Speech Communication 116 (2020) 86–97

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models

Jiang Fu, Yuya Chiba, Takashi Nose, Akinori Ito

Graduate School of Engineering, Tohoku University, Aramaki Aza-Aoba 6–6–05, Aoba-ku, Sendai, 980–8579 Japan

ARTICLE INFO

Keywords:
Automatic proficiency assessment
Non-native speech
Speech recognition
Acoustic models
Deep neural network (DNN)
Japanese learners
Computer-assisted language learning (CALL)

ABSTRACT

Speech-based computer-assisted language learning (CALL) systems should recognize the utterances of the learner with high accuracy and evaluate the language proficiency of the specific speaker with appropriate methods. In this paper, we discuss the automatic assessment of the second language (L2) for non-native speakers. There are many existing works on pronunciation evaluation by applying the goodness of pronunciation (GOP) method. This paper introduces an automatic proficiency evaluation system that combines various kinds of non-native acoustic models and native ones, such as Gaussian mixture model (GMM)-hidden Markov model (HMM) and deep neural network (DNN)-HMM. Most of existing works assume that we know the transcription of an utterance (the reference sentence) when evaluating the utterance, especially in reading and repeating tasks. To realize a reference-free proficiency evaluation, we propose a novel machine score named as the reference-free error rate (RER) to evaluate English proficiency. In our experiments, the DNN-based non-native acoustic models outperformed the traditional acoustic models on non-native speech recognition. Thus, we calculated the RER by regarding the recognition result from the DNN-based non-native acoustic model as “reference” and the result from the native acoustic model as “recognition result”. The proposed RER has high correlation with human proficiency scores, which indicates the effectiveness of RER for automatically estimating the proficiency. By combining the RER with other machine scores such as the log-likelihood scores, we obtained high correlation (reading aloud task: $r = 0.826, p < 0.001, N = 190$; constrained interactive dialogue task: $r = 0.803, p < 0.001, N = 26$; spontaneous English conversation task: $r = 0.799, p < 0.001, N = 28$) to the human scores.

1. Introduction

The rapid development of globalization leads to an increase in the demand for the second language (L2) learners to study foreign languages. Traditional language teaching indoors contributes to timely interaction between the teachers and the students. However, such language teaching method tends to be costly and time-consuming. Due to the increasing computing power and modern information technology in the past two decades, computer-assisted language learning (CALL) system has attracted much attention in the field of language teaching, as an efficient tool for the L2 learners. The speech-based CALL system can recognize the words or sentences uttered by the L2 speakers via automatic speech recognition (ASR) technology. To detect pronunciation errors and judge pronunciation proficiency in high accuracy for the L2 learners, automatic assessment of pronunciation is an essential task in speech-based CALL system (Witt, 2012). A large number of researchers have investigated pronunciation evaluation in terms of various scoring methods ranging from phonemic domain (Franco et al., 2000; Moustroufas and Digalakis, 2007; Ito et al., 2008; Tu et al., 2018) to prosodic domain (Suzuki et al., 2008; Hönig et al., 2012; Truong et al., 2018).

There are three main types of automatic pronunciation evaluation on the phonemic side:

(a) Evaluating the overall goodness of one sentence or a pronounced specific word. Franco et al. (2000) indicated that the combination of machine evaluation scores from ASR system, e.g., phone log-posterior probability scores based on Gaussian mixture model (GMM)-hidden Markov model (HMM), segment duration scores and timing scores, achieved a high correlation with human rating scores in sentence-level. In addition, the correlation between the mentioned machine scores and human scores had a comparable result to human-to-human correlation when an adequate amount of evaluated speech data is available (Neumeyer et al., 2000). Speech recognition technology was also used in pronunciation evaluation system to increase human-machine score correlation (Molina et al., 2009).

(b) Evaluating goodness of a specific phoneme. L2 learners can use this system to derive a score for one arbitrary L2 phoneme they selected in their non-native speech evaluation. In this branch, Kim et al. (1997) set a series of particular phonemes to exam-

E-mail address: fujiang@ecei.tohoku.ac.jp (J. Fu).

https://doi.org/10.1016/j.specom.2019.12.002
Received 20 December 2018; Received in revised form 5 November 2019; Accepted 20 December 2019
Available online 23 December 2019

ine how well the machine scores correlate with the corresponding human scores rating of single phone utterances. To validate the automatic pronunciation evaluation results, Witt and Young (2000) developed a system for scoring pronunciation of each phoneme in an utterance and detecting pronunciation errors. The most well-known approach might be the Goodness of Pronunciation (GOP) (Witt and Young, 1997a), which is an approximation of the probability of the target phoneme.

(c) Detecting and diagnosing mispronounced phonemes. Compared with evaluating goodness of a specific phoneme, such kind of system could give a direct and accurate output on whether the pronounced phoneme is acceptable and unacceptable, or the comparison between correct phoneme and wrong one. Besides the general approach, the discriminative approach is often applied in these systems for pronunciation modeling. Franco et al. (1999) proposed a method to conduct automatic mispronunciation detection with phone segment score and a log-likelihood ratio score calculated from two dissimilar acoustic models trained by non-native speech, one from acceptable correct native-like speech and the other one from extremely non-native speech. Regarding the performance improvement in mispronunciation detection, Ito et al. (2007) developed a decision tree-based clustering method for phoneme-level classification. In addition, approaches using the deep neural network (DNN)-based acoustic models significantly improved the results of mispronunciation detection and diagnosis (Hu et al., 2015; Li et al., 2016; Mao et al., 2018; Li et al., 2018).

In this paper, we aim at improving the performance of pronunciation evaluation in type a).

Evaluation of language proficiency on L2 learners, to some extent, is an important objective as well as a difficult mission. Even in the case of traditional human evaluation, the inevitable variability in the evaluation standard of multiple teachers leads to less objective and consistent assessments. As an approach of pronunciation evaluation, the GOP is widely used in phone level pronunciation scoring of non-native speech (Witt and Young, 1997b) and utterance verification assessment (Yue et al., 2017). GOP is originally calculated using the GMM, and then a different GOP estimation method based on DNN has been proposed (Hu et al., 2013).

To improve the evaluation accuracy, the native and non-native acoustic models are used for measuring whether a specific utterance sounds native-like or non-native like (Witt and Young, 1997b; Minematsu et al., 2002; Moustroufas and Digalakis, 2007). The acoustic model in an ASR system trained from native speech database is regarded as a baseline to measure how close the given non-native speech is to the target native pronunciation, as the native acoustic model includes nearly whole phonemic features of the native language. The phoneme-level log-likelihood score and posterior probabilities generated from the native acoustic model achieved a high correlation to human score by using linear regression for non-native speech (Franco et al., 1997), and the log-posterior probabilities of aligned phonemes in the GOP are also obtained from an acoustic model trained with native speech. However, the speech from non-native speakers is more or less influenced by the accent, rhythm, or other phonemic characteristics of their mother tongue (L1). In our previous work (Fu et al., 2018), the DNN-based acoustic models trained from non-native English speech uttered by Japanese native speakers gave very high accuracy for recognizing English speech by Japanese speakers. There exists a plenty of studies that use both native and non-native acoustic models to conduct automatic pronunciation evaluation (Kawai and Hirose, 1998; Moustroufas and Digalakis, 2007; Ohkawa et al., 2009; Tu et al., 2018).

Furthermore, most systems assume that the transcription of a learners utterance can be exploited at the time of evaluation. It can be realized by letting a system specify the sentence to be read. However, the text-dependent evaluation cannot be used for evaluation on any spontaneous speech, which is essential to combine a spoken-dialog-based CALL system with pronunciation evaluation. Moustroufas and Digalakis (2007) proposed a system that used acoustic models in both L1 and L2 and compared the sentence-level likelihood scores from these acoustic models for any utterance. The objective of our system is to develop a highly accurate pronunciation evaluation system for sentence-level evaluation that can be used for any utterance, assuming both L1 and L2.

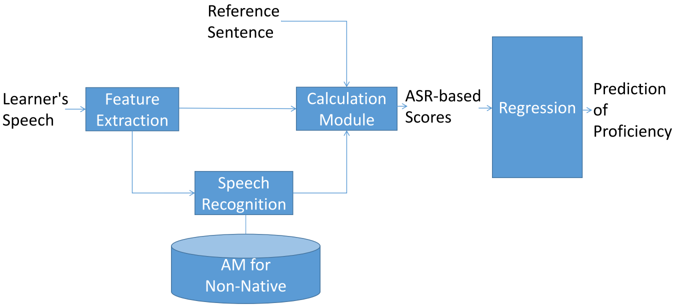

Fig. 1. Automatic proficiency evaluation systems for reading aloud and repeating tasks. Here, the reference sentence is required during the calculation step for ASR-based scores.

In this paper, for developing the reference-text-independent pronunciation scoring in a dialogue-based CALL system, we propose a novel machine score (Reference-free Error Rate, RER) that is based on speech recognition results using native and non-native acoustic models. The DNN-HMM acoustic models are utilized for improving the accuracy of evaluation. We investigate pronunciation evaluation by combining multiple machine scores including log-likelihood scores and the proposed RER.

The organization of the paper is as follows: In Section 2, we firstly compare conventional proficiency evaluation systems to our established system, and then briefly introduce the machine scores used for evaluation in our system, e.g., log-likelihood scores, Word Error Rate (WER) and RER. Experimental conditions are described in Section 3, including the detailed conditions of DNN-based acoustic modeling, the non-native and native ASR systems. Section 4 shows the experimental results of non-native speech recognition. In Section 5, the automatic assessment on three kinds of evaluation sets (reading aloud task, constrained dialogue task and English conversation task) with the proposed machine scores is discussed. Finally, discussion and conclusion are listed in Section 6.

2. Automatic English proficiency assessment

2.1. Overview of automatic proficiency assessment

The automatic proficiency assessment systems for L2 learners include a variety of tasks ranging from restricted speech to spontaneous speech (Bernstein et al., 2000; Cucchiarini et al., 2002; Zechner et al., 2009). Fig. 1 describes one of the conventional proficiency evaluation systems for reading aloud and repeating tasks (de Wet et al., 2009). At first, the learner's speech is processed with feature extraction, and sent to the speech recognizer for decoding. Then, the output of the recognizer, reference sentence and extracted features are calculated as several ASR-based scores, including the rate of speech (ROS), GOP and accuracy. Finally, these scores are sent to the regression model which produces the prediction score for the speech response. SpeechRater is a typical automatic proficiency assessment system for scoring non-native English speaker's spontaneous speech (Zechner et al., 2009), which is showed in Fig. 2. In this scoring process, the input speech is decoded into word sequence with the non-native acoustic model and forced-aligned to the native acoustic model for calculating the log-likelihood and the durations of phonemes. The linear regression model is used to predict proficiency score for the utterance of non-native speakers. The conventional

S

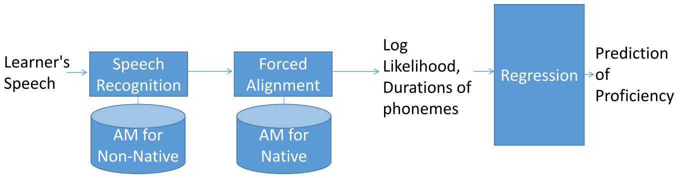

Fig. 2. Schematic diagram of a standard automatic scoring system for spontaneous speech. The input speech is decoded into word sequence with the non-native acoustic model and forced-aligned to the native acoustic model for calculating the log-likelihood and the durations of phonemes. The linear regression model is used to predict proficiency score.

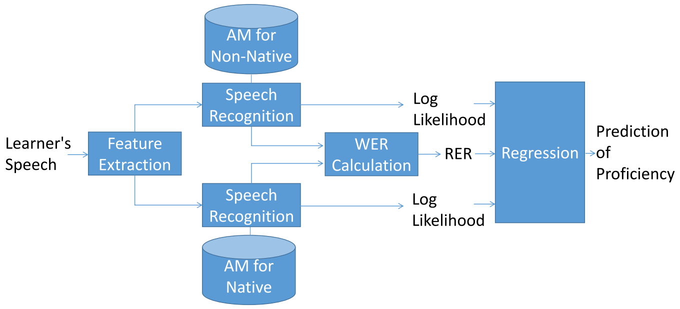

Fig. 3. Overview of the proposed proficiency evaluation system. We propose the RER as a new machine score in the regression-based prediction of proficiency, which based on the two assumed conditions: a) high recognition accuracy from non-native ASR and b) the effectiveness of evaluating language proficiency with WER from native ASR.

systems provide log-likelihood scores or other ASR-based scores as machine scores in regression-based proficiency prediction.

Log-likelihood scores are widely used for automatic pronunciation scoring as a traditional machine score (Franco et al., 1997; Cucchiarini et al., 2000; Moustroufas and Digalakis, 2007). WER could be a machine score for assessing proficiency based on its high relevance with the function of hearing and distinguishing in the ASR system. Tao et al. (2014) systematically demonstrated that WER obtained from the ASR module in an automatic assessment system played an important role in scoring. Moreover, the WER is the relatively straightforward manifestation and easily obtained among the output results from the ASR system. One disadvantage of the WER is that it requires the prepared reference transcription of the utterance. It is not always possible to prepare the reference sentences, especially in a dialogue-based CALL system.

To recognize and assess the non-native speech in a dialogue-based CALL system, we propose to utilize two acoustic models as shown in Fig. 3: one is trained with a non-native speech database and the other one with a native database. Here, we expect that the non-native ASR system recognizes non-native utterances with high accuracy regardless of the speaker's proficiency; besides, the native ASR system should give less accurate results for utterances with low proficiency. If this is true, we can distinguish the proficiency of the non-native speaker by comparing the two different recognition results from non-native and native ASR systems with the input speech. Based on this idea, we propose a new method for automatic assessment: reference-free error rate (RER).

2.2. Machine scores for automatic assessment

2.2.1. HMM-based phone log-likelihood scores

HMM-based acoustic models, trained from the speech database of native speakers, can be used to generate phonetic time alignments of the non-native speaker's speech during the decoding step by the Viterbi algorithm. For each phone segment in each sentence, the normalized log-likelihood score $\hat{l}_i$ (Franco et al., 1997) is defined as

$$\hat{l}_i = \frac{1}{d_i} \sum_{t=t_0}^{t_0+d_i-1} \log p(y_t|q_i) \qquad (1)$$

where $p(y_t|q_i)$ is the likelihood of the current observation vector y_t to the i-th phone q_i, d_i is the phones duration in frames, and t_0 is the staring frame index of the phone segment.

The phone-based log-likelihood score in one sentence L (Franco et al., 1997) is defined as

$$L = \frac{1}{N} \sum_{i=1}^{N} \hat{l}_i \qquad (2)$$

where the whole phones likelihood score is summed in the sentence over the number of phones N to obtain the average phone-based log-likelihood score.

2.2.2. Word error rate

To evaluate the accuracy of an ASR system, the word error rate (WER) is the standard measurement. Generally, we obtain the WER in the case of knowing the text spoken by the language learner as the reference transcription. The output word sequence from the ASR system is forced to align with the reference text. Assume we have I utterances in the database. $T_i^{(\mathrm{ref})}$ and $T_i^{(\mathrm{asr})}$ are the corresponding reference sentences and recognized sentences to the i-th utterance, respectively. According to the Levenshtein distance (Levenshtein, 1966), the WER is calculated as

$$WER = D(\mathbf{T}^{(\mathrm{ref})}, \mathbf{T}^{(\mathrm{asr})}) = \frac{\sum_{i=1}^{I} d_L(T_i^{(\mathrm{ref})}, T_i^{(\mathrm{asr})})}{\sum_{i=1}^{I} |T_i^{(\mathrm{ref})}|} \qquad (3)$$

where $\mathbf{T}^{(\mathrm{ref})} = (T_1^{(\mathrm{ref})}, \dots, T_I^{(\mathrm{ref})})$ and $\mathbf{T}^{(\mathrm{asr})} = (T_1^{(\mathrm{asr})}, \dots, T_I^{(\mathrm{asr})})$ are the reference sentence-set and recognized sentence-set, respectively. $D(\mathbf{X}, \mathbf{Y})$ is the WER between the sentence-sets $\mathbf{X}$ and $\mathbf{Y}$, $|T_i^{(\mathrm{ref})}|$ is the number of words in the i-th sentence, and $d_L(X, Y)$ is the Levenshtein distance between sentences X and Y.

2.2.3. Reference-free error rate

The proposed reference-free error rate (RER) is calculated by comparing the recognized text obtained from native ASR to that obtained from non-native ASR based on the Levenshtein distance. If we have I utterances of a speaker, let $\mathbf{T}^{(\mathrm{non})} = (T_1^{(\mathrm{non})}, \dots, T_I^{(\mathrm{non})})$ and $\mathbf{T}^{(\mathrm{nat})} = (T_1^{(\mathrm{nat})}, \dots, T_I^{(\mathrm{nat})})$ be the recognized sentences by the non-native and native ASR systems, respectively. Then the RER is calculated as

$$RER = D(\mathbf{T}^{(\mathrm{non})}, \mathbf{T}^{(\mathrm{nat})}). \qquad (4)$$

If the RER is high, it means that the two recognition results differ much, which happens when the proficiency of the utterance is low.

3. Experimental conditions

3.1. The databases

Four speech databases are utilized in our research:

(a) In this research, we conducted the automatic proficiency evaluation in the phonemic category of non-native speech. Therefore, regarding the human score, the segmental score of sentence utterances pronunciation was selected in ERJ (English Read by Japanese) database (Minematsu et al., 2004; Makino and Aoki, 2012). The ERJ database is widely used for many research studies on the development of the CALL systems (Ito et al., 2008; Luo et al., 2010; Minematsu et al., 2011), non-native speech recognition (Wang et al., 2017) and synthesis (Oshima et al., 2016).

Table 1
Assessment criteria for the sentences in ERJ database (Minematsu et al., 2004).

1 (Very poor)	Inaccurate, and apt to be misunderstood
2 (Poor)	Inaccurate, and considerable practice needed
3 (Fair)	Fair and common
4 (Good)	Accurate, but some practice needed
5 (Excellent)	Good and near-native speaker level

Table 2
Assessment criteria of segmental score for the dialogues in Tohoku multi-modal non-native English dialogue (TMNED) corpus (Wu et al., 2018).

1 (Very poor)	Really non-native speech
2 (Poor)	Some non-native speech
3 (Fair)	Fair and common
4 (Good)	Native-alike prounciation
5 (Excellent)	Very native-alike pronunciation

Every Japanese student was asked to be fully prepared with sufficient practice before recording. They were informed to read each given sentence as accurately as possible, and they had three chances to re-record or skip those sentences failed to be recorded during the main recording time. They can choose to challenge to re-record those failed sentences after the main recording time. English speech utterance samples spoken by 100 male and 102 female Japanese students are included in the database. Five phonetically trained American English teachers were recruited as rating experts. They were informed to listen ten sentences read by each Japanese student and give a 1–5 scale segmental score to every sentence. The 5-grade scales are showed in Table 1.

(b) As a native speech database, the TIMIT database (Garofolo et al., 1993) is widely used for research in the field of speech technology (Graves and Schmidhuber, 2005; Greff et al., 2017). The TIMIT database collected by Texas Instruments includes 630 speakers and 2343 different sentences. Each of the 630 individuals from the eight major dialect regions in the United States spoke out the given ten sentences. All the sentences were manually segmented and marked at the phoneme level.

(c) Another non-native speech database as an evaluation set in our system is Tohoku multi-modal non-native English dialogue (TMNED) corpus recorded in Tohoku university (Wu et al., 2018). For constructing this corpus, Wu et al. (2018) let the learners take English conversation with the CALL system based on scenarios learned in advance. Since university students in Japan have few opportunities to use English for conversation and communication, to a certain extent, most of them dont have much confidence to speak English out directly. To construct the corpus with high quality for research use, the learners took lots of effort to remember the content of the conversations before video recording. The original conversation text and the human transcription from the learners' speech are nearly the same. Therefore, this corpus contains restricted English speech of two different dialogue topics spoken by 13 undergraduate students (9 males and 4 females). Each dialogue covers three sentences from the system and three responding sentences from the learner. 26 dialogues were collected from this corpus. The human evaluation step of the collected speech follows the design of the ERJ database's 1–5 scale segmental score: 3 American English language teachers were employed to annotate all dialogue data with assessment criteria presented in Table 2.

(d) We constructed a new speech database named as Tohoku English conversation (TEC) corpus for the spontaneous speech evaluation. 14 (12 males and 2 females) Tohoku University students participated in the recording of the TEC corpus. Each student was asked to participate in two different conversations. Each conversation is a three-minute free English conversation with every two students. To eliminate the tension of the students during the recording process, and to better understand the dialogue, we asked participants to speak at a relatively slower pace, and in the pauses of context, try to reserve enough time for understanding, thinking and responding. The recorded conversations cover a variety of topics, such as food, movies or sports. We collected 195 different sentences spoken by those 14 students in total. 3 American English native speakers who have some teaching experience were recruited to annotate all dialogue data with 1–5 scale phonetic segmental score based on the ERJ database assessment criteria presented in Table 1.

3.2. Conditions of DNN-based acoustic modeling

Dramatic progress has been achieved in ASR over the last few years, largely due to the advances in deep learning and large databases (Hinton et al., 2012). Therefore, we applied the DNN-based acoustic models to conduct the evaluation in our system. In the process of DNN-based acoustic modeling in Kaldi toolkit (Povey et al., 2011), the training does not begin immediately with word-level transcriptions, but begins with the forced phoneme-to-audio alignments which were generated from a GMM-HMM system. Thirteen mel-frequency cepstral coefficients (MFCCs) along with the normalized energy and their first and second order derivatives are extracted in the baseline GMM-HMM system. Feature vectors are calculated every 10 ms with an overlapping analysis window of 25 ms. The recipe begins with the monophone training, then first triphone pass which comprises 2,000 regression tree leaves and 11,000 Gaussians and second triphone pass which comprises 2,500 regression tree leaves and 20,000 Gaussians are conducted. Other additional steps, frame splicing and linear discriminant analysis (LDA) transform, are performed later. Speaker adaptation is not applied to the test experiments.

As a tool for classification, basic neural networks are used, which have input nodes corresponding to the dimensions of the FBANK or MFCC features and output nodes corresponding to the states of the context-dependent triphones of the GMM-HMM system. In the step of NNET1 in Kaldi toolkit, the neural network uses the sigmoid function in hidden layers and softmax function in an output layer. Meanwhile, NNET2 uses the hyperbolic tangent function or p-norm non-linearity in hidden layers (Zhang et al., 2014).

3.3. Non-native ASR system based on the ERJ database

3.3.1. Data preparation in the ERJ database

All the sentences in the ERJ database were divided into 8 groups (S1 to S8) and the required amount of the recording in each group is about 120 sentences. Therefore, sentences in one group were read by nearly 12 male speakers and 13 female speakers, respectively. Table 3 shows the details of the sentence subsets in the ERJ database.

In order to build this non-native ASR system, only the ERJ database was used to train the acoustic models in this part. At first, all the sentences in the ERJ database were selected out and analyzed with considering the frequency of phones. We used the default phonemic symbols in the CMU pronouncing dictionary. After the repeated sentences were selected out from each sentence subset in ERJ database, each of 39 phones was counted in every modified sentence subset. Then, we checked the number of each phone in these revised sentence subsets and ensured that all the 39 phones appeared in training, development and test sets. Finally, the sets S1, S2, S4, S5, S6, and S7 were determined as the training set, S3 as the development set and S8 as the test set. Furthermore, regarding the WER of every student in ERJ database, we applied leave-one-out cross-validation (8 folds) to redefine the training set and test set for decoding based on the initial non-native ASR system. For the other two tasks of evaluating the constrained conversation speech in TMNED

Table 3
Sentence subsets in ERJ database (Minematsu et al., 2004). Totally contains eight groups (S1 to S8) and the required amount of the recording in each group is about 120 sentences.

	S1	S2	S3	S4	S5	S6	S7	S8
Number of sentences	120	123	122	123	124	123	123	122
Number of words	742	838	823	823	814	899	882	879
Speakers (male)	13	12	11	11	13	14	12	14
Speakers (female)	13	13	11	12	13	13	13	14

corpus and the free conversation speech in TEC corpus, the non-native acoustic models were trained with the whole speech of sentences in ERJ database.

3.3.2. Lexicon and language model

We modified the CMU pronouncing dictionary as the baseline pronunciation lexicon by changing all the stressed phonemes to non-stressed phonemes. One reason for this modification is that, in our CALL system, phonemic features are intended to assess non-native English speech for Japanese learners. The other reason is that Japanese speakers English pronunciation is generally different from native speakers. Therefore, if we have more data for a specific phoneme modeling, it could improve the performance of the acoustic model.

Regarding the language models in this non-native ASR system for the proficiency assessment of students in ERJ database, we selected all the sentences subsets of ERJ database and removed all repeated sentences. We developed bigram and trigram language models from these 980 different sentences. These language models are database-closed on the decision that this system is able to figure out any sentence in ERJ database as much as possible. However, the design of this language model is exclusive for ERJ database while leading to poor performance for the unknown text out of this database. Therefore, in the tasks of evaluating TMNED and TEC, we established an open trigram language model for the dialogue-based speech proficiency assessment. The open language model was trained with the content of ERJ sentences, TIMIT sentences and a variety of American English conversations from the United States Department of State[1] and the book of Conversational American English (Spears et al., 2010). We searched the language model scale among the values between 10 and 50 with a beam width of 13 to obtain the WER.

3.4. Native ASR system based on the TIMIT database

3.4.1. Data preparation in the TIMIT database

We converted all the transcriptions in TIMIT from phoneme level to word level, and meanwhile, set phonemic symbols in data preparation by using the non-stressed phonemic symbols in the CMU pronouncing dictionary. In contrast to the original separate data set as training, development and test, all of the utterances in TIMIT were used for training acoustic models in the native ASR system.

3.4.2. Language model

The TIMIT based native ASR system is the English proficiency assessment component in the CALL system. Considering the language model in this system, we decided the content including 980 sentences from ERJ and 2343 different sentences from TIMIT and established a trigram language model for evaluating each student's proficiency in ERJ database. The open language model in Section 3.3.2 was used for evaluation on TMNED and TEC.

Table 4
Settings in DNN training. Here, p-norm is the non-linearity $y = ||x||_p$, where the vector x represents a small group of inputs (Zhang et al., 2014). There is no hidden layer dimension in the p-norm networks, instead there are two parameters: p-norm input dimension/p-norm output dimension. The p-norm E.L. stands for the ensemble p-norm non-linearity, where the ensemble size is 4.

	NN11	NN21	NN22	NN23
Model type	NNET1	NNET2	NNET2	NNET2
Input feature	40FBANK	40MFCCs	40MFCCs	40MFCCs
Hidden type	sigmoid	tanh	p-norm	p-norm E.L.
Input nodes	440	360	360	360
Hidden layer	4	4	4	4
Hidden dimensions	1024	1024	1000/200	1000/200
Output nodes	3168	1551	1551	1551

Table 5
Experimental results in the development set and test set. Six kinds of acoustic models are used for decoding. The acoustic model with non-stressed phonemes performs better than that with stressed phonemes. NN23 in NNET2 series models with non-stressed phoneme training has the lowest WER in the test set as 1.97%.

	Monophone	Triphone	NN11	NN21	NN22	NN23
(non-stressed)						
DEV-WER[%]	7.09	7.83	2.99	4.99	3.13	2.81
TEST-WER[%]	6.90	5.11	2.05	3.42	2.31	1.97
(stressed)						
DEV-WER[%]	8.39	8.77	*	5.38	3.39	2.93
TEST-WER[%]	7.59	5.34	*	3.55	2.29	2.08

4. Experiments of non-native speech recognition within ERJ database

This section is an overview of our previous work using ERJ database for non-native acoustic modeling (Fu et al., 2018).

HMM-based monophone and triphone acoustic models were primarily trained as the basis of DNN-based acoustic models. The details of settings in four different DNN training methods are listed in Table 4. In which, NN11 stands for the NNET1 method and the other three NNs stand for the NNET2 methods in Kaldi toolkit. We regarded NNET1 method as a trial, experiment in NNET2 methods as the main focus. Four hidden-layer neural network is enough in most of the databases (Pan et al., 2012). MFCC feature without dimension reduction was applied to NNET2, which was slightly different with FBANK feature in NNET1 (Mohamed et al., 2012; Yoshioka et al., 2014). As this paper concentrates on the emphasis of acoustic models, we present all the results only by using the same trigram language model trained from ERJ sentences.

The recognition results within ERJ database are presented in Table 5. Compared with the results of traditional GMM-HMM acoustic models, DNN-based acoustic models significantly improve the accuracy of recognition. In addition, the acoustic model with non-stressed phonemes performed better than that with stressed phonemes, which was mentioned in Section 3.3.2. The third method with non-stressed phonemes in NNET2 series models has the lowest WER in the test set as 1.97%,

[1] https://americanenglish.state.gov

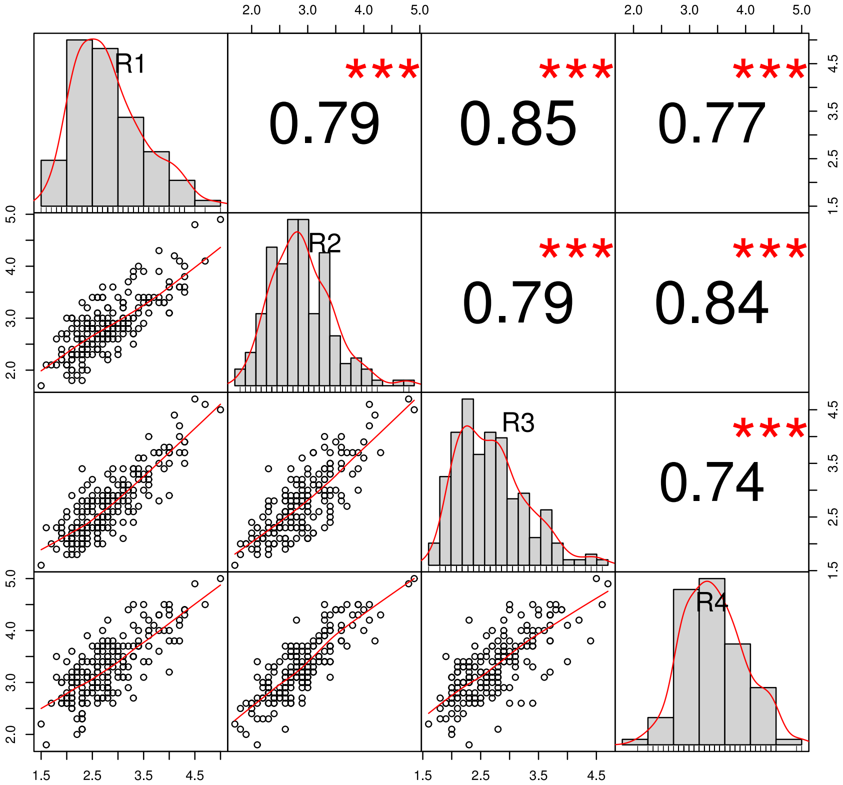

Fig. 4. The scatter plot matrix of inter-rater correlation in ERJ database. The distribution of each human-related score is displayed on the diagonal line. The two variables (one human score with the other) scatter plots are displayed with a regression line at the lower triangle, while the correlation coefficient with the level of significance as stars on the upper triangle. Each level of significance is associated with a symbol: p value (0: "***", 0.001: "**", 0.01: "*", 0.05: ".", $>$ 0.05: " "). $N = 190$.

which shows a proper potential to use the output recognized sentences as reference transcriptions.

5. Automatic assessment of English proficiency without reference transcription

5.1. Conditions in automatic assessment

Our goal is to develop a method to predict the English proficiency of a Japanese native speaker's English utterance without reference transcription. To this end, for the human scoring on the evaluation data set, we first investigated the inter-rater agreement calculated by the Pearson correlation coefficient (r). In ERJ database, one of the raters scored only male students when the other four (R1,R2,R3 and R4) scored both males and females. Therefore, we calculated the human-human agreement among the four full scoring teachers. Fig. 4 shows the inter-rater correlation matrix in ERJ database. The average correlation r among the raters is around 0.8 (N = 190). The total number of Japanese students scored by the teachers is 190 (95 males and 95 females). Additionally, we calculated the average segmental score from all of the 5 raters as the proficiency score for every Japanese student. Fig. 5 shows the histogram of proficiency scores in the ERJ database. The average value of proficiency is 2.97, and the standard deviation is 0.55. In order to explicit the relationship between the proficiency and WER obtained from native ASR system, we divided all speakers of ERJ database into three classes: LOW, MID and HIGH, where the speaker having a lower score than 2.5 were classified into LOW, those having higher score than 3.2 were HIGH, and the others were MID. The ensemble p-norm activation function in NNET2 was only selected in DNN based acoustic model, which has the same parameter with NN23 showed in Section 4. Fig. 6 shows the relationship between the WER from our established non-native ASR system and proficiency score for both GMM-HMM and DNN-HMM. These results indicate that with the proficiency score getting higher, WER of the corresponding student becomes lower.

Next, to investigate the correlation coefficients between the proficiency score and other ASR-based machine scores, we prepared various combinations of conditions. Table 6 shows the symbols to describe the condition of machine scores. A machine score is described as a combination of the training data, score type, and the acoustic model. For example, the log likelihood of GMM-HMM triphone trained from the ERJ database is denoted as "E_LL_TRI". As the calculation of RER uses acoustic models trained from both TIMIT and ERJ, in which output text

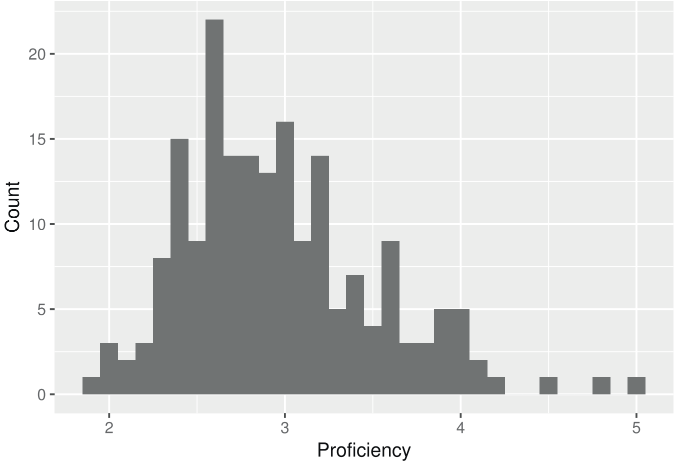

Fig. 5. Histogram of proficiency scores in ERJ database. The average value of proficiency is 2.97, and the standard deviation is 0.55. $N = 190$.

Table 6
Components of machine score combination. A machine score is described as a combination of the training data, score type and acoustic model.

Training	Score type	Acoustic model
TIMIT (T_) ERJ (E_)	Log likelihood (LL_) WER (WER_) RER (RER_)	Monophone (MONO) Triphone (TRI) NNET2 (NN)

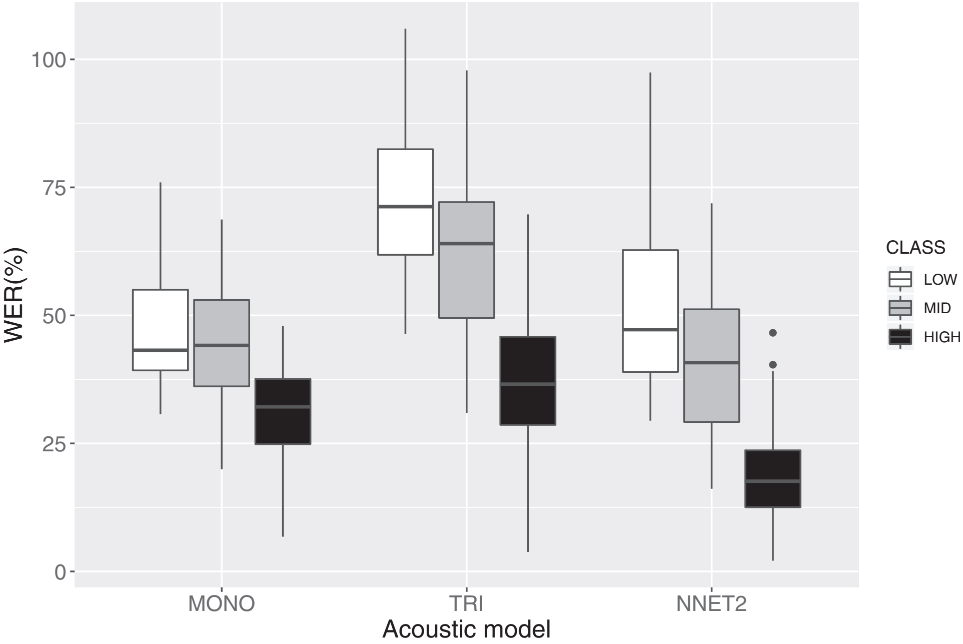

Fig. 6. Box-plots of WER from three acoustic models in the native ASR system, dependent to the proficiency score of the speakers. N = 190.

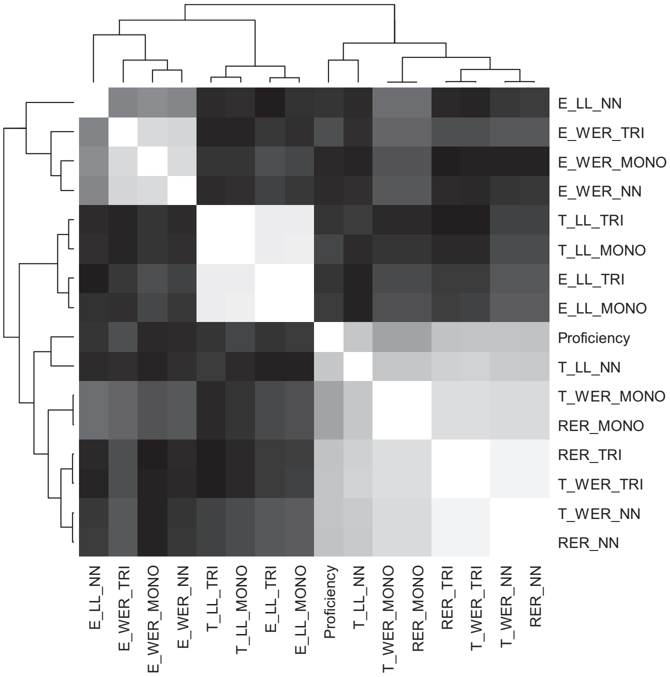

Fig. 7. Absolute correlation between all ASR-based machine scores and the proficiency score. White color means high absolute correlation and black color means low correlation.

from the DNN-based acoustic model in native ASR is regarded as the reference, RER-based scores are described without the training data part, such as “RER_MONO” or “RER_NN”.

5.2. The relationship between the proficiency and ASR-based machine scores

Fig. 7 shows the absolute correlation for all combination of machine scores and the proficiency score. This figure is a matrix of absolute correlation coefficients, and the vertical and horizontal order of scores are the same. We also show the dendrogram obtained by the hierarchical clustering conducted based on the correlation. In this figure, white color means high absolute correlation and black color means low correlation. From this figure, it is obvious that all scores are classified into three groups: the first one includes WER of ERJ-based acoustic models (E_WER_MONO, E_WER_TRI, E_WER_NN) and the log likelihood of ERJ-based NN acoustic model (E_LL_NN). The second one includes log likelihood of GMM-HMM trained from both TIMIT and ERJ (T_LL_TRI, T_LL_MONO, E_LL_TRI, E_LL_MONO). The third group includes all other scores and the proficiency score.

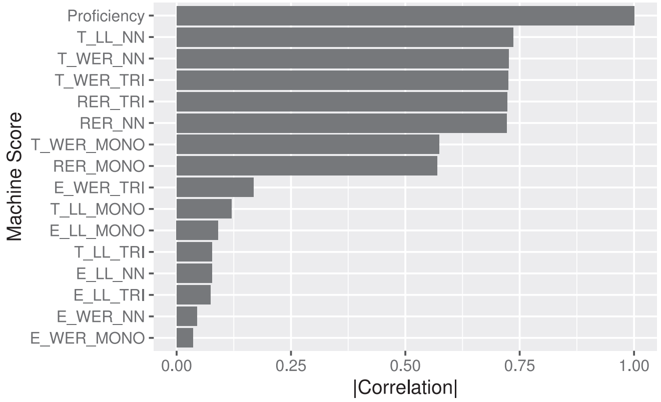

Fig. 8. Absolute correlation between the proficiency score and machine scores on ERJ database.

As presented in Fig. 7, it is obvious that the log-likelihood scores of GMM-HMM have a strong correlation with each other regardless of its training data. However, the log likelihood of DNN-HMM shows a different tendency. The log likelihood of E_LL_NN and T_LL_NN have a very low correlation ($r = -0.041, p = n.s$), which suggests that the DNN-HMM trained from different training data captures the distribution of speech differently.

Fig. 8 shows the absolute correlation between the proficiency score and all machine scores. T_LL_NN has the highest correlation ($r = 0.736, p < 0.001$), and T_WER_NN, T_WER_TRI, RER_TRI, RER_NN have almost the same values (0.721 to 0.726). The difference between

Table 7
Correlation coefficients by combining multiple machine scores. This table only shows the nine high correlation when one to nine scores were combined.

T_LL			E_LL			RER			Correlation
MONO	TRI	NN	MONO	TRI	NN	MONO	TRI	NN	
		✓							0.736, $p < 0.001$
		✓						✓	0.778, $p < 0.001$
		✓				✓		✓	0.791, $p < 0.001$
✓		✓		✓			✓		0.808, $p < 0.001$
✓	✓	✓	✓					✓	0.815, $p < 0.001$
✓	✓	✓	✓			✓		✓	0.819, $p < 0.001$
✓	✓	✓	✓		✓	✓	✓		0.820, $p < 0.001$
✓	✓	✓	✓		✓	✓	✓	✓	0.823, $p < 0.001$
✓	✓	✓	✓	✓	✓	✓	✓	✓	0.826, $p < 0.001$

Table 8
Summary of the test of the difference between two dependent correlations with one variable in common. S1 is T_LL_NN, S2 is the combination of four scores (T_LL_MONO, T_LL_TRI, T_LL_NN and E_LL_TRI) and S3 is the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). P stands for the proficiency score. $N = 190$.

Scores	Correlation	Test for difference in two correlations
S1	r(S1-P) = 0.73	r(S1-P) *vs.* r(S2-P): $z = -1.96, p = 0.049$
S2	r(S2-P) = 0.78	r(S2-P) *vs.* r(S3-P): $z = -2.93, p = 0.003$
S3	r(S3-P) = 0.82	r(S3-P) *vs.* r(S1-P): $z = -2.03, p = 0.042$

T_WER_NN and RER_NN is really small, which means the reference transcriptions are nearly the same with the output recognized text from the non-native ASR by using DNN-based acoustic model. Note that WER or RER has a negative correlation to the proficiency score because low proficiency utterance leads to larger WER.

Next, we analyzed the effect of combining multiple scores. Considering the purpose that aims to develop a method to predict a learners proficiency without a reference transcription, we excluded the WER-based scores. As a result, we used nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN). Then we examined all combination of these scores (511 combinations in total). Linear regression was used for combining those scores. Two-fold cross-validation was used for evaluation, where the data was split into even-numbered and odd-numbered samples, and then the regression coefficients were calculated from one sample set, and the correlation coefficient was calculated using the other set. Table 7 shows the result. To summarize the result, we only show the result with the highest correlation when one to nine machine scores were combined.

We conducted the test of the difference between two dependent correlations with one variable in common (Meng et al., 1992), this test was calculated based on the two-fold cross-validation basis. The summary of this test is presented in Table 8. Here, we selected three machine scores for comparison. This test shows there are significant differences among them. As shown in Fig. 8, T_LL_NN has the highest correlation as a single score. The correlation coefficient exceeds 0.8 when three scores were combined (T_LL_NN, RER_MONO, and RER_NN). The highest correlation was obtained when all machine scores were used, where the correlation coefficient was 0.826, $p < 0.001$. When we did not use RER-based scores, the best combination was T_LL_MONO + T_LL_TRI + T_LL_NN + E_LL_TRI, where the correlation coefficient was 0.786, $p < 0.001$. These results show the importance of RER for estimating proficiency. Fig. 9 shows a scatter plot of the proficiency score and the prediction score when all scores are used. This result shows that the prediction score has a good correlation with the proficiency score. The standard error was 0.31.

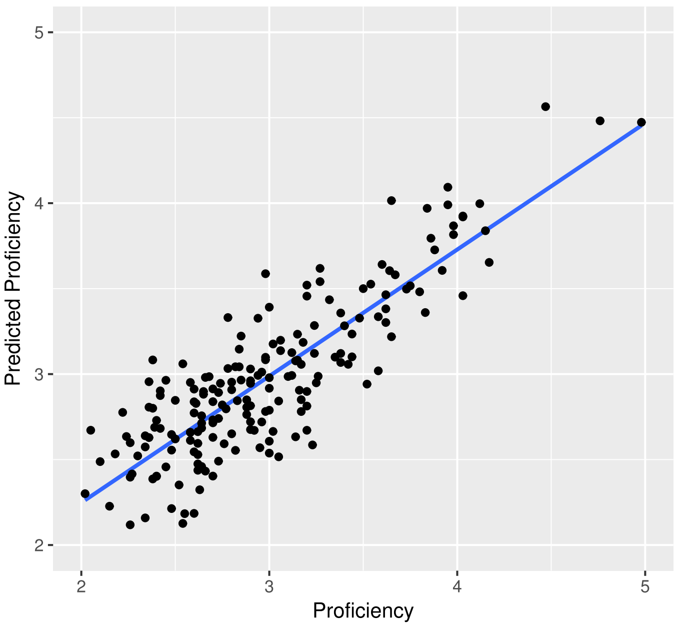

Fig. 9. Scatter plot of the proficiency and predicted proficiency on ERJ database. $N = 190$.

5.3. Automatic assessment on TMNED corpus

In order to examine the effectiveness of our proposed method on real constrained dialogue task, we conducted the experiment on the evaluation of TMNED. By using the same way in ERJ database, the average segmental score among the three human raters (H1, H2 and H3) in TMNED was calculated as proficiency score. The difference on evaluation from ERJ database is that the test unit is each of the dialogue here, while ERJ database is each of the students. The design of the test unit is based on the reason that TMNED is rather small in terms of only 13 students, and unlike reading task, the performance of the restricted conversation speech is always deviated by the different scenes. Fig. 10 shows the histogram of proficiency in TMNED. The average value of proficiency is 2.82, and the standard deviation is 0.68. Table 9 summarize the inter-rater correlation on TMNED. The correlation among the three raters is around 0.5, which means low agreement among them. We also calculated other combination scores for analysis. For example, H1H2 means the average score from human rater 1 and human rater 2, the meaning is the same with H2H3 and H1H3. The correlation coefficients (r(H1-H2H3), r(H2-H1H3) and r(H3-H1H2)) range from 0.55 to 0.58, which is higher than that of H-H. The correlation between the two-rater average score to the individual rater score (calculated in the two-rater average score) has no meaning. Each human rater has a high agreement with proficiency.

Table 9
Inter-rater correlation on TMNED corpus. Proficiency is the average segmental score among three human raters. H1H2 means the average score from human rater 1 and human rater 2. H2H3 means the average score from human rater 2 and human rater 3. H1H3 means the average score from human rater 1 and human rater 3. $N = 26$.

	H1	H2	H3	H1H2	H2H3	H1H3	Proficiency
H1	*	0.50, $p = 0.009$	0.47, $p = 0.016$	0.81, $p < 0.001$	0.56, $p = 0.003$	0.87, $p < 0.001$	0.78, $p < 0.001$
H2	0.50, $p = 0.009$	*	0.49, $p = 0.010$	0.91, $p < 0.001$	0.92, $p < 0.001$	0.58, $p = 0.002$	0.87, $p < 0.001$
H3	0.47, $p = 0.016$	0.49, $p = 0.010$	*	0.55, $p = 0.003$	0.80, $p < 0.001$	0.84, $p < 0.001$	0.77, $p < 0.001$

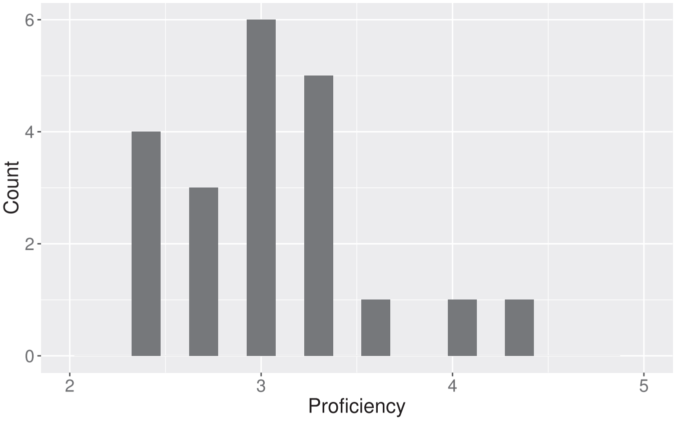

Fig. 10. Histogram of proficiency scores in TMNED corpus. The average value of proficiency is 2.82, and the standard deviation is 0.68. $N = 26$.

Table 10
Average value of WER and RER from the ASR systems by both GMM-HMM and DNN-HMM on TMNED.

	MONO	TRI	NN
WER1[%](Non-native ASR)	25.03	16.18	16.72
WER2[%](Non-native ASR)	26.60	17.45	17.72
WER1[%](native ASR)	53.70	52.20	50.96
WER2[%](native ASR)	54.82	53.34	52.88
RER[%]	54.94	55.56	55.39

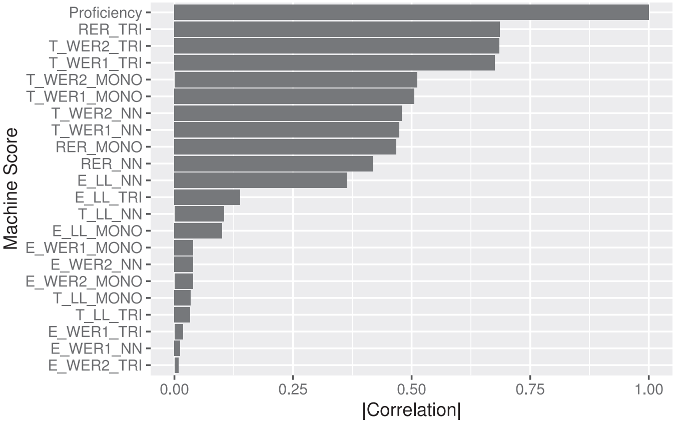

Fig. 11. Absolute correlation between the proficiency score and machine scores on TMNED corpus.

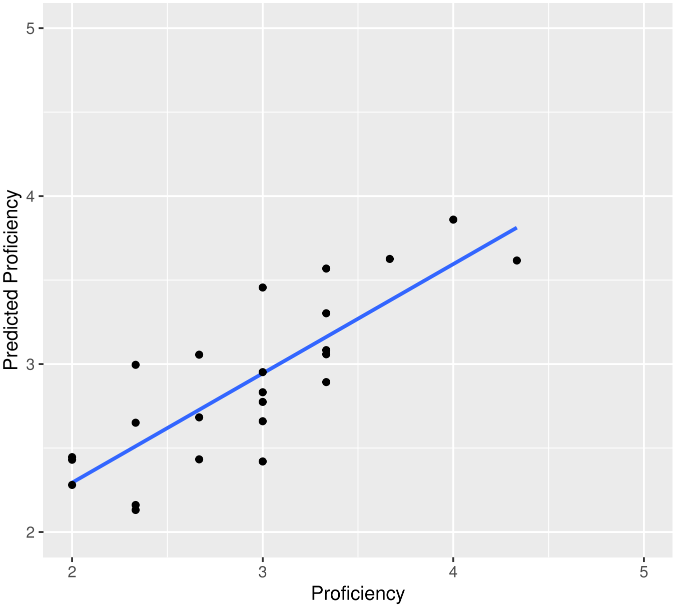

Fig. 12. Scatter plot of the proficiency and predicted proficiency on TMNED corpus. $N = 26$.

The speech from the student in each of the 26 dialogues was recognized by our non-native and native ASR systems. Since each speech in TMNED is essentially produced by memory, the human transcribed text would have more or less different places from the original reference text. Therefore, we calculated WER1 based on the comparison between the ASR output and the human transcribed reference text, and WER2 based on the comparison between the ASR output and the prompt reference text. Table 10 shows the average value of WER and RER from the output of these two ASR systems. Compared with the results of WER2, there is almost 1% improvement in WER1. Fig. 11 shows the absolute correlation between the proficiency and ASR-based machine scores for the evaluation on TMNED. Unlike the correlation results in ERJ database, this time the WER and RER scores have more agreement with proficiency than log-likelihood scores. It is positive to see that RER_TRI has the highest correlation ($r = 0.692, p < 0.001$) among all the machine scores. E_LL_NN has the highest value ($r = 0.363, p = 0.067$) among log-likelihood scores. Finally, we used nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN) for combination as the prediction score, excluding the WER-based scores. Fig. 12 shows a scatter plot of the proficiency score and the prediction score for the evaluation on TMNED. The correlation between the prediction score and the proficiency score is $0.803, p < 0.001, N = 26$. We checked the test for the difference between human-human correlation and machine-human correlation. Three different tests are set as presented in Table 11. The correlation between the prediction score and average score from two raters ranges from 0.735 to 0.798, which is significantly larger than the inter-rater correlation.

5.4. Automatic assessment on TEC corpus

In this section, TEC corpus was used for the spontaneous non-native English speech evaluation. The experimental conditions and evaluation steps are kept the same as those in the previous Section 5.3.

The proficiency score is defined as the average segmental score among the three American English native raters (A1, A2 and A3) in TEC corpus. The test unit is the speech from one student in each dialogue.

Table 11
Test for the difference between human-human correlation vs. machine-human correlation on TMNED corpus. H1 is human rater 1. H2 is human rater 12. H3 is human rater 3. H1H2 means the average score from H1 and H2. H2H3 means the average score from H2 and H3. H1H3 means the average score from H1 and H3. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). r(PS-H2H3) = 0.786, $p < 0.001$. r(PS-H1H3) = 0.798, $p < 0.001$. r(PS-H1H2) = 0.735, $p < 0.001$. $N = 26$.

Test for difference between r(human-human) and r(machine-human)			
r(H1-H2H3)	*vs.*	r(PS-H2H3):	$z = -1.89, p = 0.063$
r(H2-H1H3)	*vs.*	r(PS-H1H3):	$z = -1.85, p = 0.064$
r(H3-H1H2)	*vs.*	r(PS-H1H2):	$z = -1.67, p = 0.094$

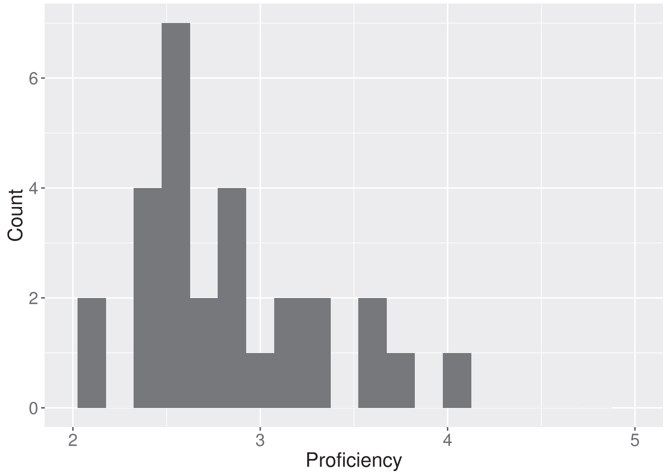

Fig. 13. Histogram of proficiency scores in TEC corpus. The average value of proficiency is 2.84, and the standard deviation is 0.48. $N = 28$.

Table 13
Average value of RER from the ASR systems by both GMM-HMM and DNN-HMM on TEC.

	MONO	TRI	NN
RER[%]	49.01	55.92	40.24

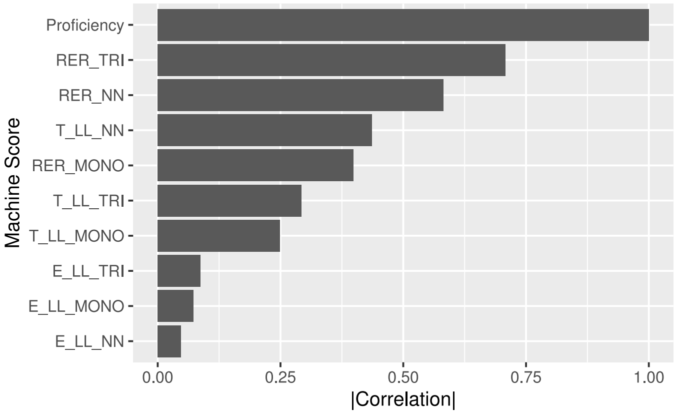

Fig. 14. Absolute correlation between the proficiency score and machine scores on TEC corpus.

Table 14
Test for the difference between human-human correlation vs. machine-human correlation on TEC corpus. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). r(PS-A2A3) = 0.768, $p < 0.001$. r(PS-A1A3) = 0.756, $p < 0.001$. r(PS-A1A2) = 0.775, $p < 0.001$. $N = 28$.

Test for difference between r(human-human) and r(machine-human)			
r(A1-A2A3)	*vs.*	r(PS-A2A3):	$z = -0.95, p = 0.340$
r(A2-A1A3)	*vs.*	r(PS-A1A3):	$z = -0.69, p = 0.491$
r(A3-A1A2)	*vs.*	r(PS-A1A2):	$z = -0.98, p = 0.326$

Fig. 13 shows the histogram of proficiency in TEC. The average value of proficiency is 2.84, and the standard deviation is 0.48. The inter-rater correlation on TEC is presented in Table 12. The correlation between every two of the three raters is around 0.6, which is a little higher than that in TMNED. The correlation coefficients (r(A1-A2A3), r(A2-A1A3) and r(A3-A1A2)) range from 0.68 to 0.69, which shows the same tendency in TMNED.

There is no gold standard reference sentence in TEC corpus, this time, we directly calculated the RER score of each test unit from the outputs of non-native and native ASR systems. Table 13 shows the average value of RER in TEC. Compared with the results in Table 10, the average value of RER in TEC with the DNN-based acoustic model is rather lower than that in TMNED. Fig. 14 shows the absolute correlation between the proficiency and ASR-based machine scores for the evaluation of TEC. RER_TRI has the highest correlation ($r = 0.708, p < 0.001$) among all the machine scores, and T_LL_NN has the highest value ($r = 0.436, p = 0.020$) among the log-likelihood scores. Nine machine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN) are combined as the predicted proficiency score. Fig. 15 shows a scatter plot of the proficiency score and the prediction score for the evaluation on TEC. The correlation between the prediction score and the proficiency score is $0.799, p < 0.001, N = 28$. We did the same test for the difference between human-human correlation and machine-human correlation as Section 5.3. Three different tests are set as presented in Table 14. The correlation between the prediction score and the average score from two raters ranges from 0.756 to 0.775. Compared with the results in Table 11, the difference is less significant. From these results, it is indicated that the proposed RER is useful and acceptable for automatic evaluation on the spontaneous non-native English conversation speech.

Table 12
Inter-rater correlation on TEC corpus. Proficiency is the average segmental score among three human raters. A1A2 means the average score from human rater 1 and human rater 2. A2A3 means the average score from human rater 2 and human rater 3. A1A3 means the average score from human rater 1 and human rater 3. $N = 28$.

	A1	A2	A3	A1A2	A2A3	A1A3	Proficiency
A1	*	0.62, $p < 0.001$	0.60, $p < 0.001$	0.90, $p < 0.001$	0.68, $p < 0.001$	0.89, $p < 0.001$	0.86, $p < 0.001$
A2	0.62, $p < 0.001$	*	0.62, $p < 0.001$	0.90, $p < 0.001$	0.89, $p < 0.001$	0.69, $p < 0.001$	0.86, $p < 0.001$
A3	0.60, $p < 0.001$	0.62, $p < 0.001$	*	0.68, $p < 0.001$	0.91, $p < 0.001$	0.90, $p < 0.001$	0.87, $p < 0.001$

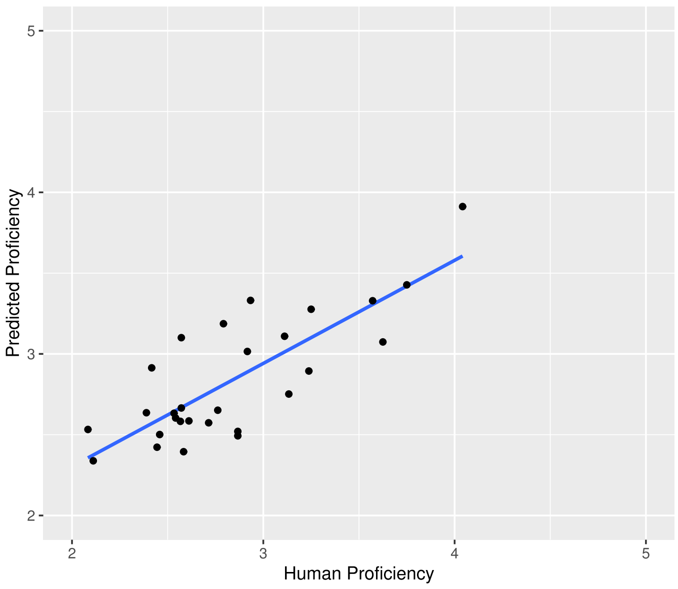

Fig. 15. Scatter plot of the proficiency and predicted proficiency on TEC corpus. $N = 28$.

6. Discussion and conclusion

This study performed an automatic proficiency evaluation system for English utterances from Japanese native speakers by utilizing both native and non-native acoustic models. To this end, we proposed the RER as a new machine score for the regression-based prediction of proficiency, based on the two assumed conditions:

(a) non-native ASR has sufficiently high recognition accuracy
(b) WER from native ASR is effective to evaluate language proficiency

Our training data include non-native speech as English read by Japanese and native American English speech; therefore, they involve much phonemic category in both L1 and L2 speakers. We applied the recent state-of-the-art DNN-based training methods for the native and non-native acoustic models. As for the test sets in this study, we prepared three different kinds of speech for evaluation: reading aloud speech, constrained interactive dialogue speech and spontaneous English conversation speech.

To evaluate the correlation between the proficiency scores and the proposed RER scores, we firstly analyzed the inter-rater correlation on ERJ database, TMNED corpus and TEC corpus. Results show the raters have very high correlation among themselves in ERJ database ($r \approx 0.8$), while not a good correlation in TMNED ($r \approx 0.5$) or TEC ($r \approx 0.6$). The speech data size of TMNED or TEC is rather smaller than that of ERJ database. Especially in TMNED corpus, the raters gave scores (three different kinds: speech-related, emotion-related and gesture-related) by directly watching the recording video. Therefore, the segmental score analyzed in our work partly influenced by other aspects from the learner in the video.

The results of the evaluation on ERJ database show the strong effectiveness of RER for language proficiency assessment. RER derived from DNN-based acoustic models already has a high correlation coefficient with the proficiency score as 0.721. On the evaluation of TEC corpus, one RER-based machine score (RER_TRI) reaches a higher correlation to proficiency than WER-based machine scores. We conducted the linear regression prediction of proficiency with combined machine scores. In all experiments, when all the machine scores (excluding WER-based scores) were combined, the correlation coefficient reached to extremely high with the proficiency score (ERJ database: $r = 0.826, p < 0.001, N = 190$; TMNED corpus: $r = 0.803, p < 0.001, N = 26$; TEC corpus: $r = 0.799, p < 0.001, N = 28$), which means this prediction method could improve human-to-machine correlation significantly.

It is common to use both non-native and native acoustic models in an automatic assessment language proficiency system, some related works are listed in Section 1. To obtain the RER, it costs double efforts to establish two ASR systems, compared with only one ASR. It also needs multi-models to decode the speech several times in terms of the RER's calculation. For the automatic assessment on non-native speech without knowing reference sentences, RER shows its superiority, especially on the experiments of TMNED and TEC, where the log-likelihood scores are rather poorly related with proficiency scores. To a CALL system, speech recognition and automatic pronunciation assessment for L2 learners are vital tasks. The former needs L2 acoustic models and the latter needs L1 acoustic models. Hence, RER obviously bridges the gap between the two different things. The RER demonstrated a farsighted meaning to the evaluation of spontaneous non-native speech in the dialogue-based CALL system.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Number JP17H00823.

References

Bernstein, J., De Jong, J., Pisoni, D., Townshend, B., 2000. Two experiments on automatic scoring of spoken language proficiency. In: Proc. InSTIL, pp. 57–61.
Cucchiarini, C., Strik, H., Boves, L., 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. Speech. Commun. 30 (2–3), 109–119.
Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners fluency: comparisons between read and spontaneous speech. J. Acoust. Soc. Am. 111 (6), 2862–2873.
Franco, H., Neumeyer, L., Digalakis, V., Ronen, O., 2000. Combination of machine scores for automatic grading of pronunciation quality. Speech Commun. 30 (2–3), 121–130.
Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for language instruction. In: Proc. ICASSP, 2, pp. 1471–1474.
Franco, H., Neumeyer, L., Ramos, M., Bratt, H., 1999. Automatic detection of phone-level mispronunciation for language learning. In: Proc. Eurospeech, pp. 851–854.
Fu, J., Chiba, Y., Nose, T., Ito, A., 2018. Evaluation of English speech recognition for Japanese learners using DNN-based acoustic models. In: Pan, J.-S., Ito, A., Tsai, P.-W. and Jain, L. C. (Eds.) Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing. Springer, pp. 93–100.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. DARPA TIMIT Acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report 93.
Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 18 (5–6), 602–610.
Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: A search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28 (10), 2222–2232.
Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.
Hönig, F., Batliner, A., Nöth, E., 2012. Automatic assessment of non-native prosody annotation, modelling and evaluation. In: Proc. IS ADEPT, pp. 21–30.
Hu, W., Qian, Y., Soong, F.K., 2013. A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL). In: Proc. Interspeech, pp. 1886–1890.
Hu, W., Qian, Y., Soong, F.K., Wang, Y., 2015. Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Commun. 67, 154–166.
Ito, A., Lim, Y.-L., Suzuki, M., Makino, S., 2007. Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree. Acoust. Sci. Technol. 28 (2), 131–133.
Ito, A., Tsutsui, R., Makino, S., Suzuki, M., 2008. Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system. In: Proc. Interspeech, pp. 2819–2822.
Kawai, G., Hirose, K., 1998. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. In: Proc. ICSLP, pp. 1823–1826.
Kim, Y., Franco, H., Neumeyer, L., 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proc. Eurospeech, pp. 649–652.
Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady, 10, pp. 707–710.
Li, K., Mao, S., Li, X., Wu, Z., Meng, H., 2018. Automatic lexical stress and pitch accent detection for L2 english speech using multi-distribution deep neural networks. Speech Commun. 96, 28–36.

Li, W., Siniscalchi, S.M., Chen, N.F., Lee, C.-H., 2016. Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling. In: Proc. ICASSP, pp. 6135–6139.

Luo, D., Qiao, Y., Minematsu, N., Yamauchi, Y., Hirose, K., 2010. Regularized-MLLR speaker adaptation for computer-assisted language learning system. In: Proc. Interspeech, pp. 594–597.

Makino, T., Aoki, R., 2012. English read by Japanese phonetic corpus: an interim report. Res. Lang. 10 (1), 79–95.

Mao, S., Li, X., Li, K., Wu, Z., Liu, X., Meng, H., 2018. Unsupervised discovery of an extended phoneme set in l2 English speech for mispronunciation detection and diagnosis. In: Proc. ICASSP, pp. 6244–6248.

Meng, X.-L., Rosenthal, R., Rubin, D.B., 1992. Comparing correlated correlation coefficients. Psychol. Bull. 111 (1), 172.

Minematsu, N., Kurata, G., Hirose, K., 2002. Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition. In: Proc. ICSLP, pp. 529–531.

Minematsu, N., Okabe, K., Ogaki, K., Hirose, K., 2011. Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database. In: Proc. Interspeech, pp. 1481–1484.

Minematsu, N., Tomiyama, Y., Yoshimoto, K., Shimizu, K., Nakagawa, S., Dantsuji, M., Makino, S., 2004. Development of English speech database read by Japanese to support CALL research. In: Proc. Int. Cong. Acoust., 1, pp. 557–560.

Mohamed, A.-r., Hinton, G., Penn, G., 2012. Understanding how deep belief networks perform acoustic modelling. In: Proc. ICASSP, pp. 4273–4276.

Molina, C., Yoma, N.B., Wuth, J., Vivanco, H., 2009. Asr based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech Commun. 51 (6), 485–498.

Moustroufas, N., Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. Comput. Speech Lang. 21 (1), 219–230.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M., 2000. Automatic scoring of pronunciation quality. Speech Commun. 30 (2–3), 83–93.

Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A., Makino, S., 2009. A speaker adaptation method for non-native speech using learners native utterances for computer-assisted language learning systems. Speech Commun. 51 (10), 875–882.

Oshima, Y., Takamichi, S., Toda, T., Neubig, G., Sakti, S., Nakamura, S., 2016. Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic characteristics. IEICE Trans. Inf. Syst. 99 (12), 3132–3139.

Pan, J., Liu, C., Wang, Z., Hu, Y., Jiang, H., 2012. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling. In: Proc. ISCSLP, pp. 301–305.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit. In: Proc. ASRU. IEEE Signal Processing Society.

Spears, R.A., Birner, B.J., Kleinedler, S.R., Nisset, L., 2010. Conversational American English. McGraw-Hill Education.

Suzuki, M., Konno, T., Ito, A., Makino, S., 2008. Automatic evaluation system of english prosody based on word importance factor. J. Syst. Cybern. Inf. 6 (4), 83–90.

Tao, J., Evanini, K., Wang, X., 2014. The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system. In: Proc. Spoken Language Technology Workshop (SLT). IEEE, pp. 294–299.

Truong, Q.-T., Kato, T., Yamamoto, S., 2018. Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours. In: Proc. Interspeech, pp. 2186–2190.

Tu, M., Grabek, A., Liss, J., Berisha, V., 2018. Investigating the role of L1 in automatic pronunciation evaluation of L2 speech. In: Proc. Interspeech, pp. 1636–1640.

Wang, X., Kato, T., Yamamoto, S., 2017. Phoneme set design based on integrated acoustic and linguistic features for second language speech recognition. IEICE Trans. Inf. Syst. 100 (4), 857–864.

de Wet, F., Van der Walt, C., Niesler, T., 2009. Automatic assessment of oral language proficiency and listening comprehension. Speech Commun. 51 (10), 864–874.

Witt, S., Young, S., 1997. Computer-assisted pronunciation teaching based on automatic speech recognition. Language Teaching and Language Technology Groningen, The Netherlands, 25–35.

Witt, S., Young, S.J., 1997. Language learning based on non-native speech recognition. In: Proc. Eurospeech, pp. 633–636.

Witt, S.M., 2012. Automatic error detection in pronunciation training: Where we are and where we need to go. In: Proc. IS ADEPT, pp. 1–8.

Witt, S.M., Young, S.J., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Commun. 30 (2–3), 95–108.

Wu, H., Chiba, Y., Nose, T., Ito, A., 2018. Analyzing effect of physical expression on English proficiency for multimodal computer-assisted language learning. In: Proc. Interspeech, pp. 1746–1750.

Yoshioka, T., Chen, X., Gales, M.J., 2014. Impact of single-microphone dereverberation on DNN-based meeting transcription systems. In: Proc. ICASSP, pp. 5527–5531.

Yue, J., Shiozawa, F., Toyama, S., Yamauchi, Y., Ito, K., Saito, D., Minematsu, N., 2017. Automatic scoring of shadowing speech based on DNN posteriors and their DTW. In: Proc. Interspeech, pp. 1422–1426.

Zechner, K., Higgins, D., Xi, X., Williamson, D.M., 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Commun. 51 (10), 883–895.

Zhang, X., Trmal, J., Povey, D., Khudanpur, S., 2014. Improving deep neural network acoustic models using generalized maxout networks. In: Proc. ICASSP, pp. 215–219.