Analysis and Comparison of Automatic Language Proficiency Assessment between Shadowed Sentences and Read Sentences

Dean Luo [1], *Nobuaki Minematsu* [1], *Yutaka Yamauchi* [2] *and Keikichi Hirose* [1]

[1] The University of Tokyo
[2] Tokyo International University
dean@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we investigate automatic language proficiency assessment from learners' utterances generated through shadowing and reading aloud. By increasing the degrees of difficulty of learners' tasks for each practice, we examine how the automatic scores, the conventional GOP and proposed F-GOP, change according to the cognitive loads posed on learners. We also investigate the effect and side-effect of MLLR (Maximum Likelihood Linear Regression) adaptation on shadowing and reading aloud. Experimental results show that shadowing can better reflect the learners' true proficiency than reading aloud. Global MLLR adaptation can improve the evaluation performances on reading aloud more significantly than shadowing. But the performance is still better in shadowing. Finally we show that, by selecting native utterances of adequate semantic difficulty, the evaluation performance by shadowing is even improved.

1. Introduction

Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages. Shadowing is a kind of "repeat-after-me" type exercise, but rather than waiting until the end of the phrase heard, learners are required to reproduce nearly at the same time. Although shadowing was originally designed to train simultaneous interpreters, its effects on foreign language learning have been widely recognized and being used in classrooms [1, 2, 3]. Studies show that, in shadowing, speakers can hardly imitate the presented speech only, but use language knowledge of their mother tongue unconsciously as well [4]. Thus shadowing productions can be good indicators of the learners' true language proficiency and, in [5], automatic assessment of shadowing utterances were examined.

Reading aloud has always been a popular practice to improve speaking skill in language learning. Unlike shadowing, utterances generated through reading aloud, or so-called read speech, are more stable and closer to the speaking style of the speech corpuses on which acoustic models (HMMs) are often trained. Therefore, read speech is often used for automatic pronunciation evaluation. Improving the evaluation performance on read speech is also one of the goals of our research.

In this study, we compare shadowing to the conventional practice of reading aloud and in order to examine how cognitive loads affect learners' speech, we also consider two situations of shadowing with and without text presented. With text, the difficulty of shadowing is reduced. We use Goodness of Pronunciation (GOP) based scores calculated through HMMs as automatic scores. Correlations between automatic scores and speakers' TOEIC overall proficiency scores are investigated to analyze the results based on the tasks posed on learners with various cognitive loads.

2. Automatic scoring

2.1. Goodness of Pronunciation (GOP)

Various techniques using HMMs have been tried in many studies to evaluate pronunciation. The confidence-based pronunciation assessment, which is referred to as the Goodness of Pronunciation (GOP), is often used for assessing speakers' articulation and shows good results on read speech [6, 7]. In this study, we use HMM acoustic models trained on WSJ and TIMIT corpuses to calculate GOP scores defined as follows. For each acoustic segment $O^{(p)}$ of phoneme p, GOP(p) is defined as posterior probability and it is calculated by the following log-likelihood ratio.

$$GOP(p) = \frac{1}{D_p}\log(P(p \mid O^{(p)})) \quad (1)$$

$$= \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q\in Q} P(O^{(p)} \mid q)P(q)}\right) \quad (2)$$

$$\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q\in Q} P(O^{(p)} \mid q)}\right), \quad (3)$$

where $P(p \mid O^{(p)})$ is the posterior probability that the speaker uttered phoneme p given $O^{(p)}$, Q is the full set of phonemes, and D_p is the duration of segment $O^{(p)}$. The numerator of equation 3 can be calculated by scores generated during the forced Viterbi alignment, and the denominator can be approximately attained by using continuous phoneme recognition.

Since the boundaries of phoneme p yielded from forced alignment do not necessarily coincide with the boundaries of phoneme q resulted from continuous phoneme recognition, the frame average log likelihoods of the same speech segment are often used in traditional GOP calculation [6].

2.2. Constrained use of speaker adaptation

Our previous analysis [8] has showed that global MLLR adaptation (with only 1 regression class) can improve results of pronunciation assessment on read speech from ERJ (English Read by Japanese Students) corpus [9]. The corpus contains proficiency labels rated by phonetic experts. In order to investigate the effect of MLLR adaptation on read speech (reading aloud) evaluation, we selected 42 learners (21 males and 21 females) with higher agreement among raters and a inter-learners variety of proficiency. The average phoneme

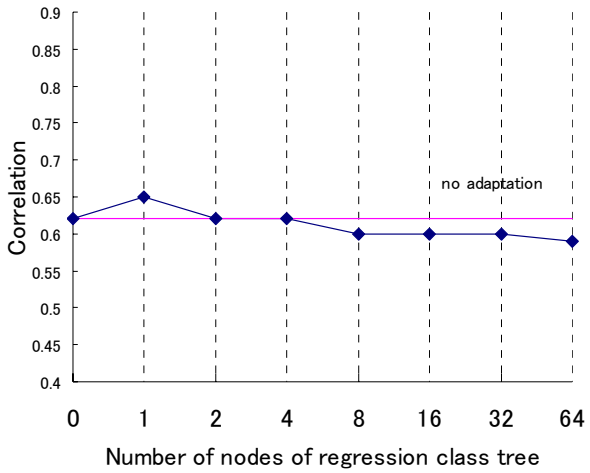

Figure 1: Correlations between GOP scores and manual scores using data from ERJ database as the number of regression classes in MLLR increases

GOP score over 30 sentences read by each learner is calculated and used as an automatic score for the learner. 60 sentence utterances of each leaner were used as adaptation data.

We investigate the correlations between GOP scores and human scores while increasing the number of the nodes of regression class tree. The results are shown in Fig 1. Here the number 0 means without adaption, and 1 represents global adaption. Global adaptation yielded the best correlation of 0.65, yet while the number of nodes of regression class tree increases from 2, the performance drops. When the number is larger than 4, the correlation is even worse than the original model. Over-adaptation of HMMs to learners tends to evaluate inadequate pronunciation as correct.

Based on the results on the above analysis, in this study, we used only 1 regression class for MLLR speaker adaptation. As mentioned previously, utterances of reading aloud were used as adaptation data.

2.3. Forced-aligned GOP (F-GOP)

Conventional GOP calculation refers to the results of both forced alignment and continuous phoneme recognition. This causes a problem as depicted in (a) of Figure 2, that there might be 3 phonemes resulting from continuous phoneme recognition, which correspond to one forced aligned phoneme p. In this case, GOP score for p is calculated using the log likelihood of p and average log likelihood of q1, q2 and q3 within the segment of p [6].

As an alternative way of calculating GOP score, we can first obtain the phoneme boundaries for phoneme p based on the result of forced alignment, and then calculate the posterior probability of that segment using equation (3) directly. We call this method Forced-aligned GOP (F-GOP). This method always refers to the boundaries of forced alignment and actually separates the calculation of GOP score into two processes, one is detecting the phoneme boundaries and the other is calculating the posterior probability for that segment. We can use different models for the two processes. We used the same data set as mentioned in 2.2 to evaluate the performance of F-GOP. We tested two different combinations of acoustic models for detecting phoneme boundaries and calculating posterior probabilities. Figure 3 shows the results of three

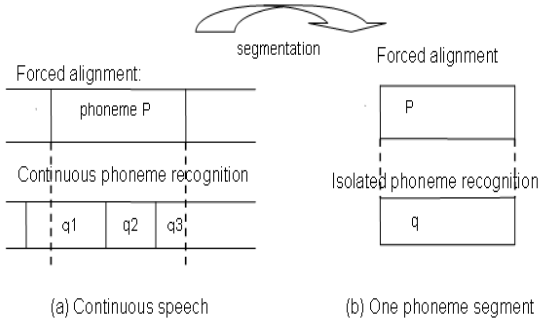

Figure 2: Forced-aligned GOP method

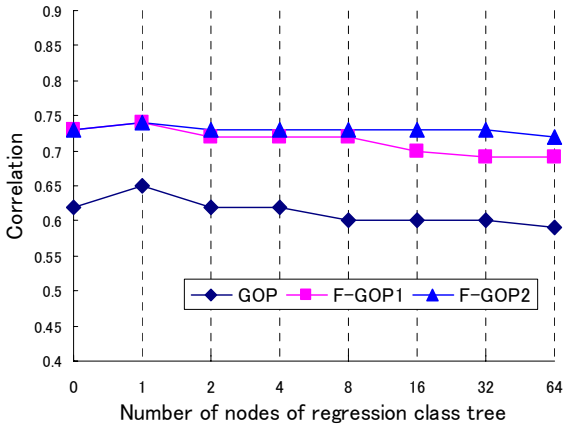

Figure 3: Correlations between human scores and Forced-aligned GOP, compared with conventional GOP

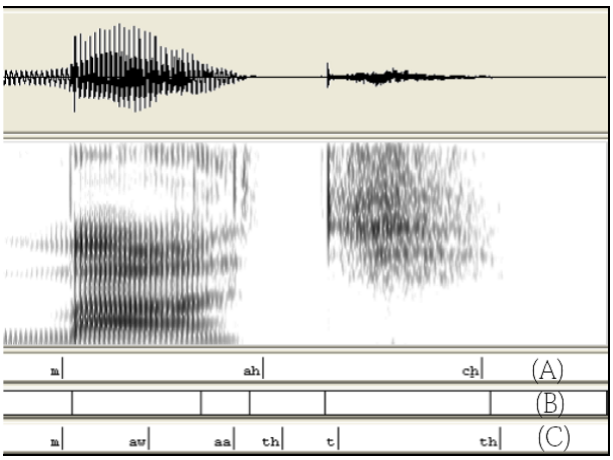

Figure 4: Phoneme segmentation results, A) forced alignment, B) unsupervised bottom-up clustering, C) continuous phoneme recognition

conditions: F-GOP1, which used the same set of models for both phoneme boundary detection and posterior probability calculation, F-GOP2, which used the adapted models (#classes $\geq$ 1) to detect phoneme forced alignment boundaries, and the original models to calculate posterior probabilities, and the conventional GOP scores.

Table1. *Subjects' TOEIC scores*

Proficiency	TOEIC scores	Average
Advanced	955, 926, 855, 832, 825, 792, 773, 752	838
Intermediate	687, 686, 668, 563, 524,	625
Beginners	496, 425, 399, 378, 252	392

As shown in Figure 3, both kinds of F-GOP outperformed the conventional GOP. We consider this is because F-GOP did not refer to the results of continuous phoneme recognition which is often unreliable. Figure 4 shows an example of phoneme segmentation results of A) forced alignment, B) unsupervised bottom-up clustering and C) continuous phoneme recognition. In this example, the result of continuous phoneme recognition is even worse than segmentation based on unsupervised clustering [10], which uses no prior knowledge at all.

F-GOP2 shows better performance than F-GOP1, especially when the number of the nodes of regression class tree is larger than 2. The only difference between F-GOP1 and F-GOP2 is that while F-GOP1 used the adapted models to calculate posterior probabilities, F-GOP2 used the original models to evaluate the same phoneme segment.

3. Experiments

3.1. Shadowing productions collection

In order to compare shadowing with reading aloud, we have designed a program to record learners' utterances in three modes with different levels of phonation difficulty: shadowing (only native model utterances are presented), reading aloud (only texts are presented), and shadowing with texts (both native model utterances and texts are presented). In shadowing and shadowing-with-text modes, learners were required to repeat at the same speed as that of the presented native utterances, but in reading-aloud mode, learners were allowed to read the presented text at his/her own pace. For each mode, the contents of presented utterances or texts were carefully selected by experts so that they contain three levels of semantic difficulty: easy, intermediate, and difficult. The subjects were instructed to first record their shadowing productions, then shadowing with text and finally reading aloud of each task with different level of semantic difficulty. Utterances under these conditions were collected from 18 Japanese learners with a variety of proficiency.

We use TOEIC (Test of English as International Communication) scores as the references of learners' overall language proficiency. The subjects' TOEIC scores are shown in table 1.

3.2. Acoustic conditions for analysis

For automatic score calculation, 39-dimensional feature vectors, consisting of 12-dimensional MFCC, log-energy, and their first and second derivatives, were extracted from utterances using a 25 ms-length window shifted every 10 ms. The CMS (cepstral mean subtraction) was applied to each utterance unit.

3.3. Automatic scores

Average phoneme GOP and F-GOP scores were calculated as automatic scores for each subject by using their utterances of shadowing, shadowing with text, and reading aloud.

Table2. *Correlations between GOP scores and TOEIC scores without adaptation*

Level of difficulty	Shadowing	Shadowing with text	Reading aloud
Easy	0.74	0.65	0.48
Intermediate	0.81	0.68	0.59
Difficult	0.71	0.67	0.61

Table3. *Correlations between GOP scores and TOEIC scores with MLLR adaptation*

Level of difficulty	Shadowing	Shadowing with text	Reading aloud
Easy	0.74	0.68	0.60
Intermediate	0.82	0.71	0.68
Difficult	0.70	0.69	0.67

The acoustic models include the original models trained on WSJ and TIMIT corpuses and the models globally adapted with a part of the subjects' utterances of reading aloud.

3.4. Comparison of shadowing, shadowing with text and reading aloud by using GOP scores

The correlations between GOP scores and TOEIC scores are shown in Table 2. In all tasks with 3 different levels of difficulty, GOP scores calculated from shadowing showed the highest correlations. The results from shadowing with text are lower than shadowing but better than reading aloud. Shadowing with the intermediate level of semantic difficulty shows the highest correlation of 0.81. This indicates that the contents of shadowing need to be carefully chosen to better measure learners' proficiency.

We then applied MLLR adaption by using a part of each learner's utterances from reading aloud to the native acoustic models. The results are shown in table 3. Although the improvement of reading aloud utterances are more significant than shadowing, automatic scores calculated from shadowing utterances still show better performances. This further confirms the advantage of shadowing over reading aloud in overall language proficiency assessment.

3.5. Correlations between F-GOP scores and TOEIC scores

Figure 5 shows the results of correlations of F-GOP scores and TOEIC by using original HMM acoustic models (without adaptation) with three different levels of difficulty: easy, intermediate and difficult, compared with GOP. Figure 6 shows the performance of F-GOP scores with/without MLLR speaker adaptation. As shown in Figure 5, although F-GOP without adaptation did not improve the scoring performances on shadowing, the improvement on read speech (reading aloud) is rather significant. We consider this might because the forced aligned boundary information F-GOP refers to is not as accurate in the case of shadowing as that of read speech. As shown in Figure 6, with MLLR adaptation, the performance of F-GOP can be further improved.

4. Discussion

In every different task, shadowing has shown better results than reading aloud. This indicates that shadowing, which poses a certain amount of cognitive load on learners, can better reflect the true language proficiency of the learners.

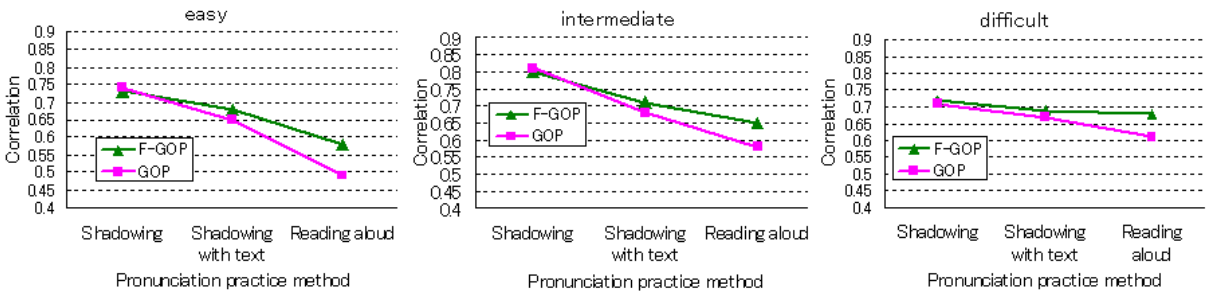

Figure 5: Performances Comparison between F-GOP and GOP without adaptation

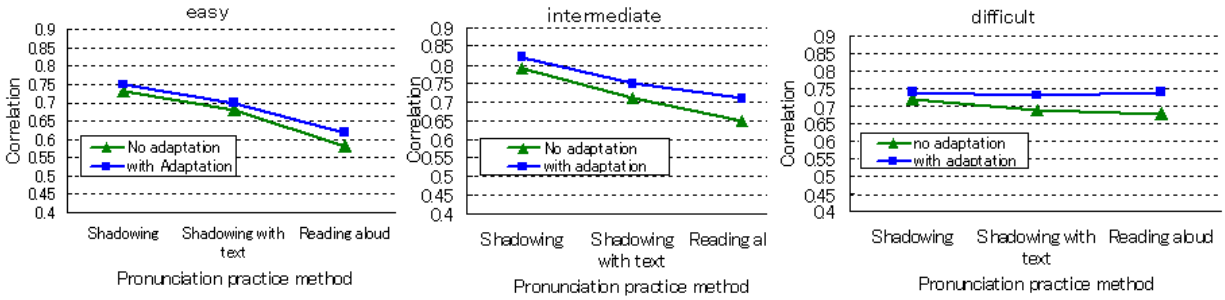

Figure 6: Effect of MLLR adaptation on the performances of F-GOP

However, MLLR adaptation, which improved the results of reading aloud significantly, did not improve the performances of shadowing evaluation as much. We considered that it is because the difference of the speaking style between shadowing and reading aloud, even by the same speaker, causes much of the mismatches between utterances generated through shadowing and the original acoustic models. The use of read speech as adaptation data can not reduce the mismatches caused by the difference of speaking style. In order to further improve the performance of shadowing evaluation, we need to address the problems caused by the speaking style of shadowing in the future.

5. Conclusions

In this paper, we compare automatic proficiency assessment results on utterances generated through three different ways of pronunciation practices: shadowing, shadowing with text, and reading aloud. Three different degrees of difficulty of the presented text or native utterances are employed to examine the effects of cognitive loads posed on learners. Experimental results show that shadowing with a proper degree of difficulty, or cognitive load, can be used to assess language learners' proficiency with the best accuracy. We also analyze the effect of MLLR adaptation on automatic scores and find out that MLLR improves the performances on reading out significantly but little improvement is found on shadowing. We are planning to investigate the change of learner's proficiency after routinely shadowing practices over a period of time.

6. References

[1] T.Hori, "Exploring Shadowing as a Method of English Pronunciation Training," A Doctoral Dissertation Presented to the Graduate School of Language Communication and Culture, Kwansei Gakuin University. 2008

[2] S.Miyake, "Cognitive processes in phrase shadowing and EFL listening," *JACET Bulletin* Tokyo: Japan Association of College English Teachers. Forthcoming

[3] H.Mochizuki, "Shadowing and English language learning," Unpublished MA thesis, Kwansei Gakuin University, 2004

[4] P.W.Nye et al., "Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English," *Journal of Phonetics*, pp.63–69, 2003

[5] D.Luo et al., "Automatic pronunciation evaluation of language learners' utterances generated through shadowing," *Proc. INTERSPEECH*, pp.2807-2810, 2008-9

[6] S.M. Witt and S.J. Young, "Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning," *Speech Communications*, 30 (2–3): pp.95-108, 2000

[7] L.Neumeyer et al., "Automatic scoring of pronunciation quality," *Speech Communications*, 30(2-3): pp.83-93, 2000

[8] D.Luo et al., "Quantitative analysis of the adverse effect of speaker adaptation on pronunciation evaluation," *Proc ASJ Spring Meeting*, pp., 2009

[9] Minematsu et al., "English Speech Database Read by Japanese Learners for CALL System Development," *Proceedings of International Conference on Language Resources and Evaluation*, pp896-903, 2002

[10] N. Shimomura et al., "Automatic segmentation of continuous speech based on time-constrained bottom-up clustering," *Proc. ASJ Autumn Meeting*, pp.353-356, 2007