

Survey and Practical Report

Promoting Listening and Speaking Development through an AI Conversation App: Challenges and Prospects from Classroom Practice

Ryuichi Sato*

Abstract

This classroom practice report integrates an AI conversation application into a first-year university listening course. During the first half of the semester, individual scaffolding was provided, while in the latter half, the focus shifted to general reminders. Students' engagement was tracked through 10 weekly usage logs and questionnaires administered at three points across the semester. Results showed that on-time task completion peaked in Week 6 (94.7%) but gradually declined thereafter, with non-completion reaching 23.7% by the end of the semester. Meanwhile, learners' evaluations of the app remained stable at a moderate level in terms of usefulness and enjoyment, and their confidence in listening to academic content—though still low—increased slightly over time ($T1 = 1.66$, $T3 = 1.78$). Open-ended responses revealed perceived benefits in phonological awareness, difficulties in securing study time, and mixed opinions regarding automatic speech recognition (ASR). These findings suggest that automatic notifications alone are insufficient to sustain engagement; however, interventions at key points—such as reiterating evaluation criteria or making small adjustments to task volume and difficulty—can support continuous learning and foster incremental gains in learners' confidence in academic listening.

[Keywords] AI English conversation app, spaced practice, automatic speech recognition (ASR), sustained learning support

1. Introduction

With the rise of AI, a fundamental question arises: Will teachers become unnecessary? Through the use of mobile devices and generative AI, learners can now practice English conversation anytime, anywhere. If such practice alone were sufficient for significant improvement, then classroom-based English instruction could be minimized. This report attempts to address this simple yet essential question by examining the case of a first-year university course.

The standpoint of this study is clear. While conversation apps offer several pedagogical advan-

* Institute for Liberal Arts and Sciences, Kyoto University

tages—such as high-variability input, immediate automated feedback through Automatic Speech Recognition (ASR), and distributed repetition—they do not automatically ensure sustained or deepened learning if learners are merely given the app. In particular, connecting phonetic and phonological awareness (e.g., linking, assimilation, stress, and intonation) to actual language use requires deliberate intervention from teachers in the form of task sequencing, feedback, and accountability mechanisms (i.e., systems in which learners report their behaviors and outcomes according to deadlines or criteria, which instructors then review and evaluate).

The purpose of this classroom practice was therefore to incorporate an AI conversation app into the course design, while combining it with accountability mechanisms such as weekly reminders and the presentation of grade incentives for task completion. The study seeks to describe how learning unfolded across the semester under these conditions and to identify, from a practitioner's perspective, which forms of pedagogical intervention were effective at different points.

The participants were 38 students enrolled in an English listening course at Kyoto University's Institute for Liberal Arts and Sciences, aimed at developing academic language skills. Their academic backgrounds spanned a wide range of disciplines, including the humanities, social sciences, engineering, medicine, and agriculture, and the cohort included both undergraduate students (Years 1–4) and master's-level students. The course was designed with a dual-layer structure: (1) outside class, learners were assigned conversation practice through the app, and (2) inside class, short oral drills and feedback sessions focused on vowels, consonants, linking, assimilation, stress, and intonation.

Evaluation of the app was conducted through three phases—diagnostic, formative, and summative. Questionnaires including Likert-scale and open-ended items were administered at the beginning, middle, and end of the semester, and interpreted alongside app usage logs (i.e., number of units completed and their continuation across the semester). The analysis combined descriptive statistics of distributions (i.e., means, medians, interquartile ranges, and percentages) with thematic coding of open-ended responses, organizing results around individual differences in usage, stability of evaluations, and learners' awareness of or difficulties with ASR.

The central issues to be clarified are twofold: (a) how does app integration transform learner engagement in English study? and (b) what forms of teacher support are necessary to sustain app usage? Based on the classroom results and discussion, this paper also addresses the principles of pedagogical intervention when introducing AI conversation apps into language learning courses.

The paper proceeds as follows: it first discusses the course context that serves as the premise of the report (Chapter 2), then examines the role of AI conversation apps in listening instruction (Chapter 3). Next, it presents the course design (Chapter 4) and the methods of data collection and analysis (Chapter 5). The findings are reported in Chapter 6, followed by a discussion of pedagogical implications and app design in Chapter 7. Finally, the paper concludes with a summary and future directions (Chapter 8).

2. Course Context: Pronunciation Instruction for Listening Improvement

The course in which this practical report was conducted aimed to improve students' listening skills, with a particular focus on the phonological aspects of English speech perception. At the first class session, students' oral proficiency was assessed using the AI conversation app's placement test. The results suggested that most students had not yet reached near-native proficiency (see Chapter 4 for details). In addition, based on prior teaching experience, the instructor was aware that many students tended to pronounce consonants and vowels with a Japanese accent and often struggled with phonological features such as assimilation, linking, stress, and intonation at the sentence level. Therefore, it was judged necessary to begin with basic pronunciation training and to direct students' attention toward units of sound, connected speech, and prosody.

The instructor argues that pronunciation and conversation practice are essential means of improving listening comprehension. This perspective is grounded in the view of interaction between perception and production. Research in phonetics and second language (L2) learning shows that auditory training—listening while discriminating sounds—improves the perceptual identification of non-native sounds and generalizes to untrained stimuli, including unfamiliar speakers and new vocabulary (Mahdi, 2024; Zhang, 2021). In addition to the research evidence, the instructor speculates—based on classroom experience—that production training can also contribute by enhancing pronunciation accuracy and making it easier for learners to extract perceptual cues. For instance, the phenomenon of linking in English, where words are connected and sounds are weakened or dropped, can become easier to perceive in fast speech (e.g., news broadcasts or online videos) after repeated practice of saying them smoothly. In this way, production practice reinforces perception. Therefore, moving back and forth between perception and production, and cycling between the two, is considered effective for improving listening skills, with micro-level phonological perception reinforced through out-of-class pronunciation practice and macro-level academic content comprehension addressed through in-class listening activities.

To pursue this bidirectional improvement in perception and production, the integration of an AI conversation app into listening instruction is considered useful. The educational implications of such an app can be summarized in three ways. First, it provides a continuous supply of varied input across speakers, accents, and speech rates. Second, its ASR-based feedback clarifies learners' performance in both segmental features and suprasegmental features, sustaining a formative cycle of listening, speaking, and listening again. Third, it promotes sustainable learning by supporting repeated and distributed practice outside the classroom. In short, an AI conversation app increases access to materials beyond the classroom and enables learners to focus on phonological features while maintaining ongoing practice.

3. The Role of AI Conversation Apps in Improving Listening Skills

As described above, pronunciation practice is closely connected to listening comprehension,

and an AI conversation app can serve to reinforce this relationship. Prior research offers several insights into how such apps can contribute to the development of listening skills.

High-variability input training (HVPT). The ability to perceive sounds improves more effectively when learners are exposed to multiple speakers, accents, and speech rates. Training under such conditions has been shown to promote the extraction of phonological categories that generalize beyond a single talker. Meta-analyses confirm these effects. For example, Zhang (2021) synthesized 18 studies ($N = 549$) and found that short-term overall listening gains were small, but perceptual identification improved moderately ($g = .56$), generalization to new talkers improved substantially ($g = .72$), and retention was strong ($g = 1.09$). Similarly, Mahdi (2024) reviewed 18 studies and reported a moderate overall effect on pronunciation attainment ($g = .77$), particularly for consonants and lexical tone, with stable gains achieved through exposure to about five to eight speakers. These findings suggest that AI conversation apps, which provide access to characters with varied accents and speech rates, are well positioned to implement the benefits of HVPT both inside and outside the classroom.

ASR-based immediate feedback. Automatic speech recognition (ASR) can establish a formative cycle of listening, speaking, and listening again by transcribing learner speech and highlighting errors. Kartushina et al. (2015) demonstrated this with French learners of Danish vowels: the experimental group received immediate feedback, listened to models, imitated, compared deviations, and repeated. Their pronunciation accuracy improved by about 17%, and perceptual discrimination also improved significantly, while the control group showed limited gains. This suggests that ASR feedback can enhance both production and perception, and AI apps can reproduce this cycle outside the classroom.

Spaced practice. AI conversation apps support distributed practice through notifications, micro-tasks, and learning logs. Spaced practice—repeating material across intervals—has been shown to outperform massed practice for long-term retention and generalization (Cepeda et al., 2006). Reviews emphasize the effectiveness of combining spacing with retrieval practice (Kang, 2016). In a meta-analysis, Sung et al. (2015) reported that mobile-assisted language learning (MALL) yielded moderate overall benefits ($g \approx 0.55$), though outcomes varied depending on device type, program duration, and learner proficiency. More recently, Mihaylova et al. (2022) reported higher overall effects for app-based learning ($g \approx 0.88$ after outlier removal), highlighting that embedding notifications, micro-tasks, and logs is key to stable outcomes. Accordingly, AI conversation apps are well suited to operationalize spaced practice in everyday learning.

Previous research has demonstrated the usefulness of AI conversation apps from the perspectives of HVPT, ASR, and spaced practice. However, it remains unclear whether simply providing an app constitutes effective instruction in the context of classroom practice. The learning effects of MALL are known to depend on factors such as duration of use, instructional goals, and learners' proficiency levels (Sung et al., 2015). It has also been shown that designs embedding notifications, micro-tasks, and learning logs help stabilize the effects of spaced practice (Mihaylova et al., 2022). What remains unclear, however, is how learners' behaviors evolve over the course of a semester and when specific interventions—such as task sequencing, feedback, or adjustments in the strength of

accountability systems—should be implemented.

The present classroom practice report aims to clarify the factors that influence sustained learning and to identify principles of effective intervention when an AI conversation app is integrated into a course. Specifically, it addresses two questions: (a) how app integration transforms learners' engagement in English study, and (b) what forms of teacher support are necessary to sustain app usage. To pursue these aims, the report focuses on three aspects:

- trajectory of app usage frequency and continuity across the semester,
- learners' evaluations of app-based learning (i.e., perceived usefulness, perceived burden, acceptance of ASR, and perceived contribution to listening) and how these evaluations change over time, and
- timing and nature of interventions that are effective in addressing signs of disengagement or stagnation.

4. Overview of the Classroom Practice

4.1. Course Context and Objectives

This classroom practice was conducted in the Active Listening course, offered as a university-wide common subject by the Institute for Liberal Arts and Sciences at Kyoto University. Active Listening is classified as one of the E-courses, subjects within the common curriculum that are designed to strengthen students' English proficiency. More specifically, it is categorized as an E3-course, which aims to enhance academic English skills. Within the E3 category are courses such as Academic Discussion, Critical Listening, and Academic Presentation, designed to strengthen both receptive skills for understanding lectures in English and productive skills for academic communication. Among these, Active Listening is particularly designed to improve listening proficiency. Although previously restricted to second-year students and above, the course became available to first-year students starting in April 2025. Consequently, it was anticipated from the outset that students with a wide range of language proficiency levels and learning objectives would enroll.

E3-courses, including Active Listening, intend to strengthen listening proficiency so that students can undertake E2-courses delivered through English Medium Instruction (EMI), as noted by Yanase (2024). With this in mind, Active Listening intends to accomplish three objectives: (1) to analytically understand the phonological features of English and apply them in listening comprehension, (2) to develop top-down processing skills that make use of background knowledge and context, and (3) to enable students to grasp the overall content of a 10–20 minute lecture in English. While adhering these objectives, the present classroom practice emphasized phonological perception as its primary goal, in light of the initial distribution of learners' proficiency levels and the author's prior teaching experience. Accordingly, the course incorporated not only top-down activities such as comprehension of English lectures and news videos but also bottom-up activities, including explicit instruction and short oral practice in vowels, consonants, linking, assimilation, stress, and intonation. Outside the classroom, students were assigned weekly shadowing tasks based on English lecture videos. In addition, the AI conversation app was introduced as an out-of-class task in order to ensure

practice volume and continuous learning through a combination of HVPT, ASR-based immediate feedback, and spaced practice.

Assessment was designed to connect in-class activities with out-of-class repetition, thereby making the listening-speaking-listening cycle visible across the semester. Grading was conducted as follows.

- a. **End-of-semester listening test (30 points):** Students listened to approximately two minutes of academic lecture audio and wrote a summary in either Japanese or English. The instructor evaluated the responses with a focus on informational accuracy.
- b. **End-of-semester speaking test (30 points):** Following the IELTS speaking format, students were evaluated by the instructor, who acted as examiner, on the basis of course content (i.e., vowels and consonants stress, intonation, linking, and fluency).
- c. **Out-of-class activity 1 Shadowing (20 points):** Students completed weekly assignments for ten weeks. Each required students to produce a transcript of a two-minute academic or business-related video, then perform shadowing after checking the transcript. Each submission earned 2 points, up to a maximum of 20 points.
- d. **Out-of-class activity 2 AI conversation app (20 points):** Based on a weekly plan (detailed in Chapter 4.4), students earned 20 points if they completed the designated lessons by the end of the semester. No partial credit was awarded. Progress was verified through the app's learning logs, including completion flags, dates and times of study, and number of units completed.

Participation in the research was voluntary, and consent or non-consent to participate had no impact on grades; assessment was conducted solely for educational purposes.

4.2. Participants

The participants in this classroom practice consisted of 38 students enrolled in either undergraduate programs or a master's program. Their fields of study were diverse, spanning literature, law, education, economics, science, medicine, engineering, and agriculture. No specific enrollment requirements were set, and the course was open to students of all years as long as the registration cap of 60 had not been reached during the elective registration period. At the time of class orientation in the first class, 53 students had begun using the app. After accounting for voluntary course withdrawal (until Week 7) and exclusion of those who did not consent to participate in the study, the final sample for analysis consisted of 38 students. Written informed consent for participation and for the use of app log data was obtained at the end of the semester, and all data were anonymized. This study was approved by the Research Ethics Review Committee of the Institute for Liberal Arts and Sciences at Kyoto University (Approval No. 2025-02) and was conducted with permission to use educational data held by the Institute. Participants were informed of the procedure for withdrawing consent, but no withdrawals had been submitted at the time of writing.

Self-reported English proficiency was based primarily on the TOEFL ITP ($n = 32$; $M = 553.7$, $SD = 31.6$, $Mdn = 553$, $IQR [535, 573]$, range $[503, 620]$). Additional scores included TOEIC L&R ($n = 3$; $635-775$, $M = 755.0$, $SD = 26.5$, $Mdn = 765$) and IELTS ($n = 1$; Overall = 7.0, L = 8.0,

R = 8.5, W = 6.0, S = 6.0). Information on first language, age, gender, and study-abroad experience was not collected. As supplementary information on learners' initial proficiency, the results of the app's built-in placement test (10 levels) were also considered. The majority of students were in the B1 band (levels 6–7) on the Japanese version of the Common European Framework of Reference for Languages (CEFR-J; Tono, 2017). The median was 7, IQR [5, 8], and the range was 2–10 ($n = 38$).

These indicators reflect overall English competence placement rather than listening-only ability and are presented solely to describe the cohort's baseline distribution. They were excluded from the main analyses of this classroom practice report for several reasons: (a) the indicators represent a composite measure of pronunciation, fluency, accuracy, expressive range, and speaking volume, and thus do not fully align with the construct of phonological perception, which is the focus of this study; (b) the scoring algorithm and the procedures used for conversion to CEFR-J levels are proprietary, with little information on their validity or reliability, making longitudinal comparison difficult to justify; (c) substantial measurement error is expected due to the lack of control over test content and testing conditions; and (d) because the measurement scale is not standardized, interpreting formative changes or learning effects based on these scores could easily lead to misinterpretation. For these reasons, the scores are reported only to describe the cohort's baseline distribution and are not used for inferences about instructional effects.

4.3. AI English Conversation App Used

This classroom practice employed the AI English conversation app SpeakBuddy, provided by SpeakBUDDY Ltd. The decision to adopt this app was based on three features: (a) its design for routine exposure to high-variability input (HVPT) through multiple speakers, accents, and speech rates; (b) the transcription of speech and pronunciation feedback through automatic speech recognition (ASR), creating a feedback cycle; and (c) support for spaced practice through weekly goals, reminders, and learning logs (SpeakBUDDY Ltd., 2025), which were considered particularly suitable for sustaining out-of-class practice and reinforcing phonological awareness as a foundation for academic listening.

First, the app contains multiple characters with different cultural backgrounds, English accents, and speech rates. This allows learners to engage continuously with diverse input rather than relying on a single speaker, thereby recreating an HVPT-style high-variability environment for out-of-class practice.

Second, learners' utterances are transcribed by ASR and evaluated for pronunciation accuracy. While corrective tasks for errors are provided later as review activities and not always in real time, the process nonetheless enables formative cycles at the task level. Each unit integrates listening to a model, reading the script, speaking, checking the result, and revising—thus facilitating the cycle of listening, speaking, and listening again.

Third, when weekly tasks are assigned, the app enables instructors to set estimated study times, schedule reminder notifications, and visualize learning logs. This makes it possible to distribute short practice tasks at high frequency and sustain continuous learning behavior. The system is also compatible with low-intensity progress management on the instructor side (i.e., weekly reminders and

grade incentives for task completion).

The sequencing of tasks was based on the app's internal level divisions (i.e., beginner, intermediate, and advanced). Each week, learners selected tasks aligned with their proficiency level and practiced on a conversation-unit basis. This framework allowed fine-tuning of task load and focus according to learners' duration of app use, target skills, and proficiency level, as suggested by previous research.

In this study, all participants accessed the app on smartphones. Data use was approved through a formal agreement with the app provider, and explicit consent was obtained for the educational and research use of participants' app usage logs and placement test results. All data were anonymized, and the handling of course assessment data and research data was kept strictly separate for storage and analysis. There was no conflict of interest: the author received no funding, benefits, or supervisory involvement from the app provider. The app license fee for this study was subsidized by the budget allocated to the Institute for Liberal Arts and Sciences at Kyoto University under its 2025 Request for the Expansion and Improvement of Liberal and General Education. The funding body had no role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to submit for publication.

4.4. App Assignments and Teacher Interventions

During each week of the semester, students were assigned the completion of specified lessons in the app as a goal, with accountability applied through the award of grade incentives for achievement. Learners could select lessons from beginner, intermediate, or advanced tracks according to their perceived progress and workload for that week, and they were informed that they could adjust the difficulty level up or down in the following week. This arrangement provided upper-level learners with room for further extension while allowing those who stagnated the opportunity to work at a lower level. The sequencing of the lessons is shown in Table 1.

The app's reminder notifications and weekly task scheduling were used on the premise of distributing short practice sessions throughout the week. Assignments were divided into smaller units so as to lower the load of each session while increasing frequency. Progress data were provided weekly to the instructor in the form of time-stamped logs, managed by a data administrator at the app provider.

Until Week 6, the instructor presented these logs to students in class. Learners who had not completed their tasks were individually notified and verbally instructed to finish them. At these times, the grading rules were reviewed, and students were reminded that assignments had to be completed by the end of the semester. They were also advised, in line with individual circumstances, that distributed practice over time was preferable to short-term cramming. Where necessary, the instructor consulted with students about their specific situations and suggested lowering the task difficulty, rescheduling practice times, or checking device environments. These interactions were kept to short, individualized prompts, with priority given to communicating the importance of continuing the work.

In Week 7, the grading rules were re-announced to the entire class, after which no further indi-

Table 1. Weekly App Lesson Schedule

W	Lesson Title	Due
1	Class Orientation	
2	Introduction of the Application	
3	Introducing Japanese Culture – Ch. 1: Let's Help Others Learn About Japan (I)	W4
4	Everyday Conversation: London – Ch. 1: A New Setting (A)	W5
5	Free choice	
6	Everyday Conversation: Los Angeles – Ch. 1: Visiting a Friend (B) Everyday Conversation: Malaysia – Ch. 1: Bustling Asia (I) Introducing Japanese Culture – Ch. 2: Exploring Japan in Depth (A)	W7
7	Everyday Conversation: Los Angeles – Ch. 2: Enjoying Good Times (B) Everyday Conversation: Malaysia – Ch. 2: A Multicultural City (I) Everyday Conversation: London – Ch. 2: Enjoying British Culture (A)	W8
8	Everyday Conversation: Los Angeles – Ch. 3: Strolling Around Los Angeles (B) Everyday Conversation: Malaysia – Ch. 3: Support Around You (I) Everyday Conversation: London – Ch. 3: A Season of Farewells (A)	W9
9	Expressing Opinions – Ch. 1: Basic Level (B) Expressing Opinions – Ch. 2: Intermediate Level (I) Expressing Opinions – Ch. 3: Advanced Level (A)	W10
10	Business – Ch. 1: An International Workplace (B) Business – Ch. 3: Struggles at Work (I) Business – Ch. 5: A Business Trip to Sydney (A)	W11
11	Business – Ch. 2: Internal Communication (B) Business – Ch. 4: Earning Customer Trust (I) Business – Ch. 6: The Best Buddies (A)	W12
12	Learning English Conversation Through Stories – Ch. 1: Cinderella (B) Idioms – Ch. 1: Intermediate Idioms (I) News English – Ch. 1: Everyday Topics (A)	W13
13	Small Talk – Ch. 1: Small Talk (B) Idioms – Ch. 2: Advanced Idioms (I) News English – Ch. 2: International Topics (A)	W14
14	No task assignment	

Note. W = assigned week; Lesson Title indicates the title of each lesson requested to complete, alphabets in parentheses indicate recommended difficulty (B = basic, I = intermediate, A = advanced). “Free choice” means students could select lessons outside the assigned set. Week 1 (“Class Orientation”) and Week 2 (“Introduction of the Application”) were briefings with no task due. “Due” indicates the assignment was due before the start of the listed week’s class. Lesson titles are translation of Japanese titles provided in the app.

vidual checking, reporting, or prompting of app usage was carried out. The plan was to establish clear rules and expectations during the first half of the semester (Weeks 1–6), supporting students’ initial engagement and self-regulation strategies through individualized advice, and then, in the latter half, to limit interventions to general announcements and reminders of evaluation criteria. This approach was intended to allow observation of the natural trajectory of app usage and to help identify at which points specific interventions might be necessary. In addition, frequent individualized prompting was expected to blur the distinction between the effects of teacher intervention and those of the app or the learners themselves. Continued individual prompting could also create unequal levels of support among students. For these reasons, standardized intervention (i.e., general announcements only)

was adopted in the latter half of the semester to ensure fairness and validity of measurement.

5. Methods

5.1. Data and Data Collection

The data for this study consisted of two sources from 38 participants: (a) app usage logs and (b) responses to three questionnaire surveys.

The app usage logs recorded each participant's completion status for the 10 weekly assignments set during the semester (i.e., Weeks 3, 4, and 6–13). For each learning unit, the start and completion times were provided weekly to the instructor by the app provider. The instructor then coded the completion status as on-time completion (1), delayed completion (2), or non-completion (0).

The questionnaires were administered in class at three time points using the same format: at the course introduction in April, in mid-May, and at the end of June. Using a five-point Likert scale, they measured learners' perceptions of the app's usefulness, perceived burden, acceptance of ASR, and perceived contribution to listening. Open-ended questions asked students to comment on their learning, difficulties, and requests for improvements to the app's design (see Appendix A). Responses were entered on smartphones. There were no missing responses for individual items, although one participant did not complete the third questionnaire. Accordingly, the number of responses at each time point is specified in the descriptive and analytical results.

5.2. Data Analysis

This study is based primarily on descriptive analysis. Because the data come from a convenience sample of a single group ($n = 38$) without randomization or a control group, and were collected in the context of classroom practice, inferential statistics intended for causal inference or population estimation are not appropriate. In addition, the main variables are weekly completion logs and Likert-scale ratings, which include temporal correlation and some missing responses. To avoid mismatches with distributional assumptions and to reduce the risk of multiple comparisons, the analysis is limited to trajectories, distributions, and representative descriptive statistics (M, SD, Mdn, IQR, percentages), together with the coding of open-ended responses. The aim is to clarify practically useful patterns and intervention timings while restraining overgeneralization.

First, to capture changes in engagement, completion rates for each of the 10 weekly assignments (Weeks 3, 4, and 6–13) were calculated: the percentage of on-time completions, delayed completions, and non-completions. These indicators of completion status were then visualized across the semester to identify potential turning points, such as plateaus or sharp declines in the on-time completion rate.

Second, to examine the types of teacher-support necessary for sustaining app usage, questionnaire responses were analyzed at three time points: T1 (April, beginning of the semester), T2 (mid-May), and T3 (end of June). For measures of perceived usefulness, enjoyment, and confidence in listening to academic English, M, SD, Mdn, IQR, and valid n were reported for each time point. Line graphs were used to display the average trajectories from T1 to T3, and dot plots were used to visu-

alize the distributions of change scores ($T2 - T1$, $T3 - T2$). Open-ended responses were inductively coded, and counts and representative examples were presented for themes related to perceived usefulness, burden, acceptance of ASR, and perceived contribution to listening. Co-occurrence patterns of codes were supplemented with simple cross-tabulations, and shifts in themes across time points were described.

After presenting these results, the trajectories of weekly completion rates, the time series of questionnaire ratings, and the patterns of open-ended responses were cross-referenced to identify when interventions were needed and what forms they should take. Specifically, analysis focused on whether declines in completion rates during the middle of the semester coincided with rising reports of burden, plateauing of usefulness, or “signs of stagnation” in open-ended responses (e.g., difficulties in securing study time, distrust of ASR, or giving up on practice). These signals were then examined to consider what types of interventions would be effective at those points. Missing data were reported with valid n for each time point.

6. Results

6.1. Trends from App Usage Logs: Trajectories of Engagement

On-time completion was initially high, peaked in Week 6 (94.7%), and then gradually declined to 55.3% by the final week. Correspondingly, non-completion increased to 23.7% by the end of the semester. Delayed completions partly offset the decline but did not prevent the overall drop (see Table 2 and Figure 1).

Table 2. Weekly completion rates for assigned app-based lessons ($N = 38$)

Week	On Time_n	Delayed_n	Missed_n	OnTime_%	Delayed_%	Missed_%
WK 3	29	9	0	76.32	23.68	0
WK 4	31	7	0	81.58	18.42	0
WK 6	36	2	0	94.74	5.26	0
WK 7	33	5	0	86.84	13.16	0
WK 8	32	5	1	84.21	13.16	2.63
WK 9	25	11	2	65.79	28.95	5.26
WK 10	27	9	2	71.05	23.68	5.26
WK 11	25	9	4	65.79	23.68	10.53
WK 12	25	10	3	65.79	26.32	7.89
WK 13	21	8	9	55.26	21.05	23.68

Note. WK3–WK13 denote the 10 assigned weeks. Percentages are computed as $n/38 \times 100$ each week. “On-Time” = completed before the weekly due date (start of that week’s class); “Delayed” = completed after the due date but before semester end; “Missed” = no completion recorded. Denominators are constant ($N = 38$); no log data were missing.

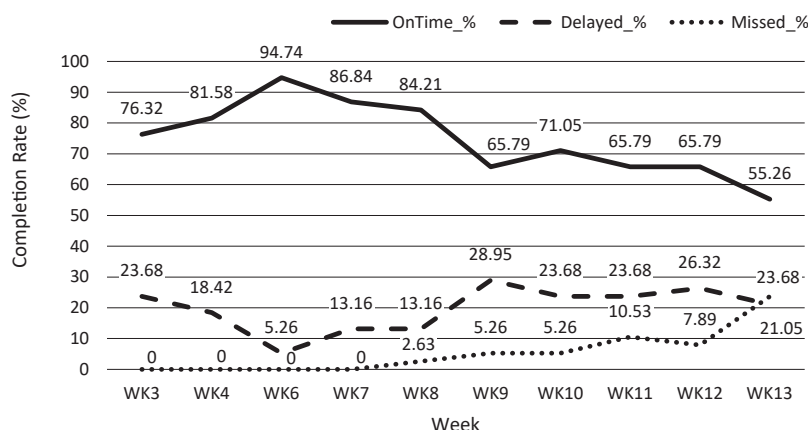


Figure 1. Weekly Completion Rates and Trajectories for Assigned App Lessons Across the Semester

Note. Weekly completion trajectories for app-based lessons across the semester ($N = 38$). The lines show the percentage of participants who completed the assigned lesson on time, completed late, or did not complete; percentage labels are printed at each point. Week labels WK3–WK13 correspond to the 10 assigned weeks; WK5 is omitted because no app assignment was set for that week (see Table 1). Rates are calculated over all 38 participants. “On-Time” = completed before the weekly due date; “Delayed” denotes completion after the weekly deadline; “Missed” = no completion recorded.

6.2. Quantitative Indicators: Perceived Usefulness, Enjoyment, and Confidence in Listening to Academic English

Perceived usefulness remained moderate, enjoyment was stable at a similar level, and confidence in listening to academic content—though low overall—showed a slight upward trend by the end of the semester. These tendencies are illustrated in Figure 2.

Although the gains were modest, a small accumulation toward the end of the semester was evident. As discussed in Chapters 2 and 3, the cycle of repeated practice and immediate feedback targeting phonological awareness (i.e., vowels and consonants, linking and assimilation, stress and intonation) appears to have supported the emergence of self-efficacy even within a short time.

In summary, the evaluation indicators for this semester can be condensed into three points: (a) perceived usefulness remained at a moderate level, (b) enjoyment was stable at a moderate level, and (c) confidence was low but showed a slight increase. The pattern of no rapid gains but no collapse is consistent with prior findings that the effectiveness of mobile-assisted learning depends on design and sustained support (Sung et al., 2015; Mihaylova et al., 2022). In particular, the light accountability mechanism of weekly reminders and grade incentives for completion may have helped prevent declines in learners’ evaluations.

6.3. Open-Ended Responses: Perceived Usefulness, Burden, Acceptance of ASR, and Perceived Contribution to Listening

Open-ended responses were collected at three time points (T1 = introduction, T2 = mid-semester, and T3 = end of semester) and inductively coded into four themes: perceived usefulness,

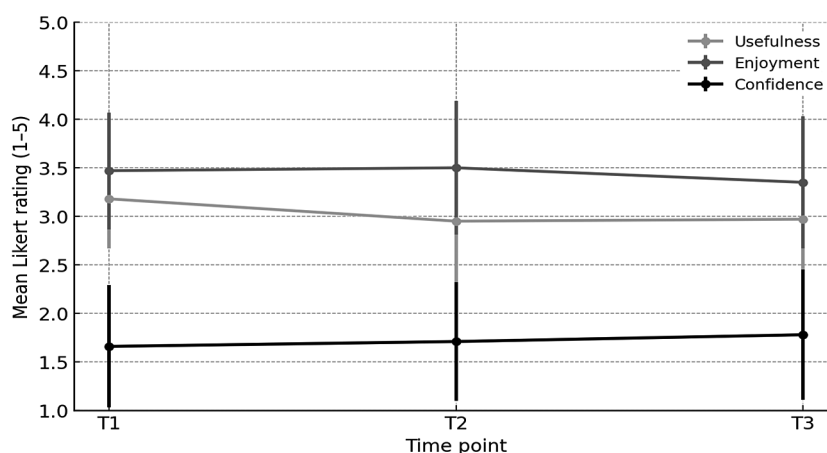


Figure 2. Trends in Perceived Usefulness, Enjoyment, and Confidence (T1-T3)

Note. Mean Likert ratings (1–5) at three time points—T1 (course start), T2 (mid-semester), and T3 (end). Error bars show ± 1 SD. Usefulness, enjoyment, and confidence correspond to composite items drawn from the questionnaire listed in Appendix A. Usefulness: T1 $M = 3.18$ ($SD = 0.51$, $n = 38$), T2 $M = 2.95$ ($SD = 0.66$, $n = 38$), T3 $M = 2.97$ ($SD = 0.73$, $n = 37$). Enjoyment: T1 $M = 3.47$ ($SD = 0.60$), T2 $M = 3.50$ ($SD = 0.69$), T3 $M = 3.35$ ($SD = 0.68$). Confidence in listening to academic content: T1 $M = 1.66$ ($SD = 0.63$), T2 $M = 1.71$ ($SD = 0.61$), T3 $M = 1.78$ ($SD = 0.67$).

perceived burden, acceptance of ASR, and perceived contribution to listening. Overall, usefulness and contribution to listening appeared frequently across the semester. Reports of burden increased in mid-semester at T2 (T1: $n = 39$; T2: $n = 119$; T3: $n = 124$), while acceptance of ASR, though less frequent, appeared consistently with both positive and negative comments (negative ASR comments; T1: $n = 1$; T2: $n = 14$; T3: $n = 13$). Below, typical content for each theme is summarized, with representative excerpts.

For perceived usefulness, students often mentioned that the app allowed short, frequent practice that made it easier to sustain learning, and that it connected with phonological instruction in class. In particular, repeated comments highlighted the cycle of applying what was learned in class—such as linking and stress—in app practice (T1: $n = 25$; T2: $n = 65$; T3: $n = 57$):

T1, P31: “Being able to break practice into short intervals during spare time was helpful. It became easier to pay attention to pronunciation points.”

T2, P11: “The design allowed small completions to accumulate, which made it easier to continue. I could see my progress each week.”

T3, P18: “I was able to notice and apply features such as linking, which we practiced in class, in the app as well. It felt usable.”

For perceived burden, students most often mentioned difficulties in securing time, balancing the app with other courses and assignments, and maintaining motivation toward the end of the semester. Reports of burden increased markedly at T2 (T1: $n = 39$; T2: $n = 119$; T3: $n = 124$). While students understood the intention of distributed practice, some reflected that they ended up completing tasks in batches instead:

T2, P03: “In weeks when other assignments piled up, I tended to put this off.”

T2, P27: “There is value in doing it every week, but continuing it was harder than I expected.”

T3, P12: “Toward the end, I ended up doing several tasks at once, so it wasn’t really distributed.”

Finally, regarding perceived contribution to listening, students reported that using the app promoted awareness of phonological features such as linking, stress, and intonation, and that this awareness carried over to listening comprehension of news and lectures (T1: $n = 2$; T2: $n = 30$; T3: $n = 30$):

T2, P06: “After practicing with attention to linking, it became easier to identify word boundaries even in fast English.”

T2, P15: “When I paid attention to stress and intonation, it was easier to catch the main points.”

T3, P30: “I was able to understand more when listening to news videos than before.”

In summary, the open-ended responses consistently highlighted the connection between perceived usefulness and perceived contribution to listening. They supported the view that phonological instruction in class—such as linking and stress—was reinforced through short, distributed practice with the app, and that this cycle promoted awareness which carried over to listening comprehension. By contrast, reports of burden were most prominent in the middle of the semester, with time management and balancing other coursework identified as the main obstacles to sustained use. Regarding ASR, students expressed both dissatisfaction with misrecognitions and appreciation for the visualization of hard-to-comprehend segments. The findings suggest that when instructors explicitly explain the reasons for errors and model possible reformulations, ASR feedback can be more readily accepted as a resource for self-monitoring. These results are consistent with the quantitative indicators in Section 6.2 (i.e., no substantial gains but stable maintenance of evaluations). From an instructional perspective, they suggest the usefulness of (a) readjusting workload and resetting goals toward the middle of the semester, (b) reiterating the significance of assignments and rescheduling distributed practice, and (c) continuing to provide guidance on how to interpret ASR feedback.

7. Discussion

7.1. Interpretation of Educational Effects

This classroom practice integrated an AI conversation app with course instruction, combining individual support in the first half of the semester with mainly class-wide reminders in the second half. Usage logs showed high engagement at the beginning, a temporary peak following the Week 6 reminder (on-time completion 94.7%), and a decline thereafter (55.3% by the final task; missed submissions 23.7%). Questionnaire results indicated stable perceptions of usefulness and enjoyment (moderate level) and a small increase in confidence for listening to academic content (T1 = 1.66, T3 = 1.78). Open-ended responses confirmed perceived benefits for listening, while burden increased mid-semester and ASR feedback was met with both dissatisfaction and appreciation. Overall, app notifications alone were insufficient; brief, well-timed teacher interventions were critical for sustaining engagement (cf. Mihaylova et al., 2022).

The Week 6 peak illustrates the effectiveness of short reminders and clear evaluation criteria. Later declines suggest the need for additional checkpoints toward the middle and end of the semester. Interventions can remain lightweight: brief whole-class reminders, selective follow-up with learners showing repeated late or missed submissions, and small in-class drills to provide quick success experiences. Such measures support persistence without undermining learner autonomy or fairness.

With respect to acceptance of ASR, students expressed both dissatisfaction—such as “the speech recognition is too strict” or “my speech is not picked up”—and appreciation, noting that it provided “a reference point for segments that are hard to understand.” This duality suggests that instruction should frame ASR not as a grading machine but as a mirror for awareness. A brief in-class demonstration of features that ASR tends to misrecognize (e.g., word-final consonants, reduction of function words, linking, and stress placement), followed by explicit guidance on how to reformulate utterances before returning to the app, can help transform dissatisfaction into actionable cues for correction. In light of this, it may also be effective for instructors themselves to use the app in class to model such demonstrations.

Practical implications for classroom instruction can also be drawn from these findings. In particular, three forms of short intervention appear useful: preparing a minimum completion pack in weeks when students face heavy workloads, so that they can realistically embed tasks into their schedules; incorporating brief phonological drills and immediate retries in class to create small but tangible success experiences; and offering checklist-based progress reviews at mid- and end-semester points. Such measures can help sustain engagement while reinforcing the app’s built-in functions of variability, immediate feedback, and distributed practice.

In summary, while the app’s functions—high-variability input, ASR, and spaced practice—are powerful, instructors still need to mediate its use by adjusting form and workload, reframing meaning during periods of stagnation, and providing visualization of progress and feedback. The peak in achievement at Week 6 symbolizes that short teacher feedback and the restatement of criteria were more effective than app notifications alone. Accordingly, four types of support appear necessary for classroom use of AI conversation apps: (1) placing planned whole-class boosts at key points in the semester, (2) providing rapid and minimal individualized follow-up only for those who need it, (3) briefly demonstrating in class how to reformulate utterances so that ASR functions as a mirror for awareness, and (4) circulating meaning and a sense of achievement by highlighting one point to notice each week and offering small words of praise.

7.2. Recommendations for App Design

Based on the findings of this classroom practice, recommendations can be made not only for how instructors provide educational support but also for how app design can be improved. Four areas emerge as priorities: making ASR feedback more transparent, supporting sustained distributed practice, providing functions for classroom integration, and bridging to academic listening.

With regard to ASR, learners need clearer indications of what was misrecognized and by how much. Highlighting errors at multiple levels, displaying recognition confidence, and allowing instant comparison of model and self-recordings would enable immediate retries and complete the short

formative cycle of listening–speaking–listening within the app. Similarly, decomposing speech into mini-drills that overlay features such as linking, stress, and intonation would reinforce the phonological awareness often noted in learners’ comments.

To sustain distributed practice, apps could automatically divide weekly assignments into short sessions across several days, adjust task load in response to recent performance, and issue timely reminders such as alerts before deadlines or notifications aligned with class schedules. These design features would replicate the Week 6 spike in completion, when whole-class restatements briefly boosted engagement.

Integration with classroom teaching also requires attention. A dashboard for instructors displaying indicators such as streaks, on-time completion rates, and consecutive non-completions would allow teachers to identify students needing support with minimal effort. Exportable logs would further improve the precision of feedback and reporting in both instructional and research contexts.

Finally, providing pathways to academic listening is critical. Supplementary tracks based on news clips or short lectures, annotated with discourse markers and phonological cues, would give learners concrete successes to support their emerging confidence in processing academic content.

Taken together, these recommendations respond directly to the findings of Chapter 6. The temporary peak and later decline in task completion, the coexistence of dissatisfaction and appreciation of ASR, and the recurring importance of phonological awareness. They are not a criticism of the particular app used in this practice but design implications applicable to similar apps more broadly. If such improvements are implemented, instructors will be able to focus on short, targeted interventions, while learners will find it easier to sustain high-frequency practice autonomously.

8. Conclusion and Future Directions

This classroom practice integrated an AI conversation app into the course, with lighter individualized support in the first half (i.e., personal prompts and restatements of evaluation criteria) and a shift to whole-class announcements in the second half. Usage logs and questionnaire data collected across the semester revealed that task completion was high in the early weeks, peaked temporarily at Week 6 following a restatement of evaluation criteria, and then gradually declined. In contrast, learners’ evaluations remained stable. Perceived usefulness and enjoyment were consistently moderate, while confidence in listening to academic content showed a slight upward trend. Open-ended responses frequently mentioned usefulness and contributions to listening, while burden increased mid-semester, and acceptance of ASR reflected a coexistence of both dissatisfaction and appreciation. Overall, the findings suggest that automatic notifications alone are insufficient to sustain momentum; short, targeted teacher interventions at key points play a decisive role.

In AI-supported English learning, instructors function much like gym trainers. They demonstrate how to use the equipment effectively, set appropriate workloads, and correct posture or form at critical moments. This view also aligns with perspectives from educational coaching, which emphasize supporting learners’ self-regulation and sustained engagement rather than constant directive instruction (van Nieuwerburgh, 2017). Accordingly, lightweight interventions—such as restating goals

and deadlines or making temporary workload adjustments—should support learner autonomy. The findings of this study showed that the Week 6 restatement raised on-time completion, while the subsequent reduction in teacher intervention coincided with increases in delays and non-completions. This supports the effectiveness of inserting milestone reminders strategically in the middle and later parts of the semester.

Providing the app alone does not naturally lead to sustained use. However, as demonstrated here, strategically placed lightweight interventions can support continuity and gradually build confidence in academic listening. From a practical standpoint, three layers of intervention appear realistic: (1) form-building in the early weeks, (2) milestone restatements in the middle and later weeks, and (3) brief individualized follow-up for the small number of learners showing declines in achievement.

On the design side, app improvements can further amplify effectiveness in classroom use. Specifically, features such as ASR confidence scoring and N-best candidate displays, visualization of phonological connections, user interfaces that enable immediate retries within short cycles, automated rescheduling of distributed practice through review cues, and teacher dashboards—for early detection of delays, ready-made feedback templates, and one-click adjustment of task difficulty—would be valuable. In addition, a modular structure allowing learners and instructors to select and apply add-ons in response to contextual needs would make classroom implementation more flexible. These suggestions are not criticisms of the app used here; rather, they are design implications drawn from observed performance patterns and learner feedback. They are broadly applicable to similar apps and envision incremental improvements through collaboration between developers and educators.

From a curricular perspective, integrating reflection and portfolios (e.g., recorded speech with notes on phonological awareness) into short, high-frequency tasks could help visualize the cycle of noticing, speaking, and reflecting. Even a single weekly note—such as which sound I corrected and how my perception changed—could promote self-regulation. For advanced learners, upward extensions (e.g., faster input, alternative accents, and expanded vocabulary) could prevent stagnation. Such practices are also transferable to other courses, including presentations, discussions, and EMI classes. For example, inserting short ASR-based feedback and rehearsal sessions before and after oral tasks could strengthen both pronunciation form and content organization.

In the future, three avenues for further investigation emerge. First, quasi-experiments within the university could compare different types of milestone interventions (e.g., restatements only vs. rescheduling plus restatements) and examine the impact of app features such as confidence displays and visualization through A/B testing or cross-app comparisons. Second, longitudinal alignment of usage logs with learner experiences (e.g., open-ended responses and portfolio recordings) could yield early indicators of disengagement or stagnation. Third, tracking learners across semesters and examining transfer effects (e.g., to EMI courses or external tests) could clarify how short-term maintenance of engagement translates into longer-term learning outcomes. AI is not an automatic growth machine, but when instructors act as trainers at critical milestones and apps are designed to sustain learning cycles, the foundations for sustainable and practical English learning can be significantly strengthened.

References

- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. <https://doi.org/10.1037/0033-2909.132.3.354>
- Kang, S. H. K. (2016). Spaced repetition promotes efficient and effective learning. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 12–19. <https://doi.org/10.1177/2372732215624708>
- Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, 138(2), 817–832.
- Mahdi, H. S. (2024). *The effectiveness of high variability phonetic training (HVPT): A meta-analysis* (ERIC Document ED635161).
- Mihaylova, M., Gorin, S., Reber, T. P., & Rothen, N. (2022). A meta-analysis on mobile-assisted language learning applications: Benefits and risks. *Psychologica Belgica*, 62(1), 252–271. <https://doi.org/10.5334/pb.1146>
- SpeakBUDDY Ltd. (2025, September 1). *AI eikaiwa Speak Buddy, Tokuchou* [Characteristics of AI English conversation Speak Buddy]. <https://app.speakbuddy.me/#characteristics>
- Sung, Y.-T., Chang, K.-E., & Yang, J.-M. (2015). How effective are mobile devices for language learning? A meta-analysis. *Educational Research Review*, 16, 68–84.
- Tono, Y. (2017). The CEFR-J and its Impact on English Language Teaching in Japan. *JACET International Convention Selected Papers*, 4, pp. 31–52. JACET.
- van Nieuwerburgh, C. (2017). *Coaching in education: Getting better results for students, educators, and parents*. London: Bloomsbury.
- Yanase, Y. (2024). Reflection on the early implementation of the new E3 Course, Active Listening: Fostering mutual learning and diverse competition among learners. *The Institute for Liberal Arts and Sciences Bulletin, Kyoto University*, 7, 23–69. https://doi.org/10.14989/ILAS_7_23
- Zhang, X. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, 88, 101067. <https://doi.org/10.1016/j.wocn.2021.101067>

Appendix

Questionnaire Items

The questionnaires were administered at three points in the semester: T1 (April, course introduction), T2 (mid-May), and T3 (late June). Unless otherwise noted, items used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Open-ended questions invited written responses.

Core Likert-Scale Items (T1–T3, repeated at all time points)

1. I can understand most English conversations.
 2. I can understand English spoken at natural speed (e.g., in movies or podcasts).
 3. I feel confident in listening to academic content in English.
 4. I can grasp the gist and details of English speech.
 5. My English classes are (or have been) overly focused on grammar.
 6. I have few opportunities to use English for real communication.
 7. I find practical, real-world English learning enjoyable.
 8. I believe listening skills useful for travel, work, or research are necessary.
-

Additional Items (T1 only)

I have used AI tools (e.g., ChatGPT, SpeakBuddy) to study English.

I think AI can help improve my listening ability.

I feel excited about using AI in this class.

What listening skills do you hope to develop in this class?

What are your expectations for AI-based English learning?

Please describe memorable experiences from past English learning.

Additional Items (T2 and T3)

- The AI conversation app helps improve my listening ability.
 - Using the AI app has increased my motivation to study English.
 - The listening skills I wanted to develop in this class have improved.
 - The AI app provides realistic practice similar to actual conversation.
 - Becoming accustomed to AI has reduced my anxiety about speaking English.
 - Using the AI app has helped me notice my weaknesses (e.g., vocabulary, grammar, pronunciation).
 - The AI-assisted class is more effective than other English classes I have taken.
 - Have your expectations for AI-assisted English learning been met? Why or why not?
 - Please describe memorable experiences from using the AI app (positive or negative).
 - What aspects of the app should be improved?
 - How has your listening ability or view of English learning changed through this class?
-

AI 英会話アプリを用いた英語学習の推進： 授業実践を通して見えてきた課題と展望

佐藤 龍一 *

要 旨

本実践は、AI 英会話アプリを大学初年次のリスニング授業に連動させ、個別の促しや全体周知といった教育的介入を行いながら、学習者のアプリ使用の実態を記述・分析した。調査には、アプリ使用ログと三時点質問紙を用い、学習者のアプリに対する評価や学習行動を通時的に分析した。結果、期限内達成は Week 6 で 94.7% に達したのち後半にかけて漸減し、未達成数は期末に 23.7% へ上昇した。一方、学習者の評価は、有用感・楽しさが中程度で安定し、学術的内容を聞く自信は低位ながら若干上向いた ($T1 = 1.66$, $T3 = 1.78$)。自由記述からは、音韻的気づきの有用感、時間確保の負担、ASR への賛否の併存が確認された。以上より、アプリの自動通知だけでは学習の継続は難しく、学期の要所で評価基準の再提示や分量・難易度の微調整といった軽度の介入を取り入れることで、継続を支えつつ学習への自信の積み上げを促せることが示唆された。

[キーワード] AI 英会話アプリ、分散学習、自動音声認識 (ASR)、継続支援

* 京都大学国際高等教育院