DNN-based GOP and Its Application to Automatic Assessment of Shadowing Speeches

Junwei Yue[1] Fumiya Shiozawa[1] Shohei Toyama[1] Yutaka Yamauchi[2] Kayoko Ito[3]
Daisuke Saito[1] Nobuaki Minematsu[1]

Abstract: Shadowing is currently one of the most popular research topics in CALL (Computer Assisted Language Learning). Our previous studies realized automatic assessment using the GOP (Goodness of Pronunciation) scores, and made a step toward automatically generating corrective feedbacks for shadowing speeches. In this study, we collected English shadowing speeches from Japanese university students. Manual scores of these speeches are given by a bilingual English teacher. Using this labeled corpus, we investigated automatic proficiency assessment using DNN (Deep Neural Network) based acoustic models. Here GOP (Goodness of Pronunciation) scores were estimated using DNN and they were compared to GMM-based GOP scores in terms of assessment performance. Further, DTW (Dynamic Time Wrapping) distances between learners' shadowed utterances and model utterances were calculated using posterior vectors. This DTW-based score was also compared to GOP-based scores. The result suggests that DNN based approach shows better performance than traditional GMM based ones. In the DTW-based comparison, language independency was also discussed.

Keywords: CALL, Shadowing, Corpus, Assessment, GOP, DNN, DTW, Language Independency

1. Introduction

Shadowing is a task which requires the speaker to repeat the played audio immediately while listening to it. It has been adopted as a practicing strategy for simultaneous interpreters since it includes not only speaking and listening, but also comprehending speech. Recently, many researches have shown that shadowing is also effective for language learning, especially for second language learning [1], [2], [3]. All of these studies suggested that shadowing could be more or at least no less effective in terms of improving speakers' language skills than traditional practicing strategies such as extensive reading, reading aloud and listening. However, learners need corrective feedbacks on their shadowing speeches. This work is usually done by language teachers so far, which requires a large amount of human resources. One of the solutions is to estimate the proficiency scores and generate corresponding feedbacks automatically. To train and evaluate estimation models, a corpus of shadowing speeches with manual scores labeled is also required.

In our previous studies, we adopted GMM-based GOP (Goodness of Pronunciation) scores as automatically estimated shadowers proficiency [4]. We also made a step toward automatic corrective feedback generation, where shadowing errors in a subset of the corpus were transcribed [5]. Here, GOP was adopted as one feature to predict proficiency scores using regression models. Previous results suggested that GMM-based GOP scores have good correlation with TOEIC scores when language proficiencies of learners are well distributed and the recorded speeches are clean. In the case that speeches are recorded with background noise and many speakers have similar language proficiencies, the correlation drops down dramatically. This could be alleviated by introducing some other features and performing regression analysis [5].

However, it has been long doubted that whether it is reasonable to adopt TOEIC scores as performance metric of language proficiency of shadowing since TOEIC tests do not contain any speaking tests until a few years ago. In addition, the size of the corpus used in our previous study [4] is not sufficient since only about 40 speakers participated in those experiments. Thus, in this study, we collected English shadowing speeches from 125 university students for a wider examination. A bilingual English teacher manually scored these speeches by paying attention to the fact that these utterances were obtained from shadowing practices. By using these scores as the ground truth of learners real shadowing performance, DNN (Deep Neural Network) based and GMM-based GOP scores are computed. On the other hand, DTW (Dynamic Time Wrapping) distances between shadowed and model speeches are computed using DNN-based posteriors, and the results are compared with DNN-based GOP scores. Here, language independency was also discussed.

2. Corpus collection

As previously mentioned, we collected English shadowing speeches from university student learners in Japan. An online shadowing recording site was developed for this data collection. It can be used in both shadowing practice and recording. 125

[1] The University of Tokyo
[2] Tokyo International University
[3] Kyoto University

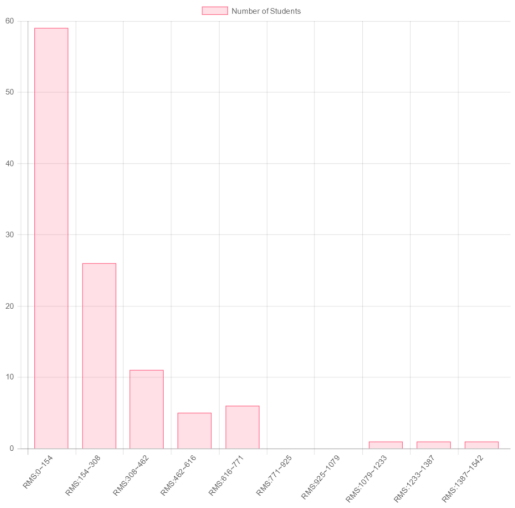


Fig. 1: An example of background noise level histogram for students from university K. X axis is the noise level and Y axis is the number of students in each noise level.

students in total participated in this recording, and they are from 3 universities, which are called K, T and A. These students are asked to shadow 50 read sentences from 4 passages without viewing the texts. Each sentence is shadowed 4 times. Students from university T and A (45 in total) recorded their speeches in the CALL (Computer Assisted Language Learning) classroom, where the background environment is quite noisy. On the other hand, students from university K (80 students) were asked to record in a quiet environment, such as private rooms at home. An instruction, including the setup of recording devices, the correct mouth position to the microphone and how to use the online recording tools, is prepared for reference.

We intentionally inserted a 1 second silence at the beginning of each model utterance, so we can assume only background noises are included in the first 1 second of each shadowed speech. Based on the sum of power in the first 1 second, noise level histograms were generated and fed back to home-recording students (from university K) for self-checking. Fig. 1 shows an example of the noise level histogram on a web page. Students can check their own noise levels on the web page, so ones with high noise levels will examine their recording environments and try to improve them the next time. This feedback increased to some degree subjects awareness of preparing a good recording condition by themselves.

3. Manual Scoring

To lay the groundwork for automatic shadowing speech estimation, we also manually scored the corpus collected in Section 2. It would be too much work to manually check all these speeches, so only 10 out of 50 sentences are picked up to be scored for each speaker. Here, only the forth recordings were adopted for manual scoring.

A bilingual English teacher assessed all these utterances. Considering the case that some shadowers are only able to shadow the beginning part of a model utterance, each utterance was divided into 2 or 3 phrases (fixed before recording) according to the length of text. Scoring was done for each phrase in these sentences. In total, the American teacher manually rated 3,375 shadowed phrases. Using these phrase-based scores, it is possible to derive sentence-level and speaker-level scores. Sentence-level manual scores are obtained by averaging phrase-level ones, and speaker-level manual scores are obtained by averaging sentence-level ones.

Table 1: Mean and standard deviation(SD) of manual scores (a) in phrase level, (b) for the three aspects.

(a) Phrase level

	Phrase A	Phrase B	Phrase C
Mean	10.5	9.8	10.2
SD	1.3	1.7	1.9

(b) Three aspects

	Segments	Prosody	Correctness
Mean	1.9	4.2	4.1
SD	0.62	0.57	0.54

Her assessment was done for the following three aspects:

- Segments (S): Goodness of producing phonemes or segments phonetically.
- Prosody (P): Goodness of realizing stress, lexical accent and phrase intonation.
- Correctness (C): How well the speaker followed the model utterance. It was examined whether the learner reproduced each given word intentionally after comprehension.

The score of each aspect ranges from 1 (worst) to 5 (best), so the full score is 15 and the worst score is 3.

Table 1(a) showed statistics about manual scores in phrase level. Phrases A, B and C indicate the beginning, middle and ending part of each sentence respectively. (3 sentences have only 2 phrases so they do not have phrase C and phrase B becomes their ending part.) Although phrase A has the highest mean score, no significant difference was found among phrase A, B and C. This is consistent with the fact that only the forth recordings were manually assessed, which means speakers have enough time to practice. Table 1(b) gives the information about detailed manual scores in terms of the three aspects. Segments mean score is only 1.9 out of 5, which is reasonable since many speakers shadowed with strong Japanese accents. On the other hand, prosody and correctness scores are relatively high, which means after practicing 3 times, speakers almost understood the meaning of sentences and could imitate them well.

To investigate the relationship between manual scores and TOEIC scores, all the participants have taken a mini TOEIC pretest. Their scores are rescaled from 0 to 100. The speaker-level manual score of a learner is compared to his/her TOEIC score. The result is plotted in Fig. 2. The correlation coefficient is only 0.44, which confirmed our doubts about TOEIC test, i.e. TOEIC test does not always represent the true shadowing performance well. Thus, later experiments are all based on manual scores, not TOEIC scores.

4. GOP

4.1 GMM-based GOP

GOP (Goodness of Pronunciation) score is an index representing the degree of clarity of speeches. It is widely used in general speech assessment tasks. Technically speaking, GOP score is just

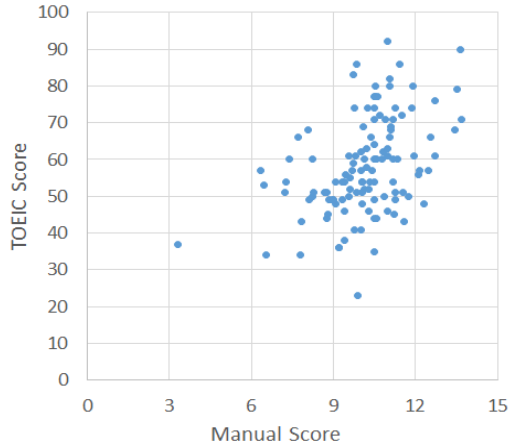


Fig. 2: Relationship between manual score and TOEIC score. Correlation coefficient = 0.44.

the posterior probability of phonemes given the utterance. It is usually defined as [6]:

$$\begin{aligned} GOP(p) &= \frac{1}{D_p}\log(P(p|O^{(p)})) \\ &= \frac{1}{D_p}\log\left(\frac{P(O^{(p)}|p)P(p)}{\sum_{q\in Q}P(O^{(p)}|q)P(q)}\right) \\ &\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)}|p)}{\max_{q\in Q}P(O^{(p)}|q)}\right), \end{aligned} \quad (1)$$

where $P(p|O^{(p)})$ is the posterior probability of phoneme p given utterance $O^{(p)}$, Q is the set of all phonemes and D_p is the duration of utterance $O^{(p)}$.

In previous studies, GMM-based acoustic models were adopted to compute GOP scores. Since it is difficult to calculate posterior probabilities directly using GMM-based models, GOP is often obtained approximately by the ratio of alignment likelihood and speech recognition likelihood. For this approximation, some accuracy lost is considered to be inevitable.

4.2 DNN-based GOP

Recent years, many studies showed that DNN based acoustic model has better recognition accuracy in many scenarios, as long as a large amount of data is provided [7]. DNN models are also considered more robust in a noisy environment, which is preferred in this research since some speeches in the corpus are recorded in a noisy CALL classroom as previously mentioned. Furthermore, DNN-based acoustic model could directly estimate the posterior probability of each frame in utterances without doing approximation. Thus, it's very natural to adopt DNN-based acoustic model to compute GOP scores. With the DNN-based model, the formula of GOP can be simplified as:

$$GOP(p) = \frac{1}{D_p}\log(P(p|O^{(p)})). \quad (2)$$

So the GOP score of an utterance/phrase could be calculated by the following steps:

(1) Align utterance/phrase with text using GMM-based model and obtain intended phoneme for each frame.
(2) Calculate the posterior probability distribution over all 3,386 kinds of context-dependent phonemes for each frame using the DNN-based model.
(3) Sum up posterior probabilities of the corresponding intended phonemes for all frames and normalize them by the duration of this utterance/phrase.

Speaker-level GOP scores can be derived by averaging utterance/phrase-level GOP scores.

4.3 DTW

The DTW is a technique that allows a non-linear mapping of one signal to another by minimizing the accumulated distance between the two [8]. The smaller the distance is, the more similar the two signals are. DTW could also be applied in measuring the distance between two sequences of posterior vectors as long as the distance between vectors are defined. Since posterior vector has the property that the sum of all elements is 1, it's possible to adopt distance metrics defined for probability distribution [9], [10]. Commonly used distance metrics are:

$$EUC(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (3)$$

$$BD(x,y) = -\log\left(\sum_{i=1}^{N}\sqrt{x_i y_i}\right) \quad (4)$$

$$KL(x,y) = \sum_{i=1}^{N} x_i * (\log x_i - \log y_i), \quad (5)$$

where x and y are two vectors which satisfy $\sum_i x_i = 1$, $\sum_i y_i = 1$, and N is their dimension.

Equation (3) is the Euclid Distance [8] between two vectors, which do not require vectors to be probability distribution. Equation (4) and Equation (5) are Bhattacharyya Distance(BD) [11] and Kullback-Leibler(KL) divergence [10] of two probability distributions, respectively. They are both commonly used indices to measure the similarity of distributions. Note that although KL-divergence is not symmetrical for x and y, it would change much if we swap x and y. In this study, all three metrics are adopted to compute the similarity between posterior sequences of model utterances and shadowed utterances.

The reason we are working on DTW is that DTW distance calculation does not require any linguistic information of utterances, such as transcripts of the utterances or their language identity. On the other hand in the GOP scoring, acoustic features of the model utterances are not needed instead, which is kind of waste. In this study, we computed the DTW distance using not only English, but also Japanese acoustic models to investigate the language independency between the language spoken and the language of the DNN model.

English and Japanese are usually considered to be very different languages, at least in terms of the pronunciations. Technically speaking, the value of each dimension of DNN-based posteriors is just the probability of a context-dependent phoneme in the acoustic model. Although English and Japanese acoustic models have nearly completely different sets of phonemes and the number of phonemes of Japanese is much smaller than that of English, as long as a large number of context-dependent phonemes is pro-

vided, it may be possible to say that most phonemes in English could be covered using Japanese acoustic models. In other words, similarity calculation using Equations (3), (4), or (5) will be similar between English models and Japanese models. This is extremely meaningful for automatic shadowing assessment of minor languages since some languages yet have not enough speech corpus to train a good acoustic model.

5. Experiment Design and Result

5.1 Acoustic Model

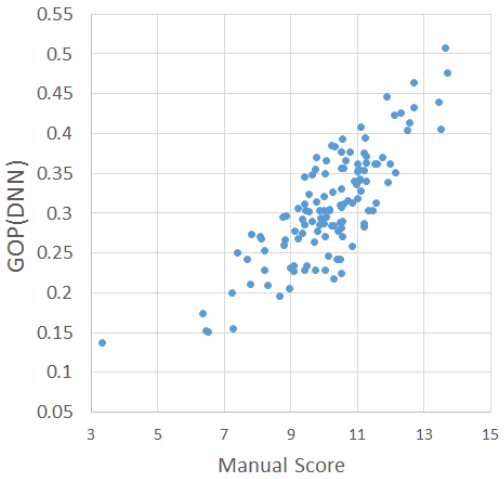
We trained three acoustic models in this experiment: Model HTK, Model KALDI_WSJ and Model KALDI_CSJ. Model HTK is a pre-trained English acoustic model using WSJ (Wall Street Journal) recipe of HTK [12]. This includes only GMM-based model and is used in previous studies. Model KALDI_WSJ is an English acoustic model trained using WSJ recipe of Toolkit KALDI [13], including both GMM-based and DNN-based models. Model KALDI_CSJ is an Japanese acoustic model trained using CSJ (Corpus of Spontaneous Japanese) recipe of KALDI, and also includes both GMM-based and DNN-based models. In addition, LDA (Linear Discriminative Analysis) dimension reduction and FMLLR (Feature space Maximum Likelihood Linear Regression) are also applied to the two KALDI models. All other settings remained unchanged, i.e. default values of the two recipes. Since CSJ has larger corpus size, the default number of tied-states of context-dependent phonemes for Model KALDI_CSJ (about 9,000 states) is about 3 times larger than Model KALDI_WSJ (about 3,000 states).

Model HTK is used as GMM model, and Model KALDI_WSJ is used as DNN model in the GOP experiments. Model KALDI_WSJ is used as English acoustic model, and model KALDI_CSJ is used as Japanese acoustic model in the DTW experiments.

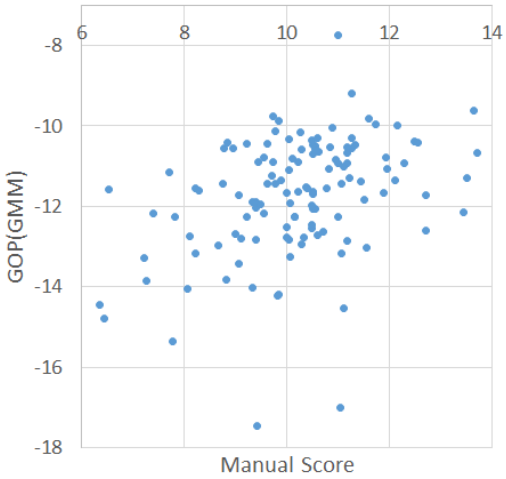
5.2 GOP

GOP scores computed using both GMM-based and DNN-based models are shown along with manual scores in Fig. 3. Speaker-level DNN-based GOP scores are obtained by averaging sentence-level posteriors, and speaker-level GMM-based ones are obtained by averaging sentence-level posteriors in log space. DNN-based GOP scores show a high correlation with manual score (with coefficient=0.82), which is consistent with the fact that DNN has better performance in speech recognition. On the other hand, GMM-based (HTK) GOP score is not as good as those in previous studies, with relatively low correlation coefficient 0.49 (Previous studies are all above 0.60). Several factors could be taken into account: Many utterances in corpus are noisy but no pre-processing like FMLLR for HTK model; Previous studies adopted TOEIC scores as target scores, which has been changed to manual scores.

Comparing to GMM-based model, adopting DNN-based model gains about a 67% relative accuracy improvement. One possible reason of higher performance of DNN-GOP is the assessment strategy of the teacher we adopted for manual rating. Her strategy might be coincident with how DNN rated shadowing speeches. We may have to examine manual scores given by other teachers.



(a) Relationship of manual scores and DNN-based GOP. Correlation Coefficient = 0.82.



(b) Relationship of manual scores and GMM-based GOP. Correlation Coefficient = 0.49.

Fig. 3: Relationship between manual scores and DNN/GMM-based GOP.

Another interesting result is that even though the correlation coefficient of GMM-based model is only 0.49, it's still higher than the one between TOEIC scores and manual scores. This means it's more adequate to use these automatic scores than taking a TOEIC test if a learner wants to know his true proficiency of English shadowing.

5.3 DTW

DTW distance is calculated between each pair of shadowing utterances and model utterances. Equations (3), (4), (5) are used as distance metrics between posterior probability vectors. The distance between frame i of model utterance and frame j of shadowed utterance is annotated as $D(i, j)$. The local path constraint for DTW is the ordinary 3-path constraint: For point (i, j), only points $(i-1, j)$, $(i, j-1)$ and $(i-1, j-1)$ are legal transitions, and the transition costs are $D(i, j)$, $D(i, j)$ and $2D(i, j)$ respectively (Fig. 4). Sentence-level DTW distances are finally normalized by the duration of corresponding model utterance. Speaker-level DTW distances are obtained by averaging sentence-level ones. Since the GMM-based model is not capable of generating poste-

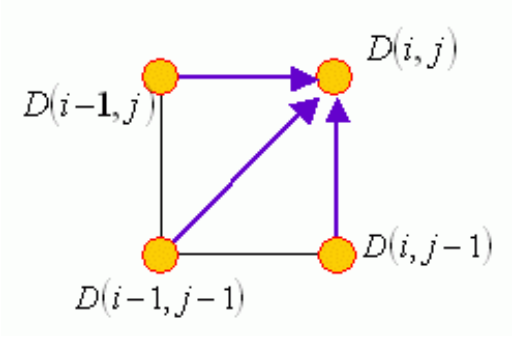


Fig. 4: The DTW local path constraints used in this study. For point (i, j), only points $(i-1, j)$, $(i, j-1)$ and $(i-1, j-1)$ are legal transitions.

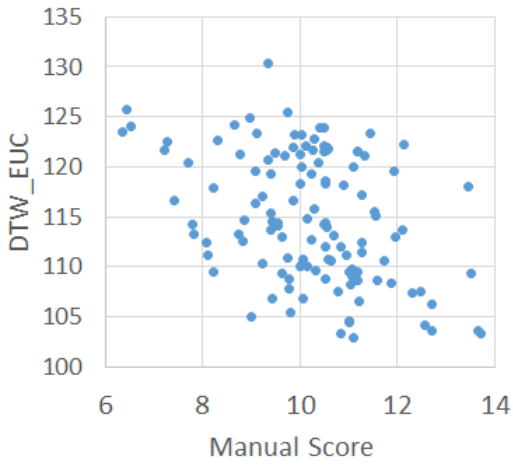
rior vectors directly, only the DNN-based model is used in this experiment.

Fig. 5 shows the relationship of manual scores and DTW distances using the English acoustic model. As previously mentioned, the dimension of posterior vectors remains default value (about 3,000). Note that the shorter DTW distance is, the more similar two signals are, so all the correlation coefficients are negative. Both BD and KL-div show promising results, with correlation coefficient -0.79 and -0.74 respectively, which are close to the GOP one, but without requiring the text. The correlation coefficients between DNN-based GOP scores and DTW distance in speaker-level are -0.64, -0.93, -0.92 using Euclid Distance, BD and KL-div as measurement respectively. Generally speaking, BD has better performance than the other two metrics.

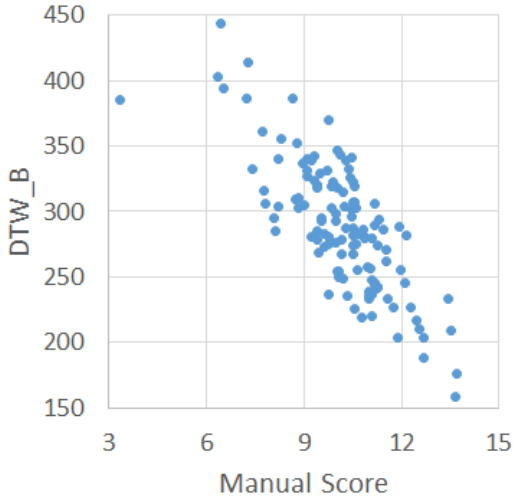
Based on the previous promising results, we made a further step to computing DTW distances using acoustic models of a different language. The Japanese acoustic model (trained using CSJ) has about 9,000 tied-states by default. However, to compare with the English acoustic model, the dimension of posterior vectors are considered to be important since it is related to the granularity of the DNN learning step. It is valuable if we could check how the posterior vector dimension is related to the correlation between DTW distances and manual scores. So in this experiment, two Japanese acoustic models were trained with the only difference that one has about 9,000 tied-states and the other has about 3,000 tied-states (which is comparable to the English model). Only BD is used as the distance metric of posterior vectors.

The relationship between Japanese-model-based DTW accumulated distances and manual scores is shown in Fig. 6, with 9,000 and 3,000 tied-states respectively. This speaker-level distance is obtained by averaging sentence-level distances of each speaker. Although shadowing speeches are recorded in English, the DTW distance generated by the 3,000 tied-states Japanese model still represents speakers' shadowing proficiencies well, with a rather high correlation coefficient -0.74, which is close to the English model based one (-0.79).

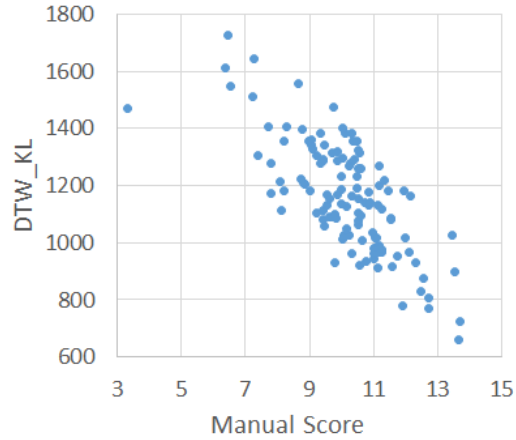
On the other hand, the 9,000 tied-states model doesn't work out very well, with only a -0.52 correlation coefficient. This confirmed our concern about a larger tied-state size doesn't mean a better performance. The reason is considered to be that different from native speakers, second language learners could hardly



(a) Relationship between manual scores and DTW distance using Euclid distance. Correlation Coefficient = -0.43.



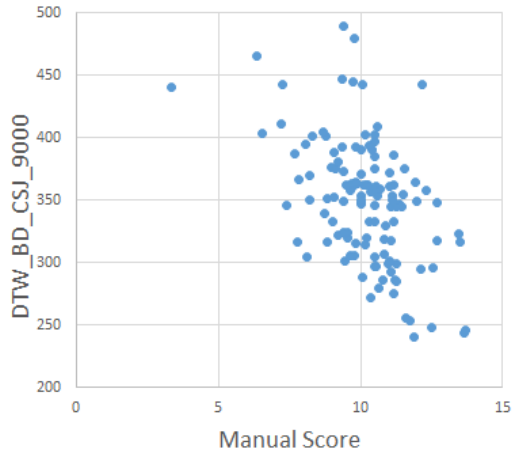
(b) Relationship between manual scores and DTW distance using BD. Correlation Coefficient = -0.79.



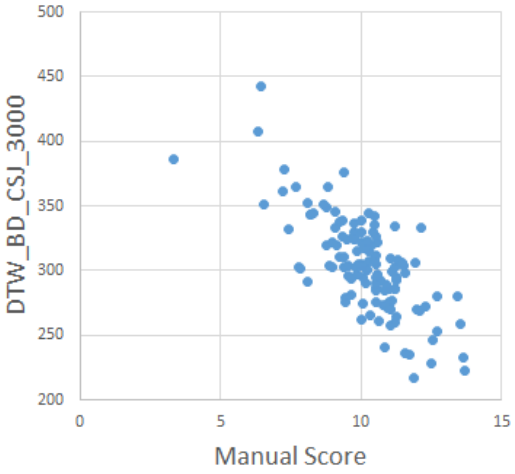
(c) Relationship between manual scores and DTW distance using KL-div. Correlation Coefficient = -0.74.

Fig. 5: Relationship between manual score and three kinds of DTW distances.

handle slight differences between two similar context-dependent phonemes. Thus many tied-states of native acoustic models become redundant, which brings bad effects on the posterior similarity estimation. Here, English speeches could be just considered as "non-native speeches" for a Japanese acoustic model, so it is



(a) DTW distance is computed using the 9,000 tied-states Japanese acoustic model. Correlation Coefficient = -0.52.



(b) DTW distance is computed using the 3,000 tied-states Japanese acoustic model. Correlation Coefficient = -0.74.

Fig. 6: Relationship between manual score and three kinds of DTW distances.

not surpursing that the 3,000 tied-state model overperformed the 9,000 one. The same phenomenon also happens when try to do speech recognition or speech assessments for non-native speakers using native mono-phone/tri-phone acoustic models. In many situations, mono-phone models have better performance than the tri-phone ones.

6. Conclusion and Future Work

In this study, we first collected English shadowing speeches from 125 university students in Japan. Then we manually scored these speeches. The result showed that the speaker-level manual scores and TOEIC scores have a rather low correlation, which confirmed our doubts about TOEIC scores. After manual scoring, both GMM-based and DNN-based GOP scores were computed. The results showed that DNN-based GOP score has higher correlation with manual scores than GMM-based one, and about 67% relative improvement was gained. In addition, the DTW distance between model utterances and shadowed utterances using the English acoustic model based on Euclid distance, BD and KL-div were also computed. A high correlation between DTW distance and manual scores was seen. Finally we changed English model to Japanese model, tried 2 different numbers of physical states and computed DTW distances again. The 3,000 tied-state Japanese model has very close performance comparing to the English one.

In the future, we are going to:

- Apply LDA dimension reduction and FMLLR to HTK GMM model to make the comparisons more consistent.
- Use regression models to improve assessment accuracy as we already did in our previous study [5]. For sentence-level or phrase-level assessment, regression models based on supervised learning will be needed.
- Try more different tied-state numbers. There should be an optimal number of physical states which maximizes the correlation coefficient between DTW distances and manual scores for both the English model and the Japanese model.
- Compute GOP scores and DTW distances for the other utterances in the collected corpus. By finding out the sentences with maximum variances, we may be able to choose the best 10 sentences for shadowing assessments automatically.

Acknowledgments This work was supported by JSPS KAKENHI Grant Numbers JP16H03084, JP16H03447, JP26240022. Thanks to all students and teachers who had participated in this experiment.

References

[1] Y. Hamada, The effectiveness of pre-and post-shadowing in improving listening comprehension skills. The Language Teacher, 38(1), 3-10, 2014.
[2] Y. Hamada, Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. Language Teaching Research, 2015.
[3] 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.
[4] 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).
[5] S. Shi, Y. Kashiwagi, S. Toyama, et al. Automatic Assessment and Error Detection of Shadowing Speech: Case of English Spoken by Japanese Learners. Interspeech 2016, 2016: 3142-3146.
[6] S. M. Witt, & S. J. Young, Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication, 30(2), 95-108, 2000.
[7] G. Hinton, L. Deng, D. Yu, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.
[8] E. Keogh, S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery, 2003, 7(4): 349-371.
[9] R. Rasipuram, M. Cernak, A. Nanchen, Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities. Proceedings of Interspeech. 2015 (EPFL-CONF-209089).
[10] R. Ullmann, R. Rasipuram, H. Bourlard, Objective intelligibility assessment of text-to-speech systems through utterance verification. Proceedings of Interspeech. 2015 (EPFL-CONF-209096).
[11] J. R. Hershey, P. A. Olsen, Variational bhattacharyya divergence for hidden markov models. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. IEEE, 2008: 4557-4560.
[12] https://www.keithv.com/software/htk/. (accessed 2017-1-23).
[13] D. Povey, A. Ghoshal, G. Boulianne, et al. The Kaldi speech recognition toolkit. IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011 (EPFL-CONF-192584).