Assessment of an Index for Measuring Pronunciation Difficulty

Katsunori Kotani
Kansai Gaidai University
kkotani@kansaigaidai.ac.jp

Takehiko Yoshimi
Ryukoku University
yoshimi@rins.ryukoku.ac.jp

Abstract

This study assesses an index for measuring the pronunciation difficulty of sentences (henceforth, pronounceability) based on the normalized edit distance from a reference sentence to a transcription of learners' pronunciation. Pronounceability should be examined when language teachers use a computer-assisted language learning system for pronunciation learning to maintain the motivation of learners. However, unlike the evaluation of learners' pronunciation performance, previous research did not focus on pronounceability not only for English but also for Asian languages. This study found that the normalized edit distance was reliable but not valid. The lack of validity appeared to be because of an English test used for determining the proficiency of learners.

1 Introduction

Research on computer-assisted language learning (CALL) has been carried out for learning the pronunciation of European languages as a foreign language such as English (Witt & Young 2002, Mak et al. 2004, Ai & Xu 2015, Liu & Hung 2016) and Swedish (Koniaris 2014). CALL research on Asian languages has considered Japanese as a foreign language (Hirata 2004) and Chinese as a foreign language (Zhao et al. 2012). The primary goal of CALL systems for the learning of foreign language pronunciation is to resolve interference from the first language of learners. For instance, a CALL system can analyze the speech in which a learner reads English sentences aloud and presents pronunciation errors that a learner must read aloud again for reducing the errors.

Even though the methods of evaluating learners' pronunciation performance have received considerable attention in previous research, the pronunciation difficulty of sentences (henceforth, pronounceability) has not been examined extensively. Given that readability and the difficulty of listening influence learners' motivation and outcomes (Hwang 2005, Lai 2015, Yoon et al. 2016), we consider that CALL for pronunciation learning should consider pronounceability in evaluating learners' pronunciation.

Pronounceability can be represented as the phonetic edit distance from reference pronunciation to a learner's expected pronunciation based on the proficiency. Phonetic edit distance can be measured using a modified version of the Levenshtein edit distance (Wieling et al. 2014) or a deep-neural-network-based classifier (Li et al. 2016).

This study measured normalized edit distance (NED) using the orthographical transcription of learners' pronunciation of reference sentences. An advantage of the NED based on orthographic transcription is the availability of data. This is because language teachers can obtain orthographical transcription without being trained for phonetic transcription.

This study measures pronounceability using multiple regression analysis considering orthographic NED as a dependent variable and the features of a sentence and a learner as independent variables. First, a corpus for multiple regression analysis is developed. This corpus includes the data for NED and the proficiency data in a score-based scale of Test of English for International Communication (TOEIC). TOEIC is a widely used English test in Asian countries, and its test score ranges from 10 to 990. In previous research (Grahma et al. 2015, Delais-Roussarie 2015, Gósy et al. 2015), proficiency was demonstrated using a point-scale such as the Common European Framework of Reference for Languages (six levels from A1 to C2).

This study assessed our phonetic learner corpus data by answering the following research questions:

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 119–124
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

- How stable is NED as a pronounceability index?
- To what extent does NED classify learners depending on their proficiency?
- How strongly does NED correlate with a learner’s proficiency?
- How accurately is NED measurable based on linguistic and learner features for pronounceability measurement?

2 Compilation of Phonetic Learner Corpus

2.1 Collection of Pronunciation Data

Our phonetic learner corpus was compiled by recording pronunciation data for English texts that learners read aloud sentence by sentence. In addition, after reading a sentence aloud, learners subjectively determined the pronounceability of sentences on a five-point Likert scale (1: easy; 2: somewhat easy; 3: average; 4: somewhat difficult; 5: difficult) (henceforth, SBJ).

The texts for reading aloud (the title of Text I is the North Wind and the Sun and that of Text II is the Boy who Cried Wolf) were selected from the texts distributed by the International Phonetic Association (International Phonetic Association 1999). Even though these texts contain only 15 sentences, they cover the basic sounds of English (International Phonetic Association 1999, Deterding 2006). This enables us to analyze which types of English sounds influence learners’ pronunciation. Deterding (2006) reported that Text I failed to cover certain sounds, such as initial and medial /z/ and syllable-initial /θ/, and then developed material that covered the English pronunciation for these sounds by rewriting a well-known fable by Aesop (Text II).

The corpus data were compiled from 50 learners of English as a foreign language at university (28 males, 22 females; mean age: 20.8 years (standard deviation, SD, 1.3)). The learners were compensated for their participation. In our sample, the mean TOEIC score was 607.7 (SD, 186.2). The minimum and maximum scores were 295 and 900, respectively.

2.2 Annotation of Pronunciation Data

Our phonetic learner corpus includes NED, the linguistic features of sentences, and learner features.

NED was derived as the Levenshtein edit distance normalized by sentence length. It reflected the differences from the reference sentences to the transcription of learners’ pronunciation due to the substitution, deletion, or insertion of letters. Before measuring the edit distance, symbols such as commas and periods were deleted and expressions were uncapitalized in the transcription and reference data.

The pronunciation was manually transcribed by a transcriber who was a native speaker of English and trained to replicate interviews and meetings but was unaccustomed to the English spoken by learners. The transcriber examined the texts before starting the transcription task. The transcriber was required to replicate learners’ pronunciation without adding, deleting, and substituting any expressions for improving grammaticality and/or acceptability (except the addition of symbols such as commas and periods).

Linguistic features were automatically derived from a sentence as follows: Sentence length was derived as the number of words in a sentence. Word length was derived as the number of syllables in a word. The number of multiple-syllable words in a sentence were derived by calculating $\sum_{i=1}^{N}(S_i - 1)$, where n was the number of words in a sentence, and S_i was the number of syllables in the i-th word (Fang 1966). This derivation eliminated the presence of 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 1 summarizes the linguistic features of the texts that learners read aloud, i.e., text length and

	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)
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 1: Linguistic features of the texts that learners read aloud.

the mean (standard deviation, *SD*) values of sentence length, word length, multiple-syllable words, and word difficulty.

Learner features were determined using the scores of TOEIC for the current or previous year. Even though TOEIC consists of listening and reading tests, it is strongly correlated with the Language Proficiency Interview, which is a well-established direct assessment of oral language proficiency developed by the Foreign Service Institute of the U.S. Department of State (Chauncey Group International 1998).

3 Properties of Phonetic Learner Corpus

Our phonetic learner corpus was compiled using the method described in Section 2, and this corpus included 750 instances (15 sentences read aloud by 50 learners). Table 2 shows the descriptive statistics for NED and SBJ in the phonetic learner corpus.

	NED	**SBJ**
Minimum	0.01	1
Maximum	0.78	5
Mean	0.15	3.03
SD	0.22	0.91
n	750	750

Table 2: Descriptive statistics of NED and SBJ.

The relative frequency distributions of NED and SBJ, in which NED was classified into five levels based on SBJ, are shown in Figure 1. The distributions are dissimilar, as the peak of NED appears at pronounceability level 2 ("somewhat easy") while that of SBJ appears at pronounceability level 3 ("average"). If NED appropriately accounts for learners' pronounceability, learners appear to overvalue pronounceability. On the contrary, if NED fails to explain pronounceability, learners appear to undervalue pronounceability. This provides a solution for the improvement of NED.

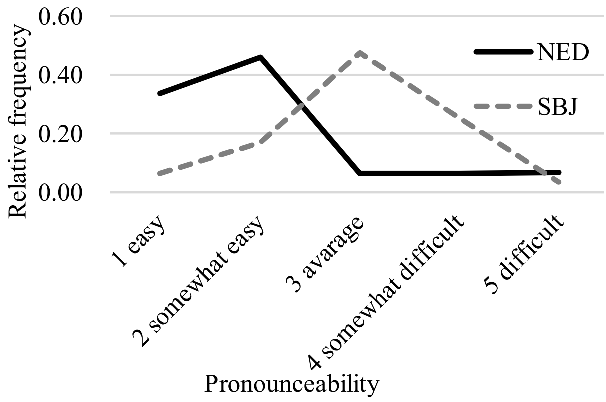

Figure 1: Distribution of NED and SBJ.

4 Assessment of NED as a Pronounceability Index

In Sections 4.1, 4.2, and 4.3, research questions 1–3 are assessed using the classical test theory (Brown 1996). The fourth question is answered in Section 4.4.

4.1 Reliability of NED

The reliability of NED was examined through internal consistency in terms of Cronbach's α (Cronbach 1970). Internal consistency refers to whether NED demonstrates similar results for sentences with similar pronounceability. Cronbach's α is a reliability coefficient 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 is the number of items (sentences 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 reliability coefficient ranges from 0 (absence of reliability) to 1 (absolute reliability), and empirical satisfaction is achieved with values above 0.8.

As reliability depends on the number of items, the reliability 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 of NED and SBJ are shown in Table 3.

	NED	**SBJ**
Text I	0.72	0.80
Text II	0.82	0.91
Text I & II	0.86	0.92

Table 3: Cronbach α coefficient of NED and SBJ.

The reliability coefficient of NED exceeded the value required for empirical satisfaction (α = 0.8) in Text II and Texts I & II. Hence, NED is partially reliable as a pronounceability index. However, NED demonstrated lower reliability compared to SBJ. This suggests that NED should be improved through modification.

4.2 Construct Validity of NED

Construct validity was examined from the viewpoint of distinctiveness. If NED appropriately reflects learners' proficiency, NED should demonstrate a statistically significant difference ($p < 0.01$) among learners at different proficiency levels. Our phonetic learner corpus data were classified into three levels based on the TOEIC scores below 490 (beginner level) ($n = 240$), below 730 (intermediate level) ($n = 240$), and 730 or above (advanced level) ($n = 270$).

Table 4 shows the mean (SD) values of NED and SBJ for the three levels. The distinctiveness of NED was investigated using ANOVA. ANOVA showed statistically significant differences between the three levels of learners for SBJ (F (2, 747) = 10.13, $p < 0.01$) but not for NED (F (2, 747) = 0.55, $p > 0.01$). NED failed to demonstrate construct validity depending on TOEIC-based proficiency.

	Beginner level	**Intermediate level**	**Advanced level**
NED	0.13 (0.21)	0.12 (0.22)	0.11 (0.21)
SBJ	3.15 (0.95)	3.13 (0.92)	2.83(0.83)

Table 4: Descriptive statistics of NED and SBJ

4.3 Criterion-related Validity of NED

Criterion-related validity was examined from the viewpoint of the correlation with learners' proficiency in terms of TOEIC scores. NED should reflect learners' proficiency because pronounceability should depend on learners' proficiency. Then, the correlation between NED and TOEIC scores and between SBJ and TOEIC scores was examined.

NED exhibited weaker correlation with TOEIC scores ($r = –0.04$) compared to SBJ ($r = –0.20$). Owing to this, NED failed to demonstrate criterion-related validity depending on TOEIC-based proficiency.

4.4 Pronounceability Measurement

Pronounceability was measured through multiple regression analysis. NED was the dependent variable, and the linguistic and learner features described in Section 2 were the independent variables. However, multiple-syllable words were not used owing to the variance inflation factor (VIF = 12.3) (Kutner et al. 2002). A significant regression equation was found (F (4, 745) = 124.15, $p < 0.01$) with an adjusted squared correlation coefficient (R^2) of 0.40, which indicates that the equation measured approximately 40% of the pronounceability.

The contribution of linguistic and learner features can be observed using standardized particle regression coefficients; the contribution increases with the absolute value of the coefficients. The standardized partial regression coefficients are summarized in Table 5. Significant contribution is observed in word difficulty but not in the other features. This result contradicts the finding of previous research, which reported the significant contribution of sentence length and word length in other modes such as readability (Crossley et al. 2017) and listening difficulty (Messerklinger 2006).

Variable	**Coefficient** *$p < 0.01$
Sentence length	–0.07
Word length	0.06
Word difficulty	0.61*
TOEIC score	–0.04

Table 5: Standardized partial regression coefficients.

The pronounceability measurement method was examined n times ($n = 750$) using a leave-one-out cross validation test, considering one instance as test data and $n – 1$ instances as training data. The measured NED exhibited moderate correlation with the observed NED ($r = 0.63$). NED demonstrated a low coefficient of determination and low predictability.

5 Conclusion

This study assessed whether NED appropriately demonstrated pronounceability for learning the pronunciation of English as a foreign language. The assessment suggests that NED is reliable (Section 4.1) but not valid (Sections 4.2 and 4.3). The results of pronounceability measurement (Section 4.4) suggest that NED was appropriately explained by the word difficulty.

In future, we will work on the improvement of pronounceability measurement in English based on NED and investigate pronounceability measurement in Asian languages as a foreign language.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers, 22300299, 15H02940, 17K18679.

References

Renlong Ai and Feiyu Xu. 2015. A system demonstration of a framework for computer assisted pronunciation training. In *Proceedings of the ACL-IJCNLP 2015 System Demonstrations*. Association for Computational Linguistics and the Asian Federation of Natural Language Processing, pages 1–6. https://doi.org/10.3115/v1/P15-4001.

James Dean Brown. 1996. *Testing in Language Programs*. Prentice-Hall, Englewood Cliffs, NJ.

Chauncey Group International. 1998. *TOEIC Technical Manual*. Chauncey Group International, Princeton, NJ.

Lee Joseph Cronbach. 1970. *Essentials of Psychological Testing 3rd edition*. Harper & Row, New York.

Scott A. Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S. McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6), pages 340–359. http://dx.doi.org/10.1080/0163853X.2017.1296264.

Elisabeth Delais-Roussarie, Fabián Santiago, and Hi-Yon Yoo. 2015. The extended COREIL corpus: First outcomes and methodological issues. In *Proceedings of Workshop on Phonetic Learner Corpora*. Individualized Feedback for Computer-Assisted Spoken Language Learning Project, pages 57–59.

Irving E. Fang. 1966. The “Easy listening formula.” *Journal of Broadcasting* 11(1), pages 63–68. https://doi.org/10.1080/08838156609363529.

Mária Gósy, Dorottya Gyarmathy, and András Beke. 2015. The development of a Hungarian-English learner speech database and a related analysis of filled pauses. In *Proceedings of Workshop on Phonetic Learner Corpora*. Individualized Feedback for Computer-Assisted Spoken Language Learning Project, pages 61–63.

Calbert Graham. 2015. Phonetic and prosodic features in automated spoken language assessment. In *Proceedings of Workshop on Phonetic Learner Corpora*. Individualized Feedback for Computer-Assisted Spoken Language Learning Project, pages 37–40.

Yukari Hirata. 2004. Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning* 17(3–4), pages 357–376. https://doi.org/10.1080/0958822042000319629.

Myung-Hee Hwang. 2005. How strategies are used to solve listening difficulties: Listening proficiency and text level effect. *English Teaching* 60(1), pages 207–226. https://doi.org/10.3968/7538.

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/S0952675700003894.

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

Christos Koniaris. 2014. An approach to measure pronunciation similiarity in second language learning using Radial Basis Function Kernel. In *Proceedings of the Third Workshop on NLP for Computer-assisted Language Learning*. The Fifth Swedish Language Technology Conference, pages 74–86.

Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. 2002. *Applied Linear Statistical Models* (5th ed.), McGrawHill/Irwin, New York.

Degang Lai. 2015. A study on the influencing factors of online learners’ learning motivation. *Higher Education of Social Science* 9(4), pages 26–30. https://doi.org/10.3968/7538.

Wei Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee. 2016. Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. The Institute of Electrical and Electronics Engineers, pages 6135–6139. https://doi.org/10.1109/ICASSP.2016.7472856.

Sze-Chu Liu and Po-Yi Hung. 2016. Teaching pronunciation with computer assisted pronunciation instruction in a technological university. *Universal Journal of Educational Research* 4(9), pages 1939–1943. https://doi.org/10.1016/S0167-6393(99)00044-8.

Brian Mak, Manhung Siu, Mimi Ng, Yik-Cheung Tam, Yu-Chung Chan, Kin-Wah Chan, Ka-Yee Leung, Simon Ho, Fong-Ho Chong, Jimmy Wong, and Jacqueline Lo. 2004. PLASER: Pronunciation learning via automatic speech recognition. In *Proceedings of HLT-NAACL Workshop on Building Educational Applications using Natural Language Processing*. Association for Computational Linguistics, pages 1–8. https://doi.org/10.3115/1118894.1118898.

Josef Messerklinger. 2006. Listenability. *Center for English Language Education Journal* 14, pages 56–70.

Martijn Wieling, Jelke Bloem, Kaitlin Mignella, Mona Timmerneister, and John Nerbonne. 2014. Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change* 4, pages 253–269. https://doi.org/10.1163/22105832-00402001.

Silke Maren Witt and Steve Young. 2000. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication* 30(2–3), pages 95–108. https://doi.org/10.1016/S0167-6393(99)00044-8.

Su-Youn Yoon, 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, pages 1–6. https://doi.org/10.18653/v1/W16-0531.

Tongmu Zhao, Akemi Hoshino, Masayuki Suzuki, Nobuaki Minematsu, and Keikichi Hirose. 2012. Automatic Chinese pronunciation error detection using SVM trained with structural features. In *Proceedings of Spoken Language Technology Workshop*. The Institute of Electrical and Electronics Engineers, pages 473–478. https://doi.org/10.1109/SLT.2012.6424270.