Automatic Pronunciation Evaluation of Language Learners' Utterances Generated through Shadowing

Dean Luo [1], *Naoya Shimomura* [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 foreign language learning, shadowing has been used as a method for improving speaking and listening ability. In this method, learners are required to repeat a presented native utterance as closely and quickly as possible. Since learners have to follow the speaking rate of the presented utterance, their pronunciation often becomes very inarticulate and unintelligible. These features of shadowing make it very difficult to build a reliable scoring system for shadowing productions. In this paper, two techniques are proposed and investigated for automatic scoring of shadowing productions. Experiments show that good correlations are found between automatic scores and TOEIC overall proficiency scores.

Index Terms: shadowing, automatic scoring, articulatory effort, goodness of pronunciation, bottom-up clustering

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 their own speech habits and language knowledge of their mother tongue unconsciously as well [4]. The adequate measurement of shadowing productions can be a good indicator of the speaker's overall language proficiency.

Most existing works on automatic pronunciation scoring have been done with HMM-based speech recognition technologies. Usually, the HMMs were trained with native and/or non-native "read" speech samples. However, in shadowing, since learners have to follow the speaking rate of the input native utterance, the speaking style of the learners is very different from "read" speech. Especially in the case of beginners, the content of the utterances generated through shadowing can be completely different from that of the presented ones. To the authors' knowledge, no automatic pronunciation scoring method has been proposed or investigated for shadowing.

In this study, we proposed two techniques for shadowing productions. One is using Goodness of Pronunciation (GOP) scores calculated through HMM-based forced alignment. In this method, for automatic scoring, the transcription of the presented utterance and the HMMs of the target language are required. The other is using a time-constrained bottom-up clustering technique. Here, only the presented utterance and the shadowed response are required. The transcription and the HMMs are not needed. Correlations between automatic scores and speakers' TOEIC overall proficiency scores have been investigated and the results are promising.

2. Alignment-based scoring technique

2.1. Goodness of Pronunciation (GOP)

Various techniques using HMM have been tried in many studies to evaluate pronunciation. The confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for assessing speakers' articulation and shows good results on read speech [5, 6]. In this study, we used HMM acoustic models trained on WSJ and TIMIT corpus to calculate GOP scores defined as follows. For each acoustic segment $O^{(p)}$ of phoneme p, *GOP(p)* is defined as posterior probability 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 an unconstrained phone loop grammar.

2.2. Estimation of the number of proficiently-pronounced words

Since the learners have to follow the native utterance, omissions and mispronunciations at word level are often found in their shadowing productions. By observing the shadowed speech database we collected, we found that two common errors occurred among learners. One is that the learners only repeat the words they understand and keep silent if they don't understand what they hear; the other is that the learners try to keep up with the presented utterance and utter whatever sounds they could make, but the shadowing

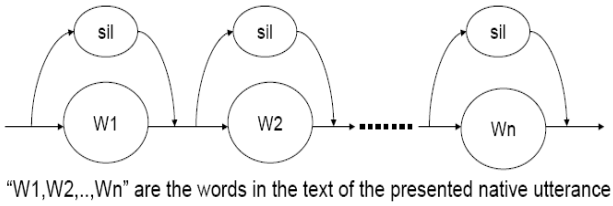

Figure 1: *Network grammar to detect omissions in shadowing*

productions are completely unintelligible. For the first typical error, we introduce a simple word-level network grammar depicted in Figure.1 to detect omissions in which words are replaced by silence. For the second typical error, we calculate average GOP score of each recognized word after introducing the network grammar, and use it as a confident score to judge whether the word is pronounced proficiently enough. Only when the average GOP score of a recognized word is above a threshold S, will it be accepted as "proficiently-pronounced word" (PPW). Thus the number of PPWs can be a good indicator of the learner's proficiency.

3. Clustering-based scoring technique

Since the learners need to immediately repeat what they hear, the speaking style in shadowing is very different from that of "read" speech. Especially in the case of beginners, their pronunciation often becomes corrupt and inarticulate. Using HMM-based alignment scoring techniques on shadowing productions might cause segmentation errors. Considering that it is desirable to build a scoring system that requires only an utterance pair: a native utterance presented to a learner and his/her utterance generated in response to the native utterance. Then, a new method is proposed here for automatic scoring of shadowing productions. The new method does not use any acoustic models such as HMMs at all, and just compares the two utterances through time-constrained bottom-up clustering.

3.1. Unsupervised phoneme segmentation based on time-constrained bottom-up clustering algorithm

We have proposed an unsupervised phoneme segmentation algorithm based on time-constrained bottom-up clustering. Here, each frame is treated as segment initially and then, acoustically similar and adjacent segments (frames) are merged into a larger segment in a greedy way. This clustering procedure stops by the condition explained below. A class of statistical measures have been used to decide which 2 segments (clusters) to be merged. Better results have been shown than other published methods [7]. In this study, we used a fast implementation of the proposed algorithm by using Ward's method.

Ward's method is hierarchical agglomerative clustering method, which searches the similarity matrix for the most similar pair of clusters and reduces the number of clusters by one through merging that pair of clusters until all clusters are merged into one [8]. The Ward objective is to find at each stage those two clusters whose merger gives the minimum increase to the total within-group error sum of squares. Suppose that adjacent speech segments p and p+1 are to be merged into a new cluster r (= $p \cup (p+1)$). If the frames are *m*-dimensional vectors $(x_1, x_2, ..., x_m)$, within-group error sum of squares $E(p)$ is defined as

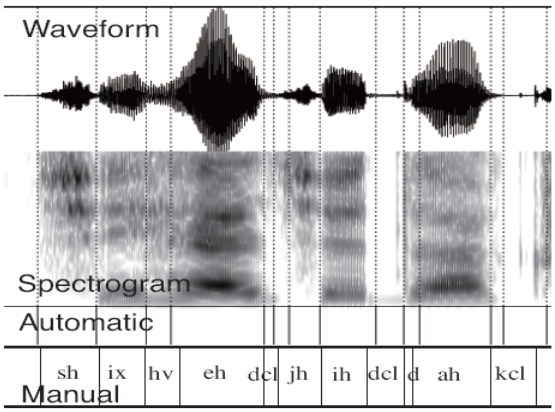

Figure 2: *An example of unsupervised phoneme segmentation*

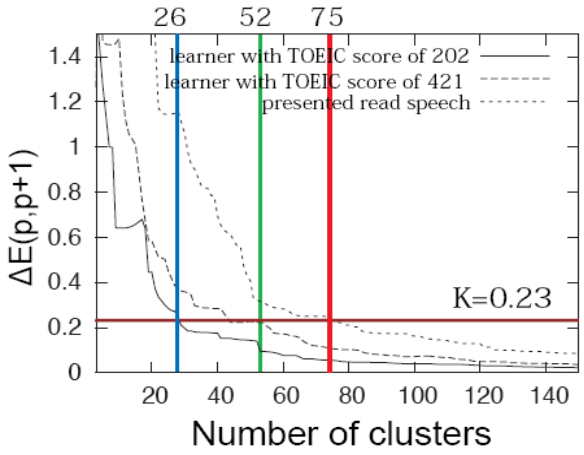

Figure 3: *Unsupervised phoneme segmentation on shadowing productions and presented native speech.*

$$E(p) = \sum_{i=1}^{n_p} \sum_{j=1}^{m} (x_{ij}^{p} - \bar{x}_{j}^{p})^2 \qquad (4)$$

where n_p is the number of samples, and $\bar{x}_j^p$ is the *j*-th element of the centroid of p. The increase of within-group error sum of square when segments p and p+1 are merged into r thus can be calculated as

$$\Delta E(p, p+1) = E(r) - \{E(p) + E(p+1)\} \qquad (5)$$

By merging adjacent segments p and p+1 with the minimum $\Delta E(p, p+1)$, we can realize bottom-up clustering of speech segments.

3.2. Stopping condition of clustering

Suppose the stage at which each segment approximately corresponds to each phoneme. Then, the next step to merge 2 segments would be merging 2 clusters that belong to different phonemes. In that case, the next merging step should yield a larger ΔE , i.e. $E(p \cup p+1) >> E(p) + E(p+1)$. Then we can set a predetermined threshold K for Δ E(p,p+1), which can be used as stopping condition of clustering.

Figure 2 shows an example of the proposed phoneme segmentation. The accuracy of automatic segmentation is fairly high compared with the manual labels. Figure 3 shows

Table1. *Subjects' TOEIC scores*

Proficiency	TOEIC scores	Average
Advanced	990, 990, 968, 955, 940, 895, 825	938
Intermediate	625, 601, 592, 581, 512, 436, 432, 427, 421	514
Beginners	395, 367, 308, 301, 289, 278, 275, 252, 202, 197, 158	275

Table2. *Acoustic conditions for analysis in clustering-based method*

sampling	16bit / 16kHz
window	Hamming / 16 ms length and 10 ms shift
parameters	MCEP (1～12)
threshold	K = 0.23

the segmentation results on a presented read speech sample and shadowing productions of 2 learners with TOEIC scores of 421 and 202 in response to that presented sample. Vertical axis is $\Delta E(p, p+1)$, and horizontal axis is the number of clusters. The threshold K was set to be 0.23. By examining the results of segmentation on these utterances, it is clear that even with the same linguistic content, the more distinctly an utterance is spoken, the more segments can be found when the clustering stops.

[9] shows that HMMs trained with "read" speech have larger distances between themselves compared to those trained with "spontaneous" speech. This is because, in read speech, each sound is generated with better articulation and distinction. [10] also shows that differences between HMMs can reflect the degree of articuratory efforts made in preparing the training data. Based on these considerations, in the Figure 3, we can say that the larger the number of segments is, the larger the articulatory efforts are made in shadowing.

4. Experiments

4.1. Shadowing database collection

In order to evaluate the proposed techniques, we collected a database of shadowing productions from 27 speakers, in which there are 7 language teachers, 9 intermediate learners and 11 beginners. The subjects' overall proficiency scores measured by TOEIC (Test of English as International Communication) are shown in Table 1. The presented utterance recorded by a native speaker of English contains 21 sentences and its topic was carefully chosen to be familiar to Japanese learners but the utterances themselves had never been presented to any of the subjects before. All the sentences were presented to the subjects sequentially at the rate of 140 wpm (words per minute), and the subjects were instructed to repeat as closely and as quickly as possible. The subjects' shadowing productions in response to the presented utterances were recorded in the environment of classroom.

4.2. Acoustic conditions for analysis

For alignment-based analysis, 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.

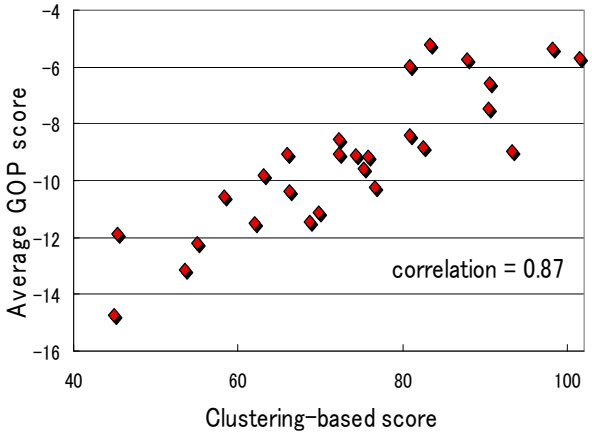

Figure 4: *Correlation between GOP scores and scores derived form clustering-based segmentation*

The acoustic conditions for analysis for clustering-based automatic segmentation are shown in Table 2.

4.3. Automatic scoring

For alignment-based automatic scoring, we first calculated GOP score of each phoneme in every shadowing production and normalized it by the number of phonemes that occur. Then we used the scheme described in section 2.2 to calculate the number of proficiently-pronounced words (NPPW). The average word-level GOP score of the shadowing production by the Japanese language teacher with highest TOEIC score has been used as the word-level confidence threshold S to judge if each word was pronounced proficiently. We use average GOP score of each speaker and NPPW for alignment-based automatic scoring.

For clustering-based automatic scoring, we used the threshold that has yielded the best phoneme segmentation result on TIMIT corpus, which has been proved to be valid even on Japanese database in our previous work [11] as the stopping condition of clustering. We then calculate the number of segments of the presented native utterance (N_p), and the number of segments of shadowing production (N_s) in response to the presented speech. The automatic score is defined as

$$S_u = \frac{N_s}{N_p} \times 100$$

4.4. Comparison of alignment-based and clustering-based scoring methods

As we mentioned in previous sections, alignment-based GOP scores are widely used to evaluate speakers' accuracy, and our proposed clustering-based automatic scores can be a good indicator of speakers' articulatory efforts. We compare both GOP scores and clustering-based automatic scores of utterances generated through shadowing and found a very high correlation of 0.87 between both scores. The result is shown in Figure 4. This further shows the validity of our unsupervised scoring technique, which is language-independent and does not need any acoustic models or transcriptions.

Table3. *Correlations between automatic scores and TOEIC scores*

method	GOP	NPPW	Clustering-based
correlation	0.82	0.84	0.72

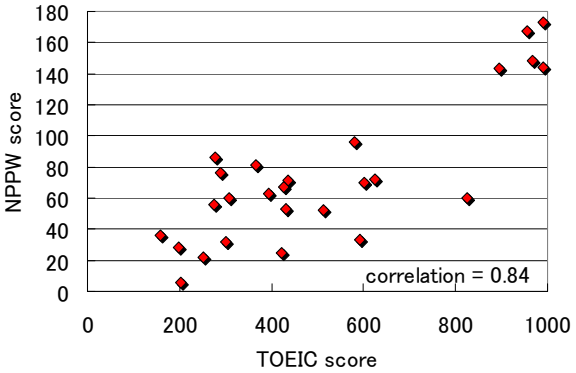

Figure 5: *Correlation between NPPW scores and TOEIC scores.*

4.5. Correlations between automatic scores and TOEIC scores

We investigated alignment-based automatic scores (GOP scores and NPPW scores), and scores derived from unsupervised automatic segmentation, and their relationship with TOEIC overall proficiency. The results are shown in Table 3. Figure 5 shows the correlation between NPPW scores and TOEIC scores, and Figure 6 shows the correlation between clustering-based scores and TOEIC scores.

Strong correlations between automatic scores and overall proficiency scores have been found. The best result is by using NPPW with a correlation of 0.84. The proposed clustering-based scoring technique also results with a rather high correlation of 0.72.

This result shows that the alignment-based scoring outperformed the clustering-based scoring technique. However, the clustering-based scoring technique does not require any acoustic models or linguistic contents of the utterances, thus with higher availability.

5. Discussion

High correlations have been found between automatic scores of the subjects' shadowing productions and their TOEIC scores. In read speech evaluation, even by using similar alignment-based GOP techniques, much lower correlations between machine and human scores are reported in recently published studies [12, 13]. This might be because shadowing poses a cognitive load on learners adequately and, therefore, the shadowing productions may reflect the learners' "true" proficiency level rather precisely. We are planning to collect read speech database of the same speakers and compare their read speech and shadowed speech.

6. Conclusions

In this paper, for automatic scoring, we have proposed the alignment-based and clustering-based scoring techniques for utterances generated through shadowing. We described how to implement these techniques and compared them with each other. Evaluation experiment results show that automatic scores have strong correlation with learners' overall language proficiency. Comparison of scores derived from both techniques shows that the proposed language-independent, easy-implementable clustering-based scoring technique is still available for evaluation of utterances generated through shadowing. We are planning to integrate both techniques and develop a hybrid evaluation system with more validity, reliability and practicality in the near future.

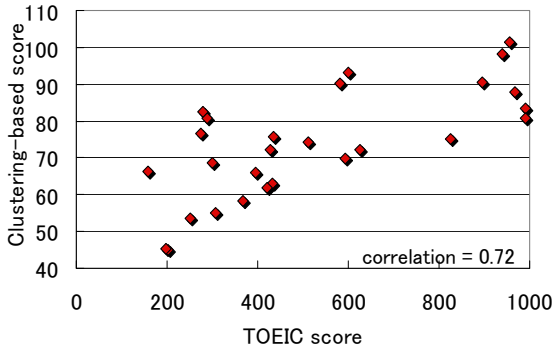

Figure 6: *Correlation between scores derived from clustering-based technique and TOEIC scores.*

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

[7] Y. Qiao, N. Shimomura, N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," Proc. ICASSP, pp.3989-3992, 2008

[8] C.Hervada-Sala et al., "A program to perform Ward's clustering method on several regionalized variables," Computers & Geosciences 30(2004), pp.881-886, 2004

[9] M. Nakamura et al., "Acoustic and linguistic characterization of spontaneous speech," Proc. Int. Workshop on Speech Recognition and Intrinsic Variations, pp.3–8, 2006

[10] N. Minematsu et al., "Para-linguistic characterization of spontaneous speech," Proc. ICASSP, vol.1, pp.261-264, 2006

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

[12] Abhishek Chandel et al., "Sensei: Spoken Language Assessment for CALL Center Agents," Proc. ASRU, pp.711-716, 2007

[13] J. Zheng et al., "Generalized Segment Posterior Probability for Automatic Mandarin Pronunciation Evaluation," Proc. ICASSP, vol.4, pp.201-204, 2007