Received: 10 August 2023 | Revised: 27 November 2023 | Accepted: 20 December 2023

DOI: 10.4218/etrij.2023-0322

SPECIAL ISSUE

ETRI Journal WILEY

AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation

Byung Ok Kang[1] | Hyung-Bae Jeon[1,2] | Yun Kyung Lee[3]

[1]Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

[2]Tutorus Labs Inc., Daejeon, Republic of Korea

[3]Soundustry Inc., Daejeon, Republic of Korea

Correspondence
Byung Ok Kang, Integrated Intelligence Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: bokang@etri.re.kr

Funding information
This work is funded by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT), Grant/Award Number: 2019-0-00004.

Abstract
This paper presents the development of language tutoring systems for non-native speakers by leveraging advanced end-to-end automatic speech recognition (ASR) and proficiency evaluation. Given the frequent errors in non-native speech, high-performance spontaneous speech recognition must be applied. Our systems accurately evaluate pronunciation and speaking fluency and provide feedback on errors by relying on precise transcriptions. End-to-end ASR is implemented and enhanced by using diverse non-native speaker speech data for model training. For performance enhancement, we combine semisupervised and transfer learning techniques using labeled and unlabeled speech data. Automatic proficiency evaluation is performed by a model trained to maximize the statistical correlation between the fluency score manually determined by a human expert and a calculated fluency score. We developed an English tutoring system for Korean elementary students called EBS AI Peng-Talk and a Korean tutoring system for foreigners called KSI Korean AI Tutor. Both systems were deployed by South Korean government agencies.

KEYWORDS
automatic speech recognition, computer-assisted language learning, end-to-end recognition, language tutoring, spoken proficiency evaluation

1 | INTRODUCTION

Artificial intelligence (AI) is considered the cornerstone of the Fourth Industrial Revolution, and extensive research is being conducted to prepare for a future society bolstered by its benefits. Because AI is projected to substantially influence future education, leading nations are strategizing education in the AI era. Traditional educational paradigms must be adjusted, particularly in the aftermath of the COVID-19 (coronavirus disease) pandemic. Online classes are widely implemented to encourage students' self-directed learning. Hence, AI-based educational services, such as those catering to foreign language instruction, are garnering increasing attention [1, 2].

This study aimed to develop an AI-based language tutor capable of listening, speaking, and teaching akin to a native language speaker. The development of these tutoring systems involves several technologies. First, spontaneous speech recognition by non-native speakers is essential for audio perception. Second, automatic proficiency evaluation is required to assess foreign language skills and provide feedback, enabling learners to refine their language abilities. Third, dialogue processing of information, including task-oriented and open-domain conversations, is required to facilitate free-flowing conversations.

Numerous research studies have reported methods for automatically assessing proficiency in non-native speech [3–10]. GenieTutor is an English tutoring system for non-native speakers that asks questions, recognizes speech, and evaluates learners' spoken English skills [3–5]. However, it relies on conventional bidirectional long short-term memory plus hidden Markov model (BiLSTM-HMM)-based hybrid automatic speech recognition (ASR) [11], which generally exhibits a lower speech recognition performance than state-of-the-art end-to-end ASR [12–16]. Additionally, it uses a limited set of features for the proficiency evaluation model. SpeechRater is an automatic English speech evaluation system used in the assessment of the Educational Testing Service (ETS) Test of English as a Foreign Language Practice Online [6]. It evaluates various aspects of proficiency, including fluency, pronunciation, repeated speech, grammar, and comprehension. However, SpeechRater focuses on assessments rather than on English tutoring. Fluency, intonation, rhythm, and pronunciation, which pertain to delivery proficiency, are automatically assessed, whereas language use proficiency, including vocabulary and grammar, as well as comprehension-related proficiency, is assessed by human raters. ELSA Speak is an online platform designed to improve English pronunciation for individuals whose native language is not English [7, 8]. However, it primarily focuses on pronunciation, intonation, fluency, and grammar rather than on providing comprehensive instructions in language aspects such as conversational skills. Furthermore, available studies and products have primarily focused on English tutoring, while research and development of Korean language tutoring remains limited [9]. Except for the KSI Korean AI Tutor described in this paper, no Korean language tutoring service is commercially available.

This paper focuses on research to improve the performance of ASR and automatic proficiency evaluation for non-native speakers as well as the development of Korean and English language tutoring systems. Language tutoring systems should automatically recognize spontaneous speech from non-native speakers. ASR is the most important function in speaking learning systems. Speech errors in aspects such as articulation, grammar, and expression are common among non-native speakers. To address this problem, we implement and improve end-to-end ASR by incorporating a variety of speech data from non-native speakers into model training. To increase performance, we combine semisupervised and transfer learning techniques that use both labeled and unlabeled speech log data. Furthermore, we propose a novel method for training text-to-speech (TTS)-generated data that targets unseen or infrequent content in speech log data. To evaluate pronunciation and fluency and then provide feedback, native speakers' rubrics should be learned to evaluate speaking and compare it with native speakers' pronunciation. Pronunciation assessment relies on the statistical correlation between fluency scores manually determined by experts and scores calculated using an evaluation model. For the proficiency evaluation model, we extracted 122 fluency evaluation features after optimization and compared the neural network model with a conventional linear regression model.

The contributions of this paper are as follows. (i) We introduce sophisticated language tutoring systems for non-native speakers that leverage end-to-end ASR and enhanced automatic proficiency assessment. Compared with GenieTutor, which uses BiLSTM-HMM-based ASR and a limited set of features for proficiency evaluation, our system demonstrates a substantial performance improvement. (ii) To improve the ASR performance for non-native speakers, we propose a combined semisupervised and transfer learning method using both labeled and unlabeled speech log data. By incorporating the proposed training method with TTS-generated data, the proposed method improves the error rate reduction (ERR) by 55.7% compared with a baseline end-to-end ASR model. (iii) The proposed language tutoring systems are successfully commercialized as an English tutoring service for Korean elementary students and a Korean tutoring service for foreigners deployed by South Korean government agencies. To the best of our knowledge, this is the first Korean language tutoring service to incorporate AI techniques.

The remainder of this paper is organized as follows. In Section 2, we briefly discuss a similar system. In Section 3, we describe the baseline end-to-end ASR system implemented using a transformer-based encoder–decoder architecture along with a speech corpus used for training the baseline system for English or Korean language tutoring. In Section 4, we detail the proposed approach for language tutoring. First, we describe the proposed training method aimed at achieving high ASR performance of the language tutoring systems designed for non-native speakers. We also discuss the proposed assessment technology to measure language proficiency and provide feedback. In Section 5, we present the experimental environment and report experimental results of the proposed methods aimed to improve the system performance. In addition, we introduce the AI PengTalk language learning service of the Korean Educational Broadcasting System (EBS) and Korean AI Tutor language learning service of the King Sejong Institute (KSI), which are implemented using the proposed solutions. In

Section 6, we discuss the findings of this research. Finally, Section 7 provides conclusions and outlines directions for future work.

2 | RELATED WORK

GenieTutor is a computer-based English tutoring system intended to enhance English proficiency [3–5]. It recognizes English learner's responses to questions provided, assesses their appropriateness, automatically detects and corrects grammatical errors, evaluates the spoken English proficiency, and provides educational feedback. The system comprises two learning stages: think and talk and look and talk. In the think-and-talk stage, various topics are available, each comprising multiple fixed-role-play dialogues. English learners select a learning topic and engage in conversations with the system based on the chosen roleplay scenario. In the look-and-talk stage, English learners select a picture and describe it. After the learner finishes the conversations on the selected topic or describes the selected picture, the system calculates the intonation curves, sentence stress patterns, and word pronunciation scores. It then offers comprehensive feedback.

GenieTutor uses hybrid BiLSTM-HMM-based ASR [11] to recognize speech from a learner and performs forced alignment to obtain a time-aligned phonemic sequence. Using ASR results and time-aligned phonemic sequences, including word, phoneme, sentence, and time alignment data, a proficiency evaluation module selects 50 meaningful features per evaluation criterion and trains the evaluation model. BiLSTM-HMM-based ASR generally exhibits a lower speech recognition performance than end-to-end ASR and achieves a negligible performance improvement when learning using large amounts of training data.

To address the problems of GenieTutor, we implement state-of-the-art end-to-end ASR and improve the performance by incorporating diverse non-native speakers' speech data into model training. A large amount of speech log data was obtained through an English tutoring service developed using the proposed method, and a small amount of the data was transcribed to obtain labeled data. For additional performance improvement, we used transfer learning combining supervised learning on labeled data and semisupervised learning on unlabeled speech data, thereby exploiting all the available speech log data. In this study, the proficiency evaluation model for an English tutoring system was trained using additional 122 fluency evaluation features. The model was then compared with the GenieTutor model, which contains 50 fluency evaluation features.

3 | BASELINE END-TO-END ASR SYSTEM AND SPEECH CORPUS

In this section, we describe the baseline end-to-end ASR system and speech corpus used in model training for English and Korean language tutoring.

3.1 | Baseline end-to-end ASR system

We establish an end-to-end ASR system using a transformer-based encoder–decoder framework [12–16]. By applying end-to-end ASR, we can achieve a substantially higher performance than that of a model that uses conventional BiLSTM-HMM-based ASR [11]. Comparative experimental results are presented in Section 5. The adopted transformer-based end-to-end model was trained using the ESPnet end-to-end speech processing toolkit [17]. Most hyperparameters of the transformer were set to the default toolkit values. The encoder comprises two convolutional layers, a linear projection layer and a positional encoding layer, followed by 12 self-attention blocks that included layer normalization. The decoder comprises six blocks of self-attention and encoder–decoder attention. Each transformer layer uses 2048-dimensional feedforward networks and four attention heads, each with a dimension of 256. Training was performed using the Adam optimization and warm-up steps over 100 epochs without early stopping. Decoding was performed using an in-house ASR system that was developed by modifying the ESPnet toolkit [18, 19]. Figure 1 shows the flowchart of the transformer-based end-to-end ASR system.

Posterior probability $P_\theta(\mathbf{y}|\mathbf{x})$ can be decomposed as follows:

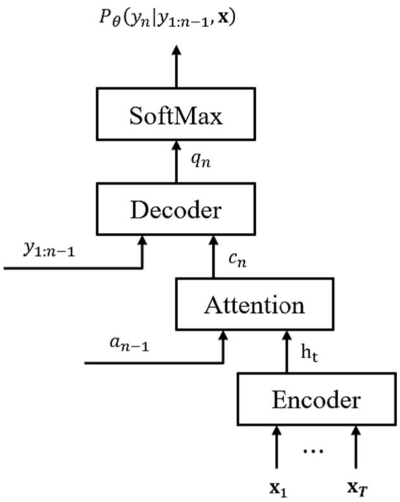


FIGURE 1 Flow of transformer-based end-to-end ASR.

TABLE 1 Length of transcribed speech data.

	Non-native speakers (h)	Total (h)
English	5000	17 000
Korean	5000	10 000

$$P_\theta(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} P_\theta(\boldsymbol{y}_n\,|\,\boldsymbol{y}_{1:n-1},\mathbf{x}), \tag{1}$$

where $\boldsymbol{y}_{1:n-1}$ represents subsequence $\{y_1, y_2, \cdots, y_{n-1}\}$ and $P_\theta(\boldsymbol{y}_n\,|\,\boldsymbol{y}_{1:n-1},\mathbf{x})$ is computed by the encoder–decoder network as

$$\mathbf{H} = \text{Encoder}(\mathbf{x}), \tag{2}$$

$$c_n = \text{Attention}(a_{n-1}, \mathbf{H}), \tag{3}$$

$$P_\theta(\boldsymbol{y}_n\,|\,\boldsymbol{y}_{1:n-1},\mathbf{x}) = \text{Softmax}(\text{Decoder}(c_n, \boldsymbol{y}_{1:n-1})), \tag{4}$$

where $\mathbf{H}$ is a series of state vectors from the top layer of the encoder for a given $\mathbf{x}$, $\mathbf{a}_n$ is an attention weight vector, and $\mathbf{c}_n$ is a context vector synthesized from all encoder outputs $\mathbf{H}$ via an attention mechanism [20, 21].

3.2 | Speech corpus

For high-performance ASR, the acoustic model must be enhanced by collecting and transcribing abundant and diverse non-native speaker speech data for training. When training data are scarce, the deep-learning acoustic model exhibits a performance improvement proportional to the availability of the training data. The lengths of the transcribed speech used to train the end-to-end ASR in the English tutoring system targeting local elementary students and the Korean tutoring system for foreigners are listed in Table 1.

4 | PROPOSED APPROACH

This section describes the proposed approach. First, we elucidate the training methods employed for performance enhancement. Specifically, we describe the proposed semisupervised transfer learning method with labeled/unlabeled speech data and training on the generated TTS data to obtain unseen samples. Finally, we introduce the proposed method for automatic proficiency evaluation.

4.1 | Semisupervised transfer learning for language tutoring

The proposed transfer learning method combines supervised learning with labeled data and semisupervised learning with unlabeled speech data [22–25]. Specifically, we consider an English tutoring system to be deployed, and thousands of hours of unlabeled data obtained from the language service log are used in semisupervised transfer learning. The loss function for training the unlabeled speech log data is given by

$$\mathcal{L}_u = \sum_{u=1}^{U} \mathbb{1}(q_u \geq \tau) H(\hat{y}_u, \boldsymbol{P_\theta}(y|\mathbf{x}_u)), \tag{5}$$

$$\hat{y}_u = \text{argmax}_y(q_u), \tag{6}$$

$$q_u = \boldsymbol{P}_\theta(y|\mathbf{x}_u), \tag{7}$$

where H represents the cross-entropy loss, $\boldsymbol{P}_\theta(y|\mathbf{x}_u)$ is the posterior probability, $\hat{y}_u$ represents the pseudo-label, which is generated from transcriptions for unlabeled speech data using a pretrained ASR model and assigned to unlabeled data $\mathbf{x}_u$, q_u is the confidence score associated with this pseudo-label, and τ is a scalar hyperparameter for confidence thresholding. Hyperparameter τ controls the pseudo-label pruning and is determined experimentally from a set of possible values, $\{0.8, 0.825, 0.85, 0.875, 0.9\}$. The proposed semisupervised end-to-end ASR training method is formulated as the following general optimization problem aimed at finding the model parameters that minimize the loss function for labeled and unlabeled speech data [26–29]:

$$\mathcal{L}_{ss} = \sum_{l=1}^{L} H(y, \boldsymbol{P}_\theta(y|\mathbf{x}_l)) + \gamma \sum_{u=1}^{U} H(\hat{y}_u, \boldsymbol{P}_\theta(y|\mathbf{x}_u)), \tag{8}$$

where the first and second terms correspond to loss functions $\mathcal{L}_l$ and $\mathcal{L}_u$ for labeled and unlabeled data, respectively, and γ is a scalar hyperparameter indicating the weight of the unlabeled data loss. Figure 2 depicts the procedure of the proposed semisupervised transfer learning method using labeled and unlabeled speech log data. All learning blocks in the figure are as explained above, except for the semisupervised transfer learning block.

The proposed semisupervised method performs transfer learning on the baseline end-to-end ASR system described in Section 3.1, which is trained using a large speech corpus, as explained in Section 3.2. While inheriting the unit list corresponding to the output nodes of the

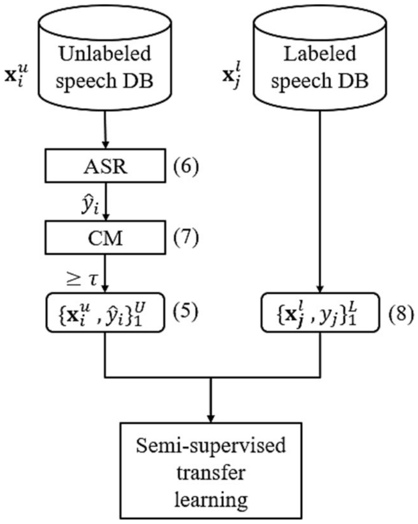


FIGURE 2 Diagram of proposed semisupervised transfer learning.

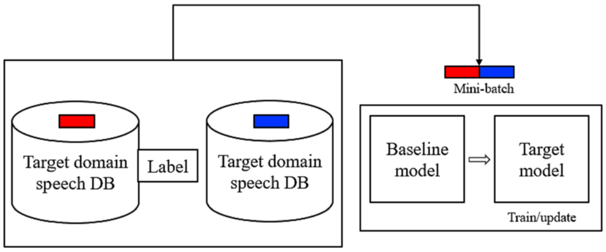


FIGURE 3 Diagram of semisupervised transfer learning block.

baseline end-to-end model, transfer learning is performed on thousands of hours using labeled and unlabeled tutoring service log speech data. However, transfer learning can lead to the overfitting on the training data of specific target domains if training is unconstrained. The trained model may then show a serious deterioration in the recognition performance in the general domain. To mitigate catastrophic forgetting caused by overfitting, we propose and apply a teacher/student-based transfer learning method. Figure 3 details the semisupervised transfer learning block [30, 31].

In the proposed transfer learning method, a baseline end-to-end ASR system trained on a large speech corpus (Section 3.2) serves as the baseline teacher model. The ASR system optimized for the target-domain speech data from the service log serves as the student model. The proposed transfer learning method determines the optimal model parameters for the target-domain speech data while simultaneously training through the transfer of the posterior distribution of the baseline model as follows:

$$\mathcal{L}_t = \sum\nolimits_{i=1}^{N} (1-\lambda) H(\mathrm{y}, \boldsymbol{P}_\theta(y|\mathbf{x}_i)) + \lambda H(\sigma(f_t(\mathbf{x}_i)), \boldsymbol{P}_\theta(y|\mathbf{x}_i)), \tag{9}$$

where $\sigma(f_t(\mathbf{x}_i))$ is a soft label from the baseline teacher model and λ is a scalar hyperparameter experimentally determined based on the number of target-domain speech samples. As a result, the target distribution is an interpolation between the empirical probability and the probability estimated from the baseline teacher model. Hence, conservative training can be resembled, in which Kullback–Leibler divergence regularization is applied [32].

The developed language tutoring system contains a vast array of content, resulting in instances of content sentences that do not appear in the log speech data. Unseen textual content not included in the speech log used for ASR model training reduces the recognition performance when spoken by learners. The conventional approach for BiLSTM-HMM-based ASR models involves building a language model (LM) with textual content data and integrating it into an ASR decoder. However, as explained in Section 5.1, applying a transformer LM trained with textual content to the end-to-end model results in a slight performance reduction. A transformer-based end-to-end model requires paired speech and textual transcriptional data for training. To obtain such unseen content, we use TTS conversion to generate the corresponding speech data and incorporate the data into model training. During this process, we alter the speech rate, pitch, and volume to augment the training data. Speech generated via TTS conversion produces utterances of unseen content and speakers from the original log speech data for model training. However, compared with the speech of actual speakers, this tends to be less natural, possibly degrading the performance of the acoustic model. To mitigate this distortion, we freeze the transformer encoder, which converts an acoustic feature sequence into a sequential representation, and focus solely on training the decoder. In addition, we include a consistent representation loss term in the training loss of sampled labeled speech data, thereby ensuring that data generated by TTS conversion, with their inevitable perturbation, are represented in a model space similar to the speech data articulated by real speakers. The consistent representation loss is given by

$$\mathcal{L}_c = \sum\nolimits_{i=1}^{M} D_{KL}[\boldsymbol{P}_\theta(y|\mathbf{x}_{i,\mathrm{REAL}}) | \boldsymbol{P}_\theta(y|\mathbf{x}_{i,\mathrm{TTS}})], \tag{10}$$

where $\mathbf{x}_{i,\mathrm{REAL}}$ represents real speech sampled from labeled target speech data and $\mathbf{x}_{i,\mathrm{TTS}}$ represents TTS-generated speech data.

4.2 | Evaluation of proficiency for language tutoring

The proposed method for automatic proficiency evaluation involves extracting fluency features to assess proficiency in various areas, training a proficiency evaluation model using these features, and automatically evaluating pronunciation proficiency [4, 9].

Figure 4 shows the overall procedure for training the proficiency evaluation model. This method calculates diverse acoustic features, including segmental features, intonation, and rate from speech pronounced by non-native speakers based on a rubric specifically designed for evaluating pronunciation proficiency. To derive these fluency attributes, the speech signals undergo transcription using the ASR system, and a forced-alignment algorithm determines time-aligned sequences of words and phonemes. Each sequence includes temporal information for individual words and phonemes along with associated acoustic scores. By leveraging these sequences, fluency features are extracted from each word and sentence across multiple dimensions. The proficiency evaluation models are then trained using the extracted fluency attributes and scores provided by expert raters. The human raters were university instructors specializing in English education and English teachers with background in education. To ensure consistent evaluations, the raters received pretraining to familiarize themselves with the evaluation rubrics. During the initial assessment period, a gold standard was established for each score, and detailed evaluation guidelines were coordinated among the raters through a calibration process. Then, two raters with high correlations within a group of multiple raters were selected, and their scores served as references. Finally, the proficiency scores were computed using the fluency features and evaluation models.

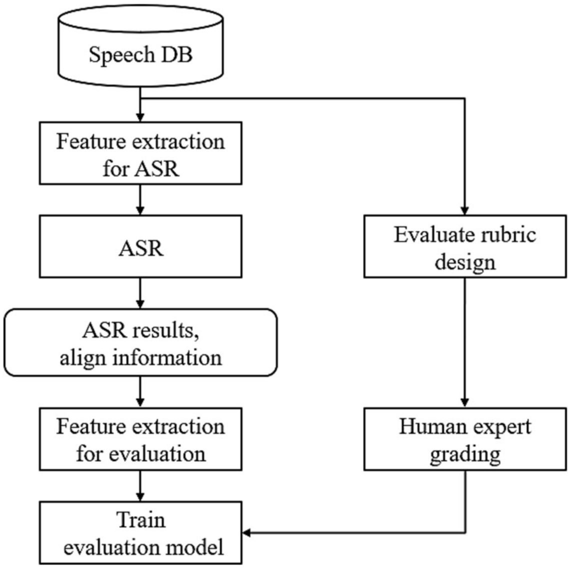

Speech DB
Feature extraction for ASR
ASR
ASR results, align information
Feature extraction for evaluation
Train evaluation model
Evaluate rubric design
Human expert grading

FIGURE 4 Diagram of training proficiency evaluation model.

Figure 5 illustrates the automatic proficiency evaluation process using a trained evaluation model. Learner speech signals were converted into text using ASR, and synchronized sequences of words and phonemes were obtained through a forced-alignment algorithm. Each time-aligned sequence provided the start and end times for each word and phoneme, as well as the acoustic scores. These sequences were used to derive fluency features from every word and sentence from multiple perspectives. Finally, the learners' pronunciation proficiency grades were estimated using a trained proficiency evaluation model.

Next, we briefly describe the components of conversation-based automatic pronunciation and fluency evaluation systems. Figure 6 presents a diagram of the developed conversational language tutoring system. This system is a language learning program based on natural

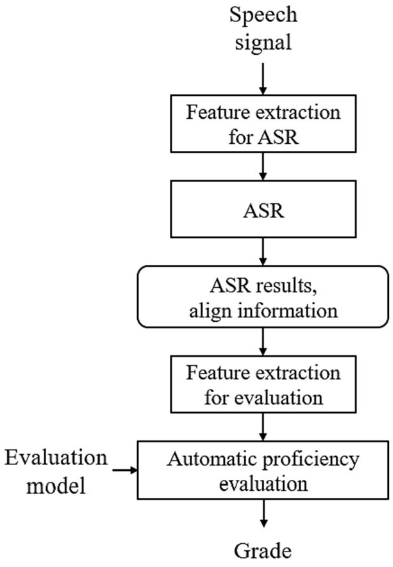

Speech signal
Feature extraction for ASR
ASR
ASR results, align information
Feature extraction for evaluation
Evaluation model
Automatic proficiency evaluation
Grade

FIGURE 5 Diagram of automatic proficiency evaluation.

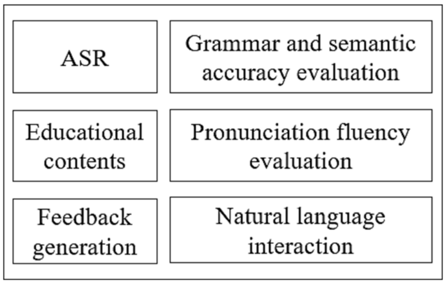

ASR
Grammar and semantic accuracy evaluation
Educational contents
Pronunciation fluency evaluation
Feedback generation
Natural language interaction

FIGURE 6 Diagram of conversational language tutoring system.

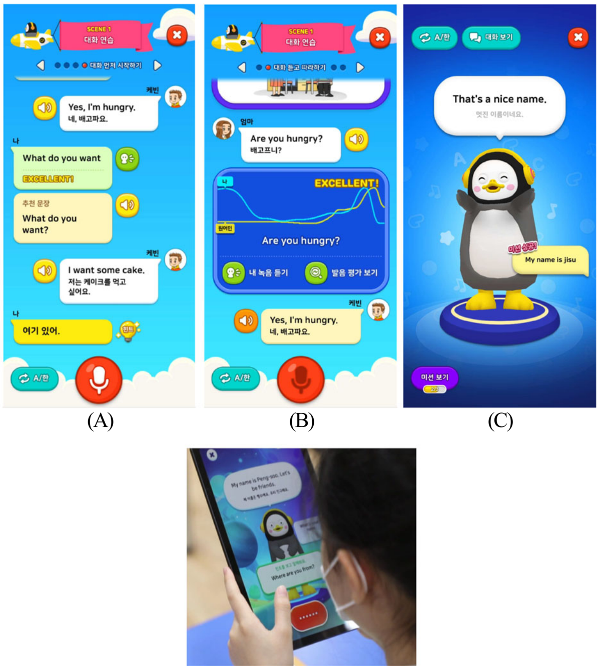
language interaction and speech technology. This allows learners to interact with the system on various topics, thereby facilitating the learning of words, sentences, expressions, and discussion skills. By providing conversational processing functions and educational content, the system offers a platform for the interactive learning of foreign language speaking. It also provides pronunciation, fluency evaluation, and scoring across various aspects of the learners' speech.

5 | EXPERIMENTAL RESULTS

5.1 | EBS AI PengTalk: English tutoring system for Korean elementary school students

EBS, with the support of the Korean Ministry of Education, has developed AI PengTalk, a free service for English learning directed to elementary school students nationwide. From the initial stages of development, ASR and automatic proficiency evaluation from our research have been implemented. Ongoing improvements in performance have led to a substantial number of active users. Figure 7 shows the user interface for learning in EBS AI PengTalk.

For efficient ASR of non-native speakers' pronunciations, the acoustic model must be enhanced by systematically collecting and transcribing a large set of diverse speech data from those speakers. As indicated in Table 1, 17 000 h of transcribed speech data was used for training the ASR system in EBS AI PengTalk. In this dataset, 5000 h of speech data was collected from non-native speakers. EBS AI PengTalk consists of six types of learning exercises: repeating words, repeating sentences, dialogue practice, expression practice, let us talk, and speaking. We constructed an evaluation set composed of approximately 1700 sentences extracted from various learning exercises in the EBS AI PengTalk pilot service logs.



케빈: Yes, I'm hungry. 네, 배고파요.

나: What do you want — EXCELLENT!

추천 문장: What do you want?

케빈: I want some cake. 저는 케이크를 먹고 싶어요.

나: 여기 있어. 힌트

(A)

엄마: Are you hungry? 배고프니?

Are you hungry?

케빈: Yes, I'm hungry. 네, 배고파요.

(B)

That's a nice name. 멋진 이름이네요.

My name is jisu

(C)

My name is Peng-soo. Let's be friends.

Where are you from?

(D)

FIGURE 7 EBS AI PengTalk language tutoring service. Screenshots of (A) dialogue exercise and scoring, (B) feedback on fluency scoring, and (C) free dialogue. (D) Language tutoring scene.

Table 2 lists the results of ASR experiments. At the onset of the language tutoring system, we applied a BiLSTM-HMM-based model trained on a large speech corpus (Section 3.2) and a domain-specific 3-g LM. The BiLSTM-HMM-based model was trained using the Kaldi open-source speech toolkit [33]. The architecture of the BiLSTM-HMM-based model comprises an input layer, five BiLSTM layers, a fully connected layer, and a softmax layer. The domain-specific 3-g LM was trained on preprocessed domain text data including educational content using the SRILM toolkit [34, 35]. The proposed BiLSTM-HMM-based model outperformed the Google API, which provides ASR for American English. The primary reason for this may be its incorporation of non-native speaker data, as described in Section 3.2. To improve the ASR performance, we developed a transformer-based end-to-end ASR model, as detailed in Section 3.1. The transformer LM comprises 16 transformer blocks, with each block containing eight attention heads of 512 dimensions and feedforward layers consisting of 2048 units. A transformer LM is incorporated with end-to-end ASR through shallow fusion [36]. Shallow fusion allows to integrate external LMs into ASR models during decoding. Specifically, the interpolated score of the ASR model and LM is calculated at each ASR decoding step and interpolated in log-linear space. By applying end-to-end ASR trained on the same data, the speech recognition performance notably improves compared with conventional BiLSTM-HMM-based ASR.

TABLE 2 Results of ASR evaluation on EBS AI PengTalk.

Model	LM	Word error rate (%)
Google API (American English)	Unknown	32.13
BiLSTM-HMM	Domain 3-g	20.38
End-to-end	–	18.54
+ Transfer learning	–	9.36
+ Transfer learning	Transformer LM	9.55
+ TTS-based data generation	–	8.21

To achieve automatic proficiency evaluation of input speech, a higher-performance ASR was necessary. To this end, we applied semisupervised transfer learning for domain transfer (Section 4.1) using 1.5 million utterances from the AI PengTalk language service log data. Of these training utterances, 45 000 utterances were labeled, and the remaining utterances were unlabeled. After semisupervised transfer learning for domain transfer to the end-to-end model, we achieved a markedly improved result, with an ERR of 49.5%. As the application of the transformer LM slightly reduced the performance, it was not implemented in the language tutoring service. AI PengTalk consisted of approximately 90 000 sentences of educational content. However, an analysis of the service log speech data revealed that most sentences were duplicated, and over 90% of the total educational content was categorized as unseen. To mitigate adverse effects of such data, we performed transfer learning on speech data generated through TTS conversion from text data inputs and sampled speech data (Section 4.1). For TTS conversion, we used a commercial tool based on the conventional unit selection synthesis from ReadSpeaker Korea. This approach resulted in an additional performance improvement with an ERR of 12.3%.

Training of the automatic proficiency evaluation model for the EBS AI PengTalk language tutoring service was performed as described in Section 4.2. A total of 7545 speech utterances pronounced by 120 elementary school students were collected and manually assessed by five expert human raters for English evaluation. The fluency evaluation scores included a holistic score and four analytic scores, namely, intonation, stress and rhythm, speech rate and pause, and segmental features. Features for fluency evaluation were extracted from each speech sample. By excluding features with zero variance, 122 features were extracted. Subsequently, the proficiency evaluation model was trained using the 122 fluency evaluation features for the 7545 utterances. The proficiency evaluation model was trained and compared using two approaches for predicting the pronunciation assessment scores of an English learner's utterances based on the feature values. Linear regression, a simple and well-established method, is widely used for automatic proficiency scoring. Additionally, we harnessed the processing power of neural networks to train the proficiency scoring model with nonlinear descriptions and enhanced accuracy. The neural network architecture included a convolutional layer with one hidden layer and three hidden units, along with a fully connected layer. Figure 8 shows the Pearson correlation coefficients for the performance evaluation of the manually collected fluency evaluation data. The Pearson correlation is a common metric to assess the effectiveness of proficiency assessment methods. The performance of the two approaches was nearly identical, but the neural network provided a slightly higher performance.

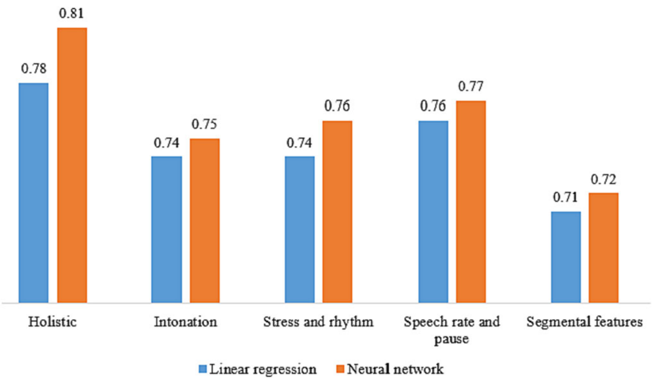


FIGURE 8 Correlation between evaluation results of human rater and proposed proficiency evaluation for five scores.

Table 3 lists a comparison between EBS AI PengTalk and other English tutoring and learning services. ETS SpeechRater, ELSA Speak, and GenieTutor were compared, and their results were retrieved from [3–8]. SpeechRater uses ASR provided by an external vendor and consistently updates its performance. However, according to a recent report, end-to-end ASR has not been implemented to date.

5.2 | KSI Korean AI Tutor: Korean tutoring system for foreigners

KSI is an educational institution that provides Korean language education to individuals outside Korea who wish to learn Korean as a foreign or second language. It is supported by the Ministry of Culture, Sports, and Tourism. Owing to recent cultural influences including K-pop, the

TABLE 3 Comparison between EBS AI PengTalk and conventional language tutoring and learning service.

Service	ASR	No. features for fluency evaluation	Dialogue based	Free to use
ETS SpeechRater	–	>100	No	No
ELSA Speak	–	–	No	No
GenieTutor	BiLSTM-HMM	50	Yes	Yes
AI PengTalk	End-to-end	122	Yes	Yes

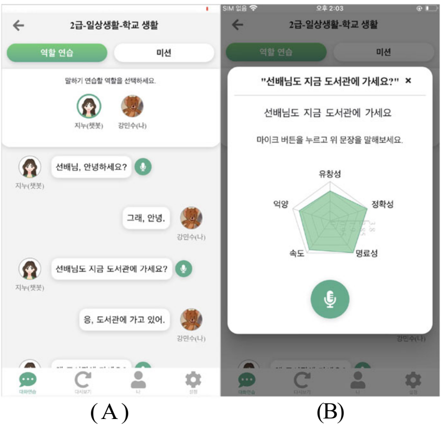
number of foreigners willing to learn Korean is rapidly increasing worldwide. The KSI Foundation has developed the Korean AI Tutor service for Korean language learning for both computer and mobile environments and implementing the developed non-native speaker ASR and proficiency evaluation technologies. Figure 9 shows the user interface of KSI Korean AI Tutor.

Similar to AI PengTalk, users can learn Korean through dialogue and receive feedback from proficiency evaluations. As listed in Table 1, the ASR system for the KSI Korean AI Tutor was trained on a corpus of 10 000 h of transcribed speech data, with 5000 h contributed by non-native speakers.

Table 4 lists the evaluation results of ASR experiments. First, a BiLSTM-based acoustic model and domain-specific 3-g LM were employed, like for AI PengTalk (Section 5.1). The experimental results demonstrated that the application of end-to-end ASR drastically improved the performance with an ERR of 23.0%, thus surpassing the performance of the conventional BiLSTM model. The architecture of each model applied in the comparative experiments, including the BiLSTM-HMM-based and end-to-end ASR models, is the same as described in Section 5.1.

6 | DISCUSSION

We focus on enhancing advanced end-to-end ASR and proficiency evaluation for non-native speakers, emphasizing the development of language tutoring and learning systems for Korean and English learners. Compared with GenieTutor, which uses BiLSTM-HMM-based ASR and limited proficiency evaluation features, the proposed approach substantially improves the performance. To improve ASR for non-native speakers, we combine semi-supervised and transfer learning techniques using both labeled and unlabeled speech log data. Incorporating TTS-generated speech during training provides a 55.7% ERR compared with the baseline end-to-end ASR model. To enhance the proficiency evaluation model, we extract additional 122 fluency evaluation features and apply a neural network model.



지누(챗봇): 선배님, 안녕하세요?

강민수(나): 그래, 안녕.

지누(챗봇): 선배님도 지금 도서관에 가세요?

강민수(나): 응, 도서관에 가고 있어.

"선배님도 지금 도서관에 가세요?"

선배님도 지금 도서관에 가세요

마이크 버튼을 누르고 위 문장을 말해보세요.

FIGURE 9 KSI Korean AI Tutor service. Screenshots of (A) dialogue exercise and (B) feedback on fluency scoring.

TABLE 4 Results of ASR evaluation on KSI Korean AI tutor.

Model	LM	Syllable error rate (%)
BiLSTM-HMM	Domain 3-g	19.34
End-to-end	–	14.90

We have successfully commercialized an English tutoring system for Korean elementary students, called EBS AI PengTalk, and a Korean tutoring system, called KSI Korean AI Tutor, for foreigners willing to speak Korean. Both systems are currently deployed by South Korean government agencies and used by many foreign language learners. Notably, this is the first instance of AI being incorporated into a Korean language tutoring service. In this study, we employed end-to-end ASR for audio perception in the language tutoring system. However, for pronunciation evaluation, we still used handcrafted features and a model that underwent intricate training. Therefore, further performance improvements can be achieved.

7 | CONCLUSIONS AND FUTURE WORK

This paper presents the development of language tutoring systems for non-native speakers using advanced end-to-end ASR and proficiency evaluation. We applied and improved a transformer-based encoder–decoder framework for end-to-end ASR. Recently, encoder frameworks such as conformers, branch formers, and e-branch formers, which combine the advantages of extracting local dependencies using convolutions and global dependencies using self-attention, have been introduced [37–39]. We are conducting research to enhance the ASR performance in language tutoring by applying models with these encoder frameworks. We applied a proficiency evaluation model trained using enhanced handcrafted features that corresponded to particular aspects of proficiency. Self-supervised learning can effectively handle various tasks in speech processing, such as ASR, emotion detection, and speaker separation [40, 41]. Hence, this learning approach captures a broad spectrum of speech characteristics and linguistic details. We are exploring a technique that leverages self-supervised learning of speech representations using models such as wav2vec 2.0 and Hubert to assess pronunciation proficiency.

ORCID

Byung Ok Kang https://orcid.org/0009-0001-8217-720X
Yun Kyung Lee https://orcid.org/0000-0002-1050-2667

REFERENCES

1. W. J. Ha and H. Choi, *Systematic review for AI-based language learning tools*, arXiv Preprint (2021), DOI 10.48550/arXiv.2111.04455.
2. Y. Gong, Z. Chen, I. H. Chu, P. Chang, and J. Glass, *Transformer-based multi-aspect multi-granularity non-native English speaker pronunciation assessment*, (IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Singapore), 2022, pp. 7262–7266.
3. O.-W. Kwon, K.-Y. Lee, Y.-H. Roh, H.-X. Huang, S.-K. Choi, Y.-K. Kim, H. B. Jeon, Y. R. Oh, Y.-K. Lee, B. O. Kang, E. Chung, J. G. Park, and Y. Lee, *GenieTutor: A computer-assisted second language learning system based on spoken language understanding*, In *Natural language dialog systems and intelligent assistants*, Springer, Cham, Switzerland, 2015, pp. 257–262.
4. Y. K. Lee and J. G. Park, *Multimodal unsupervised speech translation for recognizing and evaluating second language speech*, Appl. Sci. **11** (2021), 2642.
5. S. Bibauw, T. Francois, and P. Desmet, *Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL*, Comput. Assist. Lang. Learn. **32** (2021), 827–877.
6. L. Chen, K. Zechner, S. Y. Yoon, K. Evanini, X. Wang, A. Loukina, J. Tao, L. Davis, C. M. Lee, M. Ma, and R. Mundkowsky, *Automated scoring of nonnative speech using the SpeechRater^{SM} v. 5.0 engine*, ETS Res. Rep. Ser. (2018), 1–31.
7. A. Kholis, *Elsa speak app: automatic speech recognition (ASR) for supplementing English pronunciation skills*, Engl. Lang. Teach. **9** (2021), 1–14.
8. M. F. Sholekhah and R. Fakhrurriana, *The use of ELSA speak as a mobile-assisted language learning (MALL) towards EFL students' pronunciation*, J. Educ. Lang. Innov. Appl. Linguist. **2** (2023), 93–100.
9. Y. R. Oh, K. Y. Park, H. B. Jeon, and J. G. Park, *Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM based speech recognition*, ETRI J. **42** (2020), 761–772.
10. Y. Hayashi, Y. Kondo, and Y. Ishii, *Automated speech scoring of dialogue response by Japanese learners of English as a foreign language*, Innov. Lang. Learn. Teach. (2023), 1–15.
11. A. Graves, N. Jaitly, and A. R. Mohamed, *Hybrid speech recognition with deep bidirectional LSTM*, (IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic), 2013, pp. 273–278.
12. A. Vaswami, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, Adv. Neural Inf. Process. Syst. **30** (2017), 5998–6008.
13. S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, *Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration*, (Annual Conference of the International Speech Communication Association, Interspeech, Graz, Austria), 2019, pp. 1408–1412.
14. H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan, *Transformer-based online CTC/attention end-to-end speech recognition architecture* (Proc. IEEE International Conf. Acoustics, Speech and Signal Processing, ICASSP, Barcelona, Spain), 2020, pp. 6084–6088.
15. X. Chang, W. Zhang, Y. Qian, J. Le Roux, and S. Watanabe, *End-to-end multi-speaker speech recognition with transformer*, (IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Barcelona, Spain), 2020, pp. 6134–6138.
16. T. Hori, N. Moritz, C. Hori, and J. Le oux, *Transformer-based long context end-to-end speech recognition*, (Annual Conference of the International Speech Communication Association, Interspeech, Shanghai, China), 2020, pp. 5011–5015.
17. S. Watanabe, T. Hori, S. Karita, and others, *ESPnet: End-to-end speech processing toolkit*, (Annual Conference of the International Speech Communication Association, Interspeech, Hyderabad, India), 2018, pp. 2207–2211.
18. Y. R. Oh, K. Y. Park, and J. G. Park, *Fast offline transformer-based end-to-end automatic speech recognition for real-world applications*, ETRI J. **44** (2022), 476–490.
19. J. U. Bang, J. G. Maeng, J. Park, S. Yun, and S. H. Kim, *English–Korean speech translation corpus (EnKoST-C): construction procedure and evaluation results*, ETRI J. **45** (2023), 18–27.
20. T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, *Multichannel end-to-end speech recognition*, (34th International Conference on Machine Learning, Sydney, Australia), 2017, pp. 2632–2641.
21. T. Hori, R. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. Le Roux, *Cycle-consistency training for end-to-end speech recognition*, (IEEE International Conference Acoustics, Speech and Signal Processing, ICASSP, Brighton, UK), 2019, pp. 6271–6275.
22. L. Lamel, J.-L. Gauvain, and G. Adda, *Lightly supervised and unsupervised acoustic model training*, Comput. Speech Lang. **16** (2002), 115–129.
23. J. Ma and R. Schwartz, *Unsupervised versus supervised training of acoustic models*, (Ninth Annual Conf. International Speech Communication Association, Brisbane, Australia), 2008, pp. 2374–2377.
24. B. Li, T. N. Sainath, R. Pang, and Z. Wu, *Semi-supervised training for end-to-end models via weak distillation*, (IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Brighton, UK), 2019, pp. 2837–2841.
25. B. O. Kang, H. B. Jeon, and J. G. Park, *Speech recognition for task domains with sparse matched training data*, Appl. Sci. **10** (2020), 6155.
26. Y. Chen, W. Wang, and C. Wang, *Semi-supervised ASR by end-to-end self-training*, arXiv Preprint, (2020), DOI 10.48550/arXiv.2001.09128.
27. A. H. Liu, W. N. Hsu, M. Auli, and A. Baevski, *Towards end-to-end unsupervised speech recognition*, (2022 IEEE Spoken Language Technology Workshop, SLT, Doha, Qatar), 2022, pp. 221–228.
28. H. Chung, H. B. Jeon, and J. G. Park, *Semi-supervised training for sequence-to-sequence speech recognition using reinforcement learning*, (2020 International Joint Conference on Neural Networks, IJCNN, Glasgow, UK), 2020, pp. 1–6.
29. Y. Zhang, J. Qin, D. S. Park, W. Han, C. C. Chiu, R. Pang, Q. V. Le, and Y. Wu, *Pushing the limits of semi-supervised learning for automatic speech recognition*, arXiv Preprint, (2020), DOI 10.48550/arXiv.2010.10504.
30. C. Wang, J. Pino, and J. Gu, *Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation*, arXiv Preprint, (2020), DOI 10.48550/arXiv.2006.05474.

31. B. O. Kang, H. B. Jeon, and J. G. Park, *A study on transfer learning method for speech recognition in domains with sparse speech data*, (Winter Annual Conference of KICS, Kangwon, Republic of Korea), 2021.
32. D. Yu, K. Yao, H. Su, G. Li, and F. Seide, *KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition*, (IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada), 2013, pp. 7893–7897.
33. D. Povey, A. Ghoshal, G. Boulianne, and others, *The Kaldi speech recognition toolkit*, (IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU), 2011.
34. A. Stolcke, *SRILM—an extensible language modeling toolkit*, (Proc. International Conf. Spoken Language Process, Denver, CO, USA), 2002, pp. 901–904.
35. H. B. Jeon and S. Y. Lee, *Language model adaptation based on topic probability of latent dirichlet allocation*, ETRI J. **38** (2016), 487–493.
36. A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, *An analysis of incorporating an external language model into a sequence-to-sequence model*, (IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Calgary, Canada), 2018.
37. A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, *Conformer: convolution-augmented transformer for speech recognition*, arXiv Preprint, (2020), DOI 10.48550/arXiv.2005.08100.
38. Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, *Branchformer: parallel MLP-attention architectures to capture local and global context for speech recognition and understanding*, (International Conference on Machine Learning, Baltimore, MD, USA), 2022. pp. 17627–17643.
39. K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, *E-branchformer: branchformer with enhanced merging for speech recognition*, (2022 IEEE Spoken Language Technology Workshop, SLT, Doha, Qatar), 2023. pp. 84–91.
40. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, *wav2vec 2.0: a framework for self-supervised learning of speech representations*, Adv. Neural Inf. Proc. Syst. **33** (2020), 12449–12460.
41. W. N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, *Hubert: self-supervised speech representation learning by masked prediction of hidden units*, IEEE/ACM Trans. Audio Speech Lang. Process. **29** (2021), 3451–3460.

AUTHOR BIOGRAPHIES

Byung Ok Kang received the BS and MS degrees in electronics and electrical engineering from POSTECH, Pohang, Rep. of Korea, in 1997 and 1999, respectively. He received the PhD degree in control and robot engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2018. He worked at Samsung Electronics from 1999 to 2001 and joined the Integrated Intelligence Research Section at the Electronics and Telecommunications Research Institute (ETRI) in 2001. His research interests include automatic speech recognition, speech processing, unsupervised and semisupervised learning, and tutoring and medical artificial intelligence.

Hyung-Bae Jeon received the BS degree in electronics engineering from Yonsei University, Seoul, Republic of Korea, in 1999, MS degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2001, and PhD degree in bio and brain engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea, in 2016. In 2001, he joined the Spoken Language Processing Research Section at ETRI. In 2022, he joined Tutorus Labs, where he is now a chief research officer. His research interests include speech recognition, large language models, and artificial intelligence tutoring.

Yun Kyung Lee received the BS degree in electronics engineering, MS degree in control and instrumentation engineering, and PhD degree in control and robot engineering from Chungbuk National University, Cheongju, Republic of Korea, in 2007, 2009, and 2013, respectively. She worked at ETRI, Daejeon, Republic of Korea, from 2013 to 2022. She is currently the Chief Executive Officer of Soundustry Inc., Daejeon, Republic of Korea. Her research interests include speech processing, artificial intelligence, speech generation, language learning, pronunciation fluency evaluation, and automatic speech recognition.

How to cite this article: B. O. Kang, H.-B. Jeon, and Y. K. Lee, *AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation*, ETRI Journal **46** (2024), 48–58. DOI 10.4218/etrij.2023-0322.