Korean Journal of English Language and Linguistics, **Vol 24, November 2024, pp. 1300-1317**
DOI: 10.15738/kjell.24..202411.1300

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

The Effects of Prosody Training with AI Chatbot on the English Pronunciation Improvement of Korean EFL Learners

Sohyun Yoo · **Hyunkee Ahn** (Seoul National University)

Received: October 15, 2024
Revised: November 15, 2024
Accepted: November 23, 2024

Sohyun Yoo (first author)
PhD student, Department of English Language Education, College of Education, Seoul National University
Email: yoosh111@snu.ac.kr

Hyunkee Ahn (corresponding author)
Professor, Department of English Language Education, College of Education, Seoul National University
Email: ahnhk@snu.ac.kr

*This paper is based on the first author's master's thesis.

ABSTRACT

Yoo, Sohyun and Hyunkee Ahn. 2024. The effects of prosody training with AI chatbot on the English pronunciation improvement of Korean EFL learners. *Korean Journal of English Language and Linguistics* 24, 1300-1317.

This study investigated the impact of prosody training using pitch contour feedback on the English pronunciation of Korean elementary school students. The aim was to evaluate whether this training could lead to significant improvements in pronunciation. The study involved 18 fifth-grade students who participated in a two-week intensive program using AI PengTalk, a chatbot developed to support Korean elementary school students' English learning on a national scale. The participants practiced aligning their pitch contours with those of native English speakers using the tool. The effectiveness of the training was evaluated through AI PengTalk's automated pronunciation scoring and assessments by native English speakers, who rated pronunciation based on nativeness and intelligibility, using a 5-point Likert scale and a ratio scale, respectively. Additionally, the study explored the transferability of the training to unpracticed sentences. Although the automated scoring did not reveal statistically significant differences, assessments by native English speakers indicated notable enhancements in pronunciation regarding both nativeness and intelligibility. Furthermore, the training effect was generalized to novel sentences that were not explicitly practiced during the sessions. These findings suggest that prosody training with pitch contour feedback is a promising approach for improving English pronunciation among young Korean EFL learners.

KEYWORDS

prosody training, AI chatbot, English pronunciation, Korean elementary school learners, EFL education

1. Introduction

Pronunciation plays a vital role in effective oral communication in both first (L1) and second language (L2) contexts. Recognizing this, the significance of L2 pronunciation instruction cannot be overstated. Suprasegmentals, particularly the rhythmic and intonational aspects of speech known as prosody, are especially important in stress-timed languages like English. Studies on L2 prosody have demonstrated that explicit instruction in prosodic features can significantly improve learners' pronunciation, comprehensibility, and overall communicative competence (Chun 2002, Derwing et al. 1998, Levis and Levis 2018). Additionally, studies have shown that prosody affects listener perceptions and pronunciation assessments (Anderson-Hsieh et al. 1992, Field 2005, Hahn 2004, Kang et al. 2010), further underscoring its importance in effective communication.

In Korean school education, however, especially in middle and high schools, instruction has been focused more on written language than spoken language, leading to the neglect of pronunciation (Gu and Reynolds 2013, Hwang 2008). While elementary school education is relatively centered around spoken language, pronunciation teaching primarily focuses on phonics (Sohn 2022), which emphasizes the relationship between sounds and letters. Although phonics helps students grasp segmental sounds, suprasegmental features such as pitch, stress, and rhythm are often overlooked.

In Korea's EFL environment, where exposure to spoken English is limited, this lack of emphasis on prosody exacerbates learners' difficulties in producing natural-sounding speech. Recent advancements in artificial intelligence (AI) technology, particularly AI-based chatbots, offer a promising solution to address this gap. AI tools have facilitated English learning outside of school, enhancing exposure to the language. They have also enabled individualized instruction, which was almost impossible in a one-to-many classroom environment. Recently, AI chatbots like AI PengTalk, developed for English learning on a national scale, are increasingly being used in schools. However, their use in teaching pronunciation, especially prosody, to elementary students in EFL environments has been under-researched.

Therefore, this study aims to explore the impact of the voice-based AI chatbot, AI PengTalk, on the pronunciation of Korean elementary students, with a specific emphasis on prosody. This study aims to empirically verify the effectiveness of prosody training through the use of visual feedback. The following questions are proposed in the present research:

1) Is there a noticeable improvement in the automated pronunciation scores of Korean EFL learners after undergoing prosody training? If so, does this improvement transfer to the learners' performance on novel sentences?
2) Is there a noticeable improvement in Korean EFL learners' pronunciation as rated by native English speakers, based on the nativeness and intelligibility principles, after prosody training? If so, does this improvement transfer to the learners' performance on novel sentences?

2. Literature Review

2.1 L2 Pronunciation Instruction

Second language pronunciation research and pedagogy have been fundamentally influenced by two contrasting principles: nativeness and intelligibility. According to the nativeness principle, it is both possible and desirable to pronounce a foreign language with a native-like accent. This principle, deeply embedded in pronunciation pedagogy and research, posits that native speakers' pronunciation is the ideal standard. Consequently, foreign language learners are often encouraged to strive for this native-like pronunciation, aiming to eliminate their accents (Levis 2005).

However, achieving nativeness is typically unattainable for most adult L2 learners due to various factors, including the age of learning, limited linguistic exposure to the L2 compared to their L1, identity implications, or difficulties in perceiving and producing L2 sounds (Levis 2020). Furthermore, with the rise of English as a lingua franca amid globalization, there has been a notable shift in the focus of pronunciation instruction. Jenkins (1998) argues that for many learners, attaining a native-like accent is not the primary objective. Instead, the main motivation for learning English often lies in communicating with non-native speakers. This view is supported by researchers such as Greene (2000), who advocate for setting realistic targets in pronunciation teaching, arguing that expecting learners to completely eliminate accent transfer is both unrealistic and unreasonable. Munro and Derwing (1995) provided foundational support for this perspective by demonstrating that a strong foreign accent does not necessarily diminish the comprehensibility or intelligibility of L2 speech.

The intelligibility principle, which has emerged as a counterpoint to the nativeness principle, prioritizes clear communication over perfect pronunciation in language learning. While conceding that not all pronunciation traits are equally significant for intelligibility, this approach emphasizes that establishing intelligibility is more critical than emulating native-like pronunciation. This principle underlines the practicality of learners being understandable, even with distinct accents, and redirects language teaching towards effective communication. This shift in pronunciation pedagogy recognizes the varied realities of English use worldwide, advocating for realistic and inclusive teaching strategies that prioritize communicative effectiveness over linguistic perfection.

The concept of *intelligibility* has long been a focal point in discussions surrounding pronunciation and communication. A widely accepted definition of *intelligibility* is "the extent to which the speaker's message is actually understood by listeners" (p. 76), proposed by Munro and Derwing (1995). They differentiated *intelligibility* from *accentedness* and *comprehensibility*. While *accentedness* assesses the deviation from native speaker norms, *comprehensibility* refers to the ease with which listeners understand non-native speech. This distinction is crucial, indicating that *intelligibility* focuses on the actual understanding of the speech, whereas *comprehensibility* deals with the ease of understanding it.

Regarding measurement methods, two primary approaches are generally implemented. The first involves transcription, where listeners transcribe words or sentences they hear, a method utilized by researchers such as Bent and Bradlow (2003), Gass and Varonis (1984), and Munro et al. (2006). The second approach measures intelligibility on a scale, exemplified by Anderson-Hsieh et al.'s (1992) 7-point scale, Fayer and Krasinski's (1987) 5-point scale, and Isaacs' (2008) method of estimating the percentage of words understood.

Despite these transitioning perspectives in pronunciation pedagogy, the majority of research has primarily focused on the nativeness principle. Thomson and Derwing (2015) reviewed 75 pronunciation studies and found that most (63%) centered on accent or nativeness, while only 24% emphasized intelligibility and just 13% considered both. This predominant focus on nativeness suggests a lack of balance in incorporating both the nativeness and intelligibility principles.

2.2 Chatbot Use for English Learning

The integration of AI technologies into English language education is increasingly prevalent. In EFL contexts, chatbots are especially recommended for several reasons. Firstly, they act as interactive partners, providing students with essential practice in understanding and producing language, particularly valuable in EFL environments where opportunities for authentic language input and output are scarce. Secondly, the non-judgmental nature of AI chatbots lowers the affective filters. In Korean classrooms, where there is a strong emphasis on native-like pronunciation, chatbots offer a less intimidating setting for students, especially those with pronounced non-native accents (Cha et al. 2021). Thirdly, personalized feedback from chatbots can bridge the proficiency gap in EFL classrooms, facilitating adaptive learning. Moreover, the use of chatbots has been linked to increased motivation and positive attitude shifts toward English learning, as indicated by several studies (Jeong et al. 2021, Kim, Um, et al. 2021, Lee 2019, Oh and Back 2022, Seong 2022, Seong and Lee 2021, Woo and Kim 2023, Yoon 2022).

Given their effectiveness as conversational practice tools, AI chatbots are increasingly the subject of research and implementation in English education. Lee and Hwang (2022) conducted a meta-analysis of empirical studies on AI chatbots in EFL classrooms, uncovering a significant mean effect size of .689. This impact was notably higher in lower educational levels and when interactions were verbal. In terms of linguistic competence, four areas (listening, speaking, writing, and grammar) showed significant effect sizes, whereas reading and vocabulary did not. Cha et al. (2021) reviewed recent trends in AI chatbot research in English education, focusing on studies concerning development, analysis, and application. Most research utilized Dialogflow to practice dialogues from elementary school textbooks, either analyzing interactions technically or linguistically or focusing on chatbots' effects on English learning.

Additionally, experimental studies in Korea have tested the efficacy of chatbots in enhancing English-speaking competence. Kim (2017) and Kim, Kim, and Cha (2021) highlighted the effectiveness of both voice-based and text-based chatbots in enhancing the speaking skills of university students in a Korean EFL setting. Research has also been conducted with elementary school students. Lee (2019) found that gamification-based chatbot activities were effective in enhancing fluency, accuracy, pronunciation, and accent, with the most significant improvements observed in lower-level learners. Similarly, Chu and Min (2019) demonstrated that task-based AI chatbots facilitated more extensive communication, generating more words and sentences compared to traditional teacher-led interactions.

Within this context, AI PengTalk, a voice-based chatbot, was developed to support English-speaking practice for third to sixth-grade students in South Korea. This chatbot, a joint effort of the Ministry of Education (MOE) and the Korea Educational Broadcasting System (EBS), utilizes speech recognition and natural language processing technologies to facilitate one-on-one conversations between students and the AI.

Since its launch in 2021, AI PengTalk has been the focus of multiple studies exploring its impact on learners' attitudes toward English learning and their interaction with the chatbot (Jeong et al. 2021, Kim, Um, et al. 2021, Seong and Lee 2021). These investigations have reported that utilizing AI PengTalk significantly contributes to a positive shift in learners' attitudes toward English learning. Beyond the affective domain, several empirical studies have also assessed AI PengTalk's effectiveness in enhancing speaking skills. Oh and Back (2022) found significant improvements in vocabulary and speaking skills, especially among lower-level students. Yoon (2022) demonstrated that AI PengTalk significantly enhanced elementary students' speaking abilities, particularly in pronunciation, comprehension, and fluency, after consistent use over a semester. Additionally, Woo and Kim (2023) examined the impact of AI PengTalk on elementary students' English pronunciation and related affective factors, finding significant improvements in pronunciation abilities among fifth graders. However, in addition to using AI PengTalk, the experimental group participated in various activities, such as practicing pronunciation through presentations and conversations, recording and reviewing their presentations, sharing reflections, and writing in journals. Due to these additional activities, it is challenging to determine the specific effect of AI PengTalk alone on pronunciation improvements.

Besides its educational efficacy, AI PengTalk's automated scoring was also a focus of some research. Its automated assessment includes a holistic score and scores on four subscales: *intonation*, *segmental accuracy*, *speech rate*, and *stress/rhythm*. The holistic score, ranging from a minimum of 10 to a maximum of 100 points, is derived by integrating all subscale scores, with each subscale being scored out of a maximum of 5 points.

The automatic pronunciation assessment system of AI PengTalk was developed by ETRI (Electronics and Telecommunications Research Institute), using automatic speech recognition and a forced-alignment algorithm to generate time-aligned sequences of phonemes and words, which were then used to extract fluency features. The proficiency evaluation models were trained on scores provided by human raters and utilized these fluency features to estimate learners' pronunciation proficiency (Kang et al. 2024). Despite the growing interest, research into AI PengTalk's automatic pronunciation assessment system remains limited. A key aspect of this system, the holistic score of AI PengTalk, is not merely a sum of four subscores, and the details of its calculation are confidential. This aspect led Choi (2021) to explore which subscale factors influence AI PengTalk's automatic pronunciation scores. The findings indicated that the *stress/rhythm* subscale significantly impacts the overall pronunciation score. In another study, Park and Lee (2022) conducted a comparative study with native English speakers' assessments, finding that AI PengTalk typically assigns lower ratings for students' pronunciation accuracy and proficiency across various proficiency levels. Particularly noteworthy was AI PengTalk's more rigorous assessment of students in the higher and intermediate proficiency groups, resulting in more conservative ratings than those given to students in the lower group.

This review of empirical studies underscores the increasing interest in AI PengTalk' s role in English language education. Nevertheless, the effects of prosody training with AI PengTalk on learners' pronunciation have not yet been explored, indicating a need for further investigation in this area.

2.3 Second Language Prosody Training

Prosody instruction research encompasses various approaches, including awareness raising, perception, production, and multi-modality (Chun and Levis 2020). Awareness-raising activities help participants recognize prosodic features, while perception instruction often involves high-variability phonetic training with diverse voices in multiple linguistic environments. Production instruction includes methods that integrate visual and kinesthetic learning, and multi-modality instruction combines visual and auditory input with feedback. The computer-assisted pronunciation training (CAPT) for prosody enables multi-modality instruction by utilizing tools ranging from simple digital video to advanced programs featuring automatic speech recognition and speech signal visualization (Chun and Levis 2020). CAPT systems provide visual feedback through spectrograms, waveforms, and pitch tracings, with pitch tracings being particularly intuitive and easier to interpret.

The efficacy of pronunciation training with visual feedback, especially pitch tracing, has been empirically validated. De Bot (1983) studied the impact of visual feedback on foreign intonation learning and the role of practice time, involving 63 Dutch university students learning English intonation. The study concluded that audio-visual feedback was more effective than auditory-only feedback, while practice time was less important. Hardison (2004) explored computer-assisted prosody training's efficacy, assessing its generalization to novel sentences and segmental accuracy. In this study, visual displays of pitch contours during French prosody training with 16 American university students led to significant improvements in both prosody and segmental accuracy. Notably, these enhancements were also observed in the novel sentences, indicating the training's effective transfer to new linguistic contexts. Hirata (2004) assessed the effectiveness of computer-based pronunciation training in learning Japanese pitch and durational contrasts. Eight native English speakers participated, practicing matching the fundamental frequency (f0) contours of native models. After the training, improvements were noted in both perception and production in word-in-sentence and word-in-isolation contexts.

These experiments collectively demonstrate that pitch tracing training can significantly improve pronunciation. While existing studies have provided valuable insights, they have primarily focused on the sentence level, often involving adult learners and relying on reading tasks. Consequently, there remains an opportunity to explore how pitch tracing can be applied within discourse contexts, focusing on how sentences are articulated in connected speech. Additionally, the effects of pitch tracing on younger learners, such as elementary school students, have not been thoroughly examined, despite their distinct developmental characteristics. Furthermore, the emphasis on reading tasks may not fully capture the dynamic nature of spoken language. In response to these considerations, this study shifts the focus to spoken language, targeting elementary school students, and examines how pitch tracing influences sentence-level pronunciation within discourse contexts.

3. Method

3.1 Participants

The participants in this study were fifth-grade students from a public elementary school in Incheon, South Korea. Fifth-grade students from four classes were provided with information leaflets about the study. Participation was contingent upon the receipt of completed application forms and parental consent. To ensure voluntary participation and to avoid interference with regular classes, the study's pretest, posttest, and training sessions were conducted during the summer vacation. The participating students gathered in one classroom at the school to receive prosody training during the training sessions. Upon completion of the study, participants were rewarded with cultural gift vouchers.

All participants were native Korean speakers, with no significant exposure to English-speaking countries. The initial sample consisted of 24 students, including 16 males and 8 females. They took part in a pronunciation instruction program using the AI-based tool. Over the course of the study, 5 students (3 males and 2 females) withdrew, reducing the sample size to 19. Additionally, one female participant was excluded from the final analysis due to difficulties in accurately replicating English sentences, leaving a final sample of 18 students (13 males and 5 females) for data analysis.

3.2 Instruments

The primary experimental tool in this study was the voice-based chatbot, AI PengTalk, developed collaboratively by MOE and EBS. Among its five available modes, the *Speaking* mode was specifically chosen for this research. This mode is structured around conversational tasks between two interlocutors, typically comprising four or five sentences. Learners are first provided with context or situational explanations for each conversation. They then proceed through a three-step practice: Listen to the dialogue, Listen and repeat, and Role-play.

AI PengTalk provides two types of immediate feedback after a learner's utterance. The first is automated scoring results. Based on the holistic score, learners receive one of three comments: *EXCELLENT!*, *GOOD!*, or *TRY HARDER!*. The subscores are displayed using a five-segment bar graph. The second type of feedback is a pitch contour comparison, as depicted in Figure 1, where the native speaker's pitch contour is overlaid on the learner's production.

Figure 1. Examples of Pitch Contour Feedback

The sentences used in the experimental tasks were selected from AI PengTalk's *Speaking* mode based on three criteria. First, the difficulty level was aligned with the 2015 Revised National English Curriculum, focusing on vocabulary lists and limiting sentence length to fewer than ten words (Ministry of Education 2015). Second, sentences that overlapped with content from English textbooks for grades 3, 4, and the first semester of grade 5, which the students had already learned, were excluded, and content from the sixth-grade curriculum was also excluded to avoid advanced learning. Third, to reduce missing data caused by AI system errors or student difficulties, a pilot study with 11 fifth-grade students from a public elementary school in Siheung City was conducted. Based on the results, 20 training sentences and 10 novel sentences were selected for the main experiment (see Appendix 1 for experimental sentences).

3.3 Procedure

3.3.1 Pretest and posttest

To evaluate the effectiveness of the intervention, a pretest-posttest design was employed. In both testing phases, participants were instructed to verbally produce 20 training sentences, which would be included in the six intervention sessions. Additionally, 10 novel sentences, not included in the training, were also part of the test.

The order of the dialogues was randomly selected by the researcher and her colleagues from the system, and the context for each dialogue was explained beforehand. This approach aligns with the emphasis on discourse context in prosody research. After the context explanation, individual sentences from the dialogues were played, and participants were asked to repeat each sentence as they heard it. No visual or written aids were provided, ensuring the reliance on auditory input alone. Participants had a single opportunity to repeat each sentence, a measure taken to minimize potential researcher bias. After the completion of the six treatment sessions, the posttest was conducted the following day under the same conditions as the pretest to ensure data consistency and comparability.

3.3.2 Treatment

The prosody training, comprising six sessions each lasting 40 minutes, commenced the day following the pretest and spanned two weeks during the summer vacation. Each session involved the use of the AI PengTalk application for practicing four sentences, making up one training dialogue. Consequently, learners focused on mastering one complete dialogue per day. This method was consistently applied over five days, covering five distinct training dialogues. On the sixth and final day, a comprehensive review of all 20 sentences was conducted to address any learning discrepancies potentially arising from the sequence of instruction.

The treatment sessions utilized the AI PengTalk application. This training approach involved learners listening to sentences repeatedly and practicing their articulation. After each articulation, they received visual feedback displaying the native speaker's pitch contour overlaid on their own production, as illustrated in Figure 1. The researcher encouraged the learners to visually compare their pitch contour with that of the native speaker, aiming to continuously adjust and improve their pronunciation for closer alignment with the native model. During these sessions, the primary focus for learners was on aligning their pitch contour with that of the native speaker, with no explicit additional guidance provided on grammar, vocabulary, or other linguistic aspects.

3.3.3 Pronunciation assessment

The automated assessment was conducted via AI PengTalk application. Upon a learner's verbal response, the AI system immediately evaluates it, and the pronunciation score was displayed via the Learning Contents Management System (LCMS). Only the holistic score, not the subscale scores, was used in the analysis.

For the human assessment, three native English speakers (2 males and 1 female) were recruited to assess the pronunciation of the students' recordings. All three raters were nationals of the USA, a deliberate choice reflecting the fact that Korean elementary students predominantly learn English with an American accent. These evaluators assessed the students' pronunciation from both the pretest and posttest recordings, using distinct approaches—the nativeness and intelligibility principles. For rater reliability, rater training with a detailed explanation of the assessment rubric and collaborative scoring of sample recordings was conducted before the assessments.

In the assessment of nativeness, raters were instructed to evaluate the recordings holistically according to the rubric (Table 1). The rubric was created by synthesizing the descriptions for each subscale from AI PengTalk's scoring criteria, which are detailed in Appendix 2. By integrating these descriptors, the rubric facilitates a thorough assessment, encompassing all four elements collectively rather than separately. For the intelligibility rating, native raters used a ratio scale developed by Isaacs (2008), marking the estimated percentage of words they could immediately comprehend in each speaker's utterance without guessing, with points at 0%, 25%, 50%, 75%, and 100%.

The context descriptions for each dialogue were also provided to the human raters as suprasegmental elements can vary depending on the context and situation of utterances. To minimize potential biases and order effects, the sequence in which recordings were presented to each evaluator was randomized. In addition, to ensure thorough evaluation, the unit of assessment was not individual sentences but rather full dialogue segments. The training set consisted of five dialogues, each containing four sentences, while the novel set included two dialogues with five sentences each. Native raters evaluated a total of 252 recordings (18 students × 2 tests [pretest and posttest] × 7 dialogues).

Table 1. Rubric of Nativeness Rating

Score	Description
5	The speaker demonstrates natural intonation, accurate individual sound pronunciation, a consistent and natural speech rate, and precise stress and rhythm patterns.
4	The speaker mostly maintains accurate intonation and sound pronunciation. Minor inconsistencies in speech rate and stress and rhythm patterns are observed, but overall clarity is maintained.
3	The speaker shows noticeable irregularities in intonation and sound pronunciation. There are occasional pauses and errors affecting the speech rate, and some inconsistencies in stress and rhythm patterns.
2	The speaker frequently struggles with intonation and sound pronunciation, leading to misunderstandings. The speech rate is uneven due to frequent pausing, and there are frequent errors in stress and rhythm patterns.
1	The speaker has highly inaccurate intonation and sound pronunciation, making comprehension difficult. The speech rate is slow with frequent pauses, and there are significant errors in stress and rhythm patterns.

3.4 Data Analysis

The data were analyzed using SPSS to facilitate comprehensive statistical examinations. For research question 1, the impact of prosody training was assessed by comparing pretest and posttest scores from AI PengTalk's automated scoring system. A paired samples *t*-test was conducted to determine any significant differences attributable to the training. For research question 2, native raters evaluated the nativeness and intelligibility of students' recorded utterances from both the pretest and posttest. Given the non-normal distribution of the scores, the Wilcoxon signed-rank test was employed to identify significant shifts in student performance as perceived by the native raters.

In addition to statistical significance, effect sizes were calculated to gauge the practical significance of the observed differences. The *d* value represented the effect size for the *t*-test, and the *r* value was used for the non-parametric tests. The interpretation of these values followed Plonsky and Oswald's (2014) discipline-specific scale for effect sizes in L2 research. They suggest that for *d* values from pre-post contrasts, $d = .60$ is considered small, $d = 1.00$ medium, and $d = 1.40$ large. For *r* values, $r = .25$ is viewed as small, $r = .40$ as medium, and $r = .60$ as large.

Prior to analyzing inter-rater reliability (IRR), scores from the three raters were manually reviewed. Discrepancies exceeding two levels among raters prompted a re-scoring request. After this review, IRR was evaluated using the intra-class correlation (ICC). The interpretation of ICC values adhered to Cicchetti's (1994) benchmarks. According to these standards, IRR is considered poor for ICC values below .40, fair for values from .40 to .59, good for values between .60 and .74, and excellent for values ranging from .75 to 1.0.

Regarding data exclusion, sentences that were incomplete in either the pretest or posttest were removed before analysis. Additionally, anomalies in AI PengTalk scoring, such as unjustified zero or perfect scores, led to the exclusion of these data points to maintain the integrity of the paired samples *t*-test.

4. Results

4.1 Automated Assessment

The first research question of this study focused on understanding whether prosody training improved the pronunciation of Korean EFL learners as evaluated by AI. For this purpose, scores from the pretest and posttest were categorized into two groups: those from the training sentences ($n = 20$) and those from the novel sentences ($n = 10$). These scores were from holistic scores of AI PengTalk's automated assessment system.

As shown in Table 2, there was a slight increase in average scores from pretest to posttest for both training sentences (TS) and novel sentences (NS). The pretest average for TS was 53.18 ($SD = 10.41$), rising marginally to 54.29 ($SD = 9.46$) in the posttest. Similarly, the NS scores increased from a pretest mean of 54.17 ($SD = 9.04$) to 54.51 ($SD = 8.74$) in the posttest.

Table 2. Descriptive Statistics of the Automated Assessment Scores

Type of Sentences	Test	*N*	*Min*	*Max*	*M*	*SE*	*SD*
Training Sentences	Pretest	282	13	92	53.18	.62	10.41
	Posttest	282	12	90	54.29	.56	9.46
Novel Sentences	Pretest	151	16	81	54.17	.74	9.04
	Posttest	151	32	95	54.51	.71	8.74

Note. All numerical values were rounded to two decimal places.

A paired samples *t*-test was conducted to assess the statistical significance of these differences. For training sentences, the *t*-test revealed no statistically significant difference between the pretest and posttest scores (t = -1.12, p = .279), with a small effect size (Cohen's d = .27). Likewise, for novel sentences, there was no significant difference observed (t = -.17, p = .866), and the effect size was negligible (Cohen's d = .05).

Table 3. Results of a Paired Samples *T*-test of the Automated Assessment Scores

	Paired Differences					*t*	*df*	*p*	*d*
	M	*SD*	*SE*	95% CI					
				Lower	Upper				
TS	-1.08	4.08	.96	-3.11	.95	-1.12	17	.279	.27
NS	-1.64	4.04	.95	-2.17	1.85	-.17	17	.866	.05

Note. TS = training sentences; NS = novel sentences; CI = confidence interval. All numerical values were rounded to two decimal places, except for *p*-values rounded to the third decimal place.

4.2 Human Assessment

The second research question of this study sought to ascertain if prosody training improved the pronunciation of Korean EFL learners, as evaluated by human raters. Native English speakers rated each utterance on two distinct criteria: nativeness and intelligibility. Scores corresponding to these criteria were analyzed independently. Using ICC to determine inter-rater reliability, the nativeness score yielded a value of .836, while the intelligibility score resulted in .915. Both of these values fall within the excellent reliability range.

The scores from the three raters were averaged, and the result is shown in Table 4. The nativeness rating was examined to assess whether students' pronunciation had become more native-like. The mean score of training dialogues (TD) increased from 2.9 to 3.73, and the mean scores of novel dialogues (ND) also rose from 3.53 to 3.78. The intelligibility rating score was also analyzed to determine whether the students' pronunciation became more understandable to native English speakers. As for intelligibility, the mean TD score increased from 71.48% to 94.44%, and the mean ND score rose from 85.42% to 89.81%.

Table 4. Descriptive Statistics of the Human Rating Scores

Type	Test	*N*	*Min*	*Max*	*M*	*SE*	*SD*
Nativeness Rating							
Training Dialogues	Pretest	270	1	5	2.9	.07	.69
	Posttest	270	2	5	3.73	.05	.44
Novel Dialogues	Pretest	108	1	5	3.53	.10	.58
	Posttest	108	2	5	3.78	.08	.50
Intelligibility Rating							
Training Dialogues	Pretest	270	0	100	71.48	2.25	21.35
	Posttest	270	50	100	94.44	.80	7.60
Novel Dialogues	Pretest	108	50	100	85.42	2.26	13.57
	Posttest	108	50	100	89.81	1.82	10.93

Note. The same notes as in Table 2

Due to the deviation from normality, the Wilcoxon Signed Rank Test, a non-parametric alternative, was utilized for the subsequent analysis. Regarding the nativeness rating, with a Z-value of -7.80 and a p-value of < .001, it clearly indicates a significant improvement after the treatment. The effect size for the TD, represented by *r*, was determined to be .82, a large effect size. Similarly, for the novel dialogues, the Posttest-Pretest scores also showed a dominance of positive ranks. The *Z*-value for this set was -3.84 with a *p*-value of < .001, denoting a significant enhancement after the treatment. The effect size for ND was .64, which is also classified as large. Notably, even though the students were not specifically trained in using sentences from the novel dialogues, a meaningful improvement was observed.

A distinct difference also emerged between the pretest and posttest scores for intelligibility rating. For TD, the positive ranks were substantially elevated, yielding a *Z*-value of -7.34. Coupled with a *p*-value of < .001, there's a clear indication of a marked improvement following the intervention. The corresponding effect size for TD, quantified by *r*, stood at .77, a large effect size. The ND had a *Z*-value of -2.45 and a *p*-value of .014, signifying a noteworthy progression after the treatment. The resultant effect size for ND was .41. The improvement, albeit less pronounced than with the training sentences, was still significant, qualifying as a medium effect.

Table 5. Results of a Wilcoxon Signed Rank Test of the Human Rating Scores

Type	Ranks	Mean Rank	Sum of Ranks	*Z*	*p*	*r*
Nativeness Rating						
TD (Posttest-Pretest)	Negative	8.00	24.00	-7.80	.000***	.82
	Positive	42.77	3379.00			
ND (Posttest-Pretest)	Negative	8.50	17.00	-3.84	.000***	.64
	Positive	12.33	259.00			
Intelligibility Rating						
TD (Posttest-Pretest)	Negative	20.00	20.00	-7.34	.000***	.77
	Positive	37.24	2681.00			
ND (Posttest-Pretest)	Negative	13.50	67.50	-2.45	.014*	.41
	Positive	12.24	232.50			

Note. The same notes as in Table 3

5. Discussion

The first research question investigated the immediate effectiveness of prosody training on automated pronunciation scores for Korean EFL learners. Additionally, it sought to assess whether this training affected the learners' performance on novel sentences that were not covered in the training program.

The results of the investigation with TS showed that while there was a slight increase in the posttest scores compared to the pretest, this modest improvement was not statistically significant according to the paired samples *t*-test. The effect size (Cohen's *d* = .27) was categorized as small based on Plonsky and Oswald's (2014) benchmarks. This implies that although training can improve pronunciation, the changes may not be significant enough for AI to detect the difference. The absence of discernible improvement in the outcomes was much more evident when assessing the training's efficacy on the NS. The NS scores displayed no statistically significant improvement after training, and the effect sizes for NS were notably lower than those for TS.

The focus of this study was on evaluating pronunciation only in completely articulated sentences, leading to the exclusion of incomplete data. The lack of statistical significance in the improvements could be partially influenced by the excluded data. Notably, 34 sentences (6.3%) were incomplete in the pretest, while only 7 sentences (1.3%) remained incomplete in the posttest. This suggests that learners were able to articulate more sentences completely after the training. It's important to recognize that sentences easily completed might have already been well articulated in the pretest, leaving limited room for noticeable improvement. However, sentences that were initially difficult and hence incomplete had more potential for significant enhancement after the training. This could mean that considerable improvements might have occurred in these challenging sentences. If learners had more chances to fully articulate sentences in both the pretest and posttest, distinct improvements in pronunciation might have been observed. This is particularly true for the more challenging sentences that were excluded from the automated assessment.

More investigation was required to determine whether this lack of statistical significance was specific to automated assessments or if it also reflected the results of assessments conducted by humans. Addressing the second research question, this study investigated the impact of prosody training on Korean EFL learners' pronunciation as evaluated by native English speakers using criteria of nativeness and intelligibility. In terms of nativeness, the prosody training data showed a noticeable effect in students' pronunciation, as demonstrated by higher posttest mean scores in training dialogues than pretest mean scores. This was statistically significant, as indicated by the non-parametric Wilcoxon Signed Rank Test ($Z = -7.80$, $p < .001$), with a substantial effect size ($r = .82$).

It is likely that a distinct improvement in nativeness occurred because the learners emulated native speakers' speech as their model. Repeated practice would have brought them closer to native pronunciation. English native raters were instructed to evaluate based on several factors, such as intonation, segmental accuracy, speech rate, and stress/rhythm. It is assumed that iterative listening and repetition exercises helped learners to recognize and accurately produce segmental sounds. The other three suprasegmental elements would be naturally acquired as learners matched their pitch tracings with those of native speakers. This focused approach helped the learners adopt a more native-like pronunciation after completing the training.

For intelligibility, improvements were also observed after the training, which indicates that learners' pronunciation clarity and understandability for native English speakers were improved. These enhancements were statistically significant, as demonstrated by the Wilcoxon Signed Rank Test results, with a Z-value of -7.34 and a p-value of < .001, and a large effect size ($r = .77$). The increase in intelligibility scores after prosody training aligns with prior research (Derwing et al. 1998, Levis and Levis 2018). This demonstrates not only that prosody training assists learners in mimicking native-like pronunciation but also in producing more understandable speech.

The study further explored the impact of prosody training on novel dialogues. The experimental results showed considerable improvement after the prosody training, suggesting that the skills acquired in TD were effectively transferred to ND. In terms of nativeness, the Z-value (-3.84) and p-value (< .001) for the novel dialogues support this. With an effect size of .64, albeit lower than the training dialogues at .82, it is still considered large. Regarding the intelligibility rating of novel dialogues, similar conclusions were drawn. Students showed significant enhancement in their posttest scores, indicating their pronunciation became more understandable to native English speakers, even for untrained sentences. The effect size of .41, while smaller than the .77 observed for training sentences, was still statistically significant, particularly as the learners had not been exposed to these sentences during training. This is encouraging since it demonstrates that prosody training extends beyond trained sentences, helping the learners sound more native and intelligible in contexts they are unfamiliar with.

Overall, the findings concur with prior studies (De Bot 1983, Hardison 2004, Hirata 2004) which posit that prosody training with pitch contour feedback enhances pronunciation. Notably, while the participants in previous research were metacognitively high-functioning adults capable of interpreting visual feedback actively, this study's participants were fifth-grade elementary school students. The evident improvement of this study implies their ability to comprehend pitch contour and actively utilize it for their pronunciation enhancement. This demonstrates the potential of younger learners to benefit from contemporary pedagogical methods like chatbot-assisted visual feedback.

The improvements in both nativeness and intelligibility ratings for novel dialogues support the hypothesis that the efficacy of AI-assisted prosody training can be generalized to novel sentences. This finding aligns with Hardison's (2004) study, which showed similar generalization effects in intonation and segmental accuracy among American university students learning French prosody. Likewise, comparable outcomes were observed in fifth-grade Korean EFL students, suggesting that AI-assisted prosody instruction can be applied across different languages and age groups.

There was a clear difference between the results of the automated assessment and the human assessment. The learners' pronunciation scores did not reveal a significant improvement with a low effect size, according to the AI PengTalk's automated evaluation. However, there was a notable improvement with a substantial effect size when nativeness was assessed by human raters using the same rubric. Even though AI PengTalk's automated evaluation system was built on human rater data (ETRI 2020, Kang et al. 2024, Park 2020), such differences emerged. This can be explained by Park and Lee (2022), who found that AI PengTalk's evaluations of intermediate and advanced groups were stricter than their evaluations of beginners. In simpler terms, even when students progressed from a low to intermediate or advanced level, AI PengTalk might not have given correspondingly higher scores. The AI's conservative assessment may be the reason for the experiment's modest improvement. Additionally, Park and Lee (2022) identified some technical problems in AI PengTalk. For instance, the AI struggled to differentiate between English fricatives and plosives. The inconsistency between AI PengTalk's holistic scores and their subscores was also observed. These technological constraints may have contributed to the AI's inaccuracy in assessing the students' progress.

In conclusion, this study emphasizes the diverse advantages of prosody training within the framework of language acquisition. The training's potential to foster transferable pronunciation abilities among Korean EFL learners is highlighted by the notable improvements in nativeness and intelligibility ratings for novel dialogues. These results imply that learners are not just more successful in mimicking native-like pronunciation, but also at applying these skills to new and unpracticed situations. This ability is essential because when learners converse in English with other interlocutors, their speech is not limited to only what they have learned.

6. Conclusion

This study examined the impact of prosody training on Korean EFL learners' pronunciation skills. The study adopted both automated tool and human evaluations to assess pronunciation. The results show that chatbot-assisted prosody training has potential in English pronunciation instruction. Automated assessments displayed modest improvements in pronunciation, but they weren't statistically significant. On the contrary, human assessments revealed significant progress in learners achieving native-like pronunciation and enhancing intelligibility. The study also highlighted the transferability of prosody training. Learners effectively applied their skills acquired from training to novel sentences, indicating a profound understanding and internalization of prosodic components. This finding means that students actually learn linguistic features, not just memorizing and copying native speakers' pronunciation. The research identified the critical importance of combining automated tools and human evaluations in educational settings. Even though implementing only automated evaluations may not fully reflect the range of learners' abilities, including human assessments allows for a more perceptive assessment of students' development.

While this study highlights the effects of prosody training on Korean EFL learners, several limitations should be noted. One limitation is the short and intense training format. A longer duration might have provided a clearer understanding of learners' progress, especially for slow learners who showed less improvement. Future studies should explore the optimal duration and intensity of prosody training. Additionally, the small sample size ($n = 18$) and the homogeneity of participants (Korean EFL elementary students) limit the generalizability of these findings. Further research with a larger and more diverse group is needed to apply the results to broader educational contexts. Considering the proficiency level of the participants, the study used brief and simple sentences. This approach emphasizes the need for future research to explore more complex sentence structures and vocabularies with participants at different proficiency levels. There was also data loss due to working memory limitations, as participants listened and repeated sentences rather than reading them. This resulted in the exclusion of incomplete sentences from the analysis. Future studies should address ways to minimize such issues while avoiding scripted speech.

References

Anderson-Hsieh, J., R. Johnson and K. Koehler. 1992. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning* 42(4), 529-555.

Bent, T. and A. R. Bradlow. 2003. The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America* 114(3), 1600-1610.

Cha, S., J. Kim and S. Nam. 2021. Research trend analysis of AI chatbot in English education. *Journal of the Korea English Education Society* 20(1), 203-225.

Choi, W. 2021. An application of an AI chatbot automatic pronunciation scoring system to elementary school students. *English Language Assessment* 16(2), 167-185.

Chu, S. Y. and D. G. Min. 2019. A study of using task-based artificial intelligence (AI) chatbot for further interaction in English and the analysis of students' production. *Primary English Education* 25(2), 27-52.

Chun, D. M. 2002. *Discourse Intonation in L2: from Theory and Research to Practice*. Amsterdam: John Benjamins.

Chun, D. M. and J. M. Levis. 2020. Prosody in L2 teaching: Methodologies and effectiveness. In C. Gussenhoven and A. Chen, eds., *The Oxford Handbook of Language Prosody*, 619-630. Oxford: Oxford University Press.

Cicchetti, D. V. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 6(4), 284-290.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Laurence Erlbaum Associates.

De Bot, K. 1983. Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech* 26(4), 331-350.

Derwing, T. M., M. J. Munro and G. Wiebe. 1998. Evidence in favor of a broad framework for pronunciation instruction. *Language Learning* 48(3), 393-410.

ETRI. 2020. Server-based speech recognition technology for pronunciation assessment in foreign language learning. Available online at https://techbiz.etri.re.kr/reqstBhrcTchnlgyDc/detail.do?pblancId=12

Fayer, J. M. and E. Krasinski. 1987. Native and nonnative judgments of intelligibility and irritation. *Language Learning* 37(3), 313-326.

Field, J. 2005. Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39(3), 399-423.

Gass, S. and E. M. Varonis. 1984. The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34(1), 65-87.

Greene, G. 2000. Pedagogic priorities 1: Identifying the phonological core. In J. Jenkins, ed., *The Phonology of English as an International Language: New Models, New Norms, New Goals*, 123-163. Oxford: Oxford University Press.

Gu, S. and E. D. Reynolds. 2013. Imagining extensive speaking for Korean EFL. *Modern English Education* 14(4), 81-108.

Hahn, L. D. 2004. Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38(2), 201-223.

Hardison, D. M. 2004. Generalization of computer assisted prosody training: Quantitative and qualitative findings. *Language, Learning & Technology* 8(1), 34-34.

Hirata, Y. 2004. Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. *Computer Assisted Language Learning* 17(3-4), 357-376.

Hwang, E. 2008. Factors affecting Korean learners' English pronunciation and comprehensibility. *English Teaching* 63(4), 3-28.

Isaacs, T. 2008. Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review* 64(4), 555-580.

Jenkins, J. 1998. Which pronunciation norms and models for English as an international language? *ELT Journal* 52(2), 119-126.

Jeong, G., D. Lee, H. Lee, C. Sim and P. Hwang. 2021. Chatbots in primary English class: Focusing on EBS PengTalk. *Primary English Education* 27(4), 95-121.

Kang, B. O., H. B. Jeon and Y. K. Lee. 2024. AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. *ETRI Journal* 46(1), 48-58.

Kang, O., D. O. N. Rubin and L. Pickering. 2010. Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal* 94(4), 554-566.

Kim, H., N. Kim and Y. Cha. 2021. Is it beneficial to use AI chatbots to improve learners' speaking performance? *Journal of Asia TEFL* 18(1), 161-178.

Kim, H., H. Um, H. Oh and D. Choi. 2021. An analysis of the effectiveness of artificial intelligence English speaking program: The results of AI PengTalk pilot schools. *The Korean Journal of Educational Methodology Studies* 33(3), 563-588.

Kim, N. 2017. Effects of different types of chatbots on EFL learners' speaking competence and learner perception. *Cross-Cultural Studies* 48, 223-252.

Lee, J. and Y. Hwang. 2022. A meta-analysis of the effects of using AI chatbot in Korean EFL education. *Studies in English Language & Literature* 48(1), 213-243.

Lee, S. 2019. The effects of gamification-based artificial intelligence chatbot activities on elementary English learners' speaking performance and affective domains. *Primary English Education* 25(3), 75-98.

Levis, J. 2005. Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly* 39(3), 369-377.

Levis, J. 2020. Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation* 6(3), 310-328.

Levis, J. and G. Levis. 2018. Teaching high-value pronunciation features: Contrastive stress for intermediate learners. *The CATESOL Journal* 30(1), 139-160.

Ministry of Education. 2015. *Supplement 14: English Curriculum (Issue No. 2015-74)*. Available online at https://ncic.re.kr/mobile.dwn.ogf.inventoryList.do

Ministry of Education. 2020, December 22. *Practice Speaking English Anytime, Anywhere* [Press release].

Available online at https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&lev=0&statusYN=W&s=moe&m=020402&opType=N&boardSeq=83091

Munro, M. J. and T. M. Derwing. 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45(1), 73-97.

Munro, M. J., T. M. Derwing and S. L. Morton. 2006. The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition* 28(1), 111-131.

Oh, Y. and J. Back. 2022. Exploring the use and effectiveness of an artificial intelligence chatbot in Korean elementary school EFL classrooms: 'AI PengTalk' as a speaking learning tool. *The Journal of Modern British & American Language & Literature* 40*(1)*, 207-235.

Park, J. 2020. Development of artificial intelligence-based technology for foreign language speaking learning. *Technology & Innovation* 444(11/12) 48-52.

Park, M. and J. Lee. 2022. A study on the comparison of AI PengTalk's and English native speakers' assessment of elementary students' English pronunciation. *Journal of the Korea English Education Society* 21(4), 67-90.

Plonsky, L. and F. L. Oswald. 2014. How big is "big"? Interpreting effect sizes in L2 research. *Language Learning* 64(4), 878-912.

Seong, S. 2022. The effects of English vocabulary learning using AI PengTalk and the suggestions for its use in primary education. *Multimedia-Assisted Language Learning* 25(3), 117-145.

Seong, S. and S. Lee. 2021. Analyzing learners' and teachers' perceptions of AI PengTalk for English learning and the suggestions for its use. *Journal of Learner-Centered Curriculum and Instruction* 21(21), 915-935.

Sohn, J. 2022. The way learners of English perceive English phonemes and teaching pronunciation in primary English education in Korea. *The New Studies of English Language & Literature* 83, 195-212.

Thomson, R. I. and T. M. Derwing. 2015. The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics* 36(3), 326-344.

Woo, M. and J. Kim. 2023. Effects of AI PengTalk classes on improvement of English pronunciation abilities and affective factors of elementary students. *Modern English Education* 24, 26-44.

Yoon, T. 2022. A case study on the English learning experience of elementary school English learners with EBS AI PengTalk. *Journal of English Teaching through Movies and Media* 23(3), 27-38.

APPENDIX 1

Experimental Sentences

Type	Dialogue #	Sentence #	Sentence
Training	1	1	Can you help me, please?
		2	Sorry, I can't.
		3	Are you busy?
		4	Yes, I have a lot of homework.
	2	5	Clean up your room!
		6	Okay, I'll do it.
		7	When will you clean up?
		8	I'll do it right now.
	3	9	I finished my homework!
		10	I didn't do it yet.
		11	Can I help you?
		12	Thank you.
	4	13	What do you want to eat?
		14	Let's have pizza.
		15	What kind of pizza do you want?
		16	What about cheese pizza?
	5	17	It's snowing outside.
		18	It's snowing a lot.
		19	Let's make a snowman.
		20	Okay. Let's make a big one.
Novel	1	1	Are you hungry?
		2	Yes, I'm hungry.
		3	What do you want?
		4	I want some cake.
		5	Here you are.
	2	6	Mom, can I watch TV?
		7	Sure, you can.
		8	This is fun.
		9	Is it that fun?
		10	Yes. Let's watch it together.

APPENDIX 2

AI PengTalk's Subscale Scoring Criteria (Originally written in Korean)

Category	Score	Description
Intonation	5	The speaker uses natural intonation and pattern.
	4	The speaker uses intonation and pattern that are slightly inconsistent.
	3	The speaker has irregular and uneven intonation.
	2	The speaker frequently uses inaccurate pronunciation and intonation.
	1	The speaker's pronunciation and intonation are highly lacking.
Segmental Accuracy	5	The speaker pronounces individual sounds accurately.
	4	The speaker makes some errors in phoneme pronunciation, but it's generally fine.
	3	The speaker's inaccurate phonemes occasionally make it difficult to understand.
	2	The speaker frequently uses inaccurate phonemes, making it hard to understand.
	1	The speaker's pronunciation is so inaccurate that it's difficult to understand.
Speech Rate	5	The speaker's speech rate is accurate and natural.
	4	The speaker's speech rate is somewhat irregular and slightly awkward.
	3	The speaker's speech rate is unstable due to continuous pauses and errors.
	2	The speaker's speech rate is uneven because of frequent pausing.
	1	The speaker's speech rate is slow, and he or she frequently pauses.
Stress/Rhythm	5	The speaker accurately pronounces word and sentence stress and rhythm patterns.
	4	The speaker slightly inaccurately pronounces word and sentence stress and rhythm patterns.
	3	The speaker occasionally inaccurately pronounces word and sentence stress and rhythm patterns.
	2	The speaker frequently inaccurately pronounces word and sentence stress and rhythm patterns.
	1	The speaker highly inaccurately pronounces word and sentence stress and rhythm patterns.

Examples in: English
Applicable Languages: English
Applicable Level: All