Speech Analysis for Automatic Evaluation of Shadowing

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

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

Abstract

This paper presents acoustic analysis for the purpose of automatic evaluation of shadowing speech. We use self-checked scores of understanding, manual prosodic scores, and TOEIC scores as reference scores of learners' shadowing speech, and compare these scores with automatic scores based on acoustic features that can reflect phoneme intelligibility and prosodic fluency in terms of intonation, and rhythm. We also examine the differences of personal-best shadowing, shadowing after the transcription is shown and reading-aloud of the same contents. Experimental results show that learners' understanding of contents in shadowing affects segmental intelligibility and prosodic fluency of their shadowing productions. A multiple regression model that combines different features can better reflect learners' understanding of the contents of shadowing and other reference scores, and thus suitable for automatic evaluation of shadowing.

Index Terms: shadowing, Computer-Assisted Language Learning, prosodic evaluation, automatic scoring

1. Introduction

Shadowing is a spoken language training method that requires learners to repeat or shadow a presented native utterance as quickly and closely as possible. Since the transcription is not shown to learners, they have to focus more on pronunciations of the presented native speech and try their best to reproduce them. Studies show that shadowing can improve learners' listening and speaking skills [1].

In our previous works [2], we have proposed several automatic scoring methods for first-time shadowing, where the presented speech has not been seen or heard by the subjects before shadowing. High correlations between automatic scores of first-time shadowing and TOEIC overall proficiency scores have been found. However, we found that learners used different strategies to shadow a given native utterance. For example, some learners might focus on the contents of the presented utterance and repeat individual words with their own style of speaking. Some might focus on segmental phoneme pronunciations and others might only focus on the prosodic features yet ignoring the intelligibility of pronunciations [3, 4].

In order to further analyze segmental and prosodic features of shadowing speech, instead of first-time shadowing, more stable personal-best shadowing utterances, which are recorded after sufficient practices without the transcription, are used for our analysis. Fig. 1 shows a procedure of recording learners' utterances of shadowing and reading-aloud. This study focus on how learners' degree of understanding the contents during shadowing affects their pronunciations in shadowed utterances in terms of phoneme intelligibility and prosodic fluency. To measure learners' degree of understanding the contents, we introduce two types of scores. One is a comprehension test that contains 7 questions. Each question asks learners to choose the best answers out of 4 candidates according to the presented native speech they heard during shadowing. The other is learners' self-check of words that they do not recognize during shadowing. In this case, the transcription of the native speech is shown to the learners and, by referring to it, they are required to mark out any words they did not follow up during the personal-best shadowing. We prepare other two types of scores. We ask a language expert to rate the shadowing utterances in term of prosodic features, intonation and rhythm, and an overall prosodic score is assigned to each subject. TOEIC score is also provided from the learners. For automatic analysis, we use Goodness of Pronunciation scores as the measure for phoneme intelligibility. As for prosodic features, we focus on F0-based and power-based DTW distances between shadowed utterances and the presented native speech, utterance-level variance of F0, length of pauses and rate of speech. The relations between reference scores and automatic scores are examined.

2. Data Collection

32 subjects participated in our shadowing data collection. These subjects are Japanese learners of English from two universities in Japan and their TOEIC scores are shown in Table 1. The contents of presented speech were carefully chosen by a language expert that contains 14 sentences of an intermediate level of difficulty. The presented native speech was provided by an English teacher of native General American English speaker with normal speed but with a variety of changes in intonation. The transcription or speech was never presented to the subjects before recordings. The subjects were asked not only to pay attention to segmental pronunciations, but also to the prosodic features of the presented speech and to mimic them as closely as possible instead of speaking in their own ways.

After recording the first-time shadowing, the subjects were asked to take a comprehension test. The test is written in Japanese with seven questions related to the contents of the presented speech. For each question, the subjects need to choose the best answer out of four candidates. After the comprehension test, the subjects practiced shadowing for several times until they became familiar with the native pronunciations. Then the subjects' personal-best shadowing was recorded. After personal-best shadowing recording, the transcript was presented to the subjects and while listening to

Table 1. *Subjects' TOEIC scores.*

TOEIC scores	Number of subjects
600-800	13
400-600	11
100-400	8

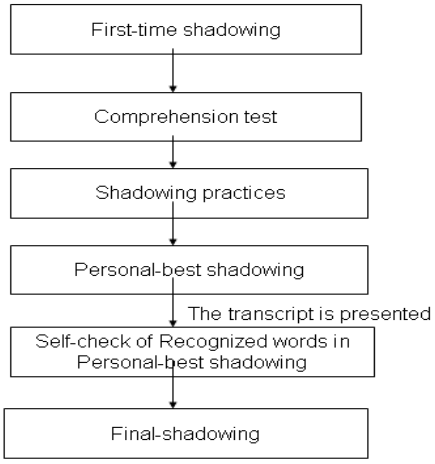

Figure 1: *Recording procedure of shadowing.*

their own recorded personal-best shadowing utterances, they were asked to mark out any words that they did not recognize during shadowing.

Now that the transcript has been shown to the subjects, we record their shadowing speech one more time for comparison with their personal-best shadowing. We will refer to this final shadowing recording as final-shadowing hereafter. Fig. 1 shows the total procedure of a sequence of recordings including a comprehension test.

3. Reference Scores

For reference scores, first, we calculate the number of words that the subjects recognized correctly during shadowing and define recognized word scores (RWS) based on the subjects' self-check results as below.

$$\mathrm{RWS} = \frac{\text{number of recognized words}}{\text{total number of words}} \times 100\% \quad (1)$$

And comprehension test scores (CTS) is defined as,

$$\mathrm{CTS} = \frac{\text{number of correct answers}}{\text{total number of questions}} \times 100\% \quad (2)$$

These two scores measure learners' degree of understanding the contents of the native utterances in different ways. RWS and CTS correspond to word-level comprehension and overall comprehension, respectively.

When learners are asked to shadow presented native utterances, they sometimes pay more attention to the prosodic aspects, intonation and rhythm, of the presented utterances. In order to measure prosodic proficiency, we ask an English education expert to rate a score for each subject based on the expert's subjective impression of that learner's prosodic fluency. We refer to this score as manually-rated prosodic score (MPS). Table 2 shows the correlations between any two of the referenced scores including TOEIC scores.

Table 2 shows the correlations of any 2 of the referenced scores including TOEIC scores.

RWS shows very high correlation with TOEIC overall proficiency scores and manual prosodic scores (MPS). This indicates that the level of word recognition during shadowing not only reflect learners' overall language proficiency but also affects prosodic fluency of shadowed utterances. The relatively low correlation between CTS and MPS might

Table 2. Correlations *between any 2 of the referenced scores.*

	RWS	CTS	TOEIC	MPS
RWS	1	0.53	0.70	0.72
CTS	0.53	1	0.73	0.54
TOEIC	0.70	0.73	1	0.72
MPS	0.72	0.54	0.72	1

indicates that it is possible to mimic prosodic features of the presented speech without comprehending the whole contents.

4. Automatic Scores

4.1. Goodness of Pronunciation Scores

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 [5]. 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.

$$GOP(p) = \frac{1}{D_p}\log(P(p \mid O^{(p)})) \quad (3)$$

$$= \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q\in Q} P(O^{(p)} \mid q)P(q)}\right) \quad (4)$$

$$\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q\in Q} P(O^{(p)} \mid q)}\right), \quad (5)$$

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 (5) 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 [5].

4.2. Scores based on prosodic measures

4.2.1. Fundamental frequency (F0)

In our experiment, F0 is extracted by using Praat, which analyzes F0 every 5 ms with 20ms frames of each utterance. The log scale values of F0 are normalized to cancel the differences due to gender. In addition, F0 pattern is smoothed with regression fitting.

[6] uses the DTW distances between native utterances and learners' read speech as measure for intonation proficiency. In the case of shadowing, the presented native speech is the only source that learners refer to during shadowing. The distances of presented native utterances and learners' shadowed ones are reasonable measure for intonation fluency.

Word-level DTW distances we use is defined as below,

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases} , \quad (6)$$

where *d(i,j)* is a local difference between normalized F0 values of the *i*-th frame of shadowed utterance and the j-th frame of the presented speech and *g(1,1)=d(1,1)*. If the speech segment of a word has *I* frames in native speech, and its corresponding segment has *J* frames in learners' shadowed speech, the DTW distance of this word is calculated by,

$$D(native, learner) = \frac{g(I,J)}{I+J} \quad . \quad (7)$$

We refer to scores calculated by Eq. (8) as F0_DTW. The smaller F0_DTW is, the closer the learner's pitch pattern is to the presented native speech.

According to [7], at utterance level, Japanese learners' pitch contours are more flat than those of native English speakers' are. Thus the variance of normalized F0 at utterance level can be used as an indicator to judge if the learners' shadowed utterances are Japanese-like or native-like.

4.2.2. Power

Power (or intensity) parameters are also extracted by Praat. DTW distances between intensity contours of learners' shadowed speech and the presented native speech calculated in the same way as mentioned in previous section. We refer to these scores as Power_DTW scores.

4.2.3. Length of pauses

Pauses are automatically detected by using a threshold-based scheme for the values of power. Durations of silence segments between words are calculated and normalized by the length of the presented utterance.

4.2.4. Rate of speech

Rate of speech (ROS) is calculated as,

$$ROS = \frac{N_{phonemes}}{D_{utterance} - D_{Silence}} \quad , \quad (8)$$

where $N_{phonemes}$ is the number of phonemes and $D_{utterance}$ is the duration of the utterance and $D_{silence}$ is the length of silence.

5. Evaluation Experiments

5.1. Correlations between automatic scores and reference scores

We investigate correlations between every automatic scores described in section 4 and referenced scores mentioned in section 3.

Table 3. *Correlations between automatic scores and RWS (Recognized word scores).*

Measures	Correlation
GOP	0.63
F0_DTW	-0.45
F0_variance	0.55
Power_DTW	-0.30
Pauses	-0.20
ROS	0.58

Table 4. *Correlation between automatic prosodic scores and MPS (manual prosodic scores).*

Measures	Correlation
GOP	0.60
F0_DTW	-0.55
F0_variance	0.49
Power_DTW	-0.30
Pauses	-0.37
ROS	0.59

Correlations between automatic scores and recognized word scores (RWS) on personal-best shadowing are shown in table 3. GOP scores, F0-based scores and ROS show better results than scores based on power or pauses.

Correlations between automatic segmental and prosodic scores and manual prosodic scores (MPS) are shown in table 4. Again, F0-based scores perform better than Power-based scores and ROS shows better result than Pauses.

5.2. Multiple regression models

We use a set of multiple regression models to combine different measures. The combined scores are given by the following equation,

$$S = \sum_{k=1}^{K} \alpha_k F_k \quad , \quad (9)$$

where F_k is the *k*-th feature score of K scores and α_k is obtained by using training data.

Here we adopted leave-one-out cross verification to estimate target scores with different features.

First, we use the 6 measures shown in table 3 to estimate RWS. The correlation between estimated scores and RWS is 0.68 which higher than any one of the features. Although the result is lower than the correlation between RWS and TOEIC or MPS (shown in table 2), the differences are not significant.

We then use the five automatic scores based on prosodic features to estimate MPS. The correlation between the estimated scores and MPS is 0.6, which is again higher than any single measure.

Considering the fact that there are no advanced learners whose TOEIC scores are higher than 800 in the subjects, the correlations we obtain are rather high.

5.3. Comparison of personal-best shadowing and final shadowing

The difference between personal-best shadowing and final shadowing is that final shadowing is done after checking the individual words in the presented native speech. We have expected that learners' pronunciation might improve significantly by checking the transcript. However, by closely

examining the MPS of both types of shadowing speech, we find that they are very similar.

Correlations between RWS and automatic scores calculated by using the data of personal-best shadowing and final shadowing are shown in table 6. Although correlations between GOP and RWS change significantly, in the case of prosodic measures, the correlations are almost the same. This indicates that knowing the contents of showing might not help learners with their prosodic fluency in shadowing.

6. Conclusions

In this paper, we analyze shadowing with automatic measures related to phoneme indelibility and prosodic fluency. We compare these automatic measures with several reference scores and propose several methods for shadowing evaluation. Experimental results show that the proposed automatic scoring methods are suitable for shadowing evaluation. Comparison of personal-best shadowing and final shadowing shows that knowing the contents before shadowing does not necessarily affect the prosodic aspects of learners' shadowed utterances. Future works include detailed comparison of shadowing with other conventional training methods, especially on prosodic aspects.

Table 6. *Correlations between automatic scores and RWS, comparing personal-best shadowing with final shadowing.*

Measures	Personal-best shadowing	Final shadowing
GOP	0.63	0.55
F0_DTW	-0.45	-0.43
F0_variance	0.55	0.56
Power_DTW	-0.3	-0.2
Pauses	-0.2	-0.25
ROS	0.58	0.56

7. 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] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, K. Hirose,"Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation," Proc. INTERSPEECH, pp.608-611 (2009-9)

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

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

[5] 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

[6] Motoyuki Suzuki, Tatsuki Konno, Akinori Ito, Shozo Makino,"Automatic Evaluation System of English Prosody Based on Word Importance Factor," Journal of Systemics, Cybernetics and Informatics, vol. 6, no.4, 2008

[7] Miwa, et al,"Analysis and comparison of the prosodic features for Japanese English and native English," IEICE technical report. Speech 101(744), 51-58, 2002-03-2