RiCL Research in Corpus Linguistics

Machine-learning classification of pronunciation difficulty for learners of English as a Foreign Language

Katsunori Kotani · Takehiko Yoshimi
Kansai Gaidai University · Ryukoku University / Japan

Abstract – This study compiled and assessed a learner corpus to measure the difficulty of pronouncing a sentence (henceforth, pronounceability). The method of measuring pronounceability is useful for computer-assisted language learning of English as a Foreign Language that employs online materials as a resource for pronunciation training. An advantage of this resource is that learners can select materials depending on their interest, a disadvantage being that pronounceability is unknown to learners. If pronounceability is automatically measured, learners can independently access materials appropriate for their proficiency levels without teachers' assistance. The pronounceability assessment demonstrated moderate reliability and partial validity when it was measured by learners' subjective judgment on a five-point Likert scale. Given the reliability and validity, this study developed a pronounceability measuring method utilizing a machine learning algorithm that automatically predicts the pronounceability of a sentence based on the linguistic features of the sentences and learners' features (i.e. learners' scores for an English proficiency test). The proposed measuring method demonstrated a higher classification accuracy (53.7 percent) than the majority class baseline (46.0 percent).

Keywords – phonetic learner corpus, English as a Foreign Language, pronunciation difficulty, Machine-learning classification

1. INTRODUCTION

Effective teaching of English as a Foreign Language (EFL) is required to ensure that learners are highly motivated (Hwang 2005; Lai 2015; Yoon et al. 2016). Learners' motivation will be sustained if they are provided with materials that are interesting and whose difficulty level is appropriate for their proficiency. A promising language resource is the internet because it offers materials covering various topics and difficulty levels from easy to difficult. The choice of proper materials is time and effort consuming. However, this task is solvable by employing a computer-assisted language learning tool that automatically selects proper materials. Previous research (Kotani et al. 2014; Xia et al. 2016,) proposed methods to select materials according to learners' proficiency by measuring the readability/listenability of materials.

This study aims to develop a method for measuring the pronunciation difficulty of materials (henceforth, pronounceability) that predicts the pronounceability based on the linguistic features of the sentences and learners' features (i.e. learners' scores for an English proficiency test). In developing a pronounceability measuring method, a classifier that predicts the pronounceability of a sentence is trained with a machine-learning algorithm, whose performance depends on the quality of a phonetic learner corpus as training data.

Subsequently, this study compiled a phonetic learner corpus where a data instance comprises a sentence read aloud by learners, learners' read-aloud speech sounds, linguistic features of read-aloud sentences, learners' features, and pronounceability. This study proposes the pronounceability determined by learners'

DOI 10.32714/ricl.06.01

subjective judgment on a five-point Likert scale. Since the reliability and validity of subjective judgment are dubious due to learners' biases, this study assesses the proposed corpus by answering the following research questions:

- How stable is pronounceability as an evaluation index?
- To what extent does pronounceability help classify learners based on English proficiency?
- How effectively does pronounceability correlate with scores representing English proficiency?
- How accurately is pronounceability measurable based on linguistic and learners' features?

2. Compilation of a phonetic learner corpus

2.1. Collection of pronunciation data

The phonetic learner corpus was compiled by recording pronunciation data for English texts that learners read aloud, sentence by sentence. After reading a sentence aloud, learners determined its pronounceability on a five-point Likert scale (1: easy; 2: somewhat easy; 3: average; 4: somewhat difficult; 5: difficult).

The texts for reading aloud were selected from those distributed by the International Phonetic Association, encompassing basic English sounds (International Phonetic Association 1999; Deterding 2006). This enables us to analyze which types of English sounds influence learners' pronunciation. These texts were originally appropriated from Aesop's Fables; the title of Text I was *The North Wind and the Sun* and that of Text II was *The Boy who Cried Wolf*. Given the texts' popularity, their contents were believed not to affect learners' subjective judgment. Deterding (2006) reported that Text I failed to encompass certain sounds such as initial and medial /z/ and syllable-initial /θ/, and subsequently developed Text II that included the English pronunciation for these sounds. Table 1 illustrates the sentences in Texts I and II and their length.

Text I	Text II
The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. (22 words)	There was once a poor shepherd boy who used to watch his flocks in the fields next to a dark forest near the foot of a mountain. (27 words)
They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. (23 words)	One hot afternoon, he thought up a good plan to get some company for himself and also have a little fun. (21 words)
Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt. (36 words)	Raising his fist in the air, he ran down to the village shouting 'Wolf, Wolf'. (15 words)
Then the Sun shone out warmly, and immediately the traveller took off his cloak. (14 words)	As soon as they heard him, the villagers all rushed from their homes, full of concern for his safety, and two of his cousins even stayed with him for a short while. (32 words)
And so the North Wind was obliged to confess that the Sun was the stronger of the two. (18 words)	This gave the boy so much pleasure that a few days later he tried exactly the same trick again, and once more he was successful. (25 words)
	However, not long after, a wolf that had just escaped from the zoo was looking for a change from its usual diet of chicken and duck. (26 words)
	So, overcoming its fear of being shot, it actually did come out from the forest and began to threaten the sheep. (21 words)
	Racing down to the village, the boy of course cried out even louder than before. (15 words)
	Unfortunately, as all the villagers were convinced that he was trying to fool them a third time, they told him, 'Go away and don't bother us again'. (27 words)
	And so the wolf had a feast. (7 words)

Table 1: Sentences (sentence length – number of words) in Texts I and II

The corpus data were compiled from 50 EFL learners at university (28 males, 22 females; mean age: 20.8 years (standard deviation *SD*=1.3), who were compensated for their participation. All learners were requested to submit valid scores from the Test of English for International Communication (TOEIC) taken in the current or previous year. In the study sample, the mean (*SD*) TOEIC score was 607.7 (186.2); the minimum score was 295 and the maximum was 900. Figure 1 presents the distribution of TOEIC scores. Half the learners appeared below 545, and the distribution followed the normal distribution according to the Kolmogorov-Smirnov test (K=0.82, p=0.25).

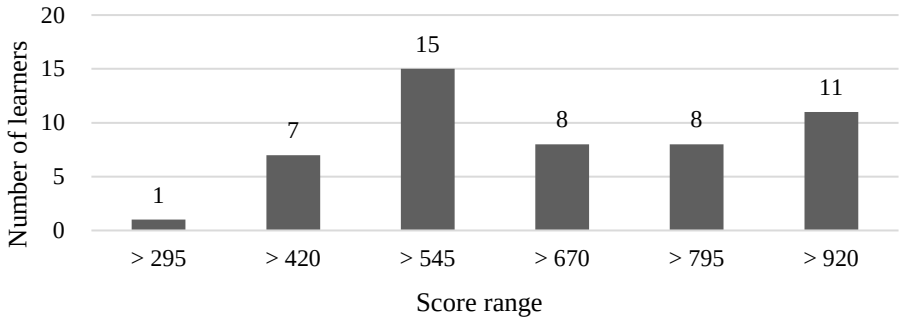

Figure 1: Distribution of TOEIC scores

2.2. Annotation of pronunciation data

In our phonetic learner corpus, a data instance comprised speech sounds of learners' reading aloud, pronounceability by learners' subjective judgment, the linguistic features of a sentence that learners read aloud, and learners' features.

Linguistic features were automatically derived from a sentence that learners read aloud as follows: sentence length (Chall and Dial 1948) was derived as the number of words in a sentence, mean word length (Chall and Dial 1948) was derived by dividing the number of syllables in a sentence by the number of words in the sentence, and the number of multiple-syllable words in a sentence (Fang 1966) was derived by calculating $\sum_{i=1}^{N}(S_i - 1)$, where N denoted the number of words in a sentence and S_i denoted the number of syllables in the i-th word, where this subtraction derivation ignored single-syllable words. Word difficulty was derived as the rate of words not listed in a basic vocabulary list (Kiyokawa 1990) relative to the total number of words in a sentence.

Table 2 summarizes the linguistic features of Texts I and II. Text length was measured in terms of sentences and words. The other linguistic features were the mean values of sentence length, mean word length, multiple-syllable word, and word difficulty.

	Text I	Text II
Text length (sentences)	5	10
Text length (words)	113	216
Sentence length (words)	22.6 (8.3)	21.6 (7.6)
Mean word length (syllables)	1.3 (0.1)	1.2 (0.1)
Multiple syllable word (syllables)	6.4 (2.8)	5.7 (3.0)
Word difficulty	0.3 (0.1)	0.2 (0.1)

Table 2: Linguistic features of Texts I and II

Texts I and II differed at the text level (i.e. text length) but demonstrated similar properties at the sentence level (i.e. sentence length, mean word length, multiple syllable word, and word difficulty). Thus, pronounceability was expected to be similar at sentence level but not at text level.

Learners' features were determined using the TOEIC scores. Although TOEIC comprises listening and reading tests, Chauncey Group International (1998) reported the strong correlation between TOEIC scores and Language Proficiency Interview results, an established direct assessment of oral language proficiency developed by the Foreign Service Institute of the U.S. Department of State. In previous research (Delais-Roussarie et al. 2015; Gósy et al. 2015; Graham et al. 2015), proficiency was demonstrated using a point-scale such as the Common European Framework of Reference for Languages (CEFR: six levels from A1 to C2). This study used TOEIC, not CEFR, because TOEIC demonstrated learners' proficiency in more detail.

2.3. Properties of pronunciation data

	Pronounceability
Minimum	1
Maximum	5
Median	4
n	750

Table 3: Descriptive statistics of pronounceability

Table 3 presents the descriptive statistics for pronounceability, while Figure 2 indicates the distribution of pronounceability in the phonetic learner corpus. The distribution failed to follow the normal distribution according to the Kolmogorov-Smirnov test (K=6.66, p<0.01). The pronounceability data were skewed to low pronounceability (i.e. difficult for pronunciation) and the peak of the pronounceability data appeared at pronounceability level 4 (i.e. somewhat difficult).

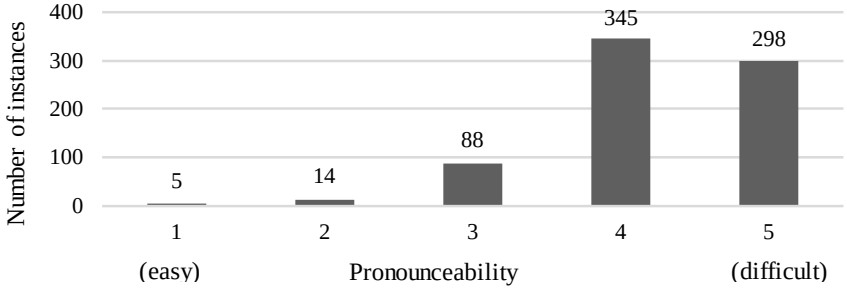

Figure 2: Distribution of pronounceability

3. Assessment of pronounceability

In Sections 3.1, 3.2, and 3.3, research questions 1–3 were assessed with the classical test theory (Brown 1996). In Section 3.4, the fourth question was answered by classifying five categories of pronounceability.

3.1. Reliability

The reliability of pronounceability was examined through internal consistency, referring to whether learners' subjective judgment demonstrates similar results for sentences with similar pronounceability. The internal consistency was tested in terms of Cronbach's α (Cronbach 1970). The Cronbach's α coefficient is defined by the following equation $\alpha = \frac{k}{k-1}\left(1 - \sum_{i=1}^{k} \frac{S_i^2}{S_T^2}\right)$, where k denotes the number of items (i.e. the number of sentences annotated with pronounceability in this study), S_i^2 is the variance associated with item i, and S_T^2 is the variance associated with the sum of all k item values. Cronbach's α is a reliability coefficient ranging from 0 (absence of reliability) to 1 (absolute reliability), and empirical satisfaction is achieved with values above 0.8.

Table 4 presents the Cronbach's α coefficients for learner groups. Learners were classified into three levels based on the TOEIC scores: below 490 (i.e. BEGinner, n=16), below 730 (i.e. INTermediate, n=16), and 730 or above (i.e. ADVanced, n=18). The Cronbach's α coefficients were derived for learners at these proficiency levels and for ALL the learners (n=50). In addition, as reliability depended on the number of items, the Cronbach's α coefficients were derived individually for each text (Text I containing 5 sentences and Text II containing 10 sentences) and jointly for both texts. The reliability coefficients exceeded the value required for empirical satisfaction (α=0.80) except in the case of Text II by INT (α=0.74). Therefore, pronounceability was reliable except for Text II by INT.

	Text I	Text II	Texts I and II
BEG	0.86	0.83	0.87
INT	0.85	0.74	0.84
ADV	0.82	0.83	0.87
ALL	0.86	0.85	0.89

Table 4: Cronbach α coefficients of pronounceability

Given the low reliability, Text II by INT was examined by excluding a sentence to identify which one decreased the reliability. Table 5 presents the Cronbach's α coefficients of Text II by INT excluding a sentence. The coefficient increased when excluding Sentence 10, indicating that this sentence decreased the reliability. This sentence was the shortest in Text II, which could explain the negative effect on the reliability scores. In a short sentence, learners' subjective judgment for pronounceability would be unstable by the presence of some word(s).

Sentence Excluded	Cronbach α	Sentence Excluded	Cronbach α
Sentence 1	0.72	Sentence 6	0.73
Sentence 2	0.68	Sentence 7	0.74
Sentence 3	0.69	Sentence 8	0.74
Sentence 4	0.71	Sentence 9	0.73
Sentence 5	0.69	Sentence 10	0.75

Table 5: Cronbach α coefficients in Text II from which one sentence is removed

3.2. Construct validity

Construct validity was examined from the perspective of distinctiveness. If pronounceability reflects learners' proficiency, it should demonstrate a statistically significant difference among learners of different proficiencies. The phonetic learner corpus data were classified into three levels based on the TOEIC scores where pronounceability was the mean values of each learner calculated by dividing the total values of pronounceability with the number of sentences (n=15) (i.e. BEG n=16, INT n=16, ADV n=18). Table 6 presents the mean (SD) values of pronounceability for learners at the three proficiency levels.

	BEG	INT	ADV
Pronounceability	4.47 (0.37)	4.32 (0.35)	3.92 (0.43)

Table 6: Mean and SD of pronounceability at the three proficiency levels

The distinctiveness of pronounceability was investigated using ANOVA, whose results indicated statistically significant differences between learners at the three levels ($F(2, 47)=9.24$, $p<0.01$). Hence, the pronounceability demonstrated the construct validity of learners at the three proficiency levels.

Tukey's HSD post hoc test demonstrated a significant difference ($p<0.01$) between BEG and ADV, but not between INT and ADV, and between BEG and INT. Hence, the pronounceability demonstrated the construct validity between BEG and ADV.

3.3.Criterion-related validity

Criterion-related validity was examined from the perspective of the correlation with learners' proficiency in terms of TOEIC scores. If pronounceability has good criterion-related validity, it should reflect learners' proficiency. Subsequently, correlation between pronounceability and the learners' TOEIC scores was examined, where pronounceability was the mean values of each learner calculated by dividing the total values of pronounceability with the number of sentences (n=15).

Figure 3 depicts a scatter plot of the correlation between pronounceability and TOEIC scores (n=50). The correlation analysis indicated moderate correlation (r=–0.51), where the higher the TOEIC scores, the lower the pronounceability (i.e. easier).

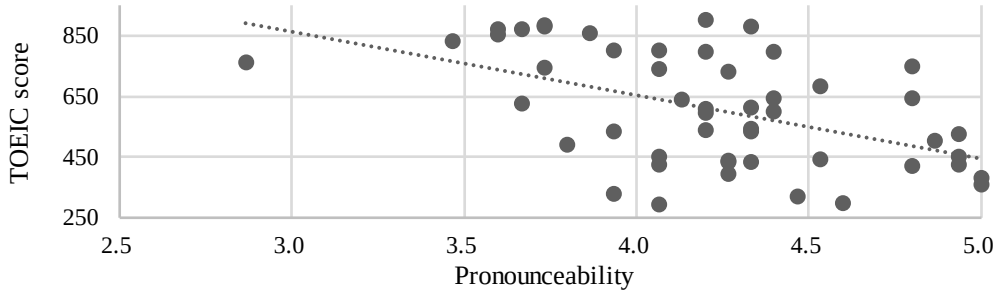

Figure 3: Scatter plot of pronounceability and TOEIC scores

3.4. Pronounceability measurement

A pronounceability measurement method was developed based on the linguistic and learners' features described in Section 2.2. Support Vector Machine, the function 'svm()' defined in the 'e1071' package of the software environment R (Meyer 2012), was employed to develop the pronounceability measurement method. The parameter setting of the function 'svm()' was default. The pronounceability measurement method was evaluated *n* times (*n*=750) employing a leave-one-out cross validation test, considering one instance as test data and *n*–1 instances as training data.

Table 7 presents the confusion matrix for the test data. The classification accuracy was 53.7 percent derived by (20+217+166)/750 in the test data, being insufficient for validating the measurement method. However, the method can still be judged as valid through a comparison with the majority class baseline 46.0 percent (i.e. (12+217+116)/750), which is defined as the percentage of instances in the most frequently occurring pronounceability in the training data. The classification entirely failed in pronounceability 1 and 2, which were misclassified into pronounceability 3 or 4. This misclassification suggests that Support Vector Machine could not properly learn pronounceability 1 and 2 due to the fewer number of instances than the pronounceability 3, 4, and 5.

Measurement method Learners	1	2	3	4	5
1	0	0	5	0	0
2	0	0	5	9	0
3	0	0	20	58	10
4	0	0	12	217	116
5	0	0	8	124	166

Table 7: Confusion matrix for the test data

Each feature's effect on classification accuracy was examined by training the pronounceability measurement method based on all features excluding a kind of features to be tested. If this feature contributes to the classification, exclusion of the feature would decrease the classification accuracy, and vice versa. Table 8 presents the classification accuracies using different features. Classification accuracies increased when excluding features of sentence length, mean word length, multiple syllable word, and decreased by excluding those of word difficulty and TOEIC scores.

Features excluded	Classification accuracies (%)
NONE	53.7
Sentence length	55.1
Mean word length	54.3
Multiple syllable word	53.9
Word difficulty	53.5
TOEIC scores	50.7

Table 8: Contribution of each feature

4. Conclusion

This study assessed whether pronounceability on learners' subjective judgment appropriately demonstrated the pronunciation difficulty of EFL learners. The assessment suggested that pronounceability was moderately reliable (Section 3.1) and partially valid (Sections 3.2 and 3.3), in that it had moderate correlation with the TOEIC scores. The pronounceability measurement results (Section 3.4) suggested that it was appropriately explained by linguistic and learners' features. Future studies should work on the development of a pronounceability measurement system, and evaluate the pronounceability in English classes.

References

Brown, James D. 1996. *Testing in language programs*. Englewood Cliffs, NJ: Prentice-Hall.

Chall, Jeanne S. and Harold E. Dial. 1948. Predicting listener understanding and interest in newscasts. *Educational Research Bulletin* 27/6: 141–153+168.

Chauncey Group International. 1998. *TOEIC technical manual*. Princeton, NJ: Chauncey Group International.

Cronbach, Lee J. 1970. *Essentials of psychological testing*. 3rd edition. New York: Harper & Row.

Delais-Roussarie, Elisabeth, Fabián Santiago and Hi-Yon Yoo. 2015. The extended COREIL corpus: first outcomes and methodological issues. In *Proceedings from the Workshop on Phonetic Learner Corpora, International Congress of the Phonetic Sciences*, 57–59.

Deterding, David. 2006. The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association* 36/2: 187–196.

Fang, Irving E. 1966. The 'Easy listening formula'. *Journal of Broadcasting* 11/1: 63–68.

Gósy, Mária, Dorottya Gyarmathy and András Beke. 2015. The development of a Hungarian-English learner speech database and a related analysis of filled pauses. In *Proceedings from the Workshop on Phonetic Learner Corpora, International Congress of the Phonetic Sciences*, 61–63.

Graham, Calbert, Andrew Caines and Paula Buttery. 2015. Phonetic and prosodic features in automated spoken language assessment. In *Proceedings from the Workshop on Phonetic Learner Corpora, International Congress of the Phonetic Sciences*, 37–40.

Hwang, Myung-Hee. 2005. How strategies are used to solve listening difficulties: listening proficiency and text level effect. *English Teaching* 60/1: 207–226.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Kiyokawa, Hideo. 1990. A formula for predicting listenability: the listenability of English language materials 2. *Wayo Women's University Language and Literature* 24: 57–74.

Kotani, Katsunori, Shota Ueda, Takehiko Yoshimi and Hiroaki Nanjo. 2014. A listenability measuring method for an adaptive computer-assisted language learning and teaching system. *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, 387–394.

Lai, Degang. 2015. A study on the influencing factors of online learners' learning motivation. *Higher Education of Social Science* 9/4: 26–30.

Meyer, David. 2012. *Support Vector Machines*. The interface to `libsvm` in package `e1071`. https://datajobs.com/data-science-repo/SVM-in-R-[David-Meyer].pdf (accessed 27 July 2018)

Xia, Menglin, Ekaterina Kochmar and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics and the Asian Federation of Natural Language Processing*, 12–22.

Yoon, Su-Youn, Yeonsuk Cho and Diane Napolitano. 2016. Spoken text difficulty estimation using linguistic features. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics and the Asian Federation of Natural Language Processing*, 1–6.

Corresponding author
Katsunori Kotani
16–1 Nakamiyahigashino-cho
Hirakata, Osaka, Japan, 573–1001
e-mail: kkotani@kansaigaidai.ac.jp

received: November 2018
accepted: December 2018