AUTOMATIC ASSESSMENT OF LANGUAGE PROFICIENCY THROUGH SHADOWING

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

[1] The University of Tokyo, [2] Tokyo International University

ABSTRACT

Shadowing is a practice that requires learners to shadow a presented native utterance as closely and quickly as possible. Learners' pronunciation in shadowing, especially in the case of beginners, often becomes inarticulate and corrupt. These features of shadowing make it very difficult to assess shadowing productions. In this paper, we investigate the automatic pronunciation scoring methods for shadowing. Three automatic scores have be proposed and compared with each other. Experiments show that good correlations are found between the automatic scores and human ratings or TOEIC overall proficiency scores.

Index Terms— shadowing, Goodness of Pronunciation, automatic scoring, unsupervised bottom-up segmentation, articulatory effort, CALL

1. INTRODUCTION

Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages [1]. 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. 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 study, we proposed three techniques for evaluating 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 acoustic models of the target language are required. Another one is based on continuous phoneme recognition, in which the acoustic models are also needed, but no transcription is required. The third method is using a time-constrained bottom-up clustering technique. Here, only the presented utterance and the shadowed response are required. The transcription and the acoustic models are not needed. Correlations between automatic scores and manually-rated scores, and correlations between automatic scores and learners' TOEIC scores have been investigated and the results are very promising.

2. EVALUATION BASED ON HMM

2.1. Goodness of Pronunciation

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 [2]. In this study, we use 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, i.e. 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 phoneme loop grammar.

2.2. Continuous phoneme recognition (CPR) score

In case of transcription not being available, we can use HMM acoustic models to conduct continuous phoneme recognition. We consider for each utterance, the less intelligible the pronunciation is, the less distinct the individual segments are in the utterance. The number of recognized phonemes per utterance can be used as an index to measure the intelligibility. Here the number of phonemes normalized by the number in the presented utterance thus can be defined as continuous phoneme recognition (CPR) score.

3. CLUSTERING-BASED SCORING TECHNIQUE

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

In previous study, we have proposed an unsupervised phoneme segmentation algorithm based on a 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 [3]. In this study, we used a fast implementation of the proposed algorithm by using Ward's method.

Ward's method is a 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 [4]. 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

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

where n_p is the number of samples of p, 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

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

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 $\Delta E(p, p+1)$, which can be used as stopping condition of clustering.

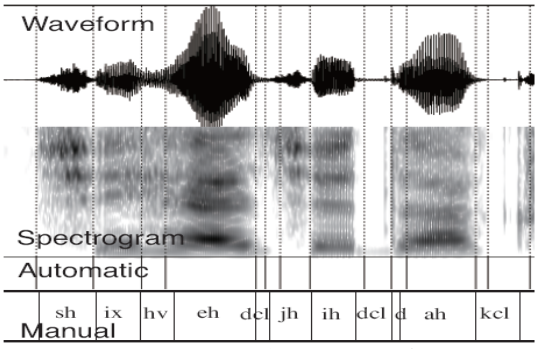

Figure 2: *An example of unsupervised phoneme segmentation*

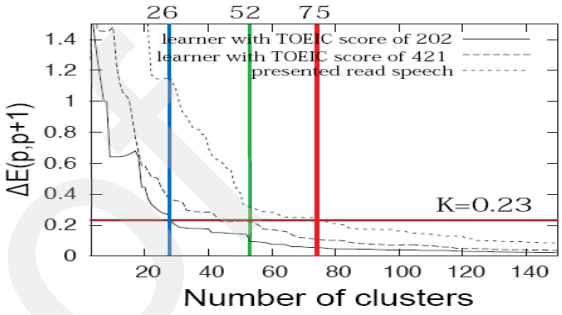

Figure 3: *Unsupervised phoneme segmentation on shadowing productions and presented native speech.*

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 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. The vertical axis is $\Delta E(p, p+1)$, and the 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 intelligibly an utterance is spoken, the more segments can be found when the clustering stops.

[5] 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. Figure 3 also shows that the larger the number of segments is, the larger the articulatory efforts are made in shadowing.

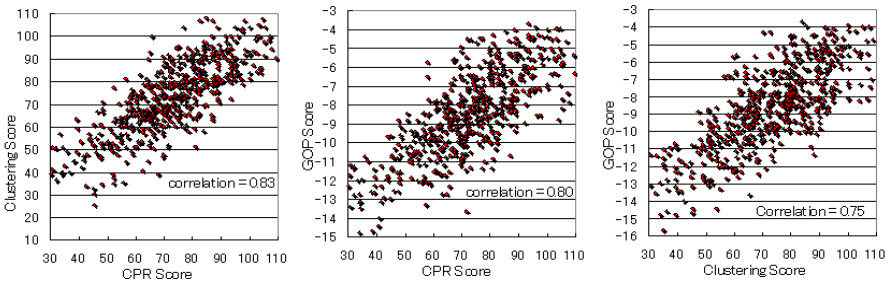

Figure 4: Utterance-level correlations between automatic scores

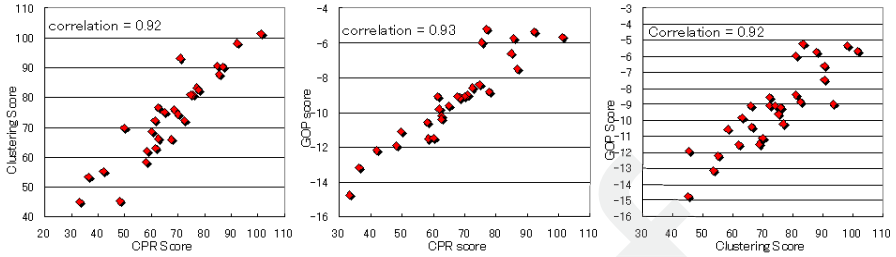

Figure 5: Speaker-level correlations between automatic scores

4. EXPERIMENTS

4.1 Shadowing database and manual assessment

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 utterances recorded by a native speaker of English contain 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.

Manual assessment was conducted by an expert in language education. Utterances of 10 sentences shadowed by 11 learners were chosen. The rater examined each utterance word by word. For each correctly pronounced word, the score would be 1. For any inserted word, the score of the word would be -1. For each partially correct word, the score would be 0.5. Thus by summing up the score of every word and normalized by the number of the words in the presented utterance, the result can be used as manual score for each shadowed utterance.

4.2. Acoustic conditions for analysis

The acoustic conditions for analysis for HMM-based evaluation are shown in Table 2. The acoustic conditions for analysis in clustering-based automatic segmentation are shown in Table 3.

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 in clustering-based method*

sampling	16bit / 16kHz
window	Hamming / 25 ms length/10 ms shift
parameters	MFCC, log-energy, and their Δ, ΔΔ

Table3. *Acoustic conditions in clustering-based method*

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

4.3. Comparison of automatic assessments

GOP score, CPR score and clustering score are supposed to play an equal role in pronunciation evaluation. To demonstrate this, we compared these 3 methods quantitatively. The correlations at utterance level and speaker level are shown in figure 4 and 5 respectively. Very high correlations have been found between any two of the three scores.

4.4. Correlations between automatic scores and manually-rated scores

The correlations at utterance-level and speaker-level are

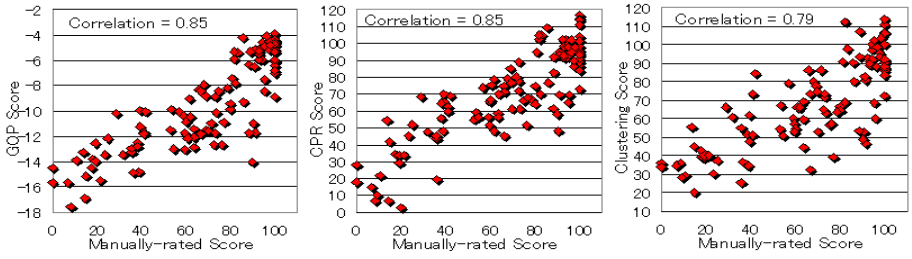

Figure 6: Utterance-level correlations between automatic scores and manually-rated scores

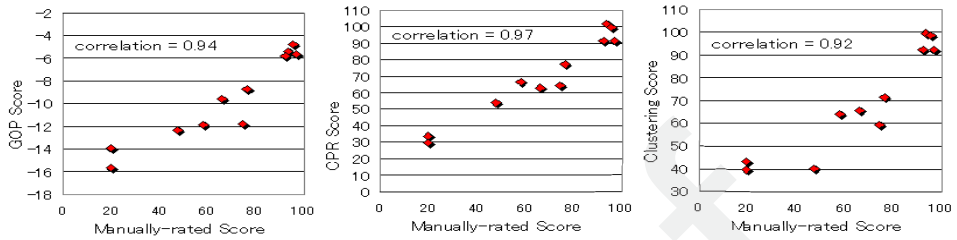

Figure 7: Speaker-level correlations between automatic scores and manually-rated scores

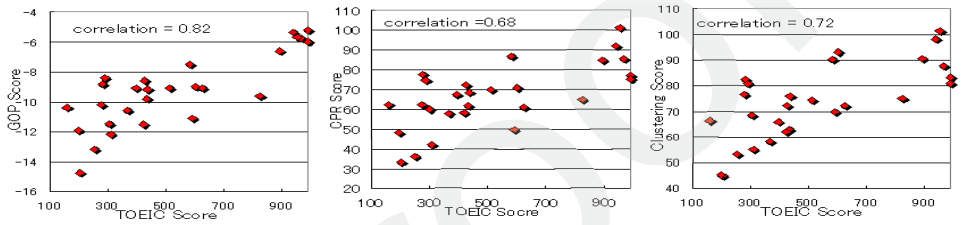

Figure 8: Correlations between automatic scores and TOEIC scores

shown in figure 6 and 7 respectively. Especially at speaker-level, very high correlations have been found.

4.5. Correlations between automatic scores and TOEIC scores

The correlations between automatic scores and TOEIC scores are shown in figure 8. GOP score shows the best correlation of 0.82 and language-independent clustering score also shows a good result of 0.72.

5. DISCUSSION

In read speech evaluation, even by using similar HMM-based GOP techniques, much lower correlations between machine and human scores were reported in recently published studies [6]. 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.

6. CONCLUSIONS

In this paper, we have proposed 3 scoring methods 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 manual scores or learners' overall language proficiency. Comparison of scores derived from different techniques shows that the proposed language-independent clustering-based scoring technique is still available for evaluation of shadowing productions. We are planning to compare and assess the shadowing productions and read speech of the same learners and build a language proficiency assessment system with more validity and reliability.

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.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

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

[4] 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

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

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