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

日本人英語発声を対象とした単語明瞭度の自動予測 ～ 特徴量とモデルに関する比較研究 ～

ポンキッティパン ティーラポン† 峯松 信明† 牧野 武彦‡ 齋藤 大輔† 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

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

あらまし 日本語訛りを有する英語音声に対して、どの単語が、米語母語話者にとって聞き取り難くなってしまうのかを自動予測することを検討している。本研究では、ERJ intelligibility データベースを用いている。これは、日本人によって発声された 800 文発声が、173 名の米語母語話者によって聴取、書き取られ、各単語毎に聴取率が定義されている。先行研究において、入力テキストあるいは入力音声に対する聴取率予測器を構築した。そこでは、入力テキストから抽出される言語的特徴や、入力音声から抽出される音声学的発音距離や単語の混同性を計算し、CART を使って聴取率の予測を行なった。本研究では、新たな特徴として韻律的特徴を検討し、また、予測モデルとしては新たに三種類のモデル（Adaboost、Random Forest、Extremely Randomized Trees）を検討した。評価実験として「非常に聞き取り難くなる単語」の同定、「やや聞き取り難くなる単語」の同定を行なった。その結果、両タスクにおいて F1 スコアが 72.74%、84.78%となり、良好な結果を得ることができた。

キーワード 明瞭度、日本人英語データベース、韻律的特徴、機械学習、 国際音声記号、第二言語、外国訛り

Automatic prediction of intelligibility of English words spoken with Japanese accents -- Comparative study of features and models used for prediction --

Teeraphon PONGKITTIPHAN† Nobuaki MINEMATSU† Takehiko MAKINO‡ Daisuke SAITO† 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

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

Abstract This study investigates automatic prediction of the words in given sentences that will be unintelligible to American listeners when they are pronounced with Japanese accents. The ERJ intelligibility database contains results of a large listening test, where 800 English sentences read with Japanese accents were presented to 173 American listeners and correct perception rate was obtained for each spoken word. By using this database, in our previous study, an intelligibility predictor was built for each word of input texts or utterances. For prediction, lexical and linguistic features were extracted from texts and pronunciation distance and word confusability were calculated from utterances. CART was used as prediction model. In this paper, new features that are related to speech prosody and three new prediction models of ensemble methods (Adaboost, Random Forest and Extremely Randomized Trees) are tested and compared to the old features and model. Finally, our new system can predict very unintelligible and rather unintelligible words with F1-scores of 72.74% and 84.78%, respectively.

Keyword spoken word intelligibility, ERJ database, prosodic features, machine learning, IPA, L2 learning, foreign accent

1. Introduction

English is the only one language used for international communication. Statistics show that there are about 1.5 billion users of English but only a quarter of them are native speakers, while the rest of them are speaking English with foreign accents [2]. This clearly indicates that foreign accented English is more globally spoken and heard than native English. Although foreign accents often cause miscommunication, native English can also 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.

A 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 and automatically extracted from input texts; 1) linguistic features, 2) lexical features and 3) features derived by considering phonological and phonotactic differences between Japanese and English. After that, by considering what seems to happen in human speech production and perception, another work of us [9] used two new features; 1) phonetic pronunciation distance and 2) word confusability extracted from actual utterances and their corresponding manually-annotated IPA transcriptions.

In this study, new features that are related to speech prosody and three new prediction models of ensemble methods (Adaboost, Random Forest and Extremely Randomized Trees) are tested and compared to the old features and model (CART). Using the results of intelligibility listening test [1], our new intelligibility predictor is trained so that it can predict which spoken words in Japanese English utterances will be unintelligible when perceived by American listeners. And, the effectiveness of prosodic features comparing to other features used in our prior work is discussed.

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 [10]. The American listeners who had no experience talking with Japanese were asked to listen to the selected utterances via a telephone line 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.

In our prior works [8][9], an expert phonetician, who is the third author of this paper, has 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. And, the same phonetician also annotated another 419 utterances spoken by one female American speaker. This corpus is called "AE-F-419", and it completely covers all the sentences used in JE-800 and AE-100, and was used as one of the correct American English pronunciation references.

The IPA transcriptions include temporal information of phone boundaries. Then, in this study, we use the transcriptions to obtain location of word boundary, which will be used to extract prosodic features at word-level. The preparation of prosodic features and all features used in our previous studies will be summarized in the next section.

3. Features Preparation

This section explains the preparation of three sets of features used in prediction experiments, shown in Table 1.

3.1. SET-1 Lexico-linguistic features

SET-1 contains lexico-linguistic features which can be directly extracted from input texts. The *1.1) lexical feature* and *1.2) linguistic features* were prepared by using the CMU pronunciation dictionary [11] and the n-gram language models trained with 15 millions words from the OANC text corpus [12]. And, the *1.3) maximum number of consecutive consonant in a word* is derived by considering

Table 1 The features used in experiments

SET-1 : Lexico-linguistic features
1.1) Lexical features for a word
▪ #phonemes in a word
▪ #consonants in a word
▪ #vowels (=#syllables) in a word
▪ forward position of 1st stress in a word
▪ backward position of 1st stress in a word
▪ forward position of 2nd stress in a word
▪ backward position of 2nd stress in a word
▪ word itself (word ID)
1.2) 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 the sentence
▪ 1-gram score of the word
▪ 2-gram score of the word
▪ 3-gram score of the word
1.3) Maximum number of consecutive consonants
SET-2 : Phonetically derived features
2.1) Phonetic pronunciation distance of a word
2.2) Word confusability of a word
SET-3 : Prosodic features
3.1) Aggregate statistic F0 and energy
3.2) Duration of word
3.3) Energy-F0-Integral

Japanese speakers' 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 in a syllable, 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 'screen' (S-K-R-IY-N) is often pronounced as (S-UH-K-UH-R-IY-N), where two UH vowels are added.

3.2. SET-2 Phonetically derived features

SET-2 are features extracted from the actual JE and AE utterances with their corresponding manually-annotated IPA transcriptions.

a) Phonetic pronunciation distance

The *2.1) phonetic pronunciation distance* is prepared by calculating the DTW-based phonetic distance between the IPA symbol sequence of an utterance in JE-800 and that of its corresponding utterances in AE-F-419. The two utterances were obtained by reading the same sentence. Here, the AE-F-419 utterance was used just as one of the correct AE utterances. This feature is designed based on our assumption 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.

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 states is defined as distance between the two phones.

The remaining 5% of IPA phones that are not included in the 153x153 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.

b) Word confusability

The *2.2) word confusability* is the number of different English words that have similar pronunciation to that of a given Japanese accented English word. From the mechanism of human speech perception and the concept of mental lexicon [13], when hearing a spoken word, humans are considered to map that sound sequence to the nearest word stored in the mental lexicon, so "*word confusability*" might be one of the critical factors affecting the intelligibility of input spoken words.

Due to the lack of phonetic pronunciation dictionaries, we rather use the CMU pronunciation dictionary 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 the 39 American phonemes used in CMU-dict are well trained using the WSJ speech corpus. Similarly to the preparation of phonetic pronunciation distance feature, the averaged Bhattacharyya distance between two corresponding states of each phoneme pair is calculated. Finally, the 39x39 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.

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 input spoken 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. Due to limit space, detailed explanation how to find the best empirical threshold can be found in [9].

3.3. SET-3 Prosodic features

SET-3 are phrase-level prosodic features. Pitch and energy are extracted over 10 msec intervals for each JE utterance, using STRAIGHT analysis [14] for F0, and HTK for energy. Duration of a word is prepared from the manually-annotated IPA transcription, mentioned in Section 2, which provides the word segmentation and time alignment. To cancel the inter-speaker variation of F0 and energy range, we use the speaker-normalized value (z-score) of pitch and energy. This SET-3 contains three subsets of features.

3.1) Aggregate statistics including mean, max, min, range, median and std. of F0 and energy of a word.

3.2) Duration of a word (msec)

3.3) Energy-F0-Integral (EFI) of a word defined in the following equation.

$$EFI = \sum_{t \in intervals} (F_t \times E_t), \quad (1)$$

where F_t and E_t are the F0 and energy extracted at time interval t.

Considering some influences of the prosodic features of left-context or right-context words on intelligibility of their central word in a given utterance, we also add the prosodic features of w_{i-1} and w_{i+1} to predict the intelligibility of w_i, where i is an index of a word in an utterance.

4. Word Intelligibility Prediction Experiment

4.1. Definition of 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). As a result, the final experimental data had 756 utterances and 5,754 spoken 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 accents 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.

4.2. Experimental design and conditions

According to preliminary experiments in our prior work, 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 then 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.

In addition to CART, in this study, we also use three new prediction models; Adaboost (AdaB) [15], Random forest (RF) [16] and Extremely Randomized Trees (ERT) [17]. These ensemble methods combine outputs from several elementary classifiers, and they are considered to be effective when a large number of features are available. On average, an ensemble method is robust than prediction of a single classifier because its variance is reduced.

Table 2 F1-scores of CART, AdaBoost, RF and ERT-based predictions [%]

	very unintelligible word				rather unintelligible word			
	CART	AdaB	RF	ERT	CART	AdaB	RF	ERT
SET 1	65.44	66.80	67.38	67.54	70.45	73.50	72.90	73.13
SET 1+3	68.01	68.13	69.22	68.91	77.59	77.41	78.63	78.94
SET 1+2	71.48	71.21	71.97	72.10	83.21	83.97	84.06	83.89
SET 1+2+3	71.66	71.68	72.59	**72.74**	84.11	84.66	**84.78**	84.70

Adaboost is one of the boosting methods designed based on the motivation that combing several weak models is able to create a powerful model. The final output of the boosted classifier is combined from the weighted sum of outputs of the other learning algorithms. Weak models are built sequentially, each of which is trained so as to reduce the errors made by a sequence of models prior to the current model. In this study, we select a tree model as a weak model to compare with other tree-based methods.

Random Forest (RF) and Extremely Randomized Trees (ERT) are two averaging algorithms specially designed for tree models. In contrast to Adaboost, several single trees are built independently and randomly. Then, prediction of the final combined model is obtained as averaged prediction of the individual trees. RF is an ensemble of unpruned trees whose randomness is given to a tree which is growing in two ways where data sampling is done differently. Slightly different from RF, ERT do not require the bagging step to construct a set of training samples for each tree because the same input training set is used to train all trees. Moreover, ERT picks each node split very extremely with random variable, while RF chooses only the best node split with the best variable.

4.3. Results and discussion

We have three sets of features as shown in Table 1, and have two levels of unintelligible words; *very unintelligible* and *rather unintelligible*. Table 2 shows the F1-scores of CART, AdaB, RF and ERT-bsaed predictions evaluated by 10 cross-validation experiments. Three of the ensemble-based predictions did give better performance than CART-based in all cases. Henceforth in this section, when an F1-score is mentioned, it refers to the best F1-score from the three ensemble-based methods or that from the four features within a single model.

As a baseline system, using only features from SET-1, the system can predict *very unintelligible words* and *rather unintelligible words* with F1-scores of 67.54% and 73.50%, respectively. From the results of our prior work, the maximum number of consecutive consonants was found to be a very effective feature which can be easily prepared only from texts.

In the case of features extracted from actual utterances, the effectiveness of SET-2 and that of SET-3 are compared by adding these two kinds of features separately to the original feature set (SET-1). From the results, we can say that, when adding SET-2 features, SET 1+2 can significantly improve the performance to 72.10% and 84.06% compared to the performance of SET 1+3 (69.22% and 78.94%). It can be firstly implied that the phonetic differences found between JE and AE are considered to be more critical factors reducing speech intelligibility than prosodic changes in JE utterances. This might be caused by the big differences in the phonological and phonotactic systems between Japanese and English.

In contrast, using only prosodic features is still effective in stress and word prominence detection, for both native and non-native English speech, whose characteristics are mostly linked with the prosodic changes in utterances [18][19][20]. It is because prosody is an important key to catch the speaker's intention and the meaning of a whole sentence. But, it is less important and contributes few benefits to our intelligibility prediction task performed at a word-level.

Finally, using all features of SET-1, SET-2 and SET-3, the prediction gave the best performance of 72.74% and 84.78%. Although Table 2 shows only F1-scores, not precision or recall, the F1-score of 84.78% was obtained as precision of 87.93% and recall of 81.85%. This claims that almost 88% 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.

It is interesting that, even if SET-2 and SET-3 are not used, our system can predict unintelligible words considerably effectively by using only features of SET-1 extracted from texts. Considering these facts, the proposed method will be able to show which words of a presentation manuscript Japanese learners should be very careful of to make their English oral presentations more intelligible,

where even actual utterances are not used for prediction.

Although, from the results of this study, prosodic features (SET-3) are shown to be not as effective as pronunciation distance and word confusability features (SET-2), the exploitation of both feature sets gave the best prediction performance. To investigate which features are effective in a real application, we are planning to collect feedback from Japanese learners of English by letting them use two kinds of predictors, separately trained with SET 1+2 and SET 1+3, then check which predictor can improve the intelligibility of their utterances more effectively. We're also interested in analyzing the prosodic pattern of JE utterances to get more meaningful and effective features, and replacing manual IPA-based features with features obtained automatically by ASR to realize to automatic prediction for practical application.

5. Conclusions

This study examines the intelligibility prediction of English words spoken by Japanese. Following our prior works using lexico-linguistic features, phonetic pronunciation distance and word confusability, we further exploit prosodic features and investigate their effectiveness by conducting comparative experiments. Defining the words that are very unintelligible and rather unintelligible to native American English listeners, the proposed method can effectively predict unintelligible words even using only the information extracted from text.

From comparative results, prosodic features did improve the prediction performance but not as effectively as phonetic pronunciation distance and word confusability features did. In the case of intelligibility prediction or word identification, the phonetic differences between AE and JE utterances are more critical and important than prosodic changes in JE utterances. Moreover, comparing the three new ensemble prediction models (Adaboost, Random Forest and Extremely Randomized Trees) to the old CART model, all of the ensemble methods did give better performance than the CART method. In the future, acoustic and phonetic information extracted automatically from ASR will be used for performance improvement and realizing practical application to support learners.

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] T. Pongkittiphan, N. Minematsu, T. Makino et al., "Improvement of intelligiblity prediction of spoken word in Japanese accented English using phonetic pronunciation distance and word confusability", Proc. O-COCOSDA, 2014.

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

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

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

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

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds",Speech Communication, Vol. 27, No. 3-4, pp. 187-207, 1999.

[15] Y. Freund, R.E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of Computer and System Sciences, Vol. 55(1), pp. 119-139, 1997.

[16] L. Breiman., "Random forests" Machine Learning, Vol. 45(1), pp. 5-32, 2001.

[17] G. Pierre et al., "Extremely randomized trees" Machine Learning, Vol. 63(1), pp. 3-42, 2006.

[18] J. Tepperman and S. Narayana., "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners", Proc. ICASSP, pp. 937-940, 2006.

[19] T. Mishra et al., "Word prominece detection using robust yet simple prosodic features", Proc. Interspeech, pp. 1864-1867 2012.

[20] W. Xia et al., "Perception and production of prominence distribution patterns of Chinese EFL Learners", Proc. Speech Prosody, 2010.