Received: 22 August 2019 | Revised: 18 November 2019 | Accepted: 23 December 2019

DOI: 10.4218/etrij.2019-0400

ORIGINAL ARTICLE

ETRI Journal WILEY

Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition

Yoo Rhee Oh | Kiyoung Park | Hyung-Bae Jeon | Jeon Gue Park

Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea

Correspondence
Yoo Rhee Oh, Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea.
Email: yroh@etri.re.kr

Funding information
This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2019-0-00004, Development of semi-supervised learning language intelligence technology and Korean tutoring service for foreigners).

This paper presents an automatic proficiency assessment method for a non-native Korean read utterance using bidirectional long short–term memory (BLSTM)–based acoustic models (AMs) and speech data augmentation techniques. Specifically, the proposed method considers two scenarios, with and without prompted text. The proposed method with the prompted text performs (a) a speech feature extraction step, (b) a forced-alignment step using a native AM and non-native AM, and (c) a linear regression–based proficiency scoring step for the five proficiency scores. Meanwhile, the proposed method without the prompted text additionally performs Korean speech recognition and a subword un-segmentation for the missing text. The experimental results indicate that the proposed method with prompted text improves the performance for all scores when compared to a method employing conventional AMs. In addition, the proposed method without the prompted text has a fluency score performance comparable to that of the method with prompted text.

KEYWORDS

automatic speech recognition (ASR) for a non-native Korean utterance, bidirectional long short–term memory (BLSTM)–based acoustic models (AMs), speech data augmentation, spoken computer-assisted language learning (CALL), spoken proficiency assessment

1 | INTRODUCTION

Owing to the rising demand for second-language learning and the advances in machine learning, there has been increase in the need for spoken computer-assisted language learning (CALL) applications [1,2]. Moreover, with the spread of Korean popular culture overseas [3], the need for Korean language learning has prompted the development of such CALL applications for non-native Korean learners. Among the spoken Korean CALL applications, this paper focuses on an automatic speech recognition (ASR)–based proficiency assessment for non-native Korean speech.

Non-native speech significantly degrades the performance of the ASR used in a spoken CALL owing to the pronunciation variabilities in non-native speech [4,5]. Consequently, numerous research results have been reported on automatic proficiency assessment methods for non-native speech that is read aloud [6–13] and for spontaneous speech [14–17]. However, there has been limited research on proficiency assessment of non-native Korean speech [18]. Moreover, most research has been focused on the analysis of pronunciation variabilities in non-native Korean speech. For instance, [19,20] analyzes the pronunciation variabilities of Korean spoken by Japanese and Chinese learners using contrastive and corpus-based statistical approaches. In addition, [21,22] examine the correlation of the proficiency and analytic scores of non-native Korean speech in a read and a spontaneously spoken corpus. Fewer studies have been conducted on automatic proficiency assessment of non-native Korean speech. That is, [20] presents an ASR system for Korean utterances by Chinese speakers by modeling a non-native pronunciation dictionary and [23] examines the performance of an oracle pronunciation assessment of Korean utterances by Chinese speakers by using manual transcription. Moreover, [24] proposes a pronunciation assessment method for non-native Korean utterance using a selection of pronunciation scoring features. In addition, [25] proposes a pronunciation training method for non-native speakers by using self-imitating feedback based on generative adversarial network (GAN).

This study aims to develop an automatic spoken proficiency assessment system for Korean utterances read aloud by non-native speakers. To this end, we present two spoken proficiency assessment methods: (a) a method with prompted text and (b) a method without prompted text. The proposed method with the prompted text is designed for utterances where the prompted text is known because the user speaks the text provided by a CALL system. The proposed method comprises (i) a speech feature extraction step, (ii) a forced-alignment step, and (iii) a linear regression-based proficiency scoring step. That is, the speech features are extracted from the input speech, and then the speech features and the prompted text are decoded into two time-aligned sequences by forced-alignment using the bidirectional long short–term memory (BLSTM)–based native acoustic model (AM) and the BLSTM-based non-native AM, respectively. The native AM is trained with a native speech corpus, whereas the non-native AM is trained with a non-native speech corpus. The latter's data are augmented using a speed perturbation method [26] and a frequency and time masking method [27], as the number of non-native data is far less than that of native data. Next, the two time-aligned sequences are converted into proficiency scoring features and then these features are fed into five linear regression–based proficiency scoring models, which output five scores (holistic impression of proficiency, segmental accuracy, phonological accuracy, fluency, and pitch and accent).

The proposed method without the prompted text is designed for utterances when the prompted text is not known because the user reads the text from his or her textbook. To replace the missing prompted text, a Korean ASR step is additionally performed before the forced-alignment step. In the Korean ASR step, the speech features of the input speech are decoded into a subword-based text by a Viterbi decoding using a BLSTM-based AM and an *n*-gram–based language model. The AM for the Korean ASR is trained with native and non-native speech corpora and its augmented data. Next, the recognized subword-based text is converted into word-like text by performing a word un-segmentation (WUS) method [28]. Subsequently, the word-like recognized text is fed into the forced-alignment step as the input text.

The contributions of this paper are summarized as follows:

A Korean proficiency assessment system: An automatic proficiency assessment system of non-native Korean speech has not yet been thoroughly presented. Therefore, this paper presents an automatic spoken proficiency assessment system for Korean utterances read by non-native speakers.

Use of BLSTM-based AMs: For better performance, BLSTM-based AMs are employed instead of the Gaussian mixture model (GMM)–based AM or the feed forward deep neural network (DNN)–based AM. From the proficiency assessment experiments, the proposed method employing BLSTM-based AMs improves the averaged correlation coefficients across five scores by 0.052 when compared to a system employing GMM-based AM. In addition, BLSTM-based AMs improve the performance for segmental accuracy by 0.016 when compared to DNN-based AMs.

Use of speech data augmentation: For better performance, a speed perturbation method [26] and a frequency and time masking method [27] are utilized. From these ASR experiments, it is found that the use of augmentation methods improves the error rate reduction (ERR) by 9.34% and 16.14% for native speech and non-native speech, respectively.

Use of WUS: The WUS method [28] is utilized for the proposed Korean proficiency assessment method without prompted text. It matches the basic unit between the subword-based text that is decoded from the speech recognition step and the word-based text that is entered at the proficiency assessment step. From the proficiency assessment experiments, the subword-based text degrades the performance by 0.015 when compared to the word-based text for the proficiency assessment step.

Performance comparison of the proposed methods

The performance of a proficiency assessment with prompted text is compared to a proficiency assessment without prompted text, which has not been performed previously. First, the proposed methods, with and without prompted text, attain comparable performance scores for fluency. As the fluency score is solely based on speaking rates, it can be concluded that its performance is robust though prompted text is not presented. Second, the proposed method without prompted text demonstrates reduced performance for the holistic, segmental accuracy, and phonological accuracy scores. Even though the ASR-decoded text is incorrect, the ASR-decoded text could achieve a high acoustic score and this could lead to a misestimation of proficiency scores that are highly related to acoustic scores.

The organization of this paper is as follows. Section 2 describes a Korean read speech corpus for a spoken proficiency assessment and Section 3 presents the proposed proficiency assessment method with prompted text and the method without prompted text. Section 4 shows performance comparisons of the proposed proficiency assessment methods for Korean speech read by a non-native. Finally, we conclude our findings in Section 5.

2 | A KOREAN READ SPEECH CORPUS FOR A PROFICIENCY ASSESSMENT

The corpus of Korean read speech for the spoken proficiency assessment consists of 2500 utterances by 50 non-native speakers and 500 utterances by 10 native speakers. Each speaker utters 50 sentences and the speech data are recorded at a rate of 16 kHz. In particular, all utterances were carefully recorded so as not to have reading errors such as insertions or deletions. The non-native speakers were Asians from China, Japan, Cambodia, Vietnam, and the Philippines. The gender and spoken language proficiency levels were evenly distributed among the speakers. Moreover, the 50 sentences were designed for the spoken language proficiency assessment and contained various phonological rules, of which a detailed description is provided in [21].

Each non-native utterance was annotated by four human experts for the five proficiency area scores: holistic impression of proficiency (s_{holistic}), segmental accuracy (s_{segment}), phonological accuracy ($s_{\text{phonology}}$), fluency (s_{fluency}), and pitch and accent (s_{pitch}). Each score was measured on a scale of 1–5 with 1, 2, 3, 4, and 5 indicating very poor, poor, acceptable, good, and perfect, respectively. Accordingly, each utterance from the non-native corpus was annotated with 20 scores. Conversely, each native utterance was not annotated but was recorded as a 5 (top score). The detailed description of the proficiency score is found in [21].

Throughout this study, a proficiency score (s_t^i) of the i-th utterance for the score type (t) is calculated as an average as follows:

$$s_t^i = \frac{\sum_{r=1}^{4} s_t^{i,r}}{4} \tag{1}$$

where t represents either holistic, segment, phonology, or fluency and pitch. $s_t^{i,r}$ is the score of the i-th utterance for type t by the r-th human expert. Furthermore, we use a 10-fold cross-validation test method owing to the small size of the speech assessment corpus.

3 | PROPOSED SPOKEN PROFICIENCY ASSESSMENT METHODS WITH AND WITHOUT PROMPTED TEXT

Figure 1 shows the overall procedure of the proposed automatic proficiency assessment methods with and without prompted text for Korean speech read aloud by a non-native speaker. The dotted areas of the figure indicate additional components for the proposed method without prompted text.

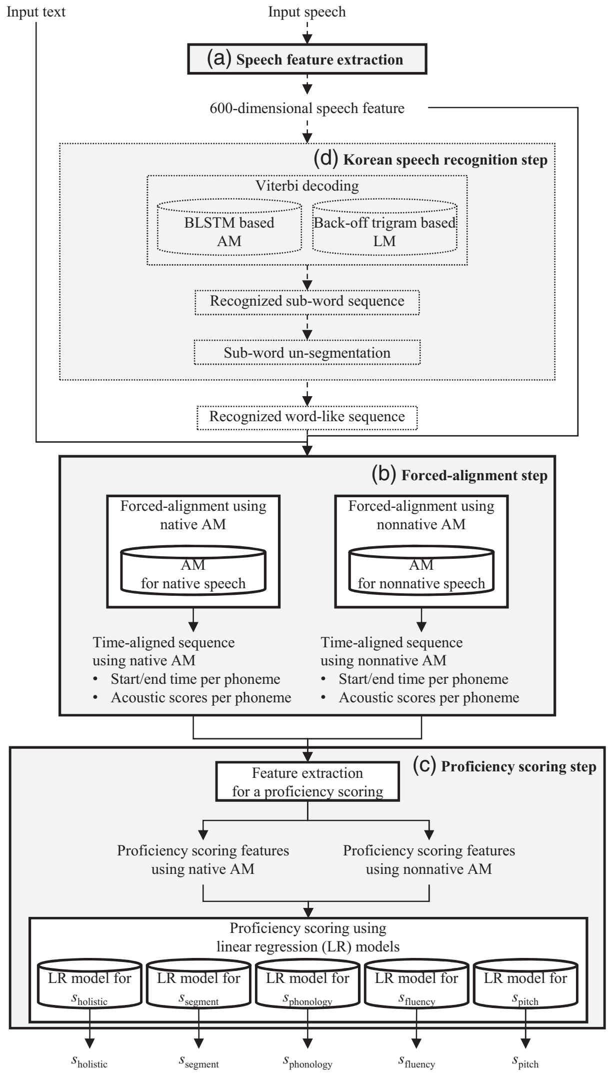
Input text
Input speech
(a) Speech feature extraction
600-dimensional speech feature
(d) Korean speech recognition step
Viterbi decoding
BLSTM based AM
Back-off trigram based LM
Recognized sub-word sequence
Sub-word un-segmentation
Recognized word-like sequence
(b) Forced-alignment step
Forced-alignment using native AM
AM for native speech
Forced-alignment using nonnative AM
AM for nonnative speech
Time-aligned sequence using native AM
• Start/end time per phoneme
• Acoustic scores per phoneme
Time-aligned sequence using nonnative AM
• Start/end time per phoneme
• Acoustic scores per phoneme
(c) Proficiency scoring step
Feature extraction for a proficiency scoring
Proficiency scoring features using native AM
Proficiency scoring features using nonnative AM
Proficiency scoring using linear regression (LR) models
LR model for s_{holistic}
LR model for s_{segment}
LR model for $s_{\text{phonology}}$
LR model for s_{fluency}
LR model for s_{pitch}
s_{holistic}
s_{segment}
$s_{\text{phonology}}$
s_{fluency}
s_{pitch}

FIGURE 1 Comprehensive procedure for the proposed automatic proficiency assessment methods with and without prompted text for Korean speech read aloud by a non-native. First, the proposed method with prompted text consists of (a) the speech feature extraction step, (b) the forced-alignment step, and (c) the proficiency scoring step. The proposed method without prompted text consists of (a) the speech feature extraction step, (d) the Korean ASR step, (b) the forced-alignment step, and (c) the proficiency scoring step. The additional components of the proposed method without prompted text are depicted using a dotted line

The proposed method with prompted text consists of (a) the speech feature extraction step, (b) the forced-alignment step, and (c) the proficiency scoring step. In the speech feature extraction step, we initially extracted the 40-dimensional log mel filterbanks for each 10 ms analysis frame of the speech input. Next, we stacked the features of the previous seven frames and the subsequent seven frames, which resulted in a 600-dimensional speech feature for each 10 ms analysis frame. In the forced-alignment step, the speech features and prompted text were decoded using a forced-alignment algorithm using the BLSTM-based native AM and the BLSTM-based non-native AM, respectively, which resulted in two time-aligned sequences. Each time-aligned sequence contained: (a) the start and end time for each word of the prompted text, (b) the start and end time for each phoneme of the prompted text, (c) the acoustic scores of the null hypothesis (H_0) [29], and (d) the acoustic scores of the alternative hypothesis (H_1). In the proficiency scoring step, the proficiency scoring features were extracted from two time-aligned sequences. Then, the five proficiency scores were calculated using the proficiency scoring features and the five linear regression–based proficiency scoring models. A proficiency score, s_{t}, was calculated with a corresponding linear regression–based proficiency scoring model, $\mathcal{R}_t$, by performing the following equation:

$$s_t = \sum_{i=1}^{\|F\|} c_t^{(i)} f^{(i)} + b_t \tag{2}$$

where t is one of holistic, segment, phonology, fluency, or pitch, and $\|F\|$ and $f^{(i)}$ indicate the number of features and the i-th feature value, respectively. In addition, $c_t^{(i)}$ and b_t indicate the i-th coefficient and the bias of $\mathcal{R}_t$.

When a user reads aloud not the prompted text of a CALL system but his or her own text, the proposed method without prompted text consisting of (a) the speech feature extraction step, (b) the Korean ASR step, (c) the forced-alignment step, and (d) the proficiency scoring step can be used. In the speech feature extraction step, we extracted a 600-dimensional speech feature for each frame, similar to the proposed method with prompted text. In the Korean ASR step, the speech features were decoded into the subword-based text by a Viterbi decoding algorithm using the BLSTM-based AM and an n-gram–based language model. Then, the subword text was converted into the word-like text utilizing a WUS method [28]. In the forced-alignment step, the speech features and the recognized word-like text were decoded into the two time-aligned sequences. This step is similar to that of the proposed method with prompted text except for the input text. In the proficiency scoring step, we calculated the five scores, similarly as for the proposed method with the prompted text. Briefly, the main differences between the proposed methods with and without the prompted text are: (a) use of the Korean ASR step and (b) the input text of the forced-alignment step.

The following subsections provide detailed descriptions of the BLSTM-based AMs used in the forced-alignment step, speech data augmentation, the proficiency scoring features and models, as well as BLSTM-based AMs and language model for the Korean ASR step.

3.1 | BLSTM-based native and non-native acoustic models of the forced-alignment step

This subsection explains the BLSTM-based native AM and the BLSTM-based non-native AM that were used in the forced-alignment step. The proposed method uses the BLSTM-based native AM to determine whether a speech input is similar to native speech and uses the BLSTM-based non-native AM to obtain more accurate time information for the prompted or recognized text.

The BLSTM-based native AM consists of one input layer, five BLSTM layers, a fully connected layer, and a soft-max layer. It was trained with 670k clean and native Korean utterances (738 hours) using a Kaldi toolkit [30]. Because any noisy or augmented speech could increase the confusion of similarity detection between speech input and native speech, no data augmentation was applied. Meanwhile, the BLSTM-based non-native AM has the same structure as the native AM and was trained with 347k non-native Korean utterances (529.4 hours) and with the augmented non-native speech data. The augmented speech data included five additional copies provided by performing speed perturbation [26] and frequency and time masking [27], which are described in Section 3.2.

3.2 | Speech data augmentation for the BLSTM-based acoustic models

This subsection explains two speech data augmentation methods that were used to train the BLSTM-based AMs, whereas data augmentation techniques are commonly utilized for acoustic modeling to increase the quantity of training data and to reduce overfitting [26].

A speed perturbation method [26] was used with speech factors of 0.9 and 1.1 by using a SoX tool [31]. In addition, a frequency and time masking method was adopted from [27] and modified as shown in Algorithm 1. As shown in the pseudocode, the main divergence from the work of [27] is that the parameters of [27] were randomly selected by using the parameters, r_F, r_T, $m_{F,\max}$, and $m_{T,\max}$, which were set empirically as 0.15, 0.15, 2, and 2, in this study. Training data have increased sixfold after speed perturbation followed by the frequency and time masking.

Algorithm 1 Pseudocode for the frequency and time masking method [27]

```
$\upsilon$: No. mel frequency channels
$\tau$: Total time steps

// Tunable parameters of [27]
$F$: A frequency mask parameter
$T$: A time mask parameter
$m_F$: No. frequency masks to be applied
$m_T$: No. time masks to be applied

// Tunable parameters of this study
$r_F$: A ratio to be consecutively masked over frequency, $\upsilon$
$r_T$: A ratio to be consecutively masked over time, $\tau$
$m_{F,\max}$: The max. no. frequency masks to be applied
$m_{T,\max}$: The max. no. time masks to be applied

// Obtain the mean value for masking
for each Mel frequency channels do
    Obtain mean value over time
end for

// Apply a frequency masking
$F$ = a random between $[1, \upsilon \times r_F]$
$m_F$ = a random between $[1, m_{F,\max}]$
for $m$=1 to $m_F$ do
    $f$ = a random between $[0, F]$
    $f_0$ = a random between $[0, \upsilon - f)$
    Set the masking value over $[f_0, f_0 + f)$
end for

// Apply a time masking
$T$ = a random between $[1, \tau \times r_T]$
$m_T$ = a random between $[1, m_{T,\max}]$
for $m$=1 to $m_T$ do
    $t$ = a random between $[0, T]$
    $t_0$ = a random between $[0, \tau - t)$
    Set the masking value over $[t_0, t_0 + t)$
end for
```

TABLE 1 The frequency or duration related features [32]

Item	Description
Numwds	No. of non-silence words
Numsil	No. of silence
Longpfreq	No. of longp
Wdpchk	Mean of (no. of words of chk)
Wdpchkmeandev	Mean deviation of (no. of words of chk)
Globsegdur	Dur. of segment
uttsegdur	Dur. of speech segment
Segdur	Dur. of speech segment without pauses
Secpchk	Mean of (dur. of chk)
Longpmn	Mean of (dur. of longp)
Secpchkmeandev	Mean deviation of (dur. of chk)
Silmeandev	Mean deviation of (dur. of silences)
Longpmeandev	Mean deviation of (dur. of longp)
Silstddev	Standard deviation of (dur. of silences)
Longpstddev	Standard deviation of (dur. of longp)
Wpsec	Numwds/Segdur
Wpsecutt	Numwds/uttsegdur
Silpwd	(Dur. of silences)/Numwds
Silmean	(Dur. of silences)/Numsil
Silpsec	(Dur. of silences)/Wpsec
Longpwd	(Dur. of longp)/Numwds
Tpsec	(Dur. of unique words)/Segdur
Tpsecutt	(Dur. of unique words)/uttsegdur

longp: Long pause, a silence with a duration longer than 0.2 second.
chk: Speech chunk, a speech segment that is interrupted by a long pause.
dur: Duration, which is measured in seconds.

TABLE 2 The syllabic features [17,33]

Item	Description
percent*X*	(Duration of *X*)/(Duration of phonemes)
mean*X*	Mean of (Duration of *X*)
stddev*X*	Standard derivation of (Duration of *X*)
varco*X*	Variance of (Duration of *X*)
*r*PVI*X*	$1/(\|\mathrm{X}\| - 1) \times \sum_{k=1}^{n-1} \|x_{k+1} - x_k\|$
*n*PVI*X*	$100/(\|\mathrm{X}\| - 1) \times \sum_{k=1}^{n-1} \|x_{k+1} - x_k/(x_{k+1} + x_k/2)\|$

X: one of {Vowel, Consonant, Syllable, Onset, Coda}.
x_k: Duration of the *k*-th *X*.
$\|\mathrm{X}\|$: No. of *X*.

3.3 | Proficiency scoring features and linear regression–based scoring models

This subsection explains the proficiency scoring features and the linear regression–based proficiency scoring models of the proficiency scoring step.

The proficiency scoring features were extracted from the two time-aligned sequences of the forced-alignment step and they were classified into frequency- or duration-related features, syllabic features, and acoustic features, as shown in Tables 1, 2, and 3, respectively. The frequency- or duration-related features were adopted from [32] and extracted using the duration and the number of each token. The syllabic features were extended from [17,33] and extracted using the duration of each vowel, consonant, syllable, syllable onset, and syllable coda. The acoustic features were extracted using the duration and the acoustic scores for each phoneme, for which features in the form of L_x were extended from [17].

TABLE 3 The acoustic features [17]

Item	Description
$\mathrm{AM}_{\mathrm{toks},H0}$	Acoustic score of tokens
$\mathrm{AM}_{\mathrm{toks},H1}$	Anti-model acoustic score of tokens
$\mathrm{AM}_{\mathrm{wds},H0}$	Acoustic score of non-silence tokens
$\mathrm{AM}_{\mathrm{sil},H0}$	Acoustic score of silence words
SLLR	Sentence-level log-likelihood ratio
$L_{1,X}$	$\sum_{i=1}^{\mathrm{Numwds}} L(x_i)$
$L_{2,X}$	L_1/Numwds
$L_{3,X}$	L_1/(No. of syllables)
$L_{3,X}^{*}$	L_1/(No. of phones)
$L_{4,X}$	L_1/uttsegdur
$L_{5,X}$	$\left(\sum_{i=1}^{\mathrm{Numwds}} L(x_i)/t_i\right)\Big/\mathrm{Numwds}$
$L_{6,X}$	$L_{4,X}$/R
$L_{7,X}$	$L_{5,X}$/R

X: one of {H_0, H_1, WLLR}.

R: uttsegdur/(No. of phones).

t_i: the duration of the i-th word.

For the linear regression–based proficiency scoring model, $\mathcal{R}_t$, we use the training set of the non-native and native speech assessment corpus. For each utterance, the 600-dimensional speech features were extracted and the two time-aligned sequences were obtained by the forced-alignment algorithm applied to the prompted text or the recognized word-like text. In addition, the target score of each utterance for s_t were prepared as the average value described in (1). Then, $\mathcal{R}_t$ was trained with the two time-aligned sequences and the target score of the training data using a toolkit *R* [34]. Accordingly, we obtained the five linear regression models for the five proficiency scores, s_{holistic}, s_{segment}, $s_{\mathrm{phonology}}$, s_{fluency}, and s_{pitch}.

3.4 | BLSTM-based AM and language model for the Korean ASR step

This subsection explains the BLSTM-based AM and the *n*-gram–based language model for the proposed method without prompted text. Unlike the BLSTM-based native AM and BLSTM-based non-native AM of Section 3.1, the BLSTM-based native and non-native AM used in the Korean ASR step was intended to include as many pronunciation variabilities as possible to recognize both native and non-native speech.

The BLSTM-based AM had the same structure as the BLSTM-based AMs for the forced-alignment step and it was trained with 3097k native Korean utterances (3439 hours), the 347k non-native Korean utterances, the 1000-hour augmented native speech data, and the 1000-hour augmented non-native speech data. The 1000-hour augmented native speech data were obtained by utilizing speed perturbation and frequency and time masking, and by performing a random sampling of the 1000-hour data. Moreover, the 1000-hour augmented non-native speech data were obtained in the same manner. It is noted that random sampling was used to supplement limited computational resources.

The *n*-gram–based language-model is a back-off trigram of 541k subwords, in which a subword is a commonly used unit of a language model for an agglutinative language such as Korean [28]. The language model was trained with approximately 30 GB of text data including newspapers, web-sourced texts, etc. The text data were first preprocessed using a text normalization method and a word segmentation (WS) method [28], and then the most frequent 541k subwords were obtained. Next, the back-off trigram for the 541k subwords was trained with the preprocessed text data using a SRILM toolkit [35,36].

4 | EXPERIMENTS

To evaluate the proposed automatic spoken Korean proficiency assessment methods, two metrics, Pearson's correlation and root mean square error (RMSE), were used for comparing the target score of (1) and the score of the proficiency assessment system. The Pearson correlation is a metric commonly used to evaluate the performance of proficiency assessment methods, while the RMSE is used to evaluate the performance of each score because the score distribution is not balanced. We also present the averaged values over the 10-fold cross-validation test for each metric.

TABLE 4 Comparison of the averaged Pearson correlation coefficients of the automatic spoken Korean proficiency assessment systems employing the proposed method with prompted text using the BLSTM-based AMs, using the DNN-based AMs, and using the GMM-based AMs, respectively, for the 10-fold cross-validation test of the non-native speech assessment corpus. The first row presents the range of correlation coefficients of the inter-rater scores as an upper boundary

AM	s_{holistic}	s_{segment}	$s_{\mathrm{phonology}}$	s_{fluency}	s_{pitch}
Human	0.589–0.706	0.545–0.673	0.615–0.665	0.611–0.715	0.383–0.518
GMM	0.683	0.643	0.664	0.855	0.506
DNN	0.745	0.709	0.723	0.856	0.542
BLSTM	0.752	0.726	0.735	0.857	0.540

4.1 | Performance comparison of the proposed spoken proficiency assessment method with prompted text

In this subsection, we discuss the experiments using the proposed spoken Korean proficiency assessment method with prompted text: (a) comparison of the BLSTM-, the DNN-, and the GMM-based AMs and (b) comparison when using different text pre-processing methods.

To compare the performance of the proposed proficiency assessment method with the prompted text, we developed a proficiency assessment method employing the GMM-based native AM and the GMM-based non-native AM, of which GMM-based AM was extensively used in an earlier study [17]. To this end, we trained the GMM-based native AM with the 670k clean and native Korean utterances and the GMM-based non-native AM with 347k non-native Korean utterances by using the HTK toolkit [37]. As a speech recognition feature, we used the 39-dimensional mel-frequency cepstral coefficient (MFCC) [38]. Next we developed another proficiency assessment method employing the DNN-based native and non-native AMs, as DNN-based AM is increasingly being adopted in this research area [17]. We trained the DNN-based *native* AM with the 670k clean and native Korean utterances and the DNN-based *non-native* AM with 347k non-native Korean utterances. The structure of DNN-based AMs consisted of one input layer, five DNN layers, a fully connected layer, and a soft-max layer. Except for the model structure, the training procedure was identical to the BLSTM-based AMs.

TABLE 5 Comparison of the averaged RMSEs of each score range of the automatic spoken Korean proficiency assessment systems employing the proposed method with the prompted text using the BLSTM-based AMs, the DNN-based AMs, and using the GMM-based AMs, respectively, for the 10-fold cross-validation test of the non-native speech assessment corpus

	≤1.5	≤2.5	≤3.5	≤4.5	≤5
The proposed method using the GMM-based AMs					
s_{holistic}	0.828	0.527	0.536	0.631	0.707
s_{segment}	1.014	0.678	0.608	0.735	0.948
$s_{\text{phonology}}$	1.196	0.787	0.513	0.556	0.740
s_{fluency}	0.629	0.433	0.358	0.433	0.489
s_{pitch}	1.092	0.834	0.448	0.615	0.426
The proposed method using the DNN-based AMs					
s_{holistic}	0.730	0.475	0.475	0.593	0.623
s_{segment}	0.917	0.644	0.572	0.658	0.844
$s_{\text{phonology}}$	1.009	0.719	0.476	0.519	0.679
s_{fluency}	0.618	0.431	0.356	0.433	0.465
s_{pitch}	1.104	0.792	0.429	0.611	0.419
The proposed method using the BLSTM-based AMs					
s_{holistic}	0.703	0.476	0.481	0.588	0.636
s_{segment}	0.908	0.625	0.564	0.643	0.729
$s_{\text{phonology}}$	1.115	0.798	0.434	0.599	0.392
s_{fluency}	0.601	0.433	0.359	0.432	0.469
s_{pitch}	0.952	0.700	0.482	0.511	0.654

Table 4 shows that the proficiency assessment system employing the proposed method with prompted text using the BLSTM-based AMs improved the averaged correlation across the five scores by 0.052 and 0.007 when compared to the systems using the GMM-based AMs and using the DNN-based AMs, respectively. Moreover, it can be seen from the first row of the table that the performance of the proposed system is comparable with that of inter-rater scores. It can be seen from Table 5 that the proposed proficiency assessment system improved the averaged RMSEs for most score ranges when compared to the system using the GMM-based AMs. It is also noted that the proposed system significantly improved the RMSEs for s_{segment} and $s_{\text{phonology}}$ when compared to the system using the DNN-based AMs. Therefore, it can be concluded that the BLSTM-based AMs significantly improve performance when compared to the GMM-based AMs and the DNN-based AMs.

As a preliminary test of Section 4.2, to determine why the proposed method without prompted text uses the WUS method, we developed a proficiency assessment method in which the input text of the forced-alignment step was based on a subword unit by performing the WS method used for the language model of the Korean ASR. We also developed a proficiency assessment method where the input text of the forced-alignment step was based on a word-like unit by sequentially performing the WS and WUS methods. The first and second rows of Table 6 evidence that the subword-based text degraded the averaged correlation across the five scores by 0.015 when compared to the original text. This performance degradation occurred because

TABLE 6 Comparison of the averaged Pearson correlation coefficients of the automatic spoken Korean proficiency assessment systems employing the proposed method with (a) the prompted text, (b) the subword-based prompted text by performing the WS method, and (c) the word-like prompted text by sequentially performing the WS and WUS methods, respectively, for the 10-fold cross-validation test of the non-native speech assessment corpus

Basic text unit	s_{holistic}	s_{segment}	$s_{\text{phonology}}$	s_{fluency}	s_{pitch}
Original word	0.752	0.726	0.735	0.857	0.540
Subword	0.726	0.710	0.705	0.857	0.538
Word-like	0.753	0.728	0.739	0.857	0.541

the pronunciation rules between subwords were not considered in the forced-alignment step of the proposed method. It can also be seen from the first and third rows of the table that the word-like text maintained performance when compared to the original text. Therefore, the word-like text was used by applying the WUS method to the recognized subword-based text for the proposed method without prompted text.

TABLE 7 Comparison of the syllable error rate (SyllER)s (%) of the Korean ASR systems employing (a) a BLSTM-based native AM, (b) a BLSTM-based native and non-native AM without speech data augmentation, and (c) the proposed BLSTM-based native and non-native AM with speech augmented data, respectively, for the non-native and native speech assessment corpus

Training data for ASR AM	Test data		
	Native	Non-native	Avg.
Native	3.42	36.74	20.08
Native + non-native	3.32	16.48	9.90
Native + non-native + augmentation	3.01	13.82	8.42

TABLE 8 Comparison of the averaged Pearson correlation coefficients of the automatic spoken Korean proficiency assessment systems employing the proposed method without the prompted text using the BLSTM-based AMs, the DNN-based AMs, and using the GMM-based AMs, respectively, for the 10-fold cross-validation test of the non-native speech assessment corpus

AM	s_{holistic}	s_{segment}	$s_{\text{phonology}}$	s_{fluency}	s_{pitch}
GMM	0.655	0.617	0.622	0.850	0.485
DNN	0.695	0.673	0.656	0.848	0.524
BLSTM	0.703	0.678	0.672	0.853	0.513

4.2 | Performance comparison of the proposed spoken proficiency assessment method without prompted text

In this subsection, we first discuss the experiments on the Korean ASR system and then discuss the experiments on the proposed spoken Korean proficiency assessment method without prompted text.

To evaluate the performance of the Korean ASR system of the proposed method without prompted text, we trained the three additional Korean ASR systems employing (a) a BLSTM-based native AM, which was trained with the native Korean utterances, (b) a BLSTM-based native and non-native AM that was trained with the native and non-native utterances, and (c) the proposed native and non-native BLSTM-based AM that was trained with the native and non-native utterances and their augmented data. The first row of Table 7 shows that the native AM poorly degraded the SyllER for the non-native speech when compared to the performance with native speech. It can also be seen from the second and third rows of the table that the native and non-native AMs improved the SyllERs for both native and non-native speech when compared to the native AM. Moreover, it is notable that the data augmentation methods further improved the ASR performance.

TABLE 9 Comparison of the averaged Pearson correlation coefficients of the proposed automatic spoken Korean proficiency assessment system employing the proposed method without the prompted text per ASR accuracy (Acc.) range, for the 10-fold cross-validation test of the non-native speech assessment corpus

	Proficiency score					
Acc.	holistic	segment	phonology	fluency	pitch	No. of utterance
≤40	0.534	0.342	0.257	0.717	0.452	17
≤55	0.535	0.559	0.521	0.860	0.398	85
≤70	0.390	0.438	0.482	0.817	0.319	85
≤85	0.711	0.642	0.626	0.881	0.546	264
≤100	0.728	0.684	0.681	0.863	0.587	1542

TABLE 10 Comparison of the averaged Pearson correlation coefficients of the proposed automatic spoken Korean proficiency assessment system employing the proposed method without the prompted text by excluding the utterances with confidence scores lower than a given threshold (Thr.), for the 10-fold cross-validation test of the non-native speech assessment corpus

	Proficiency score					
Thr.	holistic	segment	phonology	fluency	pitch	No. of utterance
0.40	0.705	0.679	0.673	0.855	0.513	2411
0.43	0.704	0.676	0.667	0.855	0.517	2309
0.46	0.709	0.680	0.665	0.856	0.523	2151
0.49	0.711	0.676	0.659	0.854	0.527	1919
0.52	0.721	0.682	0.660	0.860	0.536	1584
0.55	0.726	0.687	0.659	0.857	0.543	1243
0.58	0.721	0.690	0.643	0.855	0.538	891

To compare the performance of the proposed proficiency assessment method without prompted text, we developed two additional proficiency assessment methods without prompted text employing the GMM-based AMs and the DNN-based AMs, respectively, that were used in Section 4.1. From the proficiency assessment experiments without prompted text as shown in Table 8, the use of the BLSTM-based AMs improved the averaged correlation across the five scores by 0.038 and 0.005, respectively, when compared to the use of the GMM-based AMs and the use of the DNN-based AMs. However, it can be seen from the last rows of Tables 6 and 8 that the proposed method without prompted text degraded the averaged correlation across the five scores by 0.038 when compared to the proposed method with prompted text. Moreover, performance degradation could occur if a low-level utterance obtains a good proficiency score from the recognized text having a low ASR performance, by maximizing the acoustic scores. From the experiments, it can be concluded that the proposed method without prompted text provides comparable performance for s_{fluency}, while it degrades performance for other scores.

To analyze the performance degradation, we divided the non-native speech assessment corpus in terms of the ASR accuracy range and then measured the proficiency assessment performance for each division. Table 9 shows that an utterance having a good ASR performance tended to have a better proficiency assessment performance. Accordingly, we tried to improve the performance by adopting a confidence measure, which was generally used to filter out invalid utterances [17]. We excluded utterances for which confidence scores were lower than a predefined threshold and then measured the proficiency assessment performance. From Table 10, we see that the proficiency assessment performance tends to improve with a higher threshold of the confidence score, except for $s_{\text{phonology}}$. Therefore, it can be concluded that the performance of the proposed method without prompted text could be improved for s_{holistic}, s_{segment}, s_{fluency}, and s_{pitch} if a confidence measure was applied with a credible threshold.

5 | CONCLUSIONS AND FUTURE WORK

In this study, we developed an automatic spoken proficiency assessment system for Korean language utterances that were read aloud by a non-native speaker. To this end, we designed two scenarios: (a) one with prompted text and (b) the other without prompted text. The assessment method with prompted text consisted of the speech feature extraction step, the forced-alignment step using the BLSTM-based native AM and the BLSTM-based non-native AM, and the linear regression–based proficiency scoring step for the five scores. The BLSTM-based native AM was trained with data that were not augmented, while the BLSTM-based non-native AM was trained with augmented speech data by performing speed perturbation along with frequency and time masking. The five proficiency scores measured were holistic, segmental accuracy, phonological accuracy, fluency, and pitch and accent. It was shown from the proficiency assessment experiments that the method with prompted text improved performance compared to the method employing the GMM-based AMs. The method without prompted text consisted of a speech feature extraction step, a Korean ASR step, a forced-alignment step, and a proficiency scoring step. That is, the Korean ASR step was additionally performed to replace the missing prompted text. For the Korean ASR step, the augmented speech data were used to train the BLSTM-based AM of the Korean ASR. The subword based recognized text from the Korean ASR was converted into word-like text using the WUS method to improve the proficiency assessment performance. The proficiency assessment experiments showed that the proposed method without prompted text had a fluency score performance comparable to the proposed method with prompted text. It is also noted that the performance for holistic and segmental accuracy may be improved by filtering utterances for which confidence is lower than a pre-defined threshold.

In this work, we did not consider the effect of reading errors because the test corpus was selected to exclude errors. As the reading accuracy is an important factor for non-native speech in a real environment, we intend to investigate a speech proficiency assessment for a non-native speech containing reading errors such as insertion, deletion, or substitution.

ORCID

Yoo Rhee Oh https://orcid.org/0000-0002-1557-0538

REFERENCES

1. M. Eskenazi, *An overview of spoken language technology for education*, Speech Commun. **51** (2009), no. 10, 832–844.
2. J. Kannan and P. Munday, *New trends in second language learning and teaching through the lens of ICT, networked learning, and artificial intelligence*, Círculo de Lingüística Aplicada a la Comunicación **76** (2018), 13–30.
3. Y. Kim, *The Rising East Asian 'Wave': Korean Media Go Global*, D.K. Thussu (ed.), Media on the Move: Global flow and contra-flow, Routledge, 2007, pp. 233–277.
4. D. Van Compernolle, *Recognizing speech of goats, wolves, sheep and … non-natives*, Speech Commun. **35** (2001), no. 1–2, 71–79.
5. E. Y. Oh, *Developmental research on an interactive application through speech recognition technology for foreign language speaking practice in (in Korean)*, Ph.D. thesis,Seoul National University, Aug. 2017.
6. L. Neumeyer et al., *Automatic text-independent pronunciation scoring of foreign language student speech,* in Proc. Int. Conf. Spoken Language Process. (Philadelphia, PA, USA), Oct. 1996, pp. 1457–1460.
7. C. Cucchiarini, H. Strik, and L. Boves, *Automatic evaluation of Dutch pronunciation by using speech recognition technology*, in Proc. of IEEE Workshop Autom. Speech Recogn. Understanding (Santa Barbara, CA, USA), Dec. 1997, pp. 622–629.
8. L. Neumeyer et al., *Automatic scoring of pronunciation quality*, Speech Commun. **30** (2000), no. 2, 83–93.

9. H. Franco, L. Ferrer, and H. Bratt, *Adaptive and discriminative modeling for improved mispronunciation detection*, in Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (Florence, Italy), May 2014, pp. 7709–7713.
10. A. Lee and J. Glass, Mispronunciation detection without nonnative training data, in Proc. Annu. Conf. Int. Speech Commun. Association (Dresden, Germany), Sept. 2015, pp. 643–647.
11. G. Huang et al., *A evaluating model of English pronunciation for Chinese students*, in Proc. IEEE Int. Conf. Commun. Softw. Netw. (Guangzhou, China), May 2017, pp. 1062–1065.
12. Y. Xiao and F. K. Soong, *Proficiency Assessment of ESL Learner's Sentence Prosody with TTS Synthesized Voice as Reference*, in Proc. Annu. Conf. Int. Speech Commun. Association (Stockholm, Sweden), Aug. 2017, pp. 1755–1759.
13. R. Duan et al., *Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection*, in Proc. ISCA Workshop Speech Language Technol. Education (Stockholm, Sweden), June 2017, pp. 42–46.
14. K. Zechner, I.I. Bejar, and R. Hemat, *Toward an understanding of the role of speech recognition in Nonnative speech assessment*, Tech. Report RR-07-02, ETS Research, June 2007.
15. Y. Wang et al., *Towards automatic assessment of spontaneous spoken english*, Speech Commun. **104** (2018), 47–56.
16. Y. Xiao, F.K. Soong, and H. Wenping, *Paired Phone-Posteriors Approach to ESL Pronunciation Quality Assessment*, in Proc. Annu. Conf. Int. Speech Commun. Association (Hyderabad, India), Sept. 2018, pp. 1631–1635.
17. L. Chen et al., *Automated scoring of nonnative speech using the SpeechRater^{SM} v. 5.0 Engine*, Tech. Report RR-18-10, ETS Research, Apr. 2018.
18. S.H. Yang, H. Ryu, and M. Chung, *A corpus-based analysis of Korean segments produced by Chinese learners*, in Proc. Asia-Pacific Signal Inf. Process. Association Annu. Summit Conf. (Hong Kong, China), Dec. 2015, pp. 583–586.
19. H. Hong, S. Kim, and M. Chung, *A Corpus-Based Analysis of Korean Segments Produced by Japanese Learners*, in Proc. ISCA Workshop Speech Language Technol. Education (Grenoble, France), Aug. 2013, pp. 189–192.
20. S.-H. Yang, M. Na, and M. Chung, *Modeling pronunciation variations for non-native speech recognition of Korean produced by Chinese learners*, in Proc. ISCA Workshop Speech Language Technol. Education (leipzig, Germany), Sept. 2015, pp. 95–99.
21. S.-H. Yang and M. Chung, *Linguistic Factors Affecting Evaluation of L2 Korean Speech Proficiency*, in Proc. ISCA Workshop Speech Language Technol. Education (Stockholm, Sweden), June 2017, pp. 53–58.
22. S.-H. Yang and M. Chung, *Assessment of Korean Spontaneous Speech Produced by Non-Native Learners: Issues and Methodology*, in Proc, Oriental Int, Committee Co-ordination Standardisation Speech Databases Assessment (Miyazaki, Japan), May 2018.
23. S.-H. Yang and M. Chung, *Automatic Proficiency Assessment for Korean Spoken by Chinese Learners*, in Proc. Seoul Int. Conf. Speech Sci. (Seoul, Rep. of Korea), Nov. 2017, pp. 72–73.
24. H. Ryu et al., *Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection*, in Proc. Asia-Pacific Signal Inf. Process. Association Annu. Summit Conf. (Jeju, Rep. of Korea), Dec. 2016, pp. 1–6.
25. S.-H. Yang and M. Chung, *Self-imitating feedback generation using GAN for computer-assisted pronunciation training*, Computing Research Repository (CoRR), (2019), abs/1904.09407.
26. T. Ko et al., *Audio Augmentation for Speech Recognition*, in Proc. Annu. Conf. Int. Speech Commun. Association (Dresden, Germany), Sept. 2015, pp. 3586–3589.
27. D. S. Park et al., SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, in Proc. Annu. Conf. Int. Speech Commun. Association (Graz, Austria), Sept. 2019, pp. 1–7.
28. E. Chung and J. G. Park, *Sentence-chain based Seq2seq model for corpus expansion*, ETRI J. **39** (2017), no. 4, 455–466.
29. H. Jiang, *Confidence measures for speech recognition: a survey*, Speech Commun. **45** (2005), no. 4, 455–470.
30. D. Povey et al., *The Kaldi Speech Recognition Toolkit*, Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Dec. 2011.
31. SoX, Audio manipulation tool, (accessed March 12, 2020). [Online], available at: http://sox.sourceforge.net/.
32. X. Xi et al., *Automated scoring of spontaneous speech using speechratersm V1.0*, Tech. Report RR-08-62, ETS Research, Aug. 2008.
33. T.-Y. Jang, *Speech rhythm metrics for automatic scoring of english speech by Korean EFL learners*, Malsori (Speech Sounds) 1 (2008), **66**, 41–59.
34. R Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
35. A. Stolcke, *SRILM – An extensible language modeling toolkit*, in Proc. Int. Conf. Spoken Language Process, (Denver, CO, USA), Sept, 2002, pp. 901–904.
36. H.-B. Jeon and S.-Y. Lee, *Language model adaptation based on topic probability of latent dirichlet allocation*, ETRI J. **38** (2016), no. 3, 487–493.
37. S. Young et al., The HTK Book Version 3.4, Cambridge University Engineering Department, 2006.
38. S. J. Lee et al., *Intra– and inter–frame features for automatic speech recognition*, ETRI J. **36** (2014), 514–517.
39. Wikipedia contributors, Revised Romanization of Korean — Wikipedia, the free encyclopedia. 2020. [Online].

AUTHOR BIOGRAPHIES

[Correction added on 22nd April 2020, after first online publication: The images in the author biography section were matched to the wrong authors and have been corrected.]

Yoo Rhee Oh received her MS and PhD degrees in Information and Communication Engineering from the Gwangju Institute of Science and Technology (GIST), Gwangju, Rep. of Korea, in 2006 and 2011, respectively. Since 2011, she has been with the Electronics and Telecommunication Research Institute (ETRI), Daejeon, Rep. of Korea, where she is now a senior researcher. Her main research interests include speech recognition, proficiency assessment, and machine learning.

Kiyoung Park received his MS and PhD degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Rep. of Korea, in 1999 and 2003, respectively. From 2003 to 2005, he worked with Samsung Advanced Institute of Technology (SAIT), Yongin, Rep. of Korea, where he contributed to the research and development of human-machine interaction systems. Since 2005, he has been with the Electronics and Telecommunication Research Institute (ETRI), Daejeon, Rep. of Korea, where he is now a principal researcher. His main research interests include speech recognition, signal processing, and machine learning.

Hyung-Bae Jeon received his BS degree in electronics engineering from Yonsei University, Seoul, Rep. of Korea in 1999; and his MS degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea in 2001; and his PhD degree in Bio and brain engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2016. Since 2001, he has been with the Spoken Language Processing Research Section of ETRI, where he is now a principal researcher. His main research interests include speech recognition, language modeling, and machine learning.

Jeon Gue Park received his PhD degree in information and communication engineering from Paichai University, Daejeon, Rep. of Korea, in 2010. He worked for ETRI, Daejeon, Rep. of Korea as a senior researcher from 1991, Lernout and Hauspie Korea, Seoul, Rep. of Korea as a director and division head from 2000, and Donga Seetech Inc. Seoul, Rep. of Korea as a director and CTO from 2002. He rejoined ETRI in 2004 and is currently leading spoken language processing projects. His current research interests include artificial intelligence, computer-assisted language learning, spoken dialogue system, and cognitive systems.

APPENDIX A
DETAILED EXPLANATION OF THE KOREAN READ SPEECH CORPUS FOR A PROFICIENCY ASSESSMENT

This section describes the detailed explanation of the speech assessment corpus. Section A1 presents the recording setup, the speakers, and the prompts and Section A2 explains the human assessment.

Appendix A1 | Setup of the Korean read speech

Each utterance was recorded at a sampling rate of 16 kHz and 16 bits/sample using a dynamic headset microphone in

TABLE A1 Summary of the speakers of the Korean read speech corpus for a proficiency assessment

Item	Description
No. of speakers	Non-native speakers: 30 females, 20 males Native speakers: 5 females, 5 males
Nationality of Non-native speakers	China: 43 Japan: 3 Cambodia: 2 Philippines: 1 Vietnam: 1
Speaking proficiency levels of Non-native speakers	Beginner: 18 Intermediate: 16 Advanced: 16
Age of speakers	18–60

TABLE A2 Characteristics of the utterances and text sentences of the Korean read speech corpus for a proficiency assessment

Item	Description
No. of sentences	50
No. of vocabulary	297
Total no. of words	322
Avg. no. of words per sentence	6.44
Std. dev. of words per sentence	1.64
Total no. of syllables	975
Avg. no. of syllables per sentence	19.5
Std dev. of syllables per sentence	4.76
Sample sentences	제취미는책읽기입니다* 고기를 먹지 않는 사람도 많다.** 한국에 몇 번 와 봤어요?***
Total duration of utterances	5.14 h
Avg. of utterance duration	7.41 s
Std. dev. of utterance duration	2.76

*제 취미는 책 읽기입니다: My hobby is reading books, in English.

**고기를 먹지 않는 사람도 많다.: There are many people that do not eat meat, in English.

***한국에몇번와봤어요?: How many times have you been to Korea?, in English.

TABLE A3 Example of the phonetic or phonological rules for the sentence, “제 취미는 책 읽기입니다 (My hobby is reading books in English).”

Korean word and pronunciation	commonly occurring error
제/j e/	None
취미는 /ch wi m i n eu n/	Aspirated consonant error for/ch/ Rounded vowel error for/wi/
책/ch ae/	Liaison error for/ng/
읽기입니다 /g i l kk i m n i d a/	Double consonant for/kk/ Glottalization for/kk/ Nasalization for/m/

The Korean pronunciation is shown in revised Romanization [39].

a quiet office environment. The utterances had an average duration of approximately 7.41 seconds and were filed in a linear PCM audio format. Table A1 lists the speakers by gender, nationality, spoken language proficiency, and age. In total, 30 female and 20 male non-native speakers participated in the non-native speech recordings and 5 female and 5 male speakers participated in the native speech recordings. The non-native speakers were Asians from China, Japan, Cambodia, Vietnam, and the Philippines. Spoken language proficiency was evenly distributed among the speakers. Their ages ranged from 18 to 60.

Each speaker uttered the same sequence of 50 sentences. For the automatic spoken proficiency assessment, the sentences were selected from the Korean textbooks for non-native speakers and contain Korean phonetic or phonological rules that are commonly mispronounced by non-native Korean speakers. Table A2 shows the characteristics of the sentences and the utterances for the speech assessment corpus. The text sentences contained 297 commonplace expressions with an approximate average of 6 words each. Table A3 gives the phonetic or phonological rules for the sentence, “제 취미는 책 읽기입니다 (My hobby is reading books in English)." as an example.

APPENDIX A2 | Human assessment for the proficiency assessment

As mentioned in Section 2, each non-native utterance was rated by the four experts. The native Korean experts were recruited from among graduate students majoring in linguistics. To ensure better agreement among the experts on the rated scores, a predefined guideline was first presented to the experts and was then calibrated using calibration data. Besides, the experts participated in biweekly training and discussions during the rating process.

For each utterance, a holistic impression of proficiency ($s_{holistic}$) was first rated and then four analytic factors, segmental accuracy ($s_{segment}$), phonological accuracy ($s_{phonology}$), fluency ($s_{fluency}$), and pitch and accent (s_{pitch}), were evaluated on a scale of 1–5. That is, $s_{holistic}$ was rated on the overall impression without the analytic factors. $s_{segment}$ and $s_{phonology}$ were evaluated based on the number of errors that occurred in the utterance and 1, 2, 3, 4, and 5 were rated if the number of errors were 0, 1–2, 3–4, 5–6, and more than 6, respectively. $s_{fluency}$ and s_{pitch} were evaluated on the subjective impressions in terms of the corresponding factors, where 1, 2, 3, 4, and 5 were rated if the utterances were natural, slightly natural, somewhat unnatural, fairly unnatural, and very unnatural, respectively.

Table A4 shows the distribution of the five scores, $s_{holistic}$, $s_{segment}$, $s_{phonology}$, $s_{fluency}$, and s_{pitch}, for the 2500 non-native utterances, where each score was calculated by using (1). Table A5 presents the consensus for the five scores, $s_{holistic}$, $s_{segment}$, $s_{phonology}$, $s_{fluency}$, and s_{pitch}, in terms of the Pearson correlation coefficients.

TABLE A4 Percentage distribution (%) of the five scores ($s_{holistic}$, $s_{segment}$, $s_{phonology}$, $s_{fluency}$, and s_{pitch}) for the 2500 non-native utterances, where each score was calculated by using (1)

Score range	$s_{holistic}$	$s_{segment}$	$s_{phonology}$	$s_{fluency}$	s_{pitch}
0<·≤1.5	10.48	11.84	2.96	7.12	2.48
1.5<·≤2.5	34.24	25.96	18.44	24.96	21.60
2.5<·≤3.5	39.64	34.20	35.92	41.28	47.20
3.5<·≤4.5	14.28	24.96	36.28	22.04	26.24
4.5<·	1.36	3.04	6.40	4.60	2.48

TABLE A5 Comparison of the Pearson correlation coefficients of the inter-rater scores, for the 10-fold cross-validation test of the non-native speech assessment corpus

	Rater 2	Rater 3	Rater 4
$s_{holistic}$			
Rater 1	0.624	0.589	0.618
Rater 2	N/A	0.685	0.647
Rater 3	N/A	N/A	0.706
$s_{segment}$			
Rater 1	0.553	0.545	0.673
Rater 2	N/A	0.649	0.591
Rater 3	N/A	N/A	0.596
$s_{phonology}$			
Rater 1	0.650	0.639	0.623
Rater 2	N/A	0.617	0.665
Rater 3	N/A	N/A	0.615
$s_{fluency}$			
Rater 1	0.652	0.611	0.657
Rater 2	N/A	0.679	0.715
Rater 3	N/A	N/A	0.713