The relationship between word error rate in automatic speech recognition and proficiency of L2 speech

Yusuke Kondo[1], Mariko Abe[2], Yutaka Ishii[1], and Yuichiro Kobayashi[3]

[1]Waseda University, [2]Chuo University, [3]Nihon University

yusukekondo@waseda.jp

Abstract

Pronunciation specialists would agree that comfortable intelligibility (sometimes, minimal intelligibility) is an appropriate goal of pronunciation for EFL learners. To achieve such intelligibility, EFL learners should master a set of phonetic properties, which is essential for intelligibility; Not all phonetic properties should be learnt. "Not to achieve native-like pronunciation, but to obtain comfortable intelligibility" is not a new idea. Some models of pronunciation have been proposed on the basis of this idea, but we have no evidence that such models are beneficial for learners. We have not provided learners with an appropriate pronunciation model yet. In this paper, we report on word recognition rates of second language speech by automatic speech recognition, which indicates that the word recognition rate serves as an important information for building pronunciation model for second language.

Keywords

L2 pronunciation model, Automatic speech recognition

1 Introduction

Pronunciation is one the important factors in L2 learning. Spyra-Kozlowska (2014) gives some reasons why pronunciation should be taught.

- The first impression can be formed by the pronunciation.
- No efficient oral communication is possible without good pronunciation.
- Phonetic errors can lead to misunderstandings and even communication breakdowns.
- People with poor pronunciation often lack the confidence to speak up and try to say as little as possible.

However, pronunciation training/teaching is often neglected in L2 learning. Effort on pronunciation learning/teaching can bring unsatisfactory results. We usually focus on aspects of language which are teachable and learnable rather than pronunciation. Pronunciation training is often neglected in the teacher training in Japan. Some of the Japanese teachers of English lack their confidence in their pronunciation.

When we teach pronunciation, we are required to set a goal of teaching. The traditional goal is to achieve native-like pronunciation. Some might have said that the problem seemed to be what native speaker model should be chosen, but Phoneticians repeatedly mentioned “Not native-like pronunciation, but comfortable intelligibility for the goal of teaching pronunciation. Gimson and Cruttenden (2001) set the conditions for simplified form of pronunciation of English as follows.

1. It should be at least as easy, and preferably, easier, for the foreign students to learn as any natural model.
2. It should be readily intelligible to most native speakers of English.
3. It should provide a base for the learners who have acquired it to understand the major natural varieties of English.

Furthermore, Jenkins (2000) proposed a new model of pronunciation. She insisted that pronunciation syllabi are still based on native speakers’ intuition, though we have some evidence that the native-speakers’ intuitions might be inaccurate. Even though the intuitions are correct, they are based on intelligibility for native speakers not for non-native speakers. We can obtain reliable information from learner corpora. Appropriate pedagogical proposal for

pronunciation must be linked to the findings of empirical research on second language communication. However, descriptive models for L2 speech cannot be easily proposed.

In this paper, we propose that word error rates (WER) obtained in automatic speech recognition (ASR) can be one of the indices for constructing L2 speech model. We show two different research projects that adopted ASR to transcribe L2 speech and the differences in WER by the proficiency level.

2 Word Error Rate in L2 Speech

2.1 Word Error Rate

WER is one of the indices to evaluate the accuracy of ASR. The equation is below.

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitution; D, that of deletion; I, that of Insertion; and N, the total number of words. WER is edit distance divided by the total number of words.

2.2 Longitudinal Corpus of L2 Speech Corpus

There has been dramatic increase in learner corpora in the last two decades, but majority of them are cross-sectional or pseudo-longitudinal. We cannot obtain complex and unpredictable developmental patterns of L2. Furthermore, what we want to grasp is the way how each individual learner progress or regress in the process of language learning. These are the motivation of this project. The purposes of this project are to collect the same learners' task performances, to construct a longitudinal spoken corpus of Japanese EFL learners, and to describe L2 speaking performance in chronological order, not only as a whole group, but also individual basis.

Table 1. WER by Level

Level	*M*	*SD*	Min.	Max
2	0.81	0.16	0.5	1
3	0.67	0.17	0	1
4	0.64	0.13	0.27	1
5	0.5	0.14	0.1	0.81
6	0.36	0.08	0.19	0.58

We collected the data from 122 Japanese senior high school students. All the student responded to 10 different open-ended questions twice a year. We started this project in 2016, and at the moment have done with 6 data collections. We have about 7320 utterances in total (122 students x 10 questions x 2 times x 3 year). Each utterance was scored based on 9 levels of English proficiency ranging from Level 1 (Novice) to Level 9 (Advanced).

As shown in Table 1, the averages of WER decrease from the lowest level to the highest.

3 Conclusion

As for pronunciation teaching, we have at least two interpretations of the results. If learners have pronunciation training, they will obtain a higher score. So that pronunciation should be taught. On the other hand, Pronunciation will be better, if leaners can perform better. So that, Pronunciation does not need to be taught.

To build ASR system, we construct a mathematical model of speech. The results of ASR can be interpreted as the distance to the model. If the WER is high, the speech is not similar to the model, which might mean that ASR serves as a descriptive model of pronunciation. If your English cannot be recognized by ASR, you should change your pronunciation.

References

Jenkins, J. (2000) *The phonology of English as an international language: new models, new norms, new goals*, Oxford, OUP.

Gimson, A.C., & Cruttenden, A. (2001). *Gimson's pronunciation of English*, Arnold.

Szpyra-Kozlowska, J. (2014). *Pronunciation in EFL instruction: A research-based approach*. Multilingual Matters.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 18K00849 and 16H03455.