IMPROVEMENT OF INTELLIGIBILITY PREDICTION OF SPOKEN WORD IN JAPANESE ACCENTED ENGLISH USING PHONETIC PRONUNCIATION DISTANCE AND WORD CONFUSABILITY

Teeraphon Pongkittiphan[1], Nobuaki Minematsu[1], Takehiko Makino[2], Keikichi Hirose[1]

[1]Faculty of Engineering, The University of Tokyo
7-3-1 Hongo Bunkyo, Tokyo, 113-8654, Japan

[2]Faculty of Economics, Chuo University
742-1 Higashinakano Hachioji, Tokyo, 192-0351, Japan

{teeraphon,mine,hirose}@gavo.t.u-tokyo.ac.jp, mackinaw@tamacc.chuo-u.ac.jp

ABSTRACT

This study investigates intelligibility prediction of English words spoken with Japanese accent that will be unintelligible when perceived by American listeners. In our previous study using the ERJ (English Read by Japanese) intelligibility database [1], 800 English sentences spoken by 200 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 this result, a CART-based spoken words intelligibility predictor was constructed. The first two sets of features used in experiments were linguistic and lexical features derived from textual information. The third one was derived by considering phonological and phonotactic differences between Japanese and English. In this paper, we focus on two new features, 1) *phonetic pronunciation distance* calculated base on the acoustic distance between manually-annotated IPA transcriptions of Japanese English and American English, and 2) *word confusability* which is the number of English words in CMU dictionary that have phonemically similar pronunciation to that of a given Japanese accented English word. These two additional features are found to be very effective and our proposed method can predict *very unintelligible* words and *rather unintelligible* words with F1-scores of 71.48% and 83.21% (+6.04% and +12.76% improvement), respectively, if phonetic transcriptions are given.

Index Terms— Speech intelligibility, ERJ database, pronunciation distance, word confusability, CART, IPA, second language learning, foreign accent

1. INTRODUCTION

English is the only one global language for international communication. Statistics show that there are about 1.5 billion 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 become 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 are asked to repeat after their English teacher, many of them don't know well how to repeat adequately. In other words, learners do not know well what kinds of mispronunciations are more fatal to the perception of listeners.

Related study done by Saz et al. [7] uses 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. Subjective evaluation was done by letting subjects read the sentences modified by altering minimal pairs and rate how confusable each sentence is. However, this only reflects a lexical and textual confusion perceived by reading sentences not by hearing spoken utterances.

In our prior work on automatic word intelligibility prediction in Japanese accented English [8], we exploited three kinds of features which can be directly extracted from textual

information; 1) linguistic features, 2) lexical features and 3) feature derived by considering phonological and phonotactic differences between Japanese and English. In this work, we focus on a task to predict the intelligibility of spoken words by considering what seems to happen in human speech production and perception. An expert phonetician, the third author, transcribed all the utterances used in the ERJ intelligibility test [1] by using IPA symbols. Using these transcriptions, we propose two new sets of features, *"phonetic pronunciation distance"* and *"word confusability"*, to predict the spoken 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 [9]. The American listeners who had no experience talking with Japanese were asked to listen to the selected utterances via a telephone call and immediately repeat what they have 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 the sentences used in JE-800 and AE-100, and the analysis of JE-800 comparing to AE-F-419 can be done at phonetic level.

In our previous work, we investigated automatic prediction of those words by using their lexical and linguistic features that can be extracted directly from textual information. In this work, referring to actual JE-800 spoken utterances, we use phonetic information from IPA transcriptions of AE-F-419 utterances, which can be used as one reference of the correct American English pronunciations. Using this phonetic information, we then prepare the *phonetic pronunciation distance* and *word confusability* features.

3. WORD CONFUSABILITY AND PHONETIC PRONUNCIATION DISTANCE

In automatic speech recognition system, the confusability between words in the lexicon of ASR is one of the important issues, which can lower the recognition accuracy [10][11]. It will become difficult to recognize an input word if it has a large number of phonetically or phonemically similar words in the ASR lexicon. Similar to the mechanism of human speech perception and the concept of mental lexicon[12], when hearing a spoken word, we are considered to map that sound sequence to the nearest word stored in the mental lexicon. Considering these assumptions, *"word confusability"* might be one of the critical factors affecting the intelligibility of communication. And, it would become more confusing, especially for American English listeners, when they perceive the Japanese accented English that commonly has different phonological characteristics compared to American English.

Even if word confusability of an input Japanese accented word is low, it can be very unintelligible if it is pronounced differently from American English pronunciation. By using JE-800 transcriptions and AE-F-419 ones, we can define word-level *"phonetic pronunciation distance"* quantitatively. Here DTW is done between them. In the next section, we investigate how effective these two new features are in predicting word 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 measuring the phonetic differences between their IPA transcriptions. Pronunciation distance is the accumulated distance obtained from the optimal alignment of the two IPA transcriptions, which can be calculated by Dynamic Time Warping (DTW). 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.

DTW 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. For two phone HMMs, the Bhattacharyya distance between corresponding states is calculated and the averaged distance over the three

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

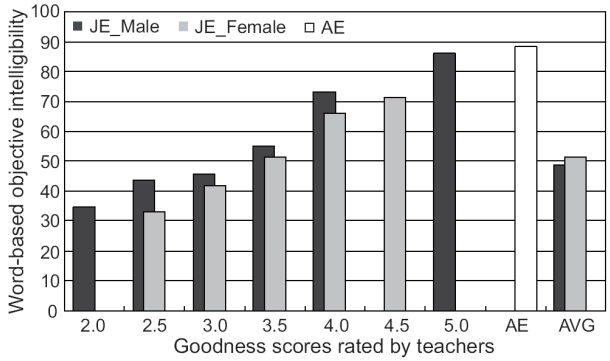

Fig. 1. Word-based correct perception rates for different learner groups.

states is defined as distance between the two phones.

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.

Shen et al. [13][14] also used this pronunciation distance matrix and the same DTW-based comparison in World Englishes pronunciation and speakers clustering tasks, and its experimental results showed that this pronunciation matrix is reliable and effective.

3.2. Preliminary analysis of the pronunciation distance

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

According to our 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 results of ERJ intelligibility listening tests, namely word-level correct perception rates for different learner groups, and words spoken by speakers with a 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 by comparing them to the correct pronunciation of AE-F-419's ones, and the obtained distance is normalized by the number of

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

Proficiency	JE-800	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

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. It draws a conclusion that the pronunciation proficiency score and the average of word-based pronunciation distance have a considerably strong correlation. And, the utterances of high-level speakers have lower phonetic pronunciation difference than those of low-level speakers.

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-F-100 and JE-M-100 have 1.497 and 1.582, respectively. These again confirm that the intelligible utterances of American speakers (AE-100) have smaller phonetic pronunciation distance and are less phonetically different from the correct pronunciation of American English.

3.3. Word confusability calculation

Due to the lack of phonetic pronunciation dictionaries, we rather use the CMU pronunciation dictionary [15] as a vocabulary lexicon containing 133k entities. In this step, we first prepare a phonemic pronunciation distance matrix, not a phonetic one. Three-state HMM-based acoustic models for each phoneme of 39 American phonemes used in CMU-dict are well trained using the WSJ speech corpus. Similarly to Section 3.1, the averaged Bhattacharyya distance between two corresponding states of each phoneme pair is calculated. Finally, the 39×39 phonemic pronunciation distance matrix is constructed.

The word confusability of each JE spoken word is basically calculated by comparing the DTW-based phonemic distance between its phonemic transcription and all the words in the CMU-dict. Note that the phonemic pronunciations of JE spoken words are prepared by converting each phone in JE's IPA transcription to the closest American English phoneme. The mapping strategy of 153 IPA phones to 39 American phonemes is carefully defined and checked by the expert phonetician.

Table 3. The features prepared for CART

[Lex] Lexical features for a word
#phonemes in a word
#consonants in a word
#vowels (=#syllables) in a word
forward position of primary stress in a word
backward position of primary stress in a word
forward position of secondary stress in a word
backward position of secondary stress in a word
word itself (word ID)
[Ling] Linguistic features for a word in a sentence
part of speech
forward position of the word in a sentence
backward position of the word in a sentence
the total number of words in a sentence
1-gram score of a word
2-gram score of a word
3-gram score of a word
[C.Con]
Maximum number of consecutive consonants in a word
[P.Dist]
Phonetic pronunciation distance of a word
[W.Conf]
Word confusability of a word

To determine the word confusability of an arbitrary word utterance is to find the total number of confusing words whose pronunciations are phonemically closer enough to that of the considered word. However, the explicit definition of threshold distance or boundary line used to distinguish between the confusing and non-confusing words is unknown. To this end, we decide to use the best empirical threshold that can maximize the prediction accuracy, which will be further discussed in Section 4.3.

4. PREDICTION OF WORD INTELLIGIBILITY

4.1. Definition of 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). As a result, the final experimental 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 0.1 to 0.3 are defined as *"rather unintelligible"*. The occupancies of very unintelligible and rather unintelligible words were 18.9% and 34.2%, respectively.

Table 4. *Precisions, recalls, and F1-scores[%]*

		Lex.Ling	Lex.Ling + C.Con	Lex.Ling + C.Con + P.Dist	Lex.Ling + C.Con + P.Dist + W.Conf
very unintel-ligible	P	60.67	74.01	78.97	82.42
	R	47.68	58.64	62.15	63.11
	F1	53.39	65.44	69.56	71.48
rather unintel-ligible	P	70.21	73.72	81.51	86.79
	R	58.66	67.46	75.44	79.91
	F1	63.92	70.45	78.36	83.21

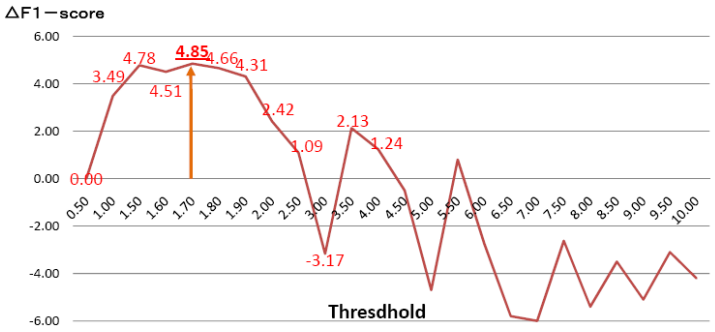

Fig. 2. Relative F1-score improvement when varying empirical threshold of [W.Conf].

4.2. Preparation of features for intelligibility 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 a binary classification was made possible by comparing the regression output to the perception rate thresholds. 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.

Table 3 summarizes five groups of features used for CART-based prediction. First, the lexical and linguistic features [Lex][Ling] were prepared by using the CMU pronunciation dictionary and the n-gram language models trained with 15 millions words from the OANC text corpus [16].

Next, the feature [C.Con], which is the maximum number of consecutive consonants in a 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 one additional UH vowel is added.

In this study, we focus on the use of two new proposed features; [P.Dist] and [W.Conf]. The feature [P.Dist] 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 [Lex], [Ling] and [C.Con] are features that can be extracted directly from text automatically. As described in Section 3, if the pronunciation of word in JE-800 utterances is phonetically different to some degree from that of AE-F-419's ones, the word will be misrecognized by native listeners.

The last feature [W.Conf], namely word confusabiliy, is the total number of confusing words that their pronunciations are phonemically nearer than *empirical threshold* to that of an JE spoken word of interest. As roughly described in Section 3.3, we use the best empirical threshold that can maximize the prediction accuracy.

4.3. Experimental results and discussion

We have five kinds of features; [Lex], [Ling], [C.Con], [P.Dist] and [W.Conf] as shown in Table 3, and have two levels of 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 [Lex] or linguistic [Ling] features, each method has low F1-scores, while combination of [Lex] and [Ling] 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.Con], 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.

After including the feature [P.Dist], 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.

To find the best *empirical threshold* for the last feature [W.Conf], we notice that the phoneme-pair distances in the 39×39 distance matrix are ranging from 0.63 to 2.7. So, the set of threshold from 0.5 to 10.0 is firstly tested to find the best empirical one. Varying the threshold, the experiments are conducted using all five features. Figure 2 shows the relative F1-score improvement compared to 78.36% of the previous case predicting rather unintelligible words. We found that the best empirical threshold is 1.70, giving the best F1-score at 71.48% and 83.21%.

The precisions in the Table 4 claim that almost 87% 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 [P.Dist] and [W.Conf] features, although no acoustic observation is used, it can detect 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 presentations more intelligible.

Use of phonetic pronunciation distance and word confusability did improve the prediction performance. The phonetic information extracted from manually-annotated IPA transcription is considered to be very reliable than textual information used in our previous study [8]. This is because our IPA transcriptions explain the actual phenomenon of continuous speech articulation in which the change of phones can be found. Furthermore, word confusability, which is derived based on the concept of mental lexicon of human speech perception and be considered as one important issue in ASR, is also found to be very effective feature. We're also interested in replacing manual IPA-based features with features obtained automatically by ASR, and adding prosodic features in future experiment.

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 and word confusability later improves the prediction performance. In the future, acoustic and phonetic information extracted automatically from ASR will be used for performance improvement.

6. 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] T. Pongkittiphan, N. Minematsu, T. Makino et al., "Automatic detection of the words that will become unintelligible through Japanese accented pronunciation of English", Proc. SLaTE, 2013.

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

[10] A. Zgank et al., "Predicting the acoustic confusability between words for a speech recognition system using Levenshtein Distance", Proc. Elektronika ir elektrotechnika, Vol.18, No.8, 2012.

[11] J. Anguita et al., "Inter-phone and inter-word distances for confusability prediction in speech recognition", Procesamiento del Lenguaje Natural 33, 2004.

[12] J. Aitchison, "Words in the mind: an introduction to the mental lexicon", Wiley-Blackwell, 2012.

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

[14] S. Kasahara et al., "Improved and robust prediction of pronunciation distance for individual-basis clustering of World Englishes pronunciation", Proc. ICASSP, 2014.

[15] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

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