INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automatic Assessment and Error Detection of Shadowing Speech: Case of English Spoken by Japanese Learners

Shuju Shi [1,2], *Yosuke Kashiwagi* [1], *Shohei Toyama* [1], *Junwei Yue* [1], *Yutaka Yamauchi* [3], *Daisuke Saito* [1]
Nobuaki Minematsu [1]

[1] The University of Tokyo, Japan
[2] Beijing Language and Culture University, P.R. China
[3] Tokyo International University, Japan

{shishuju,kashiwagi,toyama,jwyue,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, yyama@tiu.ac.jp

Abstract

Shadowing is a task where the subject is required to repeat the presented speech as s/he hears it. Although shadowing is cognitively a challenging task, it is considered as an efficient way of language training since it includes processes of listening, speaking and comprehension simultaneously. Our previous study realized automatic assessment of shadowing speech using the average of Goodness of Pronunciation (GOP) scores. But the fact that shadowing often includes broken utterances makes this approach insufficient. This study attempts to improve automatic assessment and, at the same time, give corrective feedbacks to learners based on error detection. We first manually labeled shadowing speech of 10 female and 10 male speakers and defined ten typical error types including word omission, substitution etc.. Forced alignment with adjusted grammar and GOP scores are adopted to detect word omission errors and poorly pronounced words. In the experiments, GOP scores, Word Recognition Rate (WRR), silence ratio, forced alignment log-likelihood scores, word omission rate are used to predict the overall proficiency of the individual speakers. The mean correlation coefficient between automatic scores and the speaker's TOEIC scores is 0.81, improved by 13% relatively. The detection accuracy of word omission is 73%.

Index Terms: shadowing, automatic assessment, corrective feedback

1. Introduction

Technically speaking, shadowing is a paced, high cognitive task where speakers need to immediately vocalize the presented auditory stimuli [1]. Since shadowing includes processes of speaking, listening and comprehension of speech simultaneously [2], it has been employed as a practicing strategy among simultaneous interpreters first and later was also adopted by language teachers. Recent decades have seen the effectiveness of shadowing in language learning [3-5]. [3-4] showed shadowing can improve students' listening comprehension. [3] also suggested that shadowing can enhance learners' phoneme perception ability. [5] showed that shadowing can improve learners' intonation, fluency, word pronunciation and overall pronunciation. And comparison study suggested that shadowing could be more or at least no less effective than extensive reading, reading aloud and listening in improving speaker's corresponding language skills, that is reading comprehension, speaking, and listening comprehension [4,6-7].

The reason why shadowing could benefit language learning probably has its foundation in its processing mechanism. Other than simply repeating, shadowing has shown to involve complex production-perception interaction, automatic semantic and syntactic processing [8-9], and some people even performed sophisticated error correction during shadowing [10-11]. This, plus the fact that shadowing is a combined process of speaking, listening and comprehension, suggests that analytical results of shadowing speech can represent the speakers' overall language proficiency better than those of reading speech [12].

In our previous research, we realized automatic assessment of shadowing speech using the average of Goodness of Pronunciation (GOP) [13]. The result is promising with relative high correlation coefficient between automatic scores and speakers' TOEIC scores. But shadowing speech often includes broken sentences, especially in beginners' data. This makes our previous approach insufficient. In this study, we aim to improve automatic assessment and, at the same time, give corrective feedbacks to learners based on error detection. To investigate the typical phenomena in shadowing speech, we manually labeled data of 20 speakers. Then we designed a system to address these phenomena and realized automatic assessment and error detection of shadowing speech.

2. Corpus Description

We used three sets of data in this study. Set 1 is WSJ dataset, and it includes 80 hours of speech and 37,000 utterances in total. This set is used for initial acoustic and language model training. Set 2 is the final shadowing of 163 advanced students from Kyoto University with their TOEIC simulation test score being no less than 70 (0-100). The texts used in Set 2 are all from a TOEIC simulation textbook and in total 332 passages are selected (about 2 passages/student). Students are allowed to practice as many times as they want with text before their final shadowing (without reference to text). Set-2 is used for acoustic model adaptation. Set 3 contains two subsets, Set 3_1 and Set 3_2. Both are shadowing speech from undergraduate students (Set 3_1: 37 students, Set 3_2 : 39 students) after 2-3 times practicing without reference to the text. Shadowing material used in Set 3_1 is a passage about fugu (puffer fish), which is a familiar topic for Japanese people. It has 333 words in 21 sentences. Shadowing material used in Set 3_2 is a simple conversation between a policeman and a boy who is supposed to have broken into MacDonald's house. It has 142 words in 14 sentences. Both sets in Set 3 are used for assessment and error detection. Data of 20 speakers from Set 3_1 are used for annotation. Table 1 is an overall description

http://dx.doi.org/10.21437/Interspeech.2016-915

of language proficiency by TOEIC scores in Set 3. Table 2 shows the TOEIC scores of 10 female and 10 male from Set 3_1 for annotation.

Table 1. *Language proficiency distribution in Set 3 by TOEIC scores.*

Proficiency level		TOEIC scores
low	Set 3_1	158,197,202,252,275,278,289,301,308, 367,395
	Set 3_2	226,255,311,311,325,368,396
Interme -diate	Set 3_1	421,427,432,436,512,581,592,601,608, 625, 679,
	Set 3_2	410,424,481,552,566,580,594,594,594, 608,622,636,665,677,679,679,693
high	Set 3_1	721,764,778,792,820,825,849,895,905, 905,940,955,968,990,990
	Set 3_2	707,721,721,721,722,764,778,778,792, 792,805,820,849,905,905

Table 2. *TOEIC scores of the annotated speakers.*

Gender	TOEIC score
Female	955,940,895,825,601,592,581,308,301,275
Male	990,990,968,625,436,395,367,289,278,158

3. Annotation and Result

3.1. Typical phenomena in shadowing speech

We manually annotated 20 speakers' (10 female, 10 male) shadowing speech and defined ten prototypes of phenomena or errors in shadowing. Each phenomenon, its brief description, example, and labeling norm used in our research are shown in Table 3.

Table 3. *Typical phenomena in shadowing speech.*

Name	Description and Labeling Norm
Substitution: 1)word-level 2)syllable-level	A(B)/A(<bcd>) means word A is substituted by word B or syllables <bcd>. e.g. The symptoms (sentence) e.g. expensive (<ikstin>)
Omission	A(-B) means the omission of word B e.g. had (-been) poisoned
Grammatical Errors	(sth.--sth.) defines errors that are related to tense and grammar and their combination. e.g.: Works → worked(tps--pt)
Insertion	(+B) means insertion of a word. e.g. (+the)
Repetition 1)syllable-level 2)word-level	Words are partly or fully repeated. e.g. over <+twi--> twice its e.g. very very(+1) expensive
Multi2One	A+B+..+N(=X) means a sequence of words are arranged as a cluster of syllables X. e.g. two hundred+dollars(=hudo)
Mimic	A(*) means word A is shadowed as some sound similar to the presented stimuli but the speaker actually didn't get the semantic meaning of the words.
Spoken Noise	Filled pause, e.g. <uh>, <en>, etc.
Non-spoken Noise	Noise other than spoken noise e.g. <microphone> ,<sniff>, etc.
Whispering	A(*whs) means word A is whispered because the speaker is not sure about what is presented in the stimuli.

3.2. Result of annotation

Figure 1 shows the overall result of the labeling. As can be seen, the most salient error is omission. And we further analyzed the distribution of each error type among different speaking proficiency. The result is shown in Figure 2. An overall tendency is that low level speakers tend to have more errors in each type except for non-spoken noise (NSPN). This suggests that the number of errors could serve as an indicator of the speakers' overall proficiency. In this study, for automatic assessment, we mainly focused on the error type of word omission.

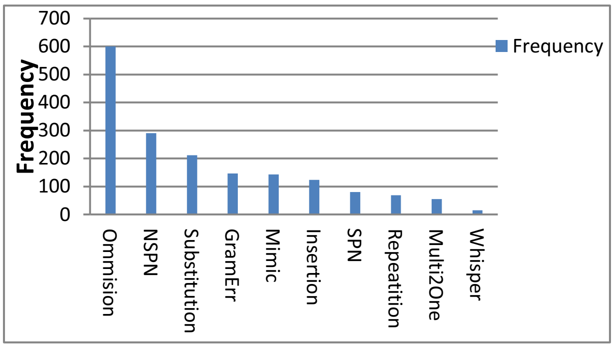

Figure 1: *Overall result of labeling, where NSPN means non-spoken noise and SPN means spoken noise.*

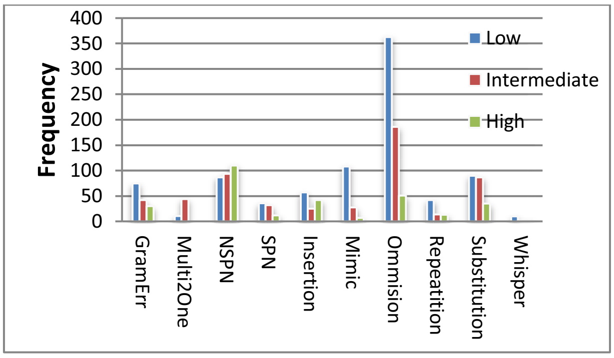

Figure 2: *Result for different proficiency levels.*

4. Design of Features for Assessment

This section explained the features and approach we used to do automatic assessment and error detection based on what we found in the last section.

As for automatic assessment, features we used include Goodness of Pronunciation (GOP) score, force-alignment likelihood score, Word Recognition Rate (WRR), word omission rate and silence ratio. The first three are to measure the pronunciation level of the speaker and the last two serve to incorporate word omission errors into the overall assessment.

4.1. Word omission detection

To detect the omitted words, we firstly trained HMM-based acoustic model using corpus Set 1, then applied Maximum A Posteriori (MAP) adaptation using corpus Set 2 and prepared the grammar where each word presented in the stimuli can be replaced by silence. A short pause can be inserted between words. Then forced-alignment was performed on the assessment data (Set 3). Figure 3 shows the grammar we used. Figure 4 is the comparison of detection results using mono-phone model, tri-phone (tri1 used mono-phone model as initial model, tri2a used results in tri1 as initial model, and tri2b used results in tri1 as initial model and LDA and MLLR for feature level normalization) model.

As Figure 4 shows, HMM-based mono-phone model achieves the highest Detection Accuracy (DA) and lowest False Rejection Rate (FRR) and thus would be used in this study.

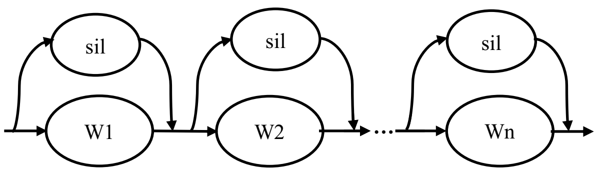

Figure 3: *Grammar for detecting word omission.*

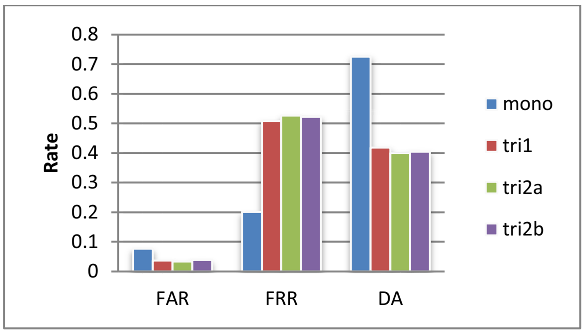

Figure 4: *Detection result of word omission.*

4.2. Design of features

4.2.1. Word recognition rate

Before performing speech recognition, the same procedure of acoustic and language model training and MAP adaptation in Section 4.1 was also done here. The recognition result using mono-phone and tri-phone based models is shown in Figure 5.

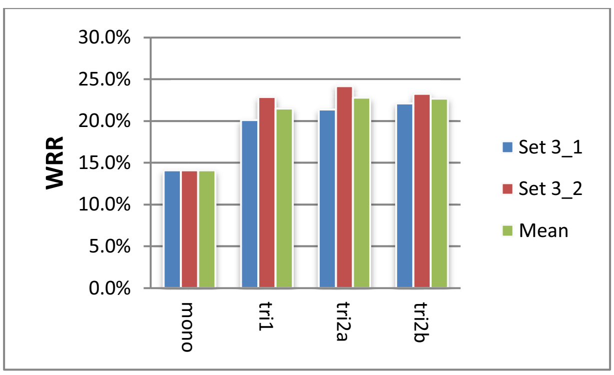

Figure 5: *Result for word recognition rate.*

HMM-based tri-phone achieved better recognition rate than mono-phone based model and the word recognition rate from tri2b are used in this experiment.

4.2.2. GOP and force alignment likelihood score

GOP is often used in assessing speakers' pronunciation proficiency level and is defined as:[14]

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

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

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

where $P(p \mid O^{(p)})$ is the posterior probability of a speaker uttering phoneme p given $O^{(p)}$, Q is the full set of phonemes, and D_p is the duration of segment $O^{(p)}$.

In this study, GOP and force alignment likelihood score is calculated based on acoustic model trained using WSJ and TIMIT [15]. For a given passage utterance, we calculated 1) GOP_P: the averaged GOP score over the presented phonemes [13], 2) Align_P: the averaged force alignment likelihood score over the presented phonemes, 3) GOP_D: the averaged GOP score over the phonemes in detected words, and 4) Align_D: the averaged force alignment likelihood score over the phonemes in detected words.

4.2.3. Word omission rate and silence ratio

Word omission rate is calculated by dividing the number of detected words in the shadowing utterance by the number of words in the corresponding native utterances. Silence ratio is calculated by dividing the duration of silence by the duration of the whole utterance.

5. Automatic Assessment

Based on the aforementioned features, automatic assessment was performed using Support Vector Regression (SVR). The kernel function used is Radial Basis Function (RBF) and optimization was done by grid search with setting of cost function c being [2^5,2^12] and parameter g being [0.01,1]. Leave-one-out cross validation was adopted to predict the target scores. To compare the performance, three sets of features are used in this experiment. Detailed information about each feature set is shown in Table 4. Table 5 shows the correlation between overall GOP score [13] and TOEIC score and that between predicted scores from each feature set with TOEIC score.

Table 4. *Features in each feature set.*

Name	Features
feature_Set1 (5 features)	GOP_D,Align_D,silence ratio, word recognition rate(wrr), word omission rate(wor)
feature_Set2 (5 features)	GOP_P, Align_P, silence ratio, wrr, wor
feaure_Set3 (7 features)	GOP_D, Align_D, GOP_P, Align_P silence ratio, wrr, wor

In this experiment feature_Set2 achieves the best performance with relative improvement of 6% and 21% on Set 3_1 and Set 3_2 respectively, compared with our previous approach using only average GOP score.

Table 5. *Result of Correlation Coefficient.*

	Set 3_1	Set 3_2
GOP_P	0.82	0.61
feature_Set1	0.86	0.61
feature_Set2	0.87	0.74
feature_Set3	0.86	0.72

6. Discussions

6.1. Annotation

6.1.1. Same error type, different strategy

Even though learners of three different proficiency levels share the same error types in our current labeling norm, the underlying mechanism is quite different. High level learners tend to maintain syntactic correctness and semantic connection. For example, in the utterance "… their hand at preparing the fish themselves", one high-level speaker mis-shadowed the word 'at' as to, meanwhile she changed 'preparing' to 'prepare'. On the other hand, errors by low level learners usually reflect their inability to catch what's in the presented stimuli or to repeat what they got correctly.

6.1.2. Female and male difference

The shadowing strategy of female and male are quite different when they encounter something they cannot comprehend. Figure 6 shows the number of spoken noise (SPN) and Mimic words. When Female learners missed the presented stimuli, they tended to keep silent or uttered some filled pause, while male learners would follow the stimuli and uttered some non-meaningful but prosodic similar sounds.

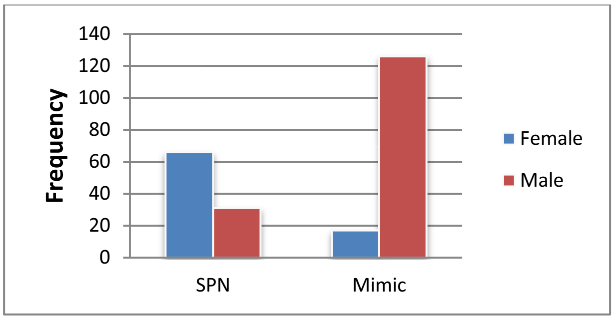

Figure 6: *Number of SPN and Mimic among female.*

These differences should be examined in our future study in both assessment and error detection approach. Also, more labeling data on different texts are needed to further investigate phenomena in shadowing speech and at the same time to constrain text and annotator bias.

6.2. Automatic Assessment

6.2.1. GOP_P v.s. GOP_D

As shown in Table 5, feature set using GOP_P and other parameters got the best correlation coefficient. In fact, we thought GOP_D would achieve better results. This is because in calculating GOP_P, it is assumed that all words are shadowed in the learner's utterance. But this is not often the case especially for low-level learners. Figure 7 shows the alignment result of using all words presented in the audio stimuli (Tier 1) and the alignment using our proposed grammar (Tier 2). Apparently, the alignment result with the new grammar is much better. The lower correlation coefficient using GOP_D compared with GOP_P might be that the former measure is only capable of estimating the pronunciation level of the detected words. Although we have considered factors like word omission rate and silence ratio, it seems more measures are needed to capture the whole picture of the speaker's overall proficiency.

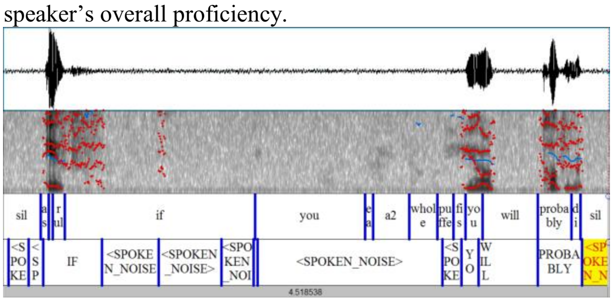

Figure 7: *Comparison of force alignment result.The original text is "As a rule, if you eat a whole puffer fish, you will probably die".*

6.2.2. Corpus dependency

In all feature sets used, the correlation coefficient is higher in corpus Set 3_1 than Set 3_2. The reason might be two-fold: 1) Range of TOEIC score in Set 3_2 (mean =616, std. =183) is smaller than that in Set 3_1 (mean =595, std. =267), and several learners share the same score; 2) difficulty level of these two text are different (Set 3_1 intermediate, Set 3_2 easy), and shadowing performance is highly related to the difficulty level of the text.

7. Conclusions

In this study, we first examined typical phenomena in shadowing speech, then realized automatic assessment and preliminary error detection. What we found are: 1) unlike reading speech, shadowing speech contains more complicated phenomena, such as omission, substitution, insertion, mimic etc.; 2) our proposed grammar with adapted acoustic model is effective in detecting word omission with a detection accuracy of 73%; 3) despite the fact that the alignment accuracy are lower in overall GOP calculation, it is more effective in predicting the learners' TOEIC score than detected-word base GOP scores, at least in our current dataset; 4) shadowing performance is dependent on the difficulty degree of materials and this fact should be considered in doing automatic assessment of shadowing speech.

In the future, we'd like to explore more complimentary measures to improve the effectiveness of detected-based GOP scores and we'd like to address other error types in shadowing speech.

8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP16H03084, JP16H03447, and JP26240022.

9. References

[1] S.Lambert, Shadowing. *Meta: Journal des traducteursMeta:/Translators' Journal*, *37*(2), 263-273, 1992.

[2] W. D. Marslen-Wilson, Speech shadowing and speech comprehension. *Speech Communication*, *4*(1-3), 55-73, 1985.

[3] Y. Hamada, The effectiveness of pre-and post-shadowing in improving listening comprehension skills. *The Language Teacher*, *38*(1), 3-10, 2014.

[4] Y. Hamada, Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. *Language Teaching Research*, 2015.

[5] K. T. Hsieh, D. A. Dong, , & L. Y. Wang, A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics*, *11*(2), 43-66, 2013.

[6] T. Nakanishi, , & A. Ueda,. Extensive reading and the effect of shadowing. *Reading in a Foreign Language*, *23*(1), 1, 2011.

[7] A. Kuramoto, O. Shiki , H. Nishida, & H. Ito, Seeking for effective instructions for reading: The impact of shadowing, text-presented shadowing, and reading-aloud tasks. *LET Kansai Chapter Collected Papers*, *11*, 13-28, 2007.

[8] H. Mitterer, & M. Ernestus, The link between speech perception and production is phonological and abstract: Evidence from the shadowing task.*Cognition*, *109*(1), 168-173, 2008.

[9] P. W. Carey, Verbal retention after shadowing and after listening.*Perception & Psychophysics*, *9*(1), 79-83, 1971.

[10] W. D. Marslen-Wilson,. Sentence perception as an interactive parallel process. *Science*, *189*(4198), 226-228，1975.

[11] W. D. Marslen-Wilson, Linguistic structure and speech shadowing at very short latencies. *Nature*,1973.

[12] D. Luo, N. Minematsu, Y. Yamauchi, & K. Hirose, Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences. In *SLaTE* (pp. 37-40), 2009.

[13] D. Luo, N. Minematsu, Y. Yamauchi, & K. Hirose, Automatic assessment of language proficiency through shadowing. *International Symposium on Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on* (pp. 1-4).

[14] S. M. Witt, & S. J. Young, Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2), 95-108, 2000.

[15] https://www.keithv.com/software/htk/