Education Tech Research Dev (2021) 69:861–884
https://doi.org/10.1007/s11423-020-09910-1

DEVELOPMENT ARTICLE

Developmental research on an interactive application for language speaking practice using speech recognition technology

Eun Young Oh[1] · **Donggil Song**[2,3]

Accepted: 18 November 2020 / Published online: 11 January 2021

Abstract
This developmental research aims to (1) examine the design and developmental process, (2) investigate the nature and structure of the application, and (3) analyze the results of expert reviews and usability tests. Twenty-five participants, including a developer, an instructional designer, Korean language educators, educational technology researchers, human–computer interaction experts, and language learners, were involved in this study. This study was conducted in the following steps: (1) formulated design principles through the literature review of language instruction and learning theories, computer-assisted language learning, speech recognition technology, human–computer interaction, and scaffolding, (2) developed a functional software prototype that adopted the formulated design principles, (3) conducted expert reviews and learner usability tests, (4) revised and updated the application through the repetitive expert reviews and learner usability tests, (5) analyzed the results of the final expert review, usability test, and log data analysis, and (6) clarified the implications of the development research. The developed application shows an approach to addressing the challenges of second language classrooms that might cause a low-level of learner's language speaking performance. This study specifically delivers knowledge about the design and developmental process of computer-assisted language learning software. This provides guidelines for educational technology researchers and practitioners who work on similar projects.

Keywords Developmental research method · Second language learning · Computer-assisted language learning · Speech recognition technology

✉ Donggil Song
donggil.song@gmail.com

Eun Young Oh
oh@rice.edu

[1] Center for Languages and Intercultural Communication, Rice University, Houston, TX, USA

[2] Instructional Systems Design and Technology, Sam Houston State University, Huntsville, TX, USA

[3] Teacher Education Center, Room 136, 1908 Bobby K Marks Drive, Huntsville, TX 77340, USA

Conversations between the instructor and learners are crucial for improving learners' speaking performance in second language (L2) learning. Through conversations, the instructor can evaluate learners' speaking abilities and offer prescriptive and timely feedback (Brown 2000; Littlewood 1981). In classrooms, however, it is almost impossible for an instructor to provide each individual learner with personalized feedback given high instructor–student ratios. Korean as a Foreign Language (KFL) classrooms have the same issue. Recently, the number of students who take KFL courses has skyrocketed due to the increased popularity of Korean cultures, such as K-pop and K-drama, around the globe. The main purpose of learning KFL is to fluently communicate with people using the Korean language (Kim 2017). However, the improvement of speaking performance has not been successful due to the aforementioned limitation of language classrooms. To tackle the issue, it is needed to utilize an automated way to support computer-learner conversations, which requires speech recognition technology.

There have been numerous positive suggestions that speech recognition technology would improve L2 learners' speaking performance (Michael 2017). However, few studies were conducted to practically utilize speech recognition technology in real situations for L2 speaking practice. The primary purpose of this study is to develop an interactive application using speech recognition technology for KFL speaking practice.

Literature review

Although improving speaking skills is essential in L2 learning, we may not expect the effectiveness of conversational interactions in classrooms, where a large number of learners are allocated to a single instructor (Chang et al. 2010). The lack of interaction in classrooms is a major cause of underperformance in L2 speaking (Petersen 2014). Along with the class size, there is a learner-related issue in this problem. Among L2 learning areas (e.g., listening, reading, writing, grammar), speaking is the component that is largely influenced by the learner's personality, culture, and participation levels (Dalby and Kewley-Port 2013; Woodrow 2006). Introverted students are reluctant and afraid to speak in class, and it has long been reported that these personality and participation issues have a negative impact on their L2 speaking performance (MacIntyre and Gardner 1989). Recent studies show similar results. In Dewaele and colleagues' studies (Dewaele and Al-Saraj 2015; Dewaele and Ip 2013), learning outcomes of students who are afraid to speak are significantly lower than those of other types of learners. Without verbal interaction with the learner, the instructor cannot properly evaluate learners' speaking skills, which leads to a lack of adequate prescriptive instructional strategies (Iwashita et al. 2008). To address this issue, L2 educators have adopted technology tools.

Second language learning and technology

The field of technology use in L2 learning, called Computer-Assisted Language Learning (CALL), was started in the 1990s. Conventionally, CALL research has focused on grammar, listening, reading, and writing (Khezrlou and Ellis 2017; Liu et al. 2002). There have been a few attempts on the use of technology for speaking practice; for example, using audio or video conferencing tools, learners have a chance to talk to different people online (e.g., Comac 2008; Volle 2005). This could be effective for speaking practice, but the practicality of synchronous meetings is still in question (Chun et al. 2016; Zhao 2013).

In the 1990s, speech recognition technology was introduced in language education. It has received attention in the CALL field as speech recognition technology has excellent potential for providing interaction for learners (Levy 2009). In the early days, it was used to help people with dysarthria (Noyes and Frankish 1992) and detect language learners' phonographic errors (Eskenazi 1996). A few years later, Ehsani and Knodt (1998) suggested two types of design guidelines of speech recognition technology in L2 learning: closed and open response designs. In the closed response design, learners' possible responses are embedded in the system in advance; for example, the learner can choose and speak one of the options that a system suggests. The recognition success rate would be stable because the system could compare the learner's speech with the predetermined responses. Even some systems in the 1990s showed over 90% success rate of recognition. On the other hand, the open response design accommodates any responses from the learners. The learners' degree of freedom is high, but recognition accuracy is not higher than that of the closed design. Derwing et al. (2000) evaluated the accuracy of speech recognition software and English native speakers. The software's accuracy was 70.75% for Spanish native speakers' English speaking, 72.45% for Cantonese native speakers' English speaking, and 90.25% for English native speakers' English speaking; and the human native-speaking evaluator showed a 99.7% accuracy. Since then, speech recognition technology has been improved by adopting machine learning algorithms (Deng and Li 2013), and different types of language education research attempts have been made, such as real-time translation and input tools (Shadiev and Huang 2016) and pronunciation training (Arora et al. 2018).

Through these studies, researchers have drawn attention to the possibility that speech recognition technology could be used effectively in L2 education because it could provide more speaking practice opportunities and reduce the learner's anxiety when speaking a language that they are learning. A few studies were conducted on interactive design, scaffolding development, and practical adoption of speech recognition technology. van Doremalen et al. (2016) evaluated a prototype of a language learning system that used speech recognition technology. Their system focuses on Dutch learners' pronunciation, morphology, and syntax exercises. The researchers reported the results of usability tests and expert reviews, which showed positive opinions about the system's performance and user-friendliness. Although they suggested future research, such as more diagnostic exercises and speech detection enhancement, practical implications for the design and development of such systems were not addressed sufficiently. In this field, there is a lack of systematic, prescriptive, and practical principles for designing a language learning support system that utilizes speech recognition technology. Although there have been studies on the implementation of speech recognition-based language learning systems focusing on the affordance of mobile devices (Ahn and Lee 2016), learner engagement (Dalim et al. 2020), and pronunciation (McCrocklin 2019), the use of speech recognition technology for designing communicative, conversational, and interactive learning systems is under-investigated. To support learners' conversations with the system, the CALL field may need to adopt interaction approaches in the field of learning technologies toward personalized and adaptive learning systems based on interactive scaffolding.

Scaffolding for second language speaking practice

Conversational interaction with corrective scaffolding is essential for L2 speaking practice (Petersen 2014). For beginner-level L2 learners, a conversation needs to be initiated by the instructor, the learner responds to it, and the instructor provides immediate

feedback through instant evaluation. To facilitate this scaffolding process, Johnston and Milne (1995) implemented a multimedia tool that offers communicative tasks. Their tool presents segmented video clips (e.g., interviews, dialogs, narratives) for a speaking topic and transcripts and contextualized grammatical explanations for scaffolding resources. The researchers reported that this tool increased students' communication exchanges by scaffolding teacher-student discourse (Johnston and Milne 1995). Although we can assume that the increased communication opportunities might enhance students' speaking skills, the subjects' speaking performance was not measured in their research. Shih (2010) investigated the impact of video-based blogging as a speaking practice task on students' speaking performance. Based on their instructor's and peers' feedback, students revised their speaking videos. It was found that the scaffolding activity was one of the crucial aspects of the students' learning process (Shih 2010). However, not all scaffolds have the same positive effect. Differently-designed scaffolds have a different impact on learners' speaking performance (e.g., Mirahmadi and Alavi 2016). Thus, for effective speaking practice feedback, scaffolding should be meticulously designed (Mesthrie 2008).

Scaffolding was first conceptualized by Wood et al. (1976) as a procedure that helps to focus on a problem and to assist in solving a problem of a child or a novice on its own. Pea (2004) summarized the guidelines for scaffolding. First, the learner must be able to identify a learning problem by themselves even if they cannot solve it on their own. Second, the learner should be aware that there is a way to solve the problem. Finally, as the learner can solve the problem, the scaffolding needs to be slowly faded out. Along with those principles, there are specific guidelines for scaffolding in L2. The most critical issue is the consideration of the conversation context. Hung and Gonzalez (2013) examined the system-learner conversation in context-centric and non-context-centric settings. In the context-centric group, the system provided scaffolds for the learner when their conversation was off-topic, whereas the system had a free conversation with the non-context-centric group. The results show that the context-centric group achieved higher performance than the control group. These results are consistent with the findings of more recent studies (e.g., Afitska 2016). This means the interaction design should be context-focused in L2 speaking practice environments. Still, few studies were conducted on incorporating speaking scaffolds in CALL, specifically when adopting speech recognition technology.

This study

The lack of research in this field is simply because it has not been long since we are able to utilize easily accessible and fully functional speech recognition technology when designing language learning systems. Reigeluth and Karnopp (2013) described three stages in the S-Curve theory by distinguishing between when developmental research with formative evaluation is needed, and when experimental research with summative evaluation of effectiveness is required. In the S-Curve theory (i.e., the graph has the shape of S; the x-axis is time, and the y-axis is productivity), at the beginning stage of an instructional technology system shows a slow increase in productivity, the middle stage shows a pattern of a rapid increase in productivity. Again it shows a trend of gradual increase in the final step. New and immature fields need the knowledge of design and development that can be used in practical settings rather than the knowledge of effectiveness that is acquired from the summative evaluation (Reigeluth and Karnopp 2013).

The topic covered in this study can be seen as an early-to-middle phase according to the S-Curve theory. Although the use of speech recognition technology for language education

was initiated in the 1990s, the field has not matured yet. At this moment, knowledge of efficient design and development is requested to maximize its effectiveness. This is when developmental research is required for the academic field (Richey et al. 2005). Accordingly, rather than examining the effectiveness of a specific design, strategy, or tool, our research question is, "how can we design and develop an interactive learning application to better support language learners' speaking practice using speech recognition technology?".

Methods

We adopted a developmental research method to design an interactive language learning application using speech recognition technology. Developmental research follows a research method that designs and develops instructional products or models focusing on ongoing growth, evolution, and change (Richey et al. 2005). Specifically, developmental research can be defined as "the systematic study of design, development and evaluation processes with the aim of establishing an empirical basis for the creation of instructional and non-instructional products and tools and new or enhanced models that govern their development" (Richey and Klein 2007, p. 1). This method is similar to Design-Based Research (DBR) in that DBR can be adopted to investigate instruction and learning in a specific context through the iterative learning/instructional design process (Design-Based Research Collective 2003). However, the developmental research method focuses more on addressing design and development processes or models than DBR does (Klein 2014). Richey and Klein (2009) divided the developmental research method into *Product & Tool Research* and *Model Research*. We followed the Product & Tool Research method, which includes analysis, design, and development steps followed by try-outs and evaluation processes; after that, the iterative process of modification and re-evaluation is conducted.

Procedure

The first step included needs analysis, learner analysis, content analysis, and context analysis. The research team reached out to a Korean language teaching institution for these analyses. It was found that KFL students are getting more focused on their speaking and communication skills. The relative importance levels are approximately 30% for speaking, 40% for listening, 20% for reading, and 10% for writing. There have been speaking practice activities in the Korean language program, such as instructor–student mock interviews, small group discussions, and student–student pair work. However, the instructors pointed out that there are always a number of students who are afraid of speaking in the classroom, which is consistent with our literature review. Most importantly, instructors mentioned that it is extremely challenging for an instructor to provide each student with individualized feedback for speaking practice. Some instructors have utilized a recording task as a speaking assignment—students are asked to record their speaking practice and submit it through email. The instructors commented that it takes a while to review their students' recordings and provide appropriate scaffolds for them. A few instructors mentioned that they attempted to use some existing tools that adopt speech recognition technology. Still, they indicated that those tools are merely offering speaking opportunities without learner-system interactions or scaffolding. The instructors appreciated that some tools present users' speech waveform, which is beneficial for correcting pronunciation. However, the lack of communicative activities could be a considerable limitation for speaking practice. These

results lead us to this development project's direction, which should be communicative and interactive through real-time scaffolding.

Second, design principles were formulated through intensive literature reviews. Third, we developed an interactive web-based application implementing the formulated design principles. Repetitive expert reviews and usability tests were performed with the rapid prototype process (i.e., the implementation of critical functions, modules, databases, and interfaces prior to full development) (Tripp and Bichelmeyer 1990). Korean language educators, instructional technology researchers, and Human–Computer Interaction (HCI) experts participated in expert reviews. Adult KFL learners participated in usability tests.

Data analysis

Given the nature of developmental research as a holistic process (Richey et al. 2005), we documented the development process from the beginning of this project. We considered this entire process as data collection and analysis, including the needs analysis, literature reviews for design principles, design and revisions, expert reviews, and usability tests. For a more specific analysis approach, expert reviews containing the experience and perspective of experts were analyzed based on the content analysis method. We followed three steps suggested by Johnson and Christensen (2008). We looked for important words or phrases in the reviews and divided them into segments. Then, the relationships between the fragmented contents were established, which were outlined by the coding process. The codes were combined into subcategories, and similar small units were aggregated into a larger theme. A total of 158 codes were combined into 50 subcategories (32 for the first review, 18 for the second). In the first expert review, 12 subcategories were from language educators, 11 from educational technology researchers, and 9 from HCI experts. In the second review, 8 subcategories were from language educators, 7 from educational technology researchers, and 3 from HCI experts. These were combined into three themes: stepwise suggestions, learning content revision, and usability enhancement.

The results of the usability tests were also analyzed, focusing on finding what to be revised based on the participants' feedback and suggestions. Then, the content of observation notes was analyzed and compared with the recorded videos of usability tests. In addition, the log data of the application was quantitatively analyzed. We checked all conversation details, including the users' correct and incorrect answers, focusing on the appropriateness of agent feedback on learners' speaking and the frequency and time information of learners' speaking attempts. The system's speech recognition error rates were calculated, and the learners' response time was analyzed. Finally, an integrated inference was carried out by analyzing all results and comparing the analyses of interviews, usability results, observations, and log data.

Results

The design process

Design principles from literature review

Fields of consideration for literature review for design principles were identified as scaffolding, interaction, and speech recognition technology. An extensive literature search was

conducted to determine the design principles and application guides. Google Scholar, the Academic Search Complete, ERIC Databases, and Research Information Sharing Service (Journal article and dissertation search platform operated by Korea Education and Research Information Service) were queried to search for literature for this study. Titles, abstracts, and keywords were searched for speech-recognition, voice-recognition, language speaking, L2 learning, speaking instruction, Korean language learning, scaffolding in language education, interactive scaffolding, interaction in language education, language learning theory, and computer-assisted language learning. As shown in Table 1, 34 design principles were formulated after the in-depth literature review. These principles were applied and implemented in practice, resulting in the development of the application.

The system framework and application design

For the speech recognition framework, HTML5 Web Speech APIs (Google Developers 2013) was utilized, which are embedded in recent web browsers. The agent's interaction algorithms were designed using server programming languages (Node.js, Typescript). The server stores subject information for conversations, questions to be asked by the agent, correct/incorrect answers in a database system (MongoDB).

The learning content was adopted from an instructional material that has been used at Seoul National University in Seoul, South Korea. The database has learners' possible answers, frequently incorrect answers, and types collected from the repetitive interviews and collaborations with KFL instructors who had more than 10-year teaching experience. Given the beginner level of the target audience, the degree of freedom was not that high; thus, the amount of initial dataset was not huge.

The content design followed the instructional theory of language speaking (Paulston and Bruder 1976): mechanical drill, meaningful drill, and communicative drill. Examples of each drill can be seen in Table 2. The mechanical drill includes learners' repetition, transformation, and application speaking practice with the goal of memorizing new patterns of typical sentences. Although the learner might not fully understand accurate meanings of what they are speaking, the process of repetition, transformation, and application contributes to the improvement of learners' conversation skills (Paulston and Bruder 1976). The meaningful drill includes the practice of structures and syntactic aspects. The learner is expected to concentrate more on meaning rather than on form. The communicative drill teaches the practical use of language for communication focusing "on what is said rather than on how it is said" (Paulston and Bruder 1976, p. 54).

The application was named *Marago*, and the agent *Yuri*. The learner accesses Marago via a web browser. After the access, the learner communicates with the agent. Computer speakers enable learners to listen to the agent's greetings and questions. The learner can check the learning content and objectives on a screen. Depending on the learner's response, the agent provides different types of scaffolds. If the agent recognizes the correct answer, the learner moves forward to the next step and continues conversations with the agent.

The first development process

We applied 34 design principles for building the system. For example, conversation videos (i.e., Principle 8) were presented with a guide script (i.e., Principle 10). The presented video shows an example of the script, and the learner can use different words and short

Table 1 Design principles and application guides

Areas	Types	Design principles	References	Application guides for this project
Scaffolding	Prerequisites	1. Make sure the learner is able to identify a problem by themselves	Pea (2004)	Before the conversation, present the contextual information and ask whether the learner identifies the problem
		2. Monitor the learners' current status	Azevedo et al. (2008), Wood et al. (1976)	Store the learner's responses, correctness, and progress in the database system
	Applications	3. Increase learner engagement	Wood et al. (1976)	Provide interesting and engaging conversational contexts for the learner
		4. Decrease the tasks' degree of freedom		Design the agent's questions that are well-aligned with learning goals
		5. Establish the learner's learning paths towards the learning goal		Present the learning goals and direction
		6. Emphasize the purpose and types of tasks		Provide videos and practical materials that show the purpose and types of tasks
		7. Avoid the learner's frustration		Reduce the difficulty of a task when the learner's speaking keeps having incorrect answers
		8. Demonstrate conversations		Present videos that show authentic conversations
		9. Design the fading of scaffolding as the learner solves a problem	Pea (2004)	Reduce scaffolds or skip grammar explanations as the learner's speaking does not have an error
		10. Provide scaffolds that reduce the learner's cognitive load so they can focus on goal-related tasks	Hmelo-Silver et al. 2007)	Present learning goals from the beginning of a conversation task and design the agent's conversation that is well-aligned with learning goals, which decreases the learner's conversation degree of freedom
	Scaffolding types	11. Conceptual scaffolding: conceptualize the problem situation for the learner to identify what is needed to achieve their learning goal	Hill and Hannafin (2001), Kim and Hannafin (2011)	Present the characteristics and contextual information of each conversation task
		12. Metacognitive scaffolding: support the learner to evaluate their current progress and achievements		Add a function that the learner can check their progress and current status
		13. Procedural scaffolding: support the learner not to waste their time finding learning materials		Present the required information on each task and the basics of each topic during the conversation

Table 1 (continued)

Areas	Types	Design principles	References	Application guides for this project
Interaction	Interaction phases	14. Provide mechanical drills for speaking practice, such as repetition, transformation, and application	Paulston and Bruder (1976)	Design mechanical drill tasks in the first phase of the program
		15. Provide meaningful drills for speaking practice, specifically the structure and syntactic		Design meaningful drill tasks in the second phase of the program
		16. Provide communicative drills so the learner can create their own meanings in authentic situations		Design communicative drill tasks in the third phase of the program
	Interaction guidelines	17. Automaticity: focus on fluency rather than grammar or linguistic forms	Brown (2000)	Focus on the learner's fluency, specifically in the third phase of the program (communicative drills)
		18. Intrinsic motivation: support the learner to be satisfied with their achievement of speaking in authentic situations		Design authentic conversation tasks, provide positive responses, and give different types of scaffolds, depending on the type of incorrect responses from the learner
		19. Strategic investment: support the learner to make a decision on how to solve a conversation problem		Support the learner to handle the situation on their own by decreasing the difficulty of each task and providing the written explanation when speaking keeps having an error
		20. Risk-taking: help the learner overcome their challenge when their meaning was not properly delivered		Provide scaffolds and give the learner the opportunities to retry their speaking when there is an error
		21. Add culture-based conversation components		Design culture-based conversation topics
		22. Add interlanguage components		Design the agent as a conversation partner and feedback provider
		23. Support communicative competence of grammar, conversation, speaking strategy, and social aspects		Design practical conversation topics in authentic language-speaking situations
	Communication	24. Add functional communication components for speaking skills in a specific situation	Littlewood (1981)	Design the first two phases (mechanical and meaningful drills) for functional communications that require the learner to practice specific skills depending on the context
		25. Provide social interaction activities using discussion and role-playing		Design the third phase of the program (communicative drills) for social interaction that includes role-playing

Table 1 (continued)

Areas	Types	Design principles	References	Application guides for this project
Speech recognition technology	System design	26. Follow the phases of the speech recognition process: speech recognition, scoring, error detection, error diagnosis, feedback presentation	Neri et al. (2003)	Develop the internal system using the phases of the speech recognition process, including speech recognition, scoring, error detection, error diagnosis, feedback presentation
	System direction	27. Use appropriate approaches between the closed response design (only receives a certain type of responses) and the open response design (received any types of responses), depending on the learning context	Ehsani and Knodt (1998)	Adopt the closed response design from the beginning and make a transition to the open response design as the learner enhances their speaking performance
		28. Provide an adequate amount of speaking practice opportunities	Eskenazi (1999)	Include different types of conversation topics and questions so the learner can practice their speaking
		29. Provide pertinent corrective feedback		Present corrective feedback on the screen through the conversation with the agent
		30. Emphasize speech components (i.e., amplitude, duration, and pitch)		Provide videos that include native speakers' actual conversation examples
		31. Provide different types of native speakers' speaking styles		Design different types of agents, including the conversation partner, restaurant server, store cashier, and taxi driver
		32. Make the learner comfortable with the speaking environment		Give enough opportunities to retry their speaking when there is an error in their sentences, so the learner does not feel fear in speaking situations
		33. Minimize speech recognition errors	Yun (2014)	Avoid short-answer questions that might increase speech recognition errors and lead the conversation that asks a long sentence within the speaking context
		34. Provide personalized speaking practice opportunities		Design adaptive learning paths, depending on the learner's performance

Table 2 Language speaking drill type and examples

Drill types	Examples*
Mechanical drill	Answer the following question using the "am going to" pattern and suggested words [Question] What are you going to do? (a book, buy) [Expected Answer] I am going to buy a book
Meaningful drill	Answer the following question [Question] What do you want? [Answer] I want ( )
Communicative drill	Answer the following question [Situation: Post Office] [Question] How may I help you? [Answer] ( )

*Translated from the Korean language

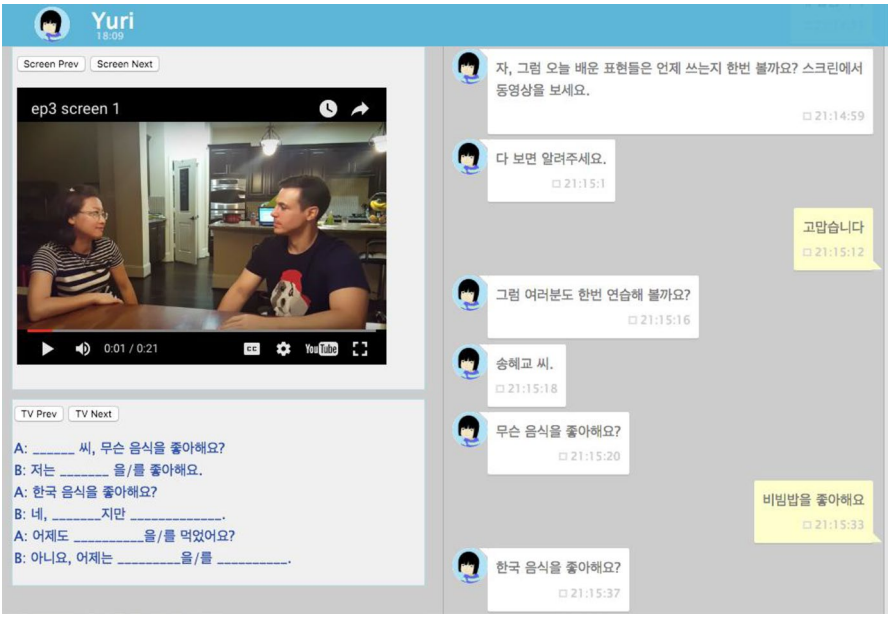

Fig. 1 Videos and guide scripts: design principles 8 and 10

sentences within this guide script, which was designed to reduce the learner's cognitive load (see Fig. 1).

First expert review

The first expert review was conducted with three groups of experts: seven Korean language educators, three educational technology researchers, and three HCI experts. The Korean language educators had masters and doctoral degrees in their field and had taught the

Korean language for more than 10 years. We adopted a focus group interview format for 3 h, including a short presentation of the research aim and context and the Marago system, and questioning sessions. The reviewers pointed out the following aspects. First, the system should adopt more individual learning supports because this tool can be mainly used for supporting individual practice, which could not be done in classrooms. Second, conversation designs should be more natural, specifically, the communicative drill tasks must use authentic and natural conversation contexts. Third, learning goals and content should be more simplified, given the target audience's levels. Specifically, more practice for mechanical drills is needed. Fourth, more engaging aspects need to be included, such as showing scores and giving rewards. Fifth, a final review function is needed so the learner can review their performance at the end of the program.

The educational technology researchers (3 professors in the United States) earned doctoral degrees and had more than 8 years of research experience in their field. Three suggestions were noted. First, there are design principles that were not clearly applied in the system. We revised the design for clarification. For example, for Principle 1 (Make sure the learner is able to identify a problem by themselves), prior to each conversation task, the system presents the contextual information of the upcoming conversation topic, and the agent asks whether the learner can handle the conversational situation or not. For Principle 9 (Design the fading of scaffolding as the learner solves a problem), the system reduces scaffolds or skips grammar explanations as the learner's speaking does not have an error, which is automatically monitored by the system. Second, instructional design and learning theories may need to be adopted, such as motivation theories, cognitive load theory, metacognition, and collaborative learning theories. Third, learners' affective aspects need to be considered; for example, remembering the learner's name and preferences and mentioning them during the conversation could be beneficial for emotional relationships.

Three HCI professors in the United States earned doctoral degrees and had more than 10 years of research experience in their field. Their reviews included the following aspects. First, the system needs to adopt interactive interface designs; for example, the agent's facial expression may need to vary rather than the one image. Second, more engaging aspects need to be included, such as the content that is more relevant to the learner (e.g., the language-specific cultural content). Third, the system needs to support social media aspects, such as a sharing feature, which could be a type of interaction rewards. The experts pointed out that current students usually share their learning activities by posting them on their social media pages (e.g., Instagram, Facebook, Twitter). Thus, Marago adopted social media share buttons so the learner can post the scripts of their conversation with the agent and speaking results on their social media pages.

First usability test

The first prototype was tested by five learners (i.e., Korean language-learning international students at Seoul National University) individually. We observed the learners when they were using Marago, and they were asked to think aloud. Five observations were noted. First, all of them showed positive opinions on the use of Marago. They like the agent's instant feedback and the system's speech recognition accuracy. All of them completed the learning steps without further assistance. Second, the system needs to give more options so the learner can control more components, such as a re-listening button when the learner could not understand a specific part or a redo option when the learner wants to speak a particular part again. Third, the display is somewhat distracting the learner because, images,

videos, learning goals, and chatting were presented on one screen. Fourth, it would be great if it had a male agent because they could learn male accents, intonations, and expressions. Fifth, shortcut key functions need to be included, such as a spacebar for speaking.

The second development process

Based on the results of the first expert review and usability test, the system was modified. We followed Richey's et al. (2014) guidelines when dealing with two issues. First, we identified the challenging aspects due to time, financial, and technological limitations and declined the impossible suggestions. For example, the Korean language experts requested the agent's natural intonations and accents; however, we could not enhance the speech quality of the current HTML5 Web Speech APIs.

We also resolved the conflict opinions between expert groups. We invited all expert groups in one place and discussed the conflicts, such as individual learning (language educators) vs. collaborative learning (educational technology researchers), mechanical and meaningful drill-focused practice (language educators) vs. communication drill-focused practice (educational technology researchers), and multi-agent approach (e.g., a conversation partner agent and a teacher agent; language educators) vs. one-agent approach (educational technology researchers). Through the discussion between groups, we reached out to the consensus on most issues; for example, the individual learning approach is appropriate given the target audience, communication drills might cause a higher degree of freedom that the system may not be able to handle without quality big data, and one-agent approach would be effective to reduce the learner's cognitive load. However, a couple of issues could not be solved, such as rewards (e.g., points, badges, balloons) vs. no-rewards (because the target audience is adult learners). For these issues, we asked ten students who were taking a Korean language course at Seoul National University to make a decision. Finally, the reward opinion was adopted; the system shows students' scores and ranking, and presents digital badges to high performers. From this process, the second prototype was developed.

Second expert reviews

The second expert review (with the same expert groups as in the first review) was conducted. After the expert groups checked whether the previous suggestions were applied, they conducted a detailed review using Marago. Three issues were identified by the Korean language educators. First, the content of each conversation topic needs to be adjusted, considering the difficulty levels. Besides, the presentation of text feedback should be enhanced, such as using red-color or underlines to give corrective feedback. Second, the system needs to include a learner evaluation function and provide the learner with the evaluation results for their additional practice. Third, the system can be used as a homework or formal evaluation tool for face-to-face courses; for example, an option for sending the learner's results to the instructor via email would be useful.

The educational technology researchers pointed out that the conversation context needs to be more relevant and authentic (i.e., Constructivism), such as a situation when the learner stops by a restaurant in South Korea to order something from the menu. Second, since a learning goal includes the learner's pronunciation improvement, the system needs to have a pronunciation practice module. Third, the learner-friendly aspects should be enhanced, such as a short orientation for using the application prior to the conversation.

The HCI experts mentioned that more cultural aspects need to be included because, in HCI research, cultural contents have been shown to increase the user's motivation. In addition, the interaction design needs to be more authentic; for example, the agent's facial expression changes need to be aligned with the conversation context.

Second usability test

The second usability test was conducted with the same participants. The results can be summarized into three aspects. First, it needs a function that the learner can hide or show the agent's text. This is because, when focusing on listening, the text might distract the learner. Second, more control options for the learner are needed, such as a task-skip button. Third, the system needs to include more real-life conversation situations.

Final development

Following the results of the second expert review and usability test, the final development was conducted. Most suggestions were applied in the final development, but a few suggestions were declined. For example, for the pronunciation practice function with sound wave graphics, the system needs to include additional technology tools with updated security options on the server-side. Given the limited budget and timeframe, these aspects were not covered in the final development.

Learning steps and the internal process of the application

The learning steps are shown in Fig. 2. In Marago, there are five learning steps. First, the learner selects their tutor and begins mechanical practice: Steps 1-A and 1-B. The learner is asked a variety of questions and is expected to answer them verbally. If the answer was incorrect, they would receive appropriate scaffolds, depending on the type of their errors. There are five types of predetermined errors—postpositional particle, vocabulary, tense, inflection, and unexpected errors. When a learner's speaking consisted of a postpositional particle error, the incorrect particle would be changed to red color, which is the first scaffold. If the learner made the same mistake, the agent would provide a grammatical explanation for the use of postpositional particles (i.e., the second scaffold). The third scaffold for the same mistake includes more detailed information of the postpositional particle with practical examples. There are more than 20 questions in this step, but when the learner speaks correctly 8 times in a row, they can skip the rest of the questions and move forward. In Step 2, the learner can choose a conversation topic among the food, fruit, and transportation options, which will be the conversation topics for the next steps. Topic videos are presented, and the learner practices conversations using the guide script used in the video. In Step 3, the learner participates in a role-play conversation depending on their chosen topic. For example, if the selected topic is transportation, the learner has a conversation with a taxi driver to get to the given destination. In Step 4, the learner learns more about Korean culture-based conversations. Finally, in Step 5, the learner reviews their performance, including their errors during the conversation in previous steps, error types, learning hours, and scores.

The internal structure of Marago is described in Fig. 3. As can be seen in the left-top box in Fig. 3, the application starts when a learner accesses through a Web browser. After the introduction and the learner's first response, Step 1 begins. When the learner responds

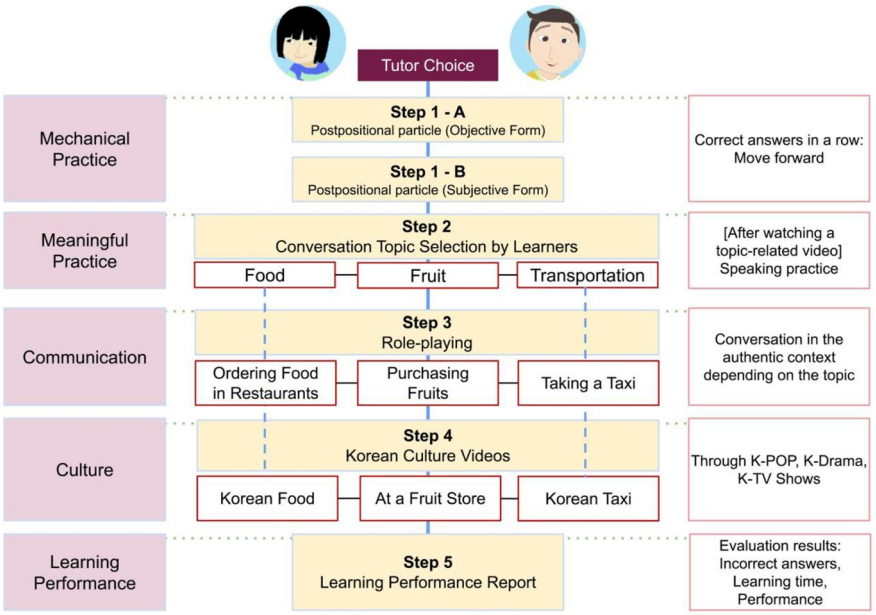

Fig. 2 Learning steps of the final product

to an agent's question, the system checks whether the response is correct. The learner moves forward if their spoken response is correct while they receive appropriate feedback from the agent and get another chance to respond correctly if their response is incorrect. This interactional information is saved in the database.

A development model

Along with the speech practice application, a development model is the product of this research (see Fig. 4): an interactive language learning application development model. We thoroughly followed the developmental research guidelines and suggestions. This procedural model includes the repetitive process of analysis, design principle formulation, content/interaction/motivation/scaffolding designs, expert reviews, usability tests, evaluation, and revision.

Final evaluation

Final expert review and usability test

The final product was reviewed by the expert groups. They confirmed that the final application was developed firmly upon the design principles and was modified based on the previous reviews. The educational technology and HCI expert groups were interested in the future use of Marago, and how much the application will be effective for speaking practice. The Korean language educator group expected that Marago would be used for supporting face-to-face courses, such as for homework and speaking evaluation/exam. In addition,

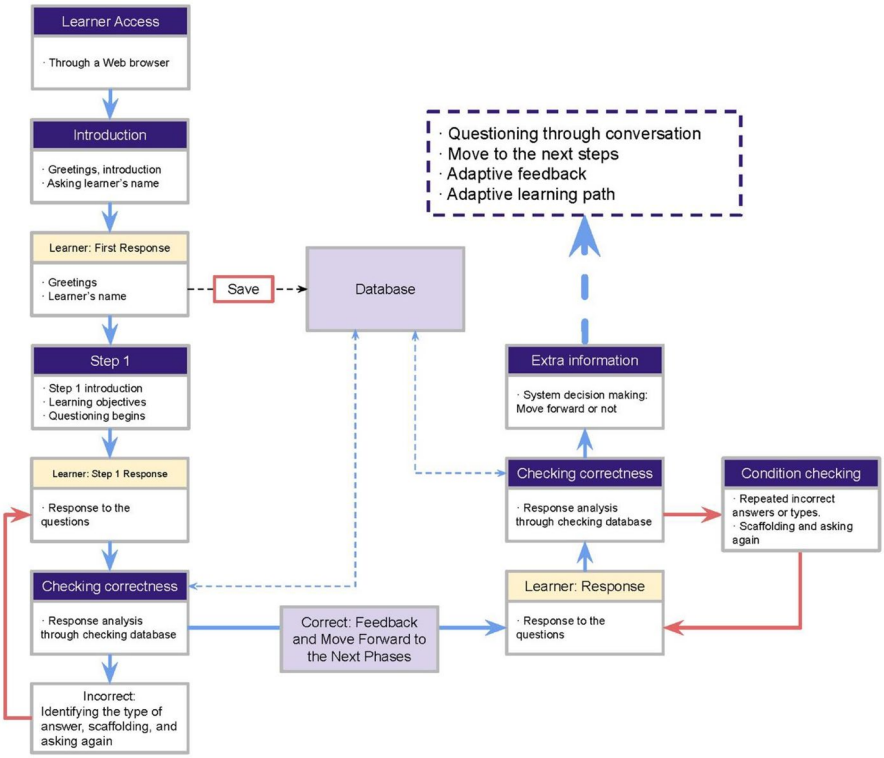

Fig. 3 The internal process of the final product

the final usability test was conducted with ten learners (five learners who participated in the previous tests and five new learners). All learners were able to successfully complete all steps of Marago without extra assistance. The participants who previously evaluated Marago confirmed that their suggestions were applied in the final development. The new learners focused on their learning when the agent responded adaptively.

Log data analysis: speech recognition error rate

To check the accuracy of the speech recognition module, we calculated speech recognition error rates using the concept of Word Error Rate (WER), which uses *Levenshtein distance* to show the difference between the actual speech by the user and the recognition by the system (Fiscus 1997). WER can be calculated as the sum of substitutions (e.g., *pace* is recognized as *face*), insertions (e.g., *SAT* is recognized as *essay tea*), and deletions (e.g., *how it works* is recognized as *how works*) is divided by the number of words spoken. However, the WER algorithm was built for English and is not well aligned with the Korean language. Also, we were not able to find any speech recognition error rate formula for the Korean language. Thus, considering the characteristics and nature of the Korean language's alphabet system, we modified WER using Korean language syllables, which can be calculated as the number of error syllables divided by all syllables.

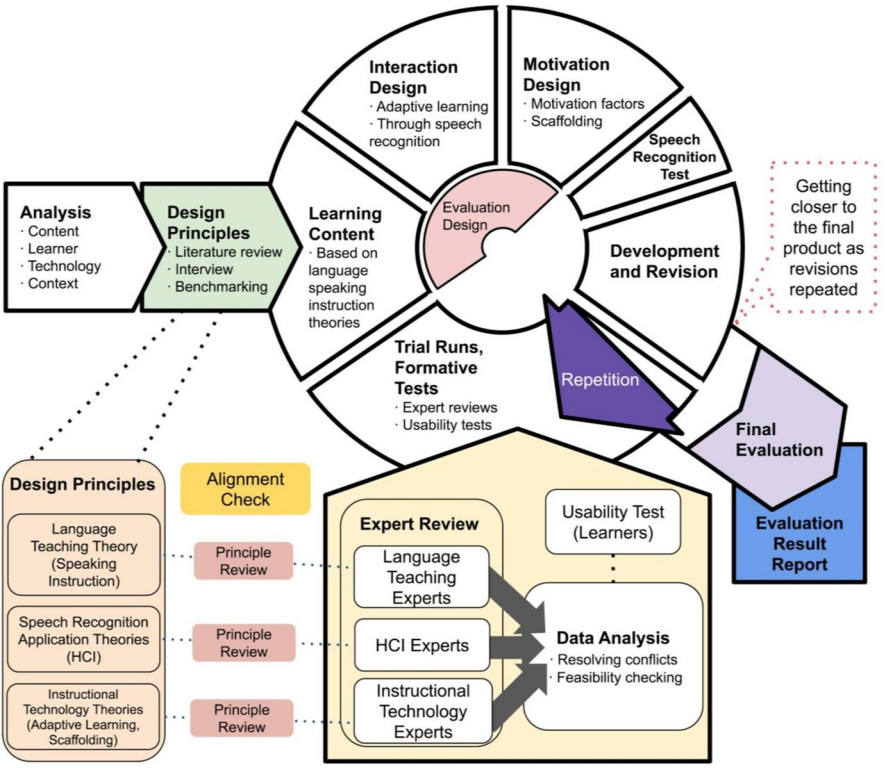

Fig. 4 An interactive language learning application development model

Table 3 Speech recognition error rates using the Korean language syllables in the final usability test

	Step 1-A	Step 1-B	Step 2	Step 3	Total
Erroneous syllables per learner. Mean (SD)	11.4 (9.42)	12.0 (5.96)	6.4 (4.38)	2.9 (2.77)	32.7 (14.17)
All syllables per learner. Mean (SD)	93.4 (20.91)	90 (16.15)	61.6 (5.34)	23.4 (2.95)	268.4 (31.42)
Error rate	12.21%	13.33%	10.39%	12.39%	12.18%

Two Korean language native speakers created a script using the video recorded in the final usability test. They transcribed what they recognized as it is. This transcript was compared with the log data script that was stored in the system. Overall, the error rate is 12.18%, as shown in Table 3. We acknowledge that this value cannot be directly compared with WER due to the different formulas, but we conducted a tentative comparison. Given the current values of reported WER in speech recognition technology, 10–18% (Negri et al. 2014) and 6–11% (Shannon 2017), we consider that 12.18% is acceptable for a speaking practice tool. There are no decreasing trends in the error rate given the similarity of

Table 4 Learners' response time during the conversation with the agent: mean (SD)—seconds

	Step 1-A	Step 1-B	Step 2	Step 3	Total
Development participants (N = 5)	10.2 (1.32)	10.1 (1.51)	14.5 (1.00)	13.3 (1.58)	10.9 (2.04)
Final evaluation only (N = 5)	13.1 (3.90)	10.2 (1.46)	14.2 (1.63)	12.4 (2.15)	12.1 (3.11)
Total (N = 10)	11.6 (3.22)	10.2 (1.47)	14.25 (1.35)	12.8 (1.94)	11.5 (2.70)

error rates between each step. We also found that there is a significant difference in the number of all syllables between learners. This is because some of them were able to skip the mechanical drills when their speaking was correct multiple times in a row. Besides the standard deviation of erroneous syllables is quite large. This might be because of a few learners' strong English accents when speaking the Korean language. Although the native Korean language speakers were able to recognize their accents correctly, the system could not. In addition, it seems that the external noise might be another reason for the large standard deviation.

Log data analysis: response time

To check whether the learner's conversation with the agent is natural, we measured the learner's response time using the log data. Due to the system setting, we measured the response time as the interval between the start point of the agent's speaking and the end of the learner's speaking; thus, the response time includes the agent's and the learner's speaking time along with the actual interval. Given the short lengths of the agent's and learners' speech (beginner-level conversations; approximately 2–3 s), we can estimate the intervals between the learners and between the steps. As shown in Table 4, the overall response time is 11.5 s, which means the actual interval might be 5.5–7.5 s.

As can be seen in Fig. 5, there is a difference between the participants who experienced Marago during the development phase and the new participants who used Marago only in the final usability test. However, this gap was closed in the middle of Step 1-A, which means the new learners needed some time to adjust themselves in using this system.

Discussion

Following the developmental research method, we reported the development process of an interactive language speaking practice application using speech recognition technology. We would like to discuss the following lessons learned through this research.

During the design and development process, we found two crucial aspects that helped our development. First, for expert reviews and usability tests, we adopted a rapid prototyping approach, which has shown its effectiveness and efficiency. In software development, it is also called rapid application development (Beynon-Davies et al. 1999). Although the software prototype was not a complete application, it was enough to show the reviewers and the learners the content and functions. We were able to reduce the development time and expenses due to the rapid prototype process. We argue that the rapid prototyping process

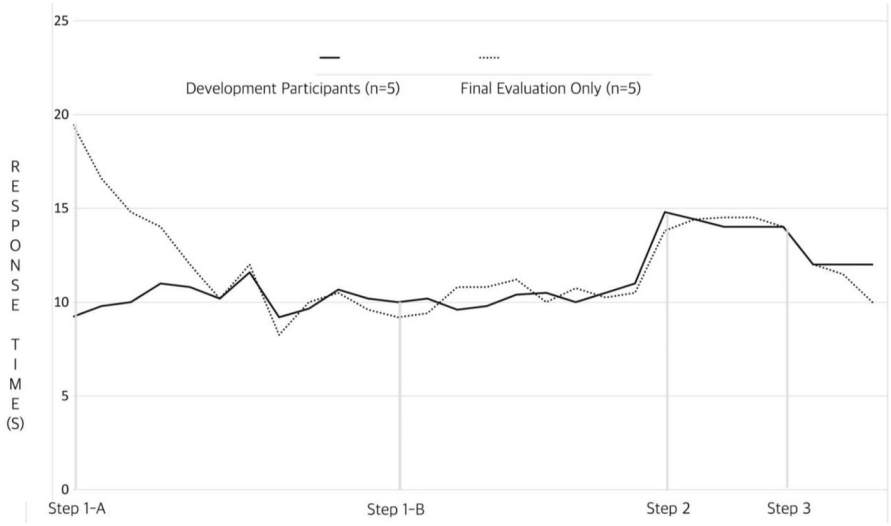

Fig. 5 Learners' response time during the conversation with the agent

could be included as an essential component in the developmental research method, specifically Product & Tool Research.

Second, the discussion process for resolving the conflicts between and within expert groups was beneficial for the quality product. Although it took time for the extra discussion, we were able to avoid unnecessary disputes and contradictions. Through the resolving discussion, the expert groups could understand the goal of the development better and reach a consensus toward an effective product. For this process, we argue that the research team should share the goals of a project and design principles clearly. The team should lead the negotiation and adjustment process to reach a consensus. During the negotiation process, it is significant to check the learning goals in each step to find a better solution. In addition, developers and technicians might need to be present in the discussion because, in many cases, expert groups' suggestions are related to the development process directly; and some of them might take a while, require an extra budget, or cannot be implemented due to technical issues. In these cases, developers and technicians can share their technological knowledge, information, and limitations of the current technology.

In this study, we also found the following challenges when using speech recognition technology for a language learning support system. First, recent speech recognition engines might be, ironically, detrimental to the learner's pronunciation improvement. We utilized Google's speech recognition engine, which has been improved through up-to-date machine learning techniques. When using the engine, we found two issues. First, the system recognized the learner's incorrect pronunciation correctly. For example, when a learner's pronunciation was *Kalbi* (i.e., the incorrect pronunciation of *Galbi*; beef/pork ribs in the Korean language), the system recognized it correctly as *Galbi*. The engine automatically and *intelligently* corrected some incorrect pronunciations. This is a significant and user-friendly development for better human–computer interaction. However, there is no way for us to provide corrective feedback for learners when the system recognizes incorrect speech correctly. We discussed this issue with the Korean language educator group, and they expressed concerns about intelligent speech recognition engines. Interestingly, some

of them showed a different perspective. They argue that since the engine is built on the actual conversation of native speakers. If the system intelligently recognized it correctly, the learner would not face any issues when communicating with native speakers. That means, if the system recognized something correctly, native speakers would be able to recognize it correctly. If the goal of practice is to improve the learner's fluent speaking, this issue might not be a problem. However, if the goal of L2 practice is precise pronunciation, we need to find and develop a new speech recognition engine that recognizes the user's speech as it is rather than the engine that is (overly) intelligent.

The other issue is similar but opposite one; the engine recognizes correct words differently or incorrectly. A similar situation was reported in a previous study (van Doremalen et al. 2016). For example, when the learner said her name, *Jeon Ji Yeon*, many times, the system recognized it as *Jeon Ji Hyeon*, which is the name of a famous Korean actress. It seems that the engine has been trained with big data that possibly contains lots of celebrities' names. Similarly, there were many cases that the system recognized differently, specifically some popular words and celebrities' names. These two issues should be further discussed in the field of language education when using speech recognition technology.

Limitations and future research

The main limitation of this research is that the learning context is limited within the field of Korean language speaking practice for beginner-level adult learners. Thus, the results of this developmental research might not be applicable to the other L2 contexts and other levels or age groups. However, as the developmental research method suggests, this research brought design and development issues to consider when designing a language speaking practice program using speech recognition technology. To address these issues and produce more generalizable knowledge, more developmental research approaches with different target learners for different languages would be needed. In addition, experimental research to evaluate the application's effectiveness is required. Finally, further educational technology research on the use of speech recognition technology as a learner-system interaction method is needed.

Conclusion

We reported the development process of an interactive language speaking practice application, including learner/context analysis, literature-based design principles, the development process, expert reviews, usability tests, the application's internal structure, and final evaluation results. Speech recognition technology has been adopted in many real-life devices as a voice-user interface. Given the need for authentic speaking practice in language education, an effective and natural communication approach between devices and humans could be utilized as a voice-user interface form, which requires reliable speech recognition technology. The field of instructional/learning technologies should be able to produce beneficial knowledge to the public regarding how to design and develop a learning support tool when using speech recognition technology. This study shows an initial step for this request.

Acknowledgements We would like to express our sincere gratitude to Dr. Ilju Rha (Seoul National University) for his guidance and direction during this research. Appreciation is also due to the expert groups for their endless support during the expert review process. We also extend our gratitude to the journal reviewers for their constructive feedback.

Author contributions EYO mainly designed the research and analyzed the data. DS mainly developed the system, supervised the implementation, and collected the data.

Funding This study is not funded by any agency.

Data availability The data will not be shared due to the confidentiality issues.

Compliance with ethical standards

Conflict of interest No potential conflict of interest was reported by the authors.

References

Afitska, O. (2016). Scaffolding learning: developing materials to support the learning of science and language by non-native English-speaking students. *Innovation in Language Learning and Teaching, 10*(2), 75–89. https://doi.org/10.1080/17501229.2015.1090993.

Ahn, T. Y., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology, 47*(4), 778–786. https://doi.org/10.1111/bjet.12354.

Arora, V., Lahiri, A., & Reetz, H. (2018). Phonological feature-based speech recognition system for pronunciation training in non-native language learning. *The Journal of the Acoustical Society of America, 143*(1), 98–108. https://doi.org/10.1121/1.5017834.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-facilitated regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56*(1), 45–72. https://doi.org/10.1007/s11423-007-9067-0.

Beynon-Davies, P., Carne, C., Mackay, H., & Tudhope, D. (1999). Rapid application development (RAD): An empirical review. *European Journal of Information Systems, 8*(3), 211–223. https://doi.org/10.1057/palgrave.ejis.3000325.

Brown, H. D. (2000). *Teaching by principles: An interactive approach to language pedagogy* (2nd ed.). San Francisco: Pearson ESL.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Educational Technology & Society, 13*(2), 13–24. https://www.jstor.org/stable/pdf/jeductechsoci.13.2.13.pdf.

Chun, D., Smith, B., & Kern, R. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal, 100*(1), 64–80. https://doi.org/10.1111/modl.12302.

Comac, L. (2008). Using audioblogs to assist English language learning. *Computer Assisted Language Learning, 21*(2), 181–198. https://doi.org/10.1080/09588220801943775.

Dalby, J., & Kewley-Port, D. (2013). Explicit pronunciation training using automatic speech recognition technology. *CALICO Journal, 16*(3), 425–445. https://www.jstor.org/stable/24147851.

Dalim, C. S. C., Sunar, M. S., Dey, A., & Billinghurst, M. (2020). Using augmented reality with speech input for non-native children's language learning. *International Journal of Human-Computer Studies, 134,* 44–64. https://doi.org/10.1016/j.ijhcs.2019.10.002.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing, 21*(5), 1060–1089. https://doi.org/10.1109/TASL.2013.2244083.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech?. *TESOL Quarterly, 34*(3), 592–603. https://www.jstor.org/stable/3587748.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8. https://doi.org/10.3102/0013189X032001005.

Dewaele, J. M., & Ip, T. S. (2013). The link between foreign language classroom anxiety, second language tolerance of ambiguity and self-rated English proficiency among Chinese learners. *Studies in Second Language Learning and Teaching, 3*(1), 47–66. https://www.ceeol.com/search/article-detail?id=189138.

Dewaele, J. M., & Al-Saraj, T. M. (2015). Foreign language classroom anxiety of Arab learners of English: The effect of personality, linguistic and sociobiographical variables. *Studies in Second Language Learning and Teaching, 5*(2), 205–228. https://www.ceeol.com/search/article-detail?id=297379.

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning & Technology, 2*(1), 45–60. https://www.learntechlib.org/p/90912/.

Eskenazi, M. (1996). Detection of foreign speakers' pronunciation errors for second language training-preliminary results. *Proceedings of the Fourth International Conference on Spoken Language* (Vol. 3, pp. 1465–1468). IEEE. https://doi.org/10.1109/ICSLP.1996.607892.

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology, 2*(2), 62–76. https://eric.ed.gov/?id=EJ577631.

Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). *Proceedings of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 347–354). IEEE.

Google Developers. (2013). Voice Driven Web Apps: Introduction to the Web Speech API. Retrieved from https://developers.google.com/web/updates/2013/01/Voice-Driven-Web-Apps-Introduction-to-the-Web-Speech-API?hl=en.

Hill, J. R., & Hannafin, M. J. (2001). Teaching and learning in digital environments: The resurgence of resource-based learning. *Educational Technology Research and Development, 49*(3), 37–52. https://doi.org/10.1007/BF02504914.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107. https://doi.org/10.1080/00461520701263368.

Hung, V. C., & Gonzalez, A. J. (2013). Context-centric speech-based human–computer interaction. *International Journal of Intelligent Systems, 28*(10), 1010–1037. https://doi.org/10.1002/int.21614.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. https://doi.org/10.1093/applin/amm017.

Johnson, R. B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches* (3rd ed.). Thousand Oaks, CA: Sage.

Johnston, J., & Milne, L. (1995). Scaffolding second language communicative discourse with teacher-controlled multimedia. *Foreign Language Annals, 28*(3), 315–329. https://doi.org/10.1111/j.1944-9720.1995.tb00801.x.

Khezrlou, S., & Ellis, R. (2017). Effects of computer-assisted glosses on EFL learners' vocabulary acquisition and reading comprehension in three learning conditions. *System, 65,* 104–116. https://doi.org/10.1016/j.system.2017.01.009.

Kim, J. I. (2017). Immigrant adolescents investing in Korean heritage language: exploring motivation, identities, and capital. *Canadian Modern Language Review, 73*(2), 183–207. https://doi.org/10.3138/cmlr.3334.

Kim, M. C., & Hannafin, M. J. (2011). Scaffolding problem solving in technology-enhanced learning environments (TELEs): Bridging research and theory with practice. *Computers & Education, 56*(2), 403–417. https://doi.org/10.1016/j.compedu.2010.08.024.

Klein, J. D. (2014). *Design and development research: A rose by another name*. Paper presented at the American Educational Research Association annual meeting, Philadelphia, PA.

Levy, M. (2009). Technologies in use for second language learning. *The Modern Language Journal, 93*(1), 769–782. https://doi.org/10.1111/j.1540-4781.2009.00972.x.

Liu, M., Moore, Z., Graham, L., & Lee, S. (2002). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990–2000. *Journal of Research on Technology in Education, 34*(3), 250–273. https://doi.org/10.1080/15391523.2002.10782348.

Littlewood, W. (1981). *Communicative language teaching: An introduction*. Cambridge: Cambridge University Press.

MacIntyre, P. D., & Gardner, R. C. (1989). Anxiety and second-language learning: Toward a theoretical clarification. *Language Learning, 39*(2), 251–275. https://doi.org/10.1111/j.1467-1770.1989.tb00423.x.

McCrocklin, S. (2019). ASR-based dictation practice for second language pronunciation improvement. *Journal of Second Language Pronunciation, 5*(1), 98–118. https://doi.org/10.1075/jslp.16034.mcc.

Mesthrie, R. (2008). Sociolinguistics and sociology of language. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp. 66–82). Malden, MA: Wiley.

Michael, C. (2017). Automated speech recognition in language learning: Potential models, benefits and impact. *Training, Language and Culture, 1*(1), 46–61. https://doi.org/10.29366/2017tlc.1.1.3.

Mirahmadi, S. H., & Alavi, S. M. (2016). The role of traditional and virtual scaffolding in developing speaking ability of Iranian EFL learners. *International Journal of English Linguistics, 6*(2), 43–56. https://doi.org/10.5539/ijel.v6n2p43.

Negri, M., Turchi, M., de Souza, J. G., & Falavigna, D. (2014). Quality estimation for automatic speech recognition. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1813–1823). https://www.aclweb.org/anthology/C14-1171.

Neri, A., Cucchiarini, C., & Strik W. (2003). Automatic speech recognition for second language learning: How and why it actually works. *Proceedings of the 15th international Conference on Phonetic Sciences* (pp. 1157–1160), Barcelona, Spain. https://www.internationalphoneticassociation.org/icphs/icphs2003.

Noyes, J., & Frankish, C. (1992). Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication, 8*(4), 297–303. https://doi.org/10.1080/07434619212331276333.

Paulston, C. B., & Bruder, M. N. (1976). *Teaching English as a second language: Techniques and procedures*. Cambridge, MA: Winthrop.

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13*(3), 423–451. https://doi.org/10.1207/s15327809jls1303_6.

Petersen, K. B. (2014). Learning theories and skills in online second language teaching and learning: Dilemmas and challenges. *Journal of the International Society for Teacher Education, 18*(2), 41–51. https://eric.ed.gov/?id=EJ1087588.

Reigeluth, C. M., & Karnopp, J. R. (2013). *Reinventing schools: It's time to break the mold*. Lanham, MD: Rowman & Littlefield.

Richey, R. C., & Klein, J. D. (2007). *Design and development research: Methods, strategies and issues*. Mahwah, NJ: Lawrence Erlbaum Associates.

Richey, R. C., & Klein, J. D. (2009). *Design and development research: Methods, strategies and issues*. New York: Routledge.

Richey, R. C., & Klein, J. D. (2014). Design and development research. In J. Spector, M. Merrill, J. Elen, & M. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 141–150). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-3185-5_12.

Richey, R. C., Klein, J. D., & Nelson, W. A. (2005). Developmental research: Studies of instructional design and development. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (2nd ed., pp. 1099–1130). New York: Lawrence Erlbaum Associates.

Shadiev, R., & Huang, Y. M. (2016). Facilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation. *Computers & Education, 98,* 130–141. https://doi.org/10.1016/j.compedu.2016.03.013.

Shannon, M. (2017). Optimizing expected word error rate via sampling for speech recognition. *Proceedings of Interspeech 2017* (pp. 3537–3541). https://doi.org/10.21437/Interspeech.2017-639.

Shih, R. C. (2010). Blended learning using video-based blogs: Public speaking for English as a second language students. *Australasian Journal of Educational Technology, 26*(6), 883–897. https://doi.org/10.14742/ajet.1048.

Tripp, S. D., & Bichelmeyer, B. (1990). Rapid prototyping: An alternative instructional design strategy. *Educational Technology Research and Development, 38*(1), 31–44. https://doi.org/10.1007/BF02298246.

van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning, 29*(4), 833–851. https://doi.org/10.1080/09588221.2016.1167090.

Volle, L. M. (2005). Analysing oral skills in a voice e-mail and online interviews. *Language Learning & Technology, 9*(3), 146–163. https://www.learntechlib.org/p/74474/.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x.

Woodrow, L. (2006). Anxiety and speaking English as a second language. *Regional Language Centre Journal, 37*(3), 308–328. https://doi.org/10.1177/0033688206071315.

Yun, J. (2014). *Analysis of Google Voice Actions' recognition of English word pronunciations by Korean young learners of English for the purpose of developing an English teaching assistant robot* [Unpublished master's thesis]. Kyungpook National University.

Zhao, Y. (2013). Recent developments in technology and language learning: A literature review and meta-analysis. *CALICO Journal, 21*(1), 7–27. https://www.jstor.org/stable/24149478.

Eun Young Oh is Lecturer of Center for Languages and Intercultural Communication, School of Humanities at Rice University. Dr. Oh earned a doctorate in Educational Technology from Seoul National University, South Korea. Her primary research focuses on computer-supported language learning environments, in particular, how to facilitate second language learning through the use of technology, such as virtual reality and speech recognition technology. Dr. Oh was Adjunct Faculty at Sam Houston State University (Instructional Systems Design and Technology), Indiana University (East Asian Languages & Cultures), and Seoul National University (Language Education Institute).

Donggil Song is Assistant Professor and Doctoral Director of Instructional Systems Design and Technology at Sam Houston State University. His lab (Einbrain Lab, www.einbrain.com) focuses on the use of artificial intelligence in education, learning analytics, adaptive learning systems, and self-regulated learning. His primary research includes the applications of virtual conversational systems in online learning environments. He holds a Ph.D. in Instructional Systems Technology from Indiana University, and an M.S. in Computer Science and Engineering and a B.A. in Religious Studies from Seoul National University (SNU), and also completed a master's program in Cognitive Science at SNU. Presently, he serves as the Managing Editor of The International Journal of Multiple Research Approaches.