reliable than the phonemic information of isolated words defined in CMU pronunciation dictionary used in our previous study [11]. This is because our IPA transcriptions explain the actual phenomenon of continuous speech articulation in which the change of phones can be found. And, 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. From the preliminary analysis, the DTW-based pronunciation distance and correct perceptions rate have a considerably strong correlation, which can be implied that the intelligible utterances have smaller phonetic pronunciation distance and less phonetically different from the correct pronunciation of American English.

Moreover, 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 automatically from ASR 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] B. Kachru, et al., “The Handbook of World Englishes”, Wiley-Blackwell, 2006.

[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/.

[11] T. Pongkittiphan, N. Minematsu, T. Makino et al., “Predicting Word Intelligibility of Japanese Accented English”, Proc.ASJ Autumn, 2013