INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Automatic Scoring of Shadowing Speech based on DNN Posteriors and their DTW

Junwei Yue[1], *Fumiya Shiozawa*[1], *Shohei Toyama*[1], *Yutaka Yamauchi*[2], *Kayoko Ito*[3], *Daisuke Saito*[1], *Nobuaki Minematsu*[1]
[1]The University of Tokyo,
[2]Tokyo International University,
[3]Kyoto University
{jwyue,shiozawa,toyama,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp, yyama@tiu.ac.jp, ito.kayoko.6m@kyoto-u.ac.jp

Abstract

Shadowing has become a well-known method to improve learners' overall proficiency. Our previous studies realized automatic scoring of shadowing speech using HMM phoneme posteriors, called GOP (Goodness of Pronunciation) and learners' TOEIC scores were predicted adequately. In this study, we enhance our studies from multiple angles: 1) a much larger amount of shadowing speech is collected, 2) manual scoring of these utterances is done by two native teachers, 3) DNN posteriors are introduced instead of HMM ones, 4) language-independent shadowing assessment based on posteriors-based DTW (Dynamic Time Warping) is examined. Experiments suggest that, compared to HMM, DNN can improve teacher-machine correlation largely by 0.37 and DTW based on DNN posteriors shows as high correlation as 0.74 even when posterior calculation is done using a different language from the target language of learning.
Index Terms: shadowing, scoring, GOP, DNN, DTW

1. Introduction

Shadowing was originally introduced as a practicing strategy for simultaneous interpreters since it includes not only speaking and listening but also comprehending input speech. Recently, research has shown that shadowing is also effective for second language learning [1, 2, 3]. However, manual scoring or error detection from shadowing utterances is very laborious and in our previous studies, we adopted HMM-based phoneme posteriors, called GOP (Goodness of Pronunciation), as automatically predicted shadowers' proficiency [4]. Further in [5], by introducing some other features extracted from shadowing utterances, we tested several regression models to improve the performance of predicting learners' TOEIC scores. It was also experimentally shown that the GOP scores of shadowing speech are better indices than those of read speech to predict the TOEIC scores [6]. In these works, however, a problem remained to be solved. Although the GOP scores were well-correlated when learners lie in a wide range of proficiency, the correlation can be lowered easily when they were in a limited range.

We can point out other problems. If learners want to know their shadowing performance, the target of prediction should not be their TOEIC scores but their shadowing scores, which are generally given manually by teachers. In addition, the size of the corpus used is not sufficient in [4, 5, 6], where only about 40 speakers participated in recording. Thus, in this study, we collect English shadowing speech from a much larger number of learners for a wider examination. After that, two native English teachers score those speech samples using three criteria, which will be used as the ground truth of learners' performance.

Technically speaking, HMM acoustic models have been replaced by DNN acoustic models in the field of speech recognition and a similar trend is seen also in CALL [7, 8]. In this work, we compare DNN posteriors with HMM posteriors in their performance of approximating teachers' scores. In many related works [5, 9, 10], GOP or posteriors are used as just one feature to predict the target score using regression models. Here, supervised training is done, but in this case, some environmental dependency will be inevitable, which will be troublesome if the trained model is applied to an unexpected environment. Then in this work, we put a technical focus on applicability or availability of DNN posteriors analytically.

Several approaches compared two utterances (native vs. non-native or natural vs. synthetic) through DTW after they were converted to sequences of posterior vectors [11, 12, 13, 14]. In shadowing scoring, a similar strategy can be used. Here, posteriors are basically phoneme posteriors but, strictly speaking, they are phoneme-state posteriors, where a few thousand sound classes, called senones, are used as labels. When the phonemes of a language are used for posterior vectors, the obtained representation will strongly depend on the language. When the senones of a language are used instead, however, the representation will have much less language dependency. In this work, DNN posteriors calculated using DNN models trained for a language, which is different from the target language of learning, are examined. If this strategy works well, a single DTW framework may be able to compare utterances in any language.

2. Corpus collection

We collected English shadowing utterances from 125 university students in Japan. An online shadowing recording site was developed for this data collection. These students are from three universities, called K, S and A in this paper. The numbers of learners are 80, 41, and 4 for K, S, and A, respectively. They were asked to shadow 55 model utterances without viewing their manuscripts. Each model utterance was shadowed 4 times. Prior to shadowing and recording, they were asked to check an instruction page to practice shadowing and recording.

3. Manual scoring

10 utterances out of the 55 model utterances were selected based on syntactic and semantic difficulty by the fourth author, who often uses shadowing practices in his English class. Learners' fourth shadowing productions for the 10 model utterances were used for manual scoring. Two American-Japanese bilingual female English teachers, who claim that their native language is American English, assessed all these utterances. Each utterance is composed of two or three phrases and scoring was done for each phrase utterance. In total, 3,375 shadowing phrases were rated. Using these phrase-based scores, it is possible to derive sentence-level and speaker-level scores. Sentence-level scores were obtained by averaging phrase-level ones, and speaker-level scores were obtained by averaging sentence-level ones.

http://dx.doi.org/10.21437/Interspeech.2017-728

Table 1: *Means and standard deviations of manual scores for each phrase position (a) and for each criterion (b).*

(a) Phrase positions

	1st	2nd	3rd
Mean	10.2	9.5	9.9
SD	1.4	1.7	1.9

(b) Three criteria

	Segments	Prosody	Correctness
Mean	2.4	3.5	3.9
SD	0.58	0.63	0.59

Their assessment was done for the following three criteria:

- Segments (S): how adequately individual segments are produced.
- Prosody (P): how adequately prosodic features, word stress and phrase intonation, are controlled.
- Correctness (C): How many words are produced.

The score for each criterion ranges from 1 (worst) to 5 (best), so the full score is 15 and the worst score is 3 in total. Scores used for later experiments are the average over the two teachers.

Table 1(a) showed statistics of the manual scores for each phrase position. Although the first phrases have the highest mean score, no significant difference was found among the phrase positions. The rated utterances are the fourth shadowing productions and it seems that the learners' shadowing performance is similar among the phrase positions. Table 1(b) gives similar statistics for each criterion. It is clearly indicated that the learners tend to shadow using Japanese phonemes. Shadowing is a task of very high cognitive load that requires online processing of listening, comprehension, and speaking. Therefore, it seems difficult for learners to produce each segment adequately based on careful control of articulators.

To investigate the relationship between the learners' manual scores and their TOEIC scores, all of them were asked to take a mini TOEIC test. The correlation between the speaker-level manual score of a learner and his/her TOEIC score was found to be as low as 0.46. This confirmed that GOP calculated from shadowing utterances can be used as one independent variable for prediction or regression of TOEIC scores but GOP is not necessarily correlated highly with TOEIC scores by itself.

4. Automatic scoring

In this section, three GOP-based methods of automatic scoring for shadowing speech are described. HMM-based GOP is the one used in our previous studies [4, 5, 6]. DNN-based GOP is also a GOP score, but computed using DNN-based acoustic models. DNN-based DTW is another method that can calculate the distance between shadowing and model utterances. As will be discussed later, the third method is expected to able to calculate the distance adequately even if the target language of learning and that used for posterior calculation are different.

4.1. HMM-based GOP [15]

The GOP score of phoneme /x/ is the posterior probability of that phoneme given acoustic observation, defined as

$$\begin{aligned} GOP(x) &= \frac{1}{D_x}\log(P(x|O^{(x)})) \\ &= \frac{1}{D_x}\log\left(\frac{P(O^{(x)}|x)P(x)}{\sum_{y\in Y}P(O^{(x)}|y)P(y)}\right) \\ &\approx \frac{1}{D_x}\log\left(\frac{P(O^{(x)}|x)}{\max_{y\in Y}P(O^{(x)}|y)}\right), \end{aligned} \quad (1)$$

where $P(x|O^{(x)})$ is the posterior probability of phoneme /x/ given observation $O^{(x)}$, Y is the full set of phonemes and D_x is the duration of $O^{(x)}$. Since it is difficult to calculate posterior probabilities directly using HMMs, GOP is often obtained approximately by the ratio of alignment likelihood and speech recognition likelihood. For this approximation, some accuracy loss is considered to be inevitable.

4.2. DNN-based GOP [7]

In recent years, many studies of speech recognition showed that DNN-based acoustic models have better accuracy in many scenarios, as long as a large amount of data is provided for training [16]. This is often explained to be because DNN-based acoustic models require neither Gaussian assumption for feature distribution nor indirect calculation of posteriors using generative models as adopted in Equation (1). Thus, it is very natural to adopt DNN models to compute GOP scores. With DNN models, the formula of GOP can be simplified as

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

By following [7], the GOP score of a shadowing phrase can be calculated through the following steps:

1. Align that phrase with its manuscript using HMMs and, for each frame, obtain the corresponding phoneme state.
2. For each frame, compute the posterior probabilities for all the phoneme states using DNN models.
3. Calculate the average posterior probability of the corresponding phoneme state over all frames of that phrase.

The speaker-level GOP can be derived by averaging phrase-level GOP scores.

4.3. DNN-based DTW

The DTW is a technique that allows a non-linear alignment of one sequence to another by minimizing the accumulated distance between the two. DTW can be applied for measuring the distance between two sequences of posterior vectors. Spectrum-based or cepstrum-based features can also be used for DTW. In this case, as is well known, the alignment path obtained between a male adult's speech and a girl's speech, for example, will surely be messy because non-linguistic differences between the two sequences will influence the alignment path drastically. This is why posterior-based DTW is examined in this paper. Since the sum of all the elements of a posterior vector is 1, we can adopt distance metrics defined for probability distribution. Commonly used distance metrics are:

$$\mathrm{EUC}(x,y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (3)$$

$$\mathrm{BD}(x,y) = -\log\left(\sum_i \sqrt{x_i y_i}\right) \quad (4)$$

$$\mathrm{KL}(x,y) = \sum_i x_i \log\left(\frac{x_i}{y_i}\right), \quad (5)$$

where x and y are two vectors satisfying $\sum_i x_i{=}1$ and $\sum_i y_i{=}1$. Equation (3) is the Euclidean Distance between two vectors, which does not require vectors to be probability distribution. Equations (4) and (5) are Bhattacharyya Distance (BD) and Kullback-Leibler (KL) divergence of two probability distributions, respectively. They are both commonly used indices

to quantify similarity of distributions. Note that although KL-divergence is not symmetric for x and y, it would not change much if we swap x and y. Here, all the three metrics are adopted to compute the similarity between the posterior vector sequence of a model utterance and that of its shadowed utterance.

Similarity of a shadowed utterance to its model one can be quantified using either of GOP or DTW, but what is a functional difference between them? In GOP, it can be said that a shadowed utterance is compared to the intended phoneme sequence underlying the model utterance through acoustic models of phonemes, which are technically implemented by HMM or DNN. Here, the model utterance is not explicitly used for comparison and only the intended phonemes are used. On the other hand in DTW, a shadowed utterance is compared to its model utterance, where the intended phonemes are not explicitly used for comparison although each frame of both utterances is converted into a phoneme-state posterior vector.

The number of phoneme states is generally as large as several thousands and such a large number of sound classes are prepared for posterior calculation. In English education in Japan, English vowels are often explained by referring to the five Japanese vowels. For example, English /ih/ is not found in the Japanese vowel system but it can be produced as intermediate vowel between Japanese /i/ and /e/. This means that /ih/ is likely to be found *acoustically* in a Japanese utterance of /ieie/ because of co-articulation. Generally speaking, the phoneme states of a language are derived so that it can cover sound variations of that language partly caused by co-articulation, and therefore the phoneme states of Japanese are expected to include /ih/-like sounds. This consideration logically leads to an idea that the language used for posterior calculation can be different from the target language of learning. A model utterance and its shadowed utterance in English could be compared adequately using a posterior calculation module trained for another language. If this idea is valid, a posterior calculation module for a language will be effectively used for utterance comparison in any language. This language independence is impossible for GOP, where the intended phonemes have to be given explicitly.

It can be said that, in the language-independent utterance comparison based on posteriors, a phoneme-state or senone is used as speaker-adapted *acoustic anchor*. Further, as class posteriors generally indicate how an input observation is similar to each class, the posterior vector of a speech frame is considered to indicate how that frame is similar to each speaker-adapted anchor. Putting it another way, that speech frame is represented relatively to the anchors, not by its spectral shape itself. This representation is very close to structural representation of speech [17, 18] and what differs between them is that posterior calculation requires a huge amount of data for training DNN models but structural representation does not, which aims at extracting speaker-invariant features from an input utterance. Detailed discussion on this is found in [18].

Posterior probabilities based on data-driven acoustic anchors can be calculated by using GMMs. In [13], GMM-based posteriors are used for utterance comparison. In this case, however, as discussed in Section 4.1, use of generative models like GMM is inevitably faced with some accuracy loss.

5. Experiments and results

5.1. Acoustic Models

We prepared four kinds of acoustic models in this experiment (Table 2). Model HMM is a set of English phoneme HMMs trained with the WSJ (Wall Street Journal) recipe of HTK [19], which were used in our previous studies. Model DNN_1, DNN_2 and DNN_3 are all trained using KALDI toolkit [20]. Model DNN_1 is English DNN-based acoustic models trained using the WSJ recipe of KALDI. Model DNN_2 and DNN_3 are Japanese acoustic models trained using the CSJ (Corpus of Spontaneous Japanese) recipe of KALDI. DNN_2 and DNN_3 are almost identical, except for the number of the output layer units, which correspond to senones. Between the HTK recipe and the KALDI recipe, the standard acoustic features are slightly different although they are MFCC-based features. While only CMVN (Cepstral Mean and Variance Normalization) is done in the former, a combination of CMN (Cepstral Mean Normalization), LDA (Linear Discriminative Analysis), and fMLLR (Feature-based Maximum Likelihood Linear Regression) are applied to the latter.

Table 2: *Settings of acoustic models.*

Name	#units of the output layer	Type	Input feature
HMM	—	English HMM	MFCC+CMVN
DNN_1	3,458	English DNN	MFCC+CMN+LDA+fMLLR
DNN_2	2,856	Japanese DNN	MFCC+CMN+LDA+fMLLR
DNN_3	9,429	Japanese DNN	MFCC+CMN+LDA+fMLLR

Models HMM and DNN_1 are used in the GOP experiment, and all the DNN models are used in the DTW experiment.

5.2. Results of GOP-based scoring

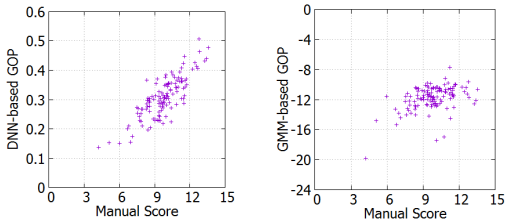
Speaker-level GOP scores computed using models HMM and DNN_1 are shown along with their corresponding speaker-level manual scores in Fig. 1. DNN-based scores are obtained by averaging posteriors directly but HMM-based scores are obtained by averaging the logarithmic values of probability densities.

DNN_1 shows a much higher correlation (cor.$=0.82$) than HMM (cor.$=0.45$), which is consistent with the fact of superiority of DNN in speech recognition. Difference in feature selection (see Table 2) may have some impact on this performance gap. But the performance of model HMM is lower than that reported in our previous studies, where correlation was generally higher than 0.60. One possible reason is a biased proficiency distribution over the learners who attended data collection. In our previous studies, learners of a wide range of proficiency were collected prior to data collection. In the current paper, about two thirds of learners are from university K, which is one of the top universities in Japan. High correlations by DNN even in the case of biased proficiency distribution will be discussed in [21].

Different from our previous study [5], no regression model is used and no supervised training is done in this experiment. Still, DNN_1 can show the scores highly correlated with the manual scores given by the two teachers. This may indicate their scoring strategies are very consistent with the scoring algorithm used in DNN_1, although more examination is needed.

5.3. Results of DTW-based scoring

The DTW distance is calculated between each pair of a shadowing utterance and its model utterance. Equations (3), (4), and (5) are used as distance metrics between two posterior vectors. Figure 2 shows the local path constraint adopted in the experiment. The sentence-level DTW distances are obtained by normalizing the accumulated distance by the duration of the corresponding model utterance. The speaker-level DTW distances are obtained



(a) DNN-GOP cor. $= 0.82$ (b) HMM-GOP cor. $= 0.45$

Figure 1: *Relationship between manual scores and DNN/HMM-based GOP.*

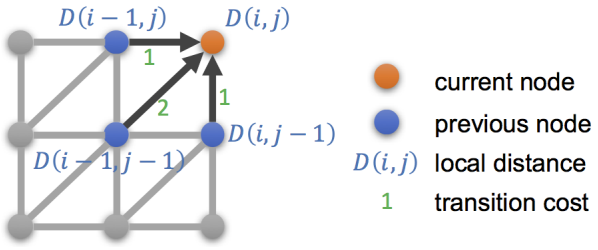


Figure 2: *An illustration of the local constraints used in DTW.*

by averaging the sentence-level ones.

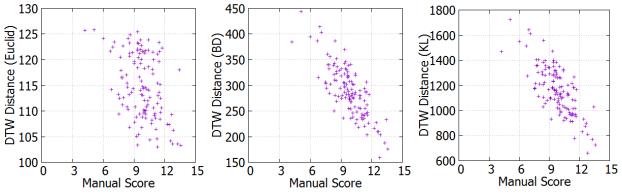
Fig. 3 shows the relationship of manual scores and DTW distances using English acoustic model (DNN_1). Both BD and KL-div show promising results and their correlations are -0.80 and -0.75, respectively. They are close to the performance of DNN-based GOP.

We made a further step to computing DTW distances using acoustic models of a different language from English. Two Japanese DNN models (DNN_2 and DNN_3) were prepared[1], where the numbers of phoneme states were about 9.4K and 3.4K. The former model was obtained by using the default configuration file of the CSJ KALDI recipe and the latter was trained so that the number of phoneme states was similar to that of DNN_1. Through comparison between the two models, it is possible to discuss the relationship between acoustic-phonetic granularity of DNN and the scoring performance.

Fig. 4 shows the relationship between DTW with Japanese DNN models (DNN_2 and DNN_3) and manual scores, where BD is used as local distance metric. The absolute value of the correlation in the case of DNN_3 is as high as 0.74, which is very close to that of DNN_1. This clearly indicates DTW with DNN posteriors have a good language independency, although some different trends may be observed if optimization of the number of phoneme states is done. From the results of DNN_2, we can say that it is a fact that the number of phoneme states has a large impact on automatic scoring.

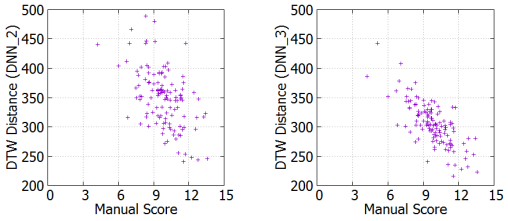
In Section 4.3, we pointed out a functional superiority of DNN-based DTW, which is language independency. Here, we can claim another one. In speech training for language learning, lines of dramas, movies, and animations are often used, so-called dubbing practice [22]. In this case, the model utterances are often expressive speech. If we use a method of GOP-based

[1]Ideally speaking, DNN models of a third language which is different from English and Japanese should have been prepared to highlight language-independence of learners' native language as well as that of the target language of learning.



(a)EUC-DTW cor. $= -0.44$ (b)BD-DTW cor. $= -0.80$ (c)KL-DTW cor. $= -0.75$

Figure 3: *Relationship between manual scores and three kinds of DTW distances.*



(a) DNN_2 cor. $= -0.55$ (b) DNN_3 cor. $= -0.74$

Figure 4: *Relationship between manual scores and two Japanese acoustic models with different numbers of senones.*

scoring, however, since a model utterance is represented just as sequence of phonemes, the expressive and para-linguistic aspect of speech has to be ignored when scores are calculated, although those aspects are very important for speech communication. Considering this constraint, GOP-based scoring will be less useful for dubbing and expressive speech training. This will be confirmed as one of our future works.

6. Conclusions and future works

In this study, we first collected English shadowing speech from 125 university students in Japan and manually scored them. Then we computed HMM-based and DNN-based GOP scores to analyze their availability for automatic scoring. The latter showed a much higher correlation with the manual scores. We also proposed a method of DTW-based comparison between a model utterance and its shadowed one. DTW distances were found to be highly correlated with manual scores. One superiority of this method over GOP-based scoring, language independency, was verified experimentally and another superiority, expressive speech assessment, was discussed theoretically.

In the future, we are going to 1) apply the acoustic condition of feature extraction used for DNN training to HMM training to make a fair comparison between them, 2) introduce regression models to achieve higher correlation as we already did in [5], and 3) examine DTW-based comparison using DNN posteriors for expressive and dubbed speech assessment.

7. Acknowledgment

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

8. References

[1] Y. Hamada, "The effectiveness of pre-and post-shadowing in improving listening comprehension skills," *The Language Teacher*, vol. 38, no. 1, pp. 3–10, 2014.

[2] ——, "Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension," *Language Teaching Research*, vol. 20, no. 1, pp. 35–52, 2016.

[3] K. T. Hsieh, D. H. Dong, and L. Y. Wang, "A preliminary study of applying shadowing technique to english intonation instruction," *Taiwan Journal of Linguistics*, vol. 11, no. 2, pp. 43–65, 2013.

[4] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, "Automatic assessment of language proficiency through shadowing," in *Chinese Spoken Language Processing, 2008. ISCSLP'08. 6th International Symposium on*. IEEE, 2008, pp. 1–4.

[5] S. Shi, Y. Kashiwagi, S. Toyama, J. Yue, Y. Yamauchi, D. Saito, and N. Minematsu, "Automatic assessment and error detection of shadowing speech: Case of english spoken by japanese learners," in *INTERSPEECH*, 2016, pp. 3142–3146.

[6] D. Luo, N. Minematsu, Y. Yamauchi, and K. Hirose, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences." in *SLaTE*, 2009, pp. 37–40.

[7] W. Hu, Y. Qian, and F. K. Soong, "An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[8] A. Metallinou and J. Cheng, "Using deep neural networks to improve proficiency assessment for children english language learners." in *INTERSPEECH*, 2014, pp. 1468–1472.

[9] E. Ribeiro, J. Ferreira, J. Olcoz, A. Abad, H. Moniz, F. Batista, and I. Trancoso, "Combining multiple approaches to predict the degree of nativeness." in *INTERSPEECH*, 2015, pp. 488–492.

[10] M. P. Black, D. Bone, Z. I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S. N. Chakravarthula, B. Xiao, M. Van Segbroeck, J. Kim *et al.*, "Automated evaluation of non-native english pronunciation quality: combining knowledge-and data-driven features at multiple time scales." in *INTERSPEECH*, 2015, pp. 493–497.

[11] R. Rasipuram, M. Cernak, A. Nanchen *et al.*, "Automatic accentedness evaluation of non-native speech using phonetic and subphonetic posterior probabilities," in *INTERSPEECH*, no. EPFL-CONF-209089, 2015.

[12] R. Ullmann, R. Rasipuram, H. Bourlard *et al.*, "Objective intelligibility assessment of text-to-speech systems through utterance verification," in *INTERSPEECH*, no. EPFL-CONF-209096, 2015.

[13] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 382–387.

[14] A. Lee and J. R. Glass, "Pronunciation assessment via a comparison-based system." in *SLaTE*, 2013, pp. 122–126.

[15] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[16] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[17] N. Minematsu, S. Asakawa, M. Suzuki, and Y. Qiao, "Speech structure and its application to robust speech processing," *New Generation Computing*, vol. 28, no. 3, pp. 299–319, 2010.

[18] F. Shiozawa, D. Saito, and N. Minematsu, "Improved prediction of the accent gap between speakers of english for individual-based clustering of world englishes," in *Spoken Language Technology (SLT) Workshop*. IEEE, 2016, pp. 129–135.

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

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[21] Y. Yamauchi, K. Ito, N. Minematsu, M. Nishikawa, A. Kawamura, and Y. Tsubota, "The relationship between accuracy improvement in automatic evaluation of l2 shadowing performances and learners overall proficiency levels," in *Proc. Annual Meeting of the Japan Association for Language Education and Technology*, 2017, to appear.

[22] D. Luo, R. Luo, and L. Wang, "Naturalness judgement of l2 english through dubbing practice," in *INTERSPEECH*, 2016, pp. 200–203.