DNN-BASED SCORING OF LANGUAGE LEARNERS' PROFICIENCY USING LEARNERS' SHADOWINGS AND NATIVE LISTENERS' RESPONSIVE SHADOWINGS

Suguru Kabashima, Yuusuke Inoue, Daisuke Saito, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo

{kabashima,inoue0124,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper investigates DNN-based scoring techniques when they are applied to two tasks related to foreign language education. One is a conventional task, which attempts to predict a language learner's overall proficiency of oral communication. For this aim, learners' shadowing utterances are assessed automatically. The other is a very new and novel task, which attempts to predict intelligibility or comprehensibility of a learner's pronunciation. In this task, native listeners' responsive shadowings are assessed. For both the tasks, similar technical frameworks are tested, where DNN-based phoneme posteriors, posteriogram-based DTW scores, ASR-based accuracies, shadowing latencies, etc are used to train regression models, which aim to predict manually rated scores. Experiments show that, in both the tasks, the correlation between the DNN-based predicted scores and the averaged human scores is higher than or at least comparable to the averaged correlation between the scores of human raters. This fact clearly indicates that our proposed automatic rating module can be introduced to language education as another human rater.

***Index Terms*—** Language learning, assessment, shadowing, comprehensibility, DNN, DTW, ASR and regression models

1. INTRODUCTION

Every language learner tries to acquire good skills of speaking, listening, writing, and reading. In this paper, the authors pay special attention to skills of speaking and listening, namely, oral proficiency. In a variety of CALL (Computer-Aided Language Learning) studies [1, 2, 3], spoken-dialogue-based systems are often designed to guide users to acquire better skills of not only speaking but also listening. Real spoken dialogues with native speakers are probably the best learning scenario to improve their oral proficiency and the above systems are designed to simulate this scenario technically.

In the current paper, we focus on another teaching/learning strategy to enhance learners' base capabilities of speaking and listening, which is shadowing [4, 5, 6]. Shadowing can be viewed as multitask training where learners are asked to listen to (and comprehend) and repeat given native utterances as simultaneously as possible. Thus, shadowing is a cognitively heavier task than simple listen-and-repeat practices. In cognitive sciences, it is explained to be *very unconsciously* that, in real conversations, native speakers mentally perform conversion from acoustics to phonological representation when listening and conversion from phonological representation to articulatory movements when speaking. Since these conversion processes are run automatically, native speakers can exploit their cognitive resources efficiently for higher-level processing such as thinking logically. Generally speaking, however, these two conversion processes are performed *consciously* by (beginning) learners and automatization of both the processes is said to be possible only by a huge number of rehearsals. Shadowing has been introduced as effective method for automatizing these processes and enhancing learners' base capabilities for oral communication [7, 8, 9, 10]. In Japan, shadowing practices are popular and imposed on learners in English classes in many middle and high schools [11], but assessment of shadowing utterances is done very rarely. This is a practical reason why the authors focus especially on shadowing.

Not only in shadowing training but also in more general pronunciation training, interactive feedback is very important to keep learners motivated to continue training. Automatization of the above two processes may depend on the quality of feedback, but what kind of corrective feedback is pedagogically valid and effective?

As far as the authors know, almost all the CALL studies for scoring or error detection of pronunciation compare learners' pronunciation with pronunciation models trained from native speakers. If phonemic or prosodic gaps to native pronunciation are detected, they are fed back to learners as errors. In this strategy, the target of pronunciation training is a native-sounding pronunciation, but a majority of teachers disagree with this strategy and they claim that the primary goal of pronunciation training is an intelligible or comprehensible enough pronunciation [12, 13]. Especially in English education, partly because English is adopted as one of the official languages in many countries and citizens in those countries speak accented English, teachers of English tend to accept accented English[1] if it is intelligible or comprehensible enough to listeners.

When one wants to provide technical supports to this practical strategy, however, a critical problem takes place. Intelligible or comprehensible enough pronunciations mean a variety of pronunciations that are accepted easily to listeners and the range of tolerance is expected to depend on individual listeners' experiences of being exposed to accented pronunciations. Further, it is difficult to *observe and define* the intelligible or comprehensible enough pronunciation because it probably exists only mentally in listeners' mind. In [15], we proposed a novel method to treat this problem adequately and a pilot and technical attempt was made successfully, which is *native listeners' responsive shadowing* of non-native utterances. Here, not learners but native listeners shadow learners' utterances and speech segments including inadequate articulatory or prosodic control in the natives' responsive shadowings are detected and used for scoring.

In this paper, after our work on automatic scoring of learners' shadowings [16] and another work of ours on automatic scoring of comprehensibility of learners' pronunciation based on natives' responsive shadowings [15], some improvements are realized by introducing regression models for both the tasks. Experiments show that, in both the tasks, the correlation between the DNN-based predicted scores and the averaged human scores is higher than or at least comparable to the averaged correlation between human raters.

[1] The term of *World Englishes* represents this policy very well [14].

2. RELATED WROKS

2.1. Intelligibility and comprehensibility [17, 18, 19, 20, 21, 15]

In applied linguistics, intelligibility and comprehensibility are defined somewhat differently [17, 18, 19]. Intelligibility indicates, for a given utterance, how accurately linguistic units such as words can be identified. Degree of intelligibility of a given utterance can be measured objectively by asking native listeners to write down that utterance word by word. Correct identification rate can represent intelligibility of that utterance. Comprehensibility of an utterance means how easily and smoothly listeners can understand the content of that utterance, often quantified using subjective questionnaires or comprehension tests imposed on listeners. Since correct comprehension often requires syntactic analysis and pragmatic analysis in addition to correct identification of words, the authors consider that comprehensibility covers intelligibility and represents more. Even if all the words of an utterance can be identified correctly but some listening (guessing) efforts are still required for comprehension, that utterance is not rated as highly comprehensible. These considerations led the authors to conclude that the target of pronunciation training for learners should be comprehensible enough pronunciation.

Objective measurement of intelligibility was made in [20, 21], where English spoken by immigrants to USA [20] and by Japanese college students [21] were presented to American English native listeners on a telephone line. The listeners were asked, after listening, not to write down but to repeat what they just heard. Their oral repetitions were transcribed word by word manually by technical staff to derive word-based intelligibility of each utterance.

In [20] and [21], good control was not made on listeners' repetition. It is highly speculated that efforts of listening and delay of repetition depended on listeners. If delay is reduced to be minimized, repetition becomes shadowing, where only small listening efforts are allowed. Since smooth shadowing is possible only with quick comprehension of presented utterances, the authors consider that results of *repetition* indicate how *intelligible* a given utterance is and that results of *shadowing* indicate how *comprehensible* it is. Then, the authors attempted to measure comprehensibility of pronunciation based on native listeners' responsive shadowings [15].

2.2. DNN-based scoring of shadowings [16, 15]

In [16], GOP (Goodness Of Pronuniciation) of learners' shadowings was compared with manually rated scores. GOP is widely used as feature indicating accuracy of articulation supposed in given utterances. GOP was tested for readings in [22] and for shadowings in [23, 24] but in these papers, GOP was calculated with HMMs (Hidden Markov Models). In [25], DNN-based GOP was proposed and in [16], it was tested firstly for shadowings.

GOP is theoretically defined as phoneme-based posterior $P(c_i|o_t)$, where o_t is a speech feature observed at time t, and c_i is phonemic class i[2] After forced alignment performed on an input utterance, the phoneme intended at time t, p_t, is obtained. Then, $P(p_t|o_t)$ is accumulated during an entire utterance. Then, the GOP score of a given utterance x is calculated as follows [16].

$$\mathrm{GOP}(x) = \frac{1}{D_x} \sum_t P(p_t|o_t), \tag{1}$$

[2]Strictly speaking, DNN-based acoustic models provide us $P(d_j|o_t)$, where d_j is a class of senones or a state in context-dependent HMMs. Here, j ranges up to several thousands. By collecting $\{d_j\}$ which belong to $c_i (d_j \in c_i)$, $P(c_i|o_t)$ is calculated as $\sum_j P(d_j|o_t)$.

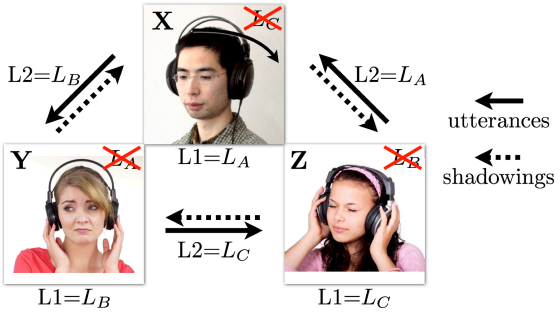

Fig. 1. Inter-learner shadowing [15]

where D_x is the frame-based duration of that utterance. DNN-based GOP is referred to as DNN-GOP. In [15], GOP was introduced not to learners' shadowings but to native listeners' responsive shadowings. GOP of natives' shadoiwngs showed a much higher correlation to listeners' (shadowers') perceived comprehensibility than GOP of learners' readings, which were presented to native shadowers.

In [16], another method for scoring with DNNs was tested, which is DNN-DTW (Dynamic Time Warping). In DNN-GOP, although the phonemic transcript of a presented model utterance for shadowing is used for forced alignment on shadowing utterances, the presented utterance itself is not compared directly with its shadowings. In DNN-DTW, the presented utterance is compared with its shadowings based on DTW after all the utterances are transformed to their posteriograms [26, 27, 28, 29]. This is because spectrogram-based DTW does not work adequately when non-linguistic gaps exist between two utterances, e.g. an adult male utterance is compared with a girl's utterance. These gaps inevitably cause unignorable acoustic mismatches between the two utterances, which often lead to inadequate DTW alignment. In [16], the posteriogram of an utterance was obtained as sequence of phoneme posterior vectors calculated through the front-end process of DNN-based ASR.

2.3. Inter-learner shadowing [15]

In [15], DNN-based GOP of native listeners' responsive shadowings showed a much higher correlation (r=0.73) to listeners' perceived comprehensibility than the GOP of learners' utterances (r=0.63) that were presented to native listeners for responsive shadowing. As pointed out in Section 1, it can be claimed that GOP of learners' utterances is a score that is suited for *native-sounding* pronunciation training and that GOP of native listeners' responsive shadowings is a score for *comprehensible enough* pronunciation training. However, the responsive shadowing approach has a critical drawback and it always requires native shadowers. A pedagogically feasible solution for this drawback was proposed conceptually in [15], which is inter-learner shadowing. Figure 1 shows inter-learner shadowing among three (groups of) learners. Learner **X**, who speaks L_A as L1 and is learning L_B but does not speak L_C, reads aloud sentences in L_B. His utterances in L_B are shadowed by learner **Y**, who speaks L_B as L1 and is learning L_C but does not speak L_A. Her utterances in L_C are shadowed by learner **Z**, who speaks L_C as L1 and is learning L_A but does not speak L_B. Her utterances in L_A are shadowed by learner **X**. Inter-learner shadowing can be viewed as a speech version of *Lang-8* [30], where any learner can support other learners and can be supported by other learners. If this infrastructure is provided for learners, a huge collection of pairs of learners' utterances and natives' shadowings will be obtained because, for any learner, natives' shadowings to his/her utterances are instructive and s/he can

know comprehensibility of pronunciation directly from the shadowings. Further, with a large enough corpus of learners' utterances and natives' shadowings, a good model will be trained that can predict natives' perceived comprehensibility from any given utterance. With this model, native shadowers will not be needed any more. The authors already started a large collection of Japanese utterances of Vietnamese learners and natives' shadowings [31].

3. IMPROVEMENTS IN SCORING LEARNERS' SHADOWINGS

3.1. Shadowing corpus collected and manually-labeled [16]

From 124 university students in Japan, the authors collected shadowing utterances in English. For this collection, a web-based shadowing recording system was developed. The students were asked to shadow 55 model utterances without viewing manuscripts, and each model utterance was shadowed four times. Ten utterances out of the 55 model utterances were selected based on syntactic and semantic difficulty, and used for investigation in this paper. Further, the fourth shadowings from the students are used in this paper. Two American teachers of English and a Canadian teacher of English assessed all the selected shadowings. Each utterance is composed of two or three phrases and scoring was done for each phrase. In total, 3,375 shadowed phrases were rated by the three teachers. Using the phrase-based scores, it is possible to derive sentence-level and speaker-level scores. A sentence-level score was obtained by averaging the phrase-level scores in that sentence, and a speaker-level score was obtained by averaging the sentence-level scores of that speaker.

The three teachers rated based on the following three criteria:

Phoneme (P): how adequately individual segments are produced?
Suprasegmental (S): how adequately prosodic control is made?
Correctness (C): How many words in a model utterance sound to be repeated as word, not as word fragment, in shadowing?

The score for each criterion ranges from 1 (worst / none) to 5 (best / all), so the full score is 15 and the worst score is 3 in total. In [16], the sum of **P+S+C** was used as human score and the average score over the three teachers was used as reference.

3.2. Features prepared to predict the manual scores

To prepare features for prediction, DNN-based ASR acoustic models were trained based on the WSJ [32] recipe of the KALDI toolkit [33], where acoustic features of MFCC (Mel Frequency Cepstrum Coefficients) were used with CMN (Cepstral Mean Normalization) and LDA (Linear Discriminant Analysis) involved. In this section, the DNN-based models are used to calculate senone posteriors and word accuracies for input shadowing utterances, where WSJ-based trigram language models are adopted as language model.

In [16], baseline GOP, defined as the average of frame-based phoneme posteriors in Equation 1 and called bGOP in this section, was compared to the human scores. The speaker-level bGOP scores were shown to be highly correlated (r=0.83) with teachers' speaker-level scores of **P+S+C**. In this paper, experiments are done using modified GOP scores, where some additional features are introduced to train regression models. Sentence-level prediction of teachers' scores as well as speaker-level prediction is also examined.

Since shadowing is always imposed with a given model utterance and the model utterance is generally read speech, the phonemic transcript of the model utterance is always available. As explained in Section 2.2, calculation of bGOP depends on usability of this phonemic transcript, which can also enable us to calculate the bGOP score for each phonemic segment in a shadowing utterance. By averaging all the phoneme-unit bGOP scores of a shadowing utterance, another version of GOP, called pGOP henceforth, is introduced for that utterance. If a model utterance covers all the kinds of phonemes of the L2, pGOP can be calculated separately for each kind of the phonemes [34]. Since some phonemes are often missing in a model utterance, we introduce pGOP for vowels and pGOP for consonants, which are referred to as vGOP and cGOP, respectively. Further, vGOP can be decomposed into vowels with primary stress (v1GOP), those with secondary stress (v2GOP), and those with no stress (v0GOP). Here, the stress level of each vowel in a word is available from the CMU pronunciation dictionary [35]. Considering that the rhythm of English is stress-timed and that of Japanese is mora-timed and that Japanese does not require alternation of stressed syllables and unstressed syllables, stress-dependent vGOP may be helpful to characterize differences well between American English and Japanese English and effective to improve the prediction performance of regression models [36].

Table 1. List of features used for prediction

GOP-based	bGOP, pGOP, vGOP, cGOP, v1GOP v2GOP, v0GOP
DTW-based	DNN-DTW
ASR-based	RS, WRR

pGOP and its variants of vGOP, cGOP, v1GOP, v2GOP, and v0GOP may be helpful for another reason. This is because they will be able to characterize teachers' strategy of rating much better. Generally speaking, vowel segments are longer than consonant segments. The frame-based average of posterior probabilities, bGOP, surely induces some biases that posterior probabilities of vowel segments are somewhat emphasized. If teachers do not have such biases when they rate shadowing utterances, pGOP and its variants will be better than bGOP. All of these features are extacted based on DNN-GOP, and by comparing a model utterance and a shadowing utterance based on DNN-DTW, another feature of average DTW distance between the two utterances is used for prediction.

In addition to the GOP-based features and the DTW-based feature, some other features are further introduced. Since shadowing is said to be a task of a high cognitive load, learners sometimes become silent or mumbles just for imitating a presented sequence of sounds. Ratio of Silence (RS), which is defined as physical and accumulated length of silent segments over that of an utterance, and Word-based automatic speech Recognition Rates (WRR) are also used for prediction. Table 1 shows a full list of the features used for prediction.

3.3. Prediction of the manual scores with regression models

3.3.1. Feature-based correlations

Correlations between a single kind of feature and the teachers' averaged manual scores, i.e. feature-based correlations, are shown in Table 2. Here, analysis was conducted in a speaker level. pGOP shows higher correlations than bGOP in every case of **P**, **S**, **C**, and **P+S+C**. Normalization in duration seems to be effective to simulate teachers' rating strategy. When vGOP and cGOP are compared, cGOP is found to be almost always better than vGOP. English has a larger number of vowels and consonants than Japanese and in Japanese English, some different vowels are often merged and pronounced as one vowel and some consonants are merged similarly. Experimentally speaking, consonant-dependent GOP is found to be more correlated with teachers' scores. It is interesting that

Table 2. Feature-based correlations with teachers' scores

features	**P**	**S**	**C**	**P+S+C**
bGOP [16]	0.74	0.83	0.71	0.83
pGOP	0.79	0.84	0.78	0.88
vGOP	0.70	0.83	0.70	0.81
cGOP	0.79	0.82	0.78	0.87
v1GOP	0.63	0.78	0.64	0.75
v2GOP	0.42	0.41	0.43	0.46
v0GOP	0.71	0.75	0.78	0.78
DNN-DTW	-0.66	-0.84	-0.69	-0.80
RS	-0.34	-0.21	-0.29	-0.30
WRR	0.79	0.81	0.71	0.84

Table 3. Model-based correlations in a speaker level

models	**P**	**S**	**C**	**P+S+C**
bGOP [16]	0.74	0.83	0.71	0.83
Lasso	0.84	0.89	0.76	0.90
SVR	0.85	0.89	0.83	0.89
Random Forest	0.77	0.84	0.79	0.86
inter-rater	0.77	0.69	0.86	0.87

Table 4. Model-based correlations in a sentence level

models	**P**	**S**	**C**	**P+S+C**
Lasso	0.68	0.73	0.65	0.77
SVR	0.70	0.73	0.68	0.78
Random Forest	0.67	0.68	0.61	0.74
inter-rater	0.58	0.54	0.74	0.75

v0GOP is better than v1GOP and v2GOP. Very small correlations of v2GOP is due to a small number of instances of vowels with secondary stress in the training data. Superiority of v0GOP to v1GOP is considered to be because of Japanese learners' poor pronunciation of unstressed vowels. Japanese learners not rarely produce every syllable as stressed syllable, often called as machine-gun rhythm English, because Japanese has no rhythmic structure comprised of alternation of stressed syllables and unstressed syllables.

DNN-DTW compares a model utterance and a shadowing utterance without referring to their phonemic transcript. Even in this case, the correlations of DNN-DTW are similar to those of vGOP. RS shows very small correlations and this is probably because we used the fourth shadowing utterances only, where three rehearsals of shadowing were allowed and silent words were rare. WRR is found to be as highly correlated with teachers' scores as bGOP. A possible problem of WRR is that the score of WRR depends on language models used. When a learner shadows model utterances A and B, the WRR scores of shadowings A and B depend on the linguistic content of A and B. Since GOP-based scoring uses a given phonemic transcript, its scores are independent of the content of utterances.

3.3.2. Model-based correlations

By combining the features prepared, three regression models of Lasso, SVR, and Random Forest were trained using `scikit` [37] to predict speaker-level and sentence-level teachers' averaged scores separately for each case of **P**, **S**, **C**, and **P+S+C**. These three models were selected after simple preliminary testing. Here, all the features in Table 1 but bGOP and v2GOP were adopted. Training and testing were carried out as 4-fold cross validation.

Table 3 shows speaker-level correlations obtained in the three models and averaged inter-rater correlations among the three teachers. The Lasso regression model shows the highest correlation of 0.90 in **P+S+C**, much higher than 0.83 obtained as feature-based correlation in [16]. This value is higher or at least comparable to the averaged inter-rater correlation of 0.87. It is the case with the other two models. This indicates that the trained regression models can work as another human rater. However, when the correlations are examined for each case of **P**, **S**, and **C**, the machine correlations are much higher in **S** but lower in **C**. When a specific type of teachers' scores is adopted, some other features should be integrated.

Table 5. The most predictive combination of three features

a) speaker level

	P	**S**	**C**	**P+S+C**
1	pGOP	DTW	pGOP	pGOP
2	WRR	vGOP	DTW	v1GOP
3	v1GOP	RS	cGOP	DTW

b) sentence level

	P	**S**	**C**	**P+S+C**
1	pGOP	DTW	DTW	DTW
2	DTW	pGOP	cGOP	cGOP
3	WRR	cGOP	RS	RS

Table 4 shows sentence-level correlations obtained in the three models and averaged inter-rater correlations among the three teachers. The SVR regression model turns out to have the highest correlation of 0.78 in **P+S+C**, which is at least comparable to the averaged inter-rater correlation of 0.75. The machine correlations in **C** are lower again than the human correlation. Why do the machine models work poorly in the case of **C**? The teachers' score of **C** indicates how many words in a model utterance sound to be repeated as word, not as word fragment, in shadowing. Even when a speech segment in shadowing which corresponds to a word is acoustically deviated from a native and normal pronunciation of that word, teachers may have found the segment to be intelligible enough and judged that the segment is produced as word. It is implied that the features used in the experiments are not sufficient enough to predict intelligibility or comprehensibility of utterances. This problem is tackled in the following section based on natives' responsive shadowing.

Table 5 shows the most predictive combinations of three features in the Lasso regression model in the eight cases of teachers' rating. It is well-known that even when a feature shows a very high feature-based correlation, if multiple features are allowed for prediction, that feature is not always selected as good feature because another feature will have a very high correlation to that feature and the other feature may be selected. Among the eight cases of teachers' rating, it is interesting that DNN-DTW is listed seven times, which is the highest among the eight features used in the experiments. Especially in the sentence level, DNN-DTW seems to be the most predictive feature. In the experiments, only a single feature was derived from DTW-based comparison but the above analysis indicates that some variants should be introduced. This is one of our future works.

4. IMPROVEMENTS IN SCORING NATIVES' RESPONSIVE SHADOWINGS

4.1. Corpus of natives' responsive shadowings [15]

Natives' responsive shadowing is examined to predict comprehensibility by adopting Japanese as L2 and using Vietnamese learners. If learners' utterances are very slow, their comprehensibility may be always high and independent of how heavily accented they are. This is why speaking rate control was introduced for speech collection.

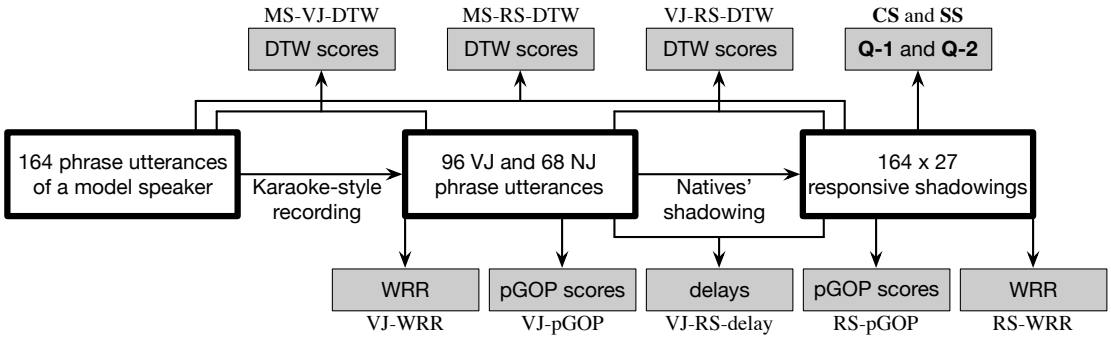

Fig. 2. Overview of the natives' responsive shadowing experiment and its features calculated for prediction

At first, an intermediate-level Japanese textbook with an audio CD was selected [38]. From the CD, ten read-aloud paragraphs were adopted. A tool of calculating readability, Jreadability [39], verified that the ten paragraphs belong to the same readability level. In addition to the professional model speaker's utterances in the CD, each phrase in the ten paragraphs was read aloud by six Vietnamese learners (three males and three females) and six native speakers (three males and three females) with their speaking rate being controlled by using a *Karaoke*-style recording program. Forced alignment was performed on the model speaker's utterances, and each phrase was shown only visually on a PC screen, where the color of text changed according to the model speaker's speaking rate. By following the text visually, a speaker read aloud each phrase. Finally, 96 Vietnamese Japanese (VJ) phrase utterances and 68 native Japanese (NJ) ones were selected and used in the experiments. They are not shared.

27 native Japanese were asked to shadow the VJ utterances and the NJ utterances. Presentation of these utterances was done in a random order through headphones. The 27 native Japanese did not have any hearing problem and went through a 10-minute simple practice of shadowing utterances. They were instructed not to imitate accented Japanese but shadow presented utterances in native Japanese.

After shadowing, two questions were always asked.

Q-1 How easily did you understand the presented utterance?

Q-2 How smoothly did you shadow the presented utterance?

A seven-degree scale was used for answering, where higher scores mean easier or smoother. The former is a comprehensibility score (**CS**) of a presented utterance and the latter is a shadowability score (**SS**) of the utterance. The two scores are expected to be highly correlated, but if strategic differences of rating are found between the two measures in some shadowers, the correlation will be low for them.

4.2. Features prepared to predict the subjective scores

To prepare features for prediction, Japanese DNN-based ASR acoustic models were trained based on the CSJ recipe of the KALDI toolkit. CSJ is the Corpus of Spontaneous Japanese [40] and it is the largest speech corpus of Japanese. The acoustic condition for training the models is the same as the condition in Section 3.2. To calculate word accuracies, CSJ-based trigram models are used.

To quantify smoothness of responsive shadowings, two kinds of speech features are focused on. One is related to accuracy of articulation and the other is to delay of shadowing. For the former, as explained in Section 3, pGOP-based features and DTW-based features as well as WRR are tested. Figure 2 shows the overview of the experiments. The thick boxes indicate three kinds of utterances: model utterances, Karaoke-style recordings of VJ and NJ, and natives' responsive shadowings. The gray boxes mean a variety of objective scores calculated automatically and two kinds of subjective scores (**Q-1** and **Q-2**) collected manually. Their abbreviations are also shown in the figure. The pGOP scores are calculated from the VJ utterances and natives' responsive shadowings. The DTW scores are calculated in three cases, which are between the model utterances and the VJ utterances, between the VJ utterances and their responsive shadowings, and between the model utterances and their native shadowings through Vietnamese learners' reading. WRR is calculated from the VJ utterances and their responsive shadowings.

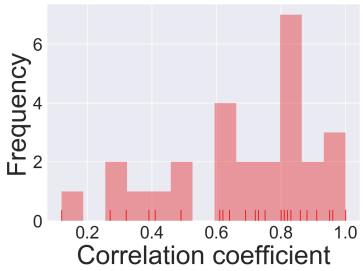

Fig. 3. The histogram of the shadowers' inter-measure correlations

As for delay of shadowing, by comparing forced alignment of a VJ utterance and that of its native responsive shadowing, the temporal gap between every pair of phoneme boundaries is obtained between the two utterances. The phoneme-based temporal gaps obtained from the two utterances were averaged to define delay of shadowing between the two utterances. Generally speaking, shadowing is performed with a delay of 1 to 2 seconds to a presented utterance.

4.3. Prediction of the subjective scores with regression models

4.3.1. Correlation between the two measures

Correlation between the two scores of **CS** and **SS** is calculated for each shadower. Their average is 0.68, which is not so high as expected. Figure 3 shows the histogram of the shadowers' correlations. Seven shadowers out of 27 show very low correlations and their average is 0.36. This is probably because of inter-measure strategic differences exhibited by the seven shadowers. To reduce these differences, prior discussion should have been done to achieve a consensus on what scores should be given to what kind of learners' utterances and to what kind of responsive shadowings. The averaged correlation among the remaining 20 shadowers is so high as 0.79. Among the 27 shadowers, there were five teachers of Japanese, and it should be noted that, after the experiments, they admitted that natives' shadowability of **SS** is a pedagogically sound and valid index for practical pronunciation training. In the following sections, it is described how well **CS** and **SS** scores can be predicted with regression models.

Table 6. Feature-based correlations with averaged **CS** and **SS**

features	**CS**	**SS**	features	**CS**	**SS**
RS-bGOP [15]	0.73	0.73	VJ-RS-delay*	0.59	0.69
VJ-bGOP [15]	0.63	0.50	MS-RS-DTW*	0.58	0.62
RS-pGOP	0.74	0.79	VJ-pGOP	0.58	0.44
RS-WRR	0.53	0.57	VJ-WRR	0.47	0.43
VJ-RS-DTW*	0.55	0.52	MS-VJ-DTW*	0.52	0.47

$*$ means that their correlations are negative. In the table, their absolute values are shown for visually easy comparison.

Table 7. Model-based correlations in a phrase level

models	**CS**	**SS**
Lasso	0.81	0.86
inter-rater	0.66	0.59

4.3.2. Feature-based correlations

As shown in Figure 2, each of the 96 VJ phrase utterances has its WRR (VJ-WRR) and pGOP (VJ-pGOP) and also has 27 responsive shadowings: 27 **CS**s and 27 **SS**s. For each shadowing, its pGOP (RS-pGOP), WRR (RS-WRR), and delay (VJ-RS-delay) are calculated. Thus, for each of the 96 VJ phrase utterances, it has five averages of **CS**, **SS**, RS-pGOP, RS-WRR, and VJ-RS-delay. As for these five variables, their over-shadower averages are used for analysis.

Table 6 shows feature-based correlations of the features examined, where those of RS-bGOP and VJ-bGOP reported in [15] are included as reference. It is clearly shown that, among the features examined, RS-related features have higher correlations than non-RS-related features of VJ-pGOP, VJ-WRR, and MS-VJ-DTW. Especially, RS-pGOP, VJ-RS-delay, MS-RS-DTW are highly expected to work effectively when **CS** and **SS** are predicted using regression models. These results indicate that, when comprehensibility of pronunciation is of interest, natives' responsive shadowings are much more informative than learners' utterances. Further, pGOP-based features and DTW-based features are better than WRR-based features (RS-pGOP and MS-RS-DTW $>$ RS-WRR, and VJ-pGOP and MS-VJ-DTW $>$ VJ-WRR). The ASR models trained only with native utterances are models optimized so as to recognize native utterances correctly. The authors can say that it is questionable whether such models can be used effectively as tolerance models of native listeners when listening to non-native utterances. If the ASR models are trained with non-native utterances, then, they will not give adequate feedback. Instead, pGOP of natives' responsive shadowings has much higher validity and usability if they are available.

4.3.3. Model-based correlations

Two Lasso regression models were trained to predict **CS** and **SS** and tested in 3-fold cross-validation. Table 7 shows the results and inter-rater correlations. The inter-rater correlation over 27 shadowers is calculated as follows. The 27 shadowers are divided into two groups of 1 and 26 shadowers. For each VJ utterance, the averaged **CS** score and **SS** score are calculated over the 26 shadowers. Then, the correlation between the remaining shadower's scores and the averaged scores is calculated. This process is run repeatedly by treating each shadower as the remaining shadower. Finally, the average of the 27 correlations is obtained and this is the inter-rater correlation in Table 7. The two regression models show much higher correlations than between-raters. Generally speaking, inter-rater correlations are higher in a speaker level, lower in a sentence level, and much lower in a phrase level. Although strict comparison is inadequate among Table 3, Table 4, and Table 7, the above tendency is found in these inter-rater correlations. However, the phrase-level machine correlations in Table 7 are between the speaker-level correlations in Table 3 and the sentence-level correlations in Table 4. These results indicate again high validity and usability of RS-based features.

5. FUTURE DIRECTIONS

In this paper, DNN-based scoring of language learners' proficiency was examined by automatically scoring learners' shadowings and native listeners' responsive shadowings. Although a variety of features were examined, almost all of them were related to the segmental aspect of speech. Well-trained regression models with the segmental features only were shown to behave well like human rater but, with some prosodic features, their performance may be improved.

When assessing learners' shadowings, if a learner shadows in a classroom situation, shadowing voices of other students are nothing but noises to that learner. Technically speaking, babble noises are the most difficult type of noise for suppression. Further, if learners are shadowing rather synchronously, shadowing voices of other students easily become the most difficult babble noise, which is synchronous babble noise. Generally speaking, since good learners tend to shadow more loudly, good learners can be said to be the most technically-difficult noise source to suppress. A novel noise suppression model should be devised for practical situations.

Very promising results were obtained about automatic prediction of comprehensibility or shadowability of pronunciation based on native listeners' responsive shadowing. After a series of experiments in this paper, we already started a larger collection of Vietnamese Japanese utterances and natives' responsive shadowings to them [31], namely, inter-learner shadowing between Vietnamese and Japanese. The authors are interested in usability of natives' shadowing performances as comprehensibility labels attached to non-native utterances. A large number of non-native speech corpora are available [41] but many of them are with speaker-level or utterance-level labels, or without them. Labels of higher temporal resolution require both expert labelers and time. pGOP scores are obtained as temporal sequence from a given non-native utterance simply by asking ordinary (non-expert) native listeners to shadow that utterance. The authors will discuss usability of temporal sequences of pGOP-based features as comprehensibility labels with higher temporal resolution.

6. CONCLUSIONS

DNN-based scoring techniques were examined in two tasks of 1) predicting a language learner's oral proficiency based on his/her shadowing utterances and 2) predicting comprehensibility of his/her pronunciation based on native listeners' responsive shadowing. In both the tasks, promising results were obtained and well-trained regression models were shown to behave like human rater. Especially, the author consider that native listeners' responsive shadowing has a very high potential because, as far as the authors know, the current work and our previous work [15] are the initial and technical attempt to deal with comprehensibility of learners' pronunciation, which may be hidden only in listeners' mind. In this sense, responsive shadowing can be said to be an easy scheme to disclose the hidden attribute of learners' pronunciation.

7. ACKNOWLEDGEMENT

This work was financially supported by JSPS or MEXT KAKENHI JP26118002, JP26240022, and JP18H04107.

8. REFERENCES

[1] Reima Karhila, Sari Ylinen, Seppo Enarvi, Kalle Palomäki, Aleksander Nikulin Olli Rantula, Vertti Viitanen, Krupakar Dhinakaran, Anna-Riikka Smolander, Heini Kallio, Katja Junttila, Maria Uther, Perttu Hämäläinen, and Mikko Kurimo, "SIAK-agame for foreign language pronunciation learning," *Proc. INTERSPEECH*, 3429–3430, 2017.

[2] Wei Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," *Proc. ICASSP*, 6135–6139, 2016.

[3] Wei Li, Kehuang Li, Sabato Marco Siniscalchi, Nancy F. Chen, and Chin-Hui Lee, "Detecting mispronunciations of L2 learners and providing corrective feedback using knowledge-guided and data-driven decision trees," *Proc. INTERSPEECH*, 3127–3131, 2016.

[4] William D. Marslen-Wilson, "Speech shadowing and speech comprehension," *Speech Communication*, 4, 1, 55–73, 1985.

[5] Holger Mitterer and Mirjam Ernestus, "The link between speech perception and production is phonological and abstract: Evidence from the shadowing task," *Cognition*, 109, 168–173, 2008.

[6] Peter. W. Carey, "Verbal retention after shadowing and after listening," *Perception & Psychophysics*, 9, 1, 79–83, 1971.

[7] Shigeru Miyake, "Cognitive processes in phrase shadowing and EFL Listening," *JACET (Japan Association of College English Teachers) Bulletin*, 48, 15–28, 2009.

[8] Yo Hamada, "The effectiveness of pre- and post-shadowing in improving listening comprehension skills," *The Language Teacher*, 38, 1, 3–10, 2014.

[9] Yo Hamada, "Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension," *Language Teaching Research*, 20, 1, 35–52, 2016.

[10] Tomoko 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.

[11] Kumi Suzuki, "Investigation on improvement of listening comprehension based on shadowing practices," Report of EIKEN BULLETIN, 2007.

[12] Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, *The handbook of World Englishes*, published by Wiley-Blackwell, 2009.

[13] Jennifer Jenkins, *World Englishes: a resource book for students,* published by Routledge, 2009.

[14] Inernational Association for World Englishes, http://www.iaweworks.org

[15] Yusuke Inoue, Suguru Kabashima, Daisuke Saito, Nobuaki Minematsu, Kumi Kanamura, and Yutaka Yamauchi, "A study of objective measurement of comprehensibility through native speakers' shadowing of learners' utterances," *Proc. INTERSPEECH*, accepted in 2018.

[16] Junwei Yue, Fumiya Shiozawa, Shohei Toyama, Yutaka Yamauchi, Kayoko Ito, Daisuke Saito, and Nobuaki Minematsu, "Automatic scoring of shadowing speech based on DNN posteriors and their DTW," *Proc. INTERSPEECH*, 1422–1426, 2017.

[17] Murray J. Munro and Tracey M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, 45, 1, 73–97, 1995.

[18] Murray J. Munro and Tracey M. Derwing, "The functional load principle in ESL pronunciation instruction: An exploratory study," *System* 34, 520–531, 2006.

[19] Tracey M. Derwing and Murray J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research,* published by John Benjamins Publishing, 2015.

[20] Jared Bernstein, "Objective measurement of intelligibility," *Proc. ICPhS*, 1581–1584, 2003.

[21] Nobuaki Minematsu, Kohji Okabe, Keisuke Ogaki, and Keikichi Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ database," *Proc. INTERSPEECH*, 1481–1484, 2011.

[22] Silke M. Witt and Steve J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, 30, 1, 95–108, 2000.

[23] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, "Automatic assessment of language proficiency through shadowing," *Proc. ISCLP*, 1–4, 2008.

[24] Luo Dean, Nobuaki Minematsu, Yutaka Yamauchi, and Keikichi Hirose, "Analysis and comparison of automatic language proficiency assessment between shadowed sentences and read sentences," *Proc. SLaTE*, 37–40, 2009.

[25] Wenping Hu, Yao Qian, and Frank K. Soong, "An improved DNN-based approach to mispronunciation detection and diagnosis of L2 learners' speech," *Proc. SLaTE*, 71–76, 2015.

[26] Ramya Rasipuram, Milos Cernak, Alexandre Nanchen and Mathew Magimai-Doss, "Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities", *Proc. INTERSPEECH*, 2015.

[27] Raphael Ullmann, Ramya Rasipuram, Mathew Magimai-Doss, and Hervé Bourlard, "Objective intelligibility assessment of text-to-speech systems through utterance verification", *Proc. INTERSPEECH*, 2015.

[28] Ann Lee and James Glass, "A comparison-based approach to mispronunciation detection", *Proc. Spoken Language Technology Workshop*, 382–387, 2012.

[29] Ann Lee and James Glass, "Pronunciation assessment via a comparison-based system", *Proc. SLaTE*, 122–126, 2013.

[30] *Lang-8*, http://lang-8.com

[31] Yuusuke Inoue, Suguru Kabashima, Daisuke Saito, and Nobuaki Minematsu, "Improvements of predicting comprehensibility scores using natives' responsive shadowings and regression models," *Proc. Autumn Meeting of Acoustical Society of Japan*, 2018 (in Japanese).

[32] Wall Street Journal Corpus, https://catalog.ldc.upenn.edu/ldc93s6a https://catalog.ldc.upenn.edu/ldc94s13a

[33] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, Karel Veselý, "The KALDI speech recognition toolkit," *Proc. ASRU*, 2011.

[34] Leonardo Neumeyer, Horacio Franco, Vassilios Digalakis, and Mitchel Weintraub, "Automatic scoring of pronunciation quality," *Speech Communication*, 30, 83–93, 2000.

[35] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[36] Michael Swan and Bernard Smith, *Learner English: A teacher's guide to interference and other problems,* Cambridge University Press, 2001.

[37] `scikit-learn`, http://scikit-learn.org/stable/index.html

[38] Mariko Matsuura, Shunsui Fukuchi, Maiko Kohno, and Kayo Yoshida, *Japanese speech training,* published by ASK publisher, 2014.

[39] Jreadability, https://jreadability.net

[40] Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara, "Spontaneous speech corpus of Japanese," *Proc. LREC*, 947–952, 2000. http://pj.ninjal.ac.jp/corpus_center/csj/

[41] Non-native speech database, https://en.wikipedia.org/wiki/Non-native_speech_database