Interrupting for Microlearning: Understanding Perceptions and Interruptibility of Proactive Conversational Microlearning Services

Minyeong Kim[*]
myk.kim.mai@hai.kangwon.ac.kr
Kangwon National University
Chuncheon, South Korea

Jiwook Lee[*]
jiwook.lee@hai.kangwon.ac.kr
Kangwon National University
Chuncheon, South Korea

Youngji Koh
youngji@kaist.ac.kr
KAIST
Daejeon, South Korea

Chanhee Lee
chanhee015@kaist.ac.kr
KAIST
Daejeon, South Korea

Uichin Lee[†]
uclee@kaist.ac.kr
KAIST
Daejeon, South Korea

Auk Kim[†]
kimauk@kangwon.ac.kr
Kangwon National University
Chuncheon, South Korea

ABSTRACT

Significant investment of time and effort for language learning has prompted a growing interest in microlearning. While microlearning requires frequent participation in 3-to-10-minute learning sessions, the recent widespread of smart speakers in homes presents an opportunity to expand learning opportunities by proactively providing microlearning in daily life. However, such proactive provision can distract users. Despite the extensive research on proactive smart speakers and their opportune moments for proactive interactions, our understanding of opportune moments for more-than-one-minute interactions remains limited. This study aims to understand user perceptions and opportune moments for more-than-one-minute microlearning using proactive smart speakers at home. We first developed a proactive microlearning service through six pilot studies (n=29), and then conducted a three-week field study (n=28). We identified the key contextual factors relevant to opportune moments for microlearning of various durations, and discussed the design implications for proactive conversational microlearning services at home.

CCS CONCEPTS

• **Human-centered computing → User interface management systems**; **Ubiquitous and mobile computing**.

KEYWORDS

Smart Speakers, Microlearning, Conversational Interaction, Opportune Moment, Interruptibility

ACM Reference Format:
Minyeong Kim, Jiwook Lee, Youngji Koh, Chanhee Lee, Uichin Lee, and Auk Kim. 2024. Interrupting for Microlearning: Understanding Perceptions and Interruptibility of Proactive Conversational Microlearning Services. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642778

1 INTRODUCTION

Language learning requires substantial investment in time and effort [54]. Such demand poses challenges for individuals who struggle to spend extra time on learning, often leading to difficulties in maintaining consistent and sustainable learning progress [18]. Thereby, microlearning is a promising solution [9, 19, 52]. While for successful microlearning, learners need to participate frequently and repetitively [18, 59], the popularity of smart speakers in homes provides a new opportunity to seamlessly integrate bite-sized learning tasks into daily routines by proactively providing them with smart speakers in daily life.

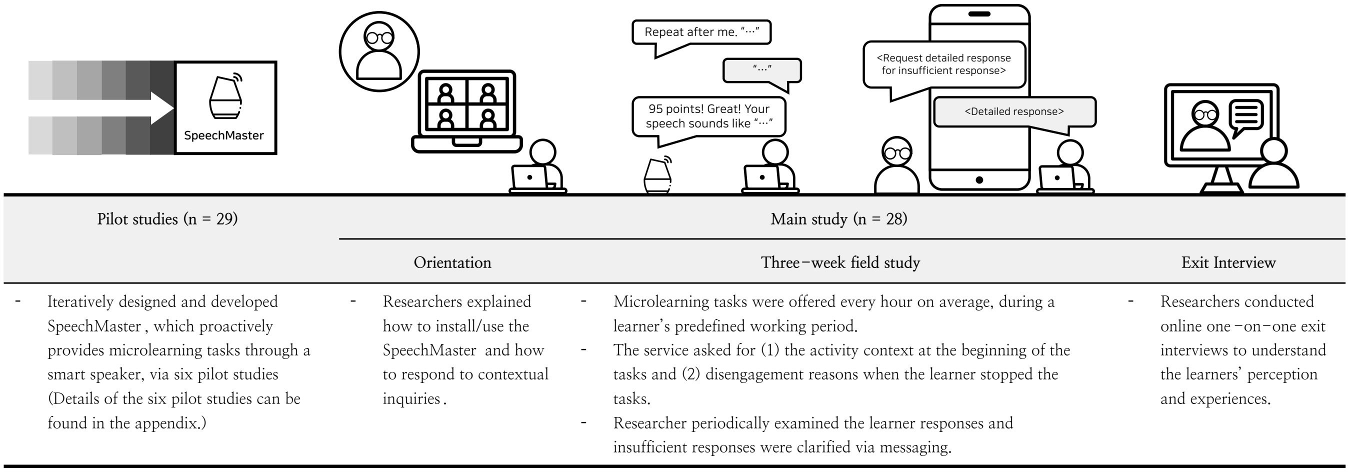
Smart speakers are typically installed in locations where users can interact easily, and they can interact with their users through verbal conversations [6, 60]. People can simultaneously perform two tasks when they involve different modalities (e.g., visual–manual tasks vs. auditory–verbal tasks) [62]. Because daily domestic tasks primarily require visual–manual operations, auditory-verbal (or conversational) microlearning can be integrated into daily domestic routines. However, proactive microlearning tasks at inappropriate times can distract users and negatively impact their experiences, leading to unfavorable outcomes (e.g., inducing stress and annoyance [25]).

Researchers in human–computer interaction are actively investigating opportune moments in which negative interruptions can be minimized [5, 25, 66]. Knowing the opportune moments at which learners are most likely to engage in microlearning tasks could minimize problems with delivery timing. To identify opportune moments, researchers have measured interruptibility or quantified the quality of being interruptible (e.g., being able to engage in microlearning) in a given context [5, 25, 66]. Prior studies have mainly considered visual–manual interactions with smartphones (e.g., notifications) [38, 46, 47, 64]. Recently, numerous studies have considered proactive conversational interactions with smart speakers in domestic settings [10, 51, 60, 61, 65].

[*] Equal contribution.
[†] Corresponding authors.

This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642778


SpeechMaster
Repeat after me. "…"
"…"
95 points! Great! Your speech sounds like "…"
<Request detailed response for insufficient response>
<Detailed response>

Pilot studies (n = 29)	Main study (n = 28)		
	Orientation	Three−week field study	Exit Interview
- Iteratively designed and developed SpeechMaster , which proactively provides microlearning tasks through a smart speaker, via six pilot studies (Details of the six pilot studies can be found in the appendix.)	- Researchers explained how to install/use the SpeechMaster and how to respond to contextual inquiries .	- Microlearning tasks were offered every hour on average, during a learner's predefined working period. - The service asked for (1) the activity context at the beginning of the tasks and (2) disengagement reasons when the learner stopped the tasks. - Researcher periodically examined the learner responses and insufficient responses were clarified via messaging.	- Researchers conducted online one−on−one exit interviews to understand the learners' perception and experiences.

Figure 1: Study procedure.

Although prior studies provide important insights, the perceptions and opportune moments of more-than-one-minute interactions with smart speakers remain underexplored. Prior studies have considered short interactions (typically less than a minute) with a smart speaker [10, 60, 61]. Further, only a hypothetical scenario of probing interruptibility by conducting an experience-sampling method survey (e.g., asking participants to respond to questionnaires) was considered, lacking useful and practical application contexts [10, 60]. Because microlearning tasks require 3–10 min of interaction with learners[2, 45], their findings are not directly applicable to proactive conversational microlearning tasks in domestic settings.

In this study, we explored user perceptions and opportune moments for a more-than-one-minute microlearning interaction with a smart speaker in domestic settings and considered three research questions: (RQ1) how learners perceive and experience proactive conversational microlearning services, (RQ2) when and for how long learners are likely to have opportune moments to engage in proactive conversational microlearning services, and (RQ3) what contextual factors are relevant to opportune moments for proactive conversational microlearning services. A single microlearning session is generally 3–10 min long for optimized learning outcomes. To identify opportune moments for microlearning, we considered its duration. We conceptualize interruptibility as the possibility of learners engaging in microlearning for a given duration.

As shown in Figure 1, through six pilot studies (n = 29), we iteratively designed and developed SpeechMaster, which proactively provides conversational microlearning tasks at random times using a smart speaker. We considered a speech-shadowing task as the learning material. To investigate when and for how long learners can engage in such microlearning tasks, we designed it to be feasible for them to dynamically control the duration of their learning sessions (e.g., stop their ongoing microlearning session at any point).

To answer the three research questions, we conducted a three-week field study with 28 learners interested in improving their English pronunciation, followed by one-on-one exit interviews. We deployed SpeechMaster in their homes for interaction log data. Our qualitative results show that while proactive microlearning tasks increase learning opportunities, learners may perceive them as less distracting when their content is relevant to them. Using the collected data, we quantitatively analyzed when and for how long learners could engage with SpeechMaster across activity, spatial, and temporal contexts prior to their engagement. Although the proactive provision of microlearning tasks at random times was non-distracting half of the time (49%), our learners did not take the learning sessions. This highlights the significance of delivering microlearning at opportune moments, as this timing can maximize the number of daily learning tasks in which learners participate. Our quantitative results show that contextual factors (i.e., activity and spatial contexts) relevant to opportune moments vary by task duration. Based on our findings, we propose design implications for proactive conversational microlearning services and enhancing learning experiences with such services.

2 RELATED WORK

2.1 Microlearning Services

Microlearning delivers several bite-sized learning tasks by breaking learning content into shorter, more manageable chunks [18]. The typical duration of a microlearning session is 3–10 min [2, 45], but it can vary from one second to exceed an hour [24]. For successful microlearning, learners must participate frequently and repetitively [18, 59]. Prior studies on microlearning have offered various domains of second language learning, such as learning vocabulary [12], expressions [16], grammar [14], listening skills [15] and pronunciation [39, 43].

Integrating bite-size learning practices into activities in mobile or computer environments can increase learning moments. For example, Trusty and Truong [58] developed ALOE, a browser extension that integrates microlearning into learners' everyday web browsing experiences by replacing English words with second language words on web pages. Dingler et al. [12] created Quicklearn, a mobile application offering microlearning through interactive push notifications. These studies considered microlearning in mobile or computer environments. However, outside mobile or computer environments, there could be more learning moments to discover.

In studies exploring interactions outside mobile or computer environments, smart speakers are commonly used [10, 51, 60, 61, 65]. They are typically positioned in fixed locations and can interact with the user through verbal conversations. Studies have shown that smart speakers are ideal platforms for educational purposes [13, 53]. Dizon et al. investigated the experiences and perceptions of interacting with a smart speaker in a second language to learn the language and found that the speaker can facilitate out-of-class self-directed language learning. Various auditory learning materials (e.g., 10 min in the English Podcast series [37]) are deliverable through smart speakers. Several edTech companies offer conversational microlearning content for commercial smart speakers (e.g., Amazon Alexa). For instance, 'Daily Dose by Innovative Language' from Amazon Alexa offers eight-minute audio lessons for learning 34 languages [33].

2.2 Opportune Moments and Contextual Factors

Delivering microlearning tasks at inappropriate moments can negatively impact user experiences. They can disrupt ongoing tasks [5], induce stress and annoyance [25], and increase human error and task-completion time [4]. Fortunately, problems with delivery timing can be minimized if we know the opportune moments at which learners are most likely to engage in microlearning tasks. To identify opportune moments, studies have measured and predicted interruptibility (the extent of being interruptible) and showed that opportune moments are closely associated with users' contextual factors, such as activity, temporal, and spatial contexts [46, 47, 64].

Activity context is a critical determining factor for opportune moments. For example, interruptions are inappropriate when users are engaged in high mental-workload (or concentration) activities, such as studying or working [10], driving [29, 30], and biking [38]. Conversely, lower mental-workload activities, such as gaming, internet surfing, or smartphone usage for leisure purposes, are considered opportune moments [10, 41]. Temporal context is another popular context associated with opportune moments. Pielot et al. [47] found that temporal context was the most important contextual factor in identifying opportune moments. Studies have also explored opportune moments in spatial contexts. Studies have considered how locations outside the home (e.g., office, restaurant, and vehicle) or indoor locations within the home impact opportune moments. For example, Nagel et al. [41] explored how contextual factors, including locations within the home (e.g., kitchen or family room), are associated with opportune moments for interactions with mobile phones, and found that location can be useful in determining such opportune moments.

2.3 Proactive Smart Speaker

To minimize the unfavorable outcomes of proactive delivery of microlearning (e.g., resumption lag [3]), it is important to understand the opportune moments for proactive conversational microlearning tasks in home environments.

Although studies have focused on visual–manual interactions with smartphones [38, 46, 47, 64], an increasing number of recent studies have investigated opportune moments for auditory–verbal (or speech) interactions in home environments [10, 51, 60, 61, 65]. For instance, Cha et al. [10] explored opportune moments for proactive smart speakers, which proactively and verbally request their users to answer a single question ("Is now a good time to talk?") for yes or no, including descriptions of their current activities. Similarly, Wei et al. [60] considered multiturn-taking conversations with proactive smart speakers. For the conversations, they provided a 4-item survey, which took 45 seconds to complete. The survey comprised three five-scale questions about availability, boredom, and mood, in addition to one question about the current activity ("What are you currently doing?").

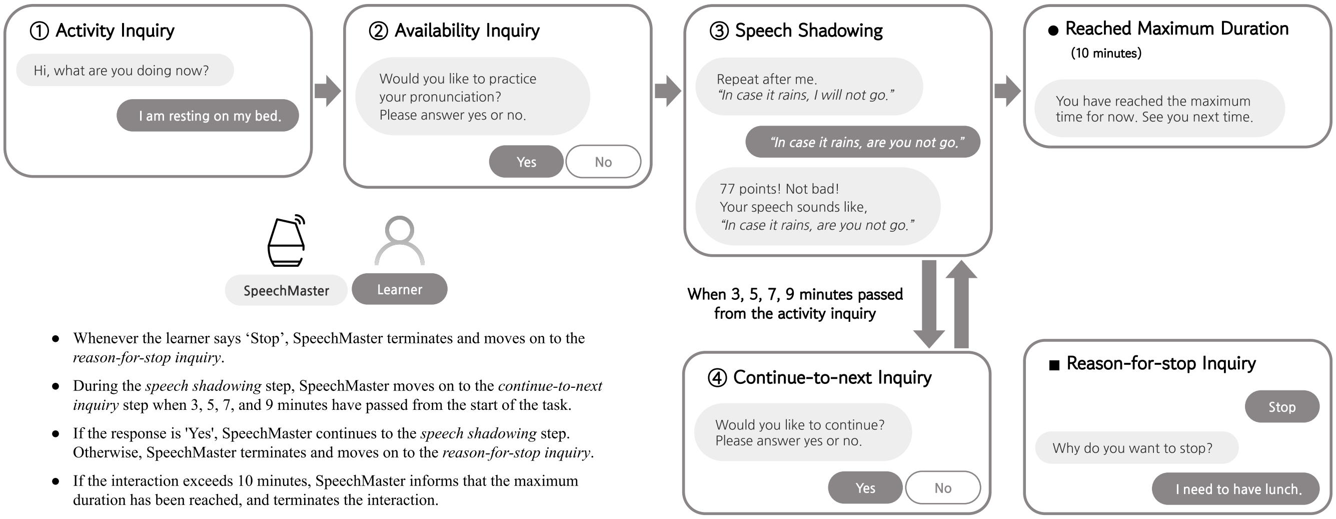
These studies provide valuable insights into the opportune moments and user perceptions of proactive smart speakers. However, their findings are not directly applicable to opportune moments in proactive conversational microlearning tasks at home. While such microlearning tasks require 3–10 min interaction [2, 45], prior studies have considered opportune moments for short interactions ($\leq$ 1 min). They only considered a hypothetical scenario of probing interruptibility by conducting an experience-sampling method (ESM) survey, which lacked utility contexts [10, 60]. ESM is a data-collection method in which participants are prompted to complete a survey at various intervals throughout the day [11]. The provision of a realistic service is explorable because it can affect user engagement and perception. Fischer et al. [17] examined content and delivery time on interruptibility using mobile SMS and found that interest, entertainment, relevance, and actionability of the content affect interruptibility.

3 ITERATIVE DEVELOPMENT OF SPEECHMASTER

We iteratively designed and developed SpeechMaster, a proactive conversational microlearning service, based on six pilot studies (n = 29). We implemented SpeechMaster in a commercial smart speaker (Google Nest Mini) and developed a speaker add-on device and related applications (triggering app, sensing app, and image deleter app) that enable the speaker to operate proactively. The procedure for each pilot study was similar to that of the main study. We conducted an online orientation and then a field study in the learners' homes. After the field study, semi-structured interviews were conducted with learners to gather in-depth feedback. The details of the six pilot studies (e.g., the procedure and important findings) can be found in Appendix A. Herein, we describe the final design of our service.

3.1 Conversational Microlearning Service

Our service provides a microlearning task consisting of five steps (see Figure 2): (1) *activity inquiry*, (2) *availability inquiry*, (3) *speech shadowing*, (4) *continue-to-next inquiry*, and (5) *reason-for-stop inquiry*. For the learning material, we considered English speech shadowing. To investigate how long learners can engage in microlearning, they must dynamically control the learning session duration (e.g., immediately stop their ongoing microlearning session). There exist diverse auditory microlearning materials for learning secondary languages. Because of the short single microlearning session (i.e., 3–10 min), microlearning materials convey a single piece


① Activity Inquiry
Hi, what are you doing now?
I am resting on my bed.
② Availability Inquiry
Would you like to practice your pronunciation? Please answer yes or no.
Yes
No
③ Speech Shadowing
Repeat after me. "In case it rains, I will not go."
"In case it rains, are you not go."
77 points! Not bad! Your speech sounds like, "In case it rains, are you not go."
● Reached Maximum Duration (10 minutes)
You have reached the maximum time for now. See you next time.
SpeechMaster
Learner
When 3, 5, 7, 9 minutes passed from the activity inquiry
④ Continue-to-next Inquiry
Would you like to continue? Please answer yes or no.
Yes
No
■ Reason-for-stop Inquiry
Stop
Why do you want to stop?
I need to have lunch.

- Whenever the learner says 'Stop', SpeechMaster terminates and moves on to the *reason-for-stop inquiry*.
- During the *speech shadowing* step, SpeechMaster moves on to the *continue-to-next inquiry* step when 3, 5, 7, and 9 minutes have passed from the start of the task.
- If the response is 'Yes', SpeechMaster continues to the *speech shadowing* step. Otherwise, SpeechMaster terminates and moves on to the *reason-for-stop inquiry*.
- If the interaction exceeds 10 minutes, SpeechMaster informs that the maximum duration has been reached, and terminates the interaction.

Figure 2: The procedure of SpeechMaster (proactive conversational microlearning service).

of information, knowledge, or concept per session (e.g., learning switch statements in Java for a single session). This limits learners from dynamically controlling the session duration. Therefore, we used a speech-shadowing task where learners performed speech shadowing on English sentences successively. This enables learners to stop their ongoing microlearning sessions immediately after completing the shadowing of a given English sentence.

To prevent a single microlearning task becoming excessively long for optimized learning [2, 45], our service limits the duration of the microlearning task to 10 min. Further, learners can end their current task by saying the *"Stop"* command. After being stopped, the service asks for the reason (i.e., reason-for-stop inquiry step; for details, see Section 3.1). SpeechMaster logs conversations with learners (e.g., activity inquiry responses). Each step is further explained below.

Activity Inquiry: Proactive speakers require a starter (or auditory cue) to indicate the start of their conversation with learners [10, 60]. The starter can take the form of a simple earcon (e.g., soft alarm sound [10]) or an utterance (e.g., greeting [20]). We considered the starter as a greeting with an activity inquiry, with the intention of collecting learners' contextual information before the conversation.

Availability Inquiry: The service then asks whether the learner wants to engage in speech shadowing. If the learner responds *"No,"* it moves to the reason-for-stop inquiry step. Otherwise, it proceeds to the speech shadowing step.

Speech Shadowing: Here, learners perform speech shadowing. SpeechMaster reads an English sentence and waits for the learner's response (or repetition of the sentence). Subsequently, it analyzes the pronunciation accuracy and offers the following sentence after providing the following feedback: (1) score (0−100), (2) compliment/encouragement, and (3) recognized pronunciation. The difficulty level of the sentences increases when the average score of the most recent 20 sentences reaches 90 points or higher. Difficulty levels are categorized based on the number of syllables and spaces within a sentence [31, 44]. At higher difficulty levels, the sentences comprise more words with longer syllables.

Continue-to-next Inquiry: Before reaching the maximum 10 min, SpeechMaster provides a continue-to-next inquiry every 2 min after learners engage in microlearning for a minimum of 3 min (i.e., providing the inquiry after 3, 5, 7, and 9 min from the start). The inquiry is given so that the learner knows how long they had been engaging in a microlearning task and to choose whether to continue learning. When the learner responds *"Yes,"* it continues to the speech shadowing step. Otherwise, it proceeds to the reason-for-stop inquiry inquiry step. On reaching the maximum duration, the service concludes.

Instead of the continue-to-next inquiry, we could also provide an auditory cue (e.g., soft alarm sound) at a certain time interval (e.g., every two minutes) to help learners be aware of their learning time, and let the learners stop the current task by saying a *"stop"* command. However, in this approach, it was difficult to distinguish whether the learners intended to terminate the task. This was because sometimes the speaker misinterpreted a learner's repeated speech shadowing sentence as a command and terminated it regardless of the learner's intention. We further reviewed this termination issue and discussed its implications in Section 8.5.2.

Reason-for-stop Inquiry: SpeechMaster asks for the reason to stop. Reason-for-stop inquiry is provided when learners say *"No"* to continue-to-next inquiry, or whenever learners say *"Stop."*

No Input Handling: When learners do not respond within eight seconds, SpeechMaster repeats the original prompt. For example, the continue-to-next inquiry step asks again whether to continue learning. Similarly, in the speech shadowing step, if learners do not repeat a given English sentence, the sentence is read out again. If learners do not respond three consecutive times, approximately 35 s, SpeechMaster self-terminates without proceeding to the reason-for-stop inquiry step.

3.2 Speaker Add-on Device

Currently, commercial smart speakers are still reactive. To make existing smart speakers to be proactive, we developed a speaker add-on device that can be attached directly on the smart speaker. It comprises a 3D-printed circular frame, an amplifier, and an external 3 W speaker, and is connected to a smartphone (see Figure 3). It secretly delivers prerecorded voice commands from the smartphone to the speaker, and the speaker is activated by listening to the voice commands.

3.3 Proactive Microlearning Triggering

We developed an application that triggers SpeechMaster by delivering prerecorded commands to a smart speaker via an add-on device. SpeechMaster is triggered randomly between 30 and 90 min (average of 60 min) within the operating hours set by the learner at home. Operating hours are set separately for weekdays and weekends.

Home Presence Detection: The triggering app periodically detects the home presence of the learners via a sensing app installed on the learners' smartphone. The sensing app analyze the MAC address of WiFi signals around the smartphone and checks whether a home WiFi signal is detected.

Voice-command Volume Adjustment and Retriggering: The app dynamically adjusts the volume level of voice commands based on the ambient noise level measured two minutes prior to providing the microlearning task. For example, when the noise level is high, the volume level is increased to ensure reliable triggering. While there was an option to start with a high volume, we were concerned that this might diminish the experience of proactive microlearning. If not triggered, the application attempts to retrigger it three times. We introduced this retriggering feature because even with the automatic volume-adjustment feature, triggering sometimes fails owing to a sudden increase in ambient noise.

Recovery Mode: When conversation with the service is terminated unintentionally, learners can resume the conversation by pressing the restart button on the smartphone screen. In our pilot studies, the unintentional terminations happened when the learners' repeated speech shadowing sentences include words (e.g., *"stop," "out," "bye"*) or expressions (e.g., *"forget about it"*) that can be recognized as generic termination commands (See Appendix A.3.1).

Contextual Data Collection: The app collects surrounding visual contextual information by capturing images at one-second intervals for two minutes before triggering a task. If the service was retriggered, it captured images from the last two minutes prior to retriggering. To protect learners' privacy, we provided an image deleter application, allowing quick review and deletion of images that learners did not want to share with the researchers. In our pilot study, learners had low privacy concerns because they had full control over the collected images. The app also collected the surrounding auditory contextual information (e.g., background sounds and conversations with the service) from two minutes before triggering until one minute after the microlearning task was completed.

4 FIELD STUDY METHODOLOGY

After iteratively developing SpeechMaster, we conducted the main study with 28 learners to answer two research questions. In this section, we describe the details of the main study. The study procedures (including the procedures of the six pilot studies) were reviewed and approved by our institution's internal ethical review board.

4.1 Participants

We recruited 28 learners from local communities. Our recruitment criteria included (1) people who were interested in improving their English pronunciation and (2) staying at home for at least four hours (excluding sleeping hours) a day and five days a week. We compensated the learners with approximately 80 USD. For our analysis, we excluded one learner's data because she withdrew during the first week due to health concerns (COVID-19). Therefore, we used the data from 27 learners in our analyses.

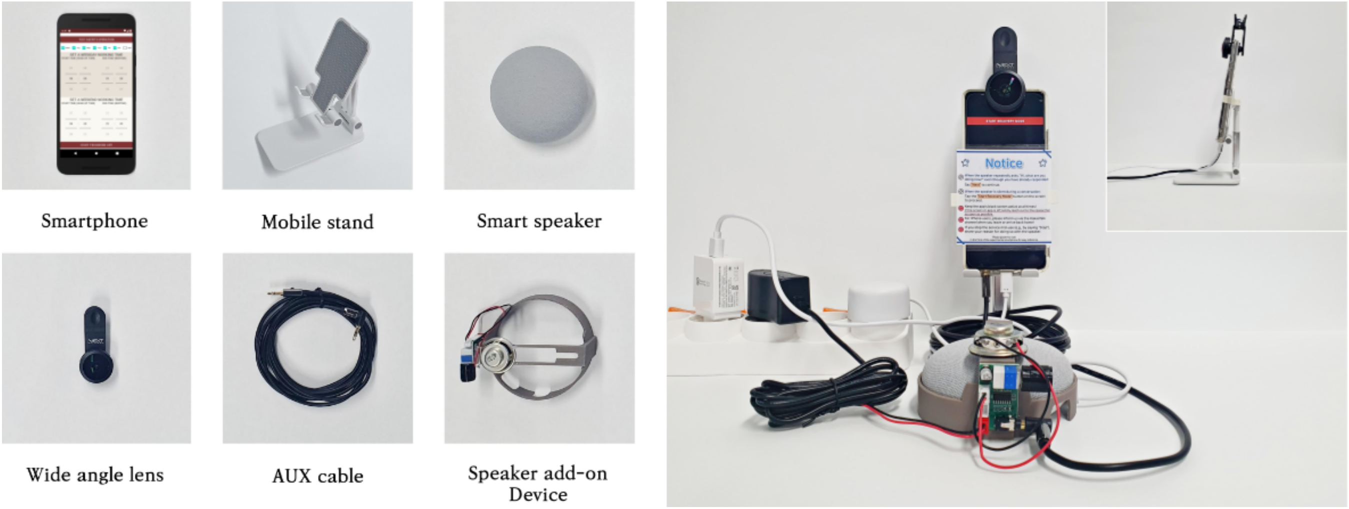


Figure 3: Experimental kit for proactive smart speaker.

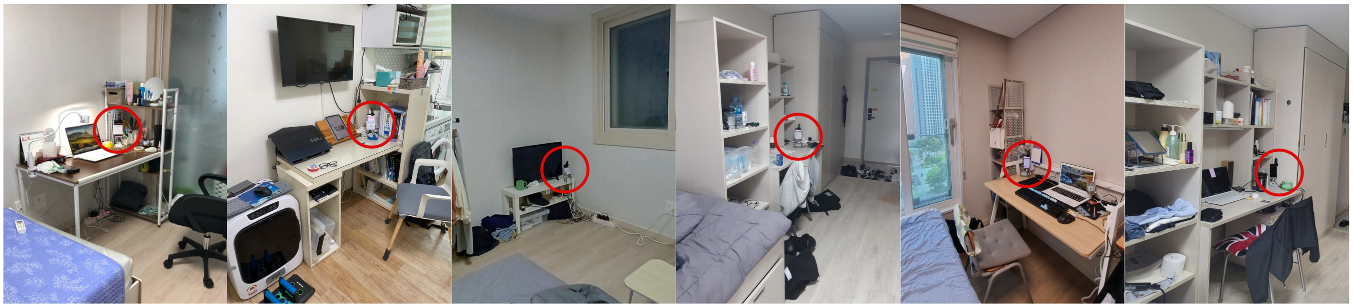


Figure 4: Configurations of living spaces with experimental kit installation.

The average age was 23.5 (SD = 9.28, range = 18–59) years, and ten learners were female (37%). As we recruited learners motivated to improve their English pronunciation, most (n = 25, 93%) were students. Others included an office worker and a homemaker. Prior studies on language learning have also considered young adults [63] because they are more interested in language learning than other age groups as part of college education and job seeking [32].

4.2 Procedure

The study began with an online orientation. On the day after the orientation, we conducted a three-week in-the-wild field study and online one-on-one exit interviews. The details of our procedure are as follows.

Learners first visited our laboratory to sign informed consent forms and received experimental kits (SpeechMaster). Before signing the form, we explained the study to them. In this orientation, we provided an overview of the study, instructions on how to install the kits and interact with the smart speaker, and guidelines on how to respond to contextual inquiries (activity inquiry and reason-for-stop inquiry). For the contextual inquiry, we asked the learners to respond in their native language. Similar to previous studies [10, 60], we asked learners to provide detailed contextual information. For the activity inquiry, they were asked to describe their current activities (e.g., what they were doing), specifying where (e.g., space within a room or house), how (if there were any tools involved), whom (if they were with someone else), and why (if there was a reason) (e.g., *"I was watching a movie on a laptop at my desk"*). For the reason-for-stop inquiry, they were asked to provide reasons to stop interacting with the service (e.g., *"I want to take a nap because I am tired."*).

On the day after the orientation, they participated in a three-week field study. Throughout the study, six researchers periodically reviewed learners' response data. For contextual inquiries, learners may not respond (i.e., no response) or respond insufficiently (i.e., incomplete response). Learners may also omit one of the activities they were simultaneously performing (e.g., not reporting *listening to music* while *taking a rest with music on*). To assess the completeness of the responses, in addition to the response data, the researchers also reviewed notable human-generated background sounds (e.g., music, in-game audio, human-to-human conversation, the hum of a vacuum cleaner, the noise of a hair dryer) by listening to the surrounding sound of corresponding audio data. This approach helped assess the potential incompleteness of detailed contextual responses. For no responses, we requested the learners to provide the reason. For incomplete responses, we requested that they supplement the responses regarding their activities at that time. The requests were sent via instant messages once daily. Similar to a previous study [10], as cues for recall, our request included response data in both the voice recordings of audio data and text format, along with the response timestamp.

After the field study, we analyzed the usage logs and conducted online one-on-one semi-structured interviews to qualitatively answer RQ1 and RQ3. The interviews lasted 90 min, on average. In the interviews, we asked the learners about (1) their perceptions and experiences with proactive conversational microlearning services (e.g., their advantages and disadvantages, and their impact on their learning and daily life) and (2) their interruptible and uninterruptible situations for engaging with the services (e.g., why and which activities, locations, and times of the day or week they found to be interruptible or uninterruptible). For qualitative analysis, we conducted thematic analysis of the transcribed interview data [8]. Four researchers carefully reviewed the transcribed texts to identify recurring codes on (RQ1) user perceptions and experiences of proactive provision of learning content and overall language learning with smart speakers and (RQ3) important contextual factors when learners perceived themselves to be interruptible (i.e., participating in microlearning tasks). Similar codes were grouped into themes. Coding and regrouping were conducted iteratively until a consensus was reached.

4.3 Speech Shadowing Contents

For the speech shadowing contents, we initially selected 12,000 English sentences, with 20 levels (e.g, Level 1: *"We came last."*, Level 10: *"I slept for another three hours."*). English sentences were selected from Tatoeba, which is a large database of English sentences and translations [57]. The levels were categorized based on a combination of syllables and space counts within a sentence [31, 44]. However, after pilot studies, we excluded 250 sentences containing words or expressions that can cause potential unintentional terminations (see Appendix A.3.1). After excluding such sentences, for the main study, we used 11,750 English sentences with 20 levels. The average number of sentences per level was 587.5 (SD = 5.58).

5 RQ1: LEARNER PERCEPTIONS AND EXPERIENCES FOR PROACTIVE CONVERSATIONAL MICROLEARNING SERVICES

Based on our interviews, we qualitatively analyzed how learners perceived and experienced proactive conversational microlearning services.

5.1 Experience of 'Proactive' Microlearning Services

Our learners mentioned that proactive microlearning services induced regular learning practices and were perceived as non-distracting.

5.1.1 Inducing Regular Learning Practices. Proactive provision was perceived as a learning reminder, thus helping learners frequently participate in microlearning sessions (or tasks). For example, P1 noted, *"If I had to manually start (SpeechMaster) before anything else, would I have used it as much? I think (that SpeechMaster) talking to me first was the best part."* It also enabled them to use their spare time to learn better, especially when their ongoing task was perceived to be less productive than microlearning. For instance, P27 commented, *"When I heard the sound (of SpeechMaster), I was usually reading webtoons with my smartphone or playing games with my laptop, [...] so I think I could make use of my spare time good enough."*

5.1.2 Non-distracting Interruption. Surprisingly, the learners perceived the provision of microlearning tasks at random times (or without considering their interruptibility) as non-distracting. A common reason for the sense of non-distraction was learners' high motivation for learning, as we recruited learners interested in improving their English pronunciation. Learners viewed the interruptions from microlearning as valuable and beneficial opportunities to improve their English. For example, P1 said, *"I don't have many opportunities to use or speak English, so it was helpful because I could have this experience of listening and speaking English."* P9 stated, *"I think it's much better (than manually starting SpeechMaster). Because I think some force is needed to keep on with the progress and study. So I think I wouldn't have done it if I did it when I wanted to do."*

Another common reason is immediate termination capability at the onset of the microlearning task. When learners were not interruptible, they had the option to promptly stop microlearning tasks at the beginning by saying the *"Stop"* command. For example, as P19 expressed, *"I never really felt distracted. [...] I could just say 'I can't do it right now' to stop the speaker at the start when I'm not available."* In addition, the response waiting time contributes to a sense of non-distraction. Our service waited for approximately 35 s to respond to the activity inquiry (see Section 3.1). Our learners reported that this short waiting time helped them prepare (e.g., completing their ongoing tasks, moving to the speaker if they were in another room) and engage in microlearning tasks. For example, P22 said, *"It asks three times, so during that time, I had a moment to pause what I was doing."*

5.2 'Post'-Experience of Microlearning

Our microlearning service was provided proactively and repeatedly during learners' daily lives. Our learners commonly reported that as the service was naturally integrated into their daily lives, it helped them manage their daily lives more constructively.

5.2.1 Helping to Establish a Productive Daily Routine. Microlearning tasks were conducted at an average interval of 60 min. The learners reported that this regularity helped them manage their daily schedules more effectively, leading to more regular and structured lifestyles. For instance, P25 noted, *"Timing is a little [random] but there is regularity. [...] So I could realize that time passed by so quickly when I was just lying on the bed."* The activity inquiry (i.e., *"Hi, what are you doing now?"*), which was provided at the start of the microlearning tasks, also contributed the learners to establish a productive daily routine by helping them become aware of their current activity, for example, as P27 stated, *"It was good to look back on what I was doing again by saying it."* Similarly, P23 said, *"Mostly, my response was lying on the bed, watching webtoons on the phone. Since I felt that this response was quite frequent, I had a bit of a feeling like, 'Ah, I should stop watching.' 'I should get up.'."* In general, the learners' awareness of their engagement in less productive activities motivated them to start or switch to more productive activities.

5.2.2 Role of Refreshing Moment during On-going Work. When microlearning tasks were provided in the middle of 'non-productive' activities, our learners performed the learning tasks as a turning point to start productive activities. For example, P20 noted, *"When I am lying down and this rang, I think it was like a catalyst that makes me do something from then on."* Similarly, P15 reported, *"After [the service] calls me out and I accomplish it when I am having a rest without any plans, I get to resume to my work and continue to focus on what I had to do by this chance. I think it was an advantage."* When microlearning tasks were provided in the middle of 'productive' activities, they performed the tasks as an opportunity for refreshment (or short break). For example, P26 stated, *"During talking with [SpeechMaster] for 10 minutes, I felt like I was having a rest, [...] and after 10 minutes, I focused on my main work again."*

6 RQ2: OPPORTUNE MOMENTS FOR PROACTIVE CONVERSATIONAL MICROLEARNING SERVICES

We quantitatively analyzed when and for how long learners were likely to engage in proactive conversational microlearning. Namely, we analyzed (1) overall usage patterns and interruptibility, and (2) interruptibility across learners' contexts (i.e., activity, spatial, and temporal contexts) prior to the provision of microlearning tasks. In Section 7, we supplement the quantitative analysis qualitatively with interviews. In Section 5, the qualitative findings suggest that offering proactive conversational microlearning services at random times (or without considering interruptibility) can be perceived as non-distracting. However, this does not negate the significance of delivering microlearning at opportune moments, as such timing can maximize learners' engagement in learning (e.g., an increase in the number of daily learning sessions engaged in by the learners).

Table 1: The number of cases across different types of termination points.

Total number of provided microlearning tasks (n = 2,937, 100%)							
No responses	Only responding to *activity / availability inquiry*	Cases where learner engage in speech shadowing (n = 1,509, 51%)					
		Stop by saying *"stop"* at any time	Stop by saying *"no"* at *continue-to-next inquiry* steps (n = 640, 22%)				End by reaching the maximum duration
	(Moment; ≤ 1 minute)		First (3 minutes)	Second (5 minutes)	Third (7 minutes)	Fourth (9 minutes)	(10 minutes)
645 (22%)	783 (27%)	107 (4%)	507 (17%)	81 (3%)	31 (1%)	21 (1%)	762 (26%)

6.1 Interruptibility Definition

To identify opportune moments, we measured how much learners invested in microlearning over a given period. We conceptualized this as interruptibility. When learners engage in a given period, we expect that they will be available for shorter microlearning sessions. Indeed, in our interview, our learners also mentioned similar opinions that, for example, *"When I used it for 10 minutes, that means I had at least 10 minutes spare time then. So, I reckon I also could've done something else that takes less than 10 minutes."* [P4].

6.2 Statistical Analyses

We conducted a series of linear mixed-model analyses to statistically compare interruptibility across various contextual conditions (e.g., interruptibility across different activity types) prior to providing the microlearning tasks. For the dependent variables, we consistently considered interruptibility except in the first analysis. In the first analysis (Table 2), we considered the percentage of cases as the dependent variable. While the independent variables varied across the analyses, details about the dependent and independent variables are available in the tables that present the statistical results in each subsection. For post-hoc comparisons, we adjusted the p-values using the Bonferroni correction. To account for the non-independence of the data, we include the learners as a random effect.

6.3 Overall Usage Patterns and Interruptibility

Overall Usage Patterns: While the learners were at home, SpeechMaster proactively provided 2,937 microlearning tasks and the learners responded to 2,292 tasks (78%). Among these tasks, learners engaged in 1,509 (51%) speech shadowing. Details of the microlearning tasks are presented in Section 3.1. As shown in Table 1, our learners participated in microlearning tasks primarily for three duration types: (1) moment (i.e., ≤ 1 min), (2) 3 minutes, and (3) 10 minutes. As shown in Table 2, our statistical analysis also confirmed that our learners participated more frequently in these specific duration types compared to others (at least $p < 0.01$). Moment interactions were 783 cases (27%) where they only responded to activity/availability inquiry, and did not perform speech shadowing. Whereas, interactions exceeding one minute (e.g., 3-minute and 10-minute interactions) were cases where the learners engaged in speech shadowing (n = 1,509, 51%).

The learners were able to end their current interactions (or microlearning tasks) by saying *"Stop"* comment at any time (stop cases = 107, 4%) or by responding *"No"* at one of the four continue-to-next inquiry steps (no cases = 640, 22%). 3-minute interactions were cases in which the learners stopped at the first continue-to-next inquiry step (n = 507, 17%). 10-minute interactions were cases where microlearning continued until the maximum duration was reached (n = 762, 26%).

When not engaging for the maximum duration, learners mainly stopped learning at one of the continue-to-next inquiry steps (22%). According to the interviews, our learners pre-established an in situ microlearning goal – whether and for how long they would engage in microlearning, and considered one of continue-to-next inquiry steps as an anchor point to stop, because these points were convenient and helped them become aware of their elapsed learning time. For instance, P27 explained *"It's good that I can know how many minutes it's been going, so I can use it for just that long."*

Overall Interruptibility: Table 3 shows the percentage of interruptible cases, which indicates how much learners are likely to engage in microlearning for a moment, 3 min, or 10 min, if the learning is provided at a random time. As shown in Table 4, our statistical analysis showed that interruptibility was highest for moment interactions ($p < 0.001$), whereas it was lowest for 10-minute interactions ($p < 0.001$). These findings indicate that in general,

Table 2: Statistical result for the number of cases across different types of termination points.

Effects	F-value	df1	df2	p-value
Duration of interactions	97.332	6	141	< 0.001

Table 3: Overall interruptibility for moment (≤ 1 min), 3-minute, 10-minute interactions.

Number of triggerings	Interruptibility for duration types		
	moment	3 minutes	10 minutes
2,937	78%	49%	25%

Table 4: Statistical result for the interruptibility across duration types (moment vs. 3 min vs. 10 min).

Effects	F-value	df1	df2	p-value
Duration types	61.065	2	78	< 0.001

learners are more likely to be interruptible as microlearning becomes shorter when microlearning is provided without considering interruptibility (i.e., providing at random times).

6.4 Interruptibility across Activity Contexts

To understand the relationship between activity contexts and interruptibility, we first categorized learners' activities prior to the provision of the microlearning task, and then analyzed interruptibility across the categories.

6.4.1 Activity Categorization. Five researchers categorized the activities according to the procedures of prior studies [10]. To enhance the precision of the activities, in addition to user response data, we utilized the surrounding sounds of audio and image data (see Section 4.2). The cases involving multiple activities (n = 116) were classified into multiple categories. For instance, the case where a learner *"was eating dinner while watching YouTube on [his] laptop at [his] desk"* was classified into both eating and using media categories.

Twelve activity categories were identified (Table 5). When microlearning tasks were provided, popular activities included using media (30%), napping/sleeping (19%), studying/working (12%), and resting (12%). These activities are primarily what learners perform at home.

6.4.2 Interruptibility across Activity Contexts. Table 6 shows interruptibility across activity types and duration types. As indicated in Table 7, both the main effects and the interaction effect were significant. The post-hoc analysis suggests that regardless of duration types, interruptibility was consistently higher when the learners were engaging in using media, resting, studying/working, self caring or returning from outside/other room. In contrast, interruptibility was consistently lower when they were performing hygiene or napping/sleeping. While the detailed distribution of interactions across activities can be found in Appendix (see Figure 5), we further discussed interruptible activities across duration types, as follows:

For moment interactions, this was discovered to be an opportune moment when learners were engaged in most activities, except for hygiene (34%) and napping/sleeping (49%). For these two activities, interruptibility was significantly lower than other activities (consistently $p < 0.01$). Interestingly, when the learners were engaged in visiting outside/other room, their interruptibility was high for moment interactions (78%), but showed a notable decrease for both 3-minute (19%) and 10-minute (10%) interactions. Our statistical analysis also confirmed that interruptibility was significantly lower for 3-minute and 10-minute interactions than moment interactions (consistently $p < 0.001$).

Table 5: Definitions and examples of activity categories.

Activity categories	Example	Number of triggerings
Using media (e.g., video gaming, internet surfing, and watching videos)	*"I'm sitting at my desk and playing games on my computer."* *"I'm lying on my bed and surfing the web on my phone."* *"I was watching TV in the family room."*	924 (30%)
Napping / sleeping	*"I was taking a nap in bed because I was tired."*	573 (19%)
Studying / working	*"I'm sitting at my desk studying my major."* *"I'm working sitting in the chair."*	375 (12%)
Resting	*"I'm just lying on bed."* *"I'm resting at my desk."*	374 (12%)
Eating	*"I'm having breakfast at the table."*	223 (7%)
Hygiene (e.g., nature's call, shower, and washing hands)	*"I was washing my face in the bathroom."*	133 (4%)
Social interaction (e.g., talking with others, chatting, and phone call)	*"I was talking with my family in the family room."* *"I'm talking to my mom on my phone in the bed."*	131 (4%)
House chores (e.g., preparing and cleaning up after a meal, cleaning, and doing laundry)	*"I was cooking my lunch in the kitchen within the family room."*	113 (4%)
Visiting outside / other room	*"I have to go out for a part-time job."*	72 (2%)
Self caring (e.g., face or body caring, changing clothes, exercise, and stretching)	*"I'm just sitting on the floor to dry my hair with a hair dryer."* *"I'm changing my clothes."*	56 (2%)
Returning from outside / other room	*"Now I came back home."* *"I was just coming into this room."*	53 (2%)
Others (e.g., pet caring and other activities not mentioned above)	*"I was caring my puppy in the family room."* *"I'm doing rap."*	26 (1%)

Table 6: Interruptibility across activity types and duration types. ≤ *1 min* = Moment.

Activity types	Number of triggerings	Interruptibility		
		≤ 1 min	3 min	10 min
Overall	2,937	78%	49%	25%
Using media	924	91%	64%	32%
Napping/sleeping	573	49%	21%	14%
Studying/working	375	95%	60%	32%
Resting	374	99%	61%	27%
Eating	223	87%	46%	25%
Hygiene	133	34%	14%	9%
Social interaction	131	81%	38%	17%
House chores	113	84%	39%	21%
Visiting outside/ other room	72	78%	19%	10%
Self caring	56	88%	55%	28%
Returning from outside/other room	53	88%	53%	33%
Others	26	70%	53%	29%

Table 7: Statistical results for the interruptibility across activity types and duration types.

Effects	F-value	df1	df2	*p*-value
Activity types	17.376	11	732	< 0.001
Duration types	51.822	2	732	< 0.001
Activity × Duration	14.153	22	732	< 0.001

For the 3-minute interactions, it was the most opportune moment when learners were engaged in using media (64%), resting (61%), studying/working (60%), self caring (55%), returning from outside/other room (53%), or others (53%). For these activities, interruptibility was significantly higher than other activities (at least $p < 0.05$). Whereas, interruptibility was significantly lower for napping/sleeping (21%), visiting outside/other room (19%), or hygiene (14%) than other activities (at least $p < 0.05$). For 10-minute interactions, it was an inopportune moment when our learners were engaged in hygiene (9%), visiting outside/other room (10%), or napping/sleeping (14%). For these activities, interruptibility was significantly lower than other activities (at least $p < 0.05$). Based on interview results, in Section 7, we discussed four contextual factors associated to learners' activities that were closely related to opportune moments for microlearning.

6.5 Interruptibility across Spatial Contexts

Within-home locations are often linked to particular activities (e.g., studying or working at a desk). To further understand the relationship between spatial context and interruptibility, we statistically analyzed interruptibility across learners' spatial contexts before the microlearning tasks.

Specifically, we statistically compared (1) interruptibilty for spaces where the speaker was installed versus non-installed (installed space vs. non-installed space), (2) interruptibilty across specific locations within the installed space and (3) interruptibilty across non-installed space. To enhance the precision of our location analysis, in addition to the user responses to the activity inquiry, we determined the learners' positions using image data. The detailed distribution of the interactions across these spatial contexts can be found in the Appendix (see Figures 6 and 7).

6.5.1 *Speaker Placement and Living Conditions.*

Most learners installed the speaker in their bedrooms (n = 24), whereas the rest installed it in their study/working rooms (n = 3). In these spaces, the learners mostly installed speakers on their desks (n = 24). Conversely, two were installed at the top of the drawer. One was installed in media furniture. Non-installed space varied depending on living conditions. For instance, 15 participants living in a one-bedroom house (i.e., one bedroom with a separate living room) or studio (i.e., everything in a single room) had no other room except for a restroom. In addition to the restroom, 12 participants living in a more-than-one-bedroom house had additional rooms (e.g., bedrooms and study/working rooms).

6.5.2 *Installed Space vs. Non-installed Space.*

Table 8 shows interruptibility when learners were at installed space and non-installed space. As shown in Table 9, both the main effects and the interaction effect were significant. The post-hoc analysis showed that regardless of duration types, our learners were more interruptible in installed space than in non-installed space (consistently $p < 0.01$). In addition, in installed space, interruptibility was higher for shorter interactions (consistently $p < 0.001$). Whereas, in non-installed space, learners were more interruptible for moment interactions than 3-minute or 10-minute interactions ($p < 0.001$).

Interestingly, even in non-installed space, almost half of the time (47%), they were interruptible – moved to the installed space to

Table 8: Interruptibility across location types (installed space vs. non-installed space) and duration types. ≤ *1 min* = Moment.

Location types	Number of triggerings	Interruptibility		
		≤ 1 min	3 min	10 min
Overall	2,937	78%	49%	25%
Installed space	2,433	85%	54%	28%
Non-installed space	504	47%	23%	13%

Table 9: Statistical result for the interruptibility across location types (installed space vs. non-installed space) and duration types.

Effects	F-value	df1	df2	*p*-value
Location types	56.788	1	153	< 0.001
Duration types	44.528	2	153	< 0.001
Location × Duration	12.947	2	153	< 0.001

Table 10: Interruptibility across specific locations within the installed space and duration types. ≤ *1 min* = Moment.

Locations within the installed space	Number of triggerings	Interruptibility		
		≤ 1 min	3 min	10 min
Overall	2,433	85%	54%	28%
Bed	1,243	76%	46%	24%
Desk	1,013	98%	66%	33%
Other locations	177	91%	42%	21%

Table 11: Statistical result for the interruptibility across specific locations within the installed space and duration types.

Effects	F-value	df1	df2	*p*-value
Locations within the installed space	36.042	2	213	< 0.001
Duration types	91.98	2	213	< 0.001
Location × Duration	7.159	4	213	< 0.001

Table 12: Interruptibility across specific locations within the non-installed space and duration types. ≤ *1 min* = Moment.

Locations within the non-installed spaces	Number of triggerings	Interruptibility		
		≤ 1 min	3 min	10 min
Overall	504	47%	23%	13%
Living room	318	68%	44%	28%
Rest room	127	30%	10%	6%
Other rooms	59	26%	26%	25%

Table 13: Statistical result for the interruptibility across specific locations within the non-installed space and duration types.

Effects	F-value	df1	df2	*p*-value
Locations within the non-installed space	15.411	2	144	< 0.001
Duration types	17.529	2	144	< 0.001
Location × Duration	8.172	4	144	< 0.001

engage in moment interactions. In our interviews, the learners commonly mentioned that when they heard the speaker sound in non-installed space, they moved to engage in microlearning tasks. In Section 7.5, based on the interview results, we discuss how the audible range (i.e., where learners can hear the speaker's sound) affects interruptibility.

6.5.3 Locations Within the Installed Space. Table 10 shows interruptibility across duration types and locations within the installed space (bed vs. desk vs. other locations). As shown in Table 11, both the main effects and the interaction effect were significant. The post-hoc analysis suggested that learners were most interruptible at the desk. Specifically, for moment interactions, our learners were more interruptible at desk or other locations (e.g., being on the floor, standing in the middle of the room, etc.) than on the bed (at least $p < 0.05$). For both 3-minute and 10-minute interactions, our learners were more interruptible at the desk than at the other locations or on the bed (at least $p < 0.05$). Given that our learners mostly installed the speaker primarily at their desk, our results suggest that learners are more likely to be interruptible when they are positioned close to the smart speaker.

During the interviews, learners commonly mentioned challenges engaging in microlearning when at a certain distance away (e.g., on a bed) from the speaker (e.g., a desk) because of the limitations in the speaker's speech recognition. In Section 7.6, based on the interview results, we discuss how the voice recognition range affects interruptibility.

6.5.4 Locations in Non-installed Space. Table 12 shows interruptibility across duration types and locations within non-installed space (living room vs. rest room vs. other rooms). As shown in Table 13, both the main effects and the interaction effect were significant. The post-hoc analysis suggested that regardless of duration types, learners were most interruptible in the living room (at least $p < 0.01$). Specifically, for moment interactions, our learners were more interruptible in the living room than the rest room or other rooms (consistently $p < 0.01$). For 3-minute and 10-minute interactions, our learners were more interruptible in the living room than the rest room (consistently $p < 0.001$). The living room was generally close to the installed space, allowing more chances to hear the speaker sound (or to be within an audible range) and move to the installed space for interactions compared to other rooms (see Section 7.5).

6.5.5 Additional Analyses. While 24 learners installed speakers in their bedrooms, the other three installed speakers in their studying/working rooms. In addition, among 24 learners, one learner did not install the speaker on his desk. We conducted additional analyses after excluding four learners with different speaker placement contexts. However, the results are similar to the original results.

6.6 Interruptibility across Temporal Contexts

Home routines tend to be repeated periodically across the days of the week or hours of the day [48]. To further understand the relationship between temporal context and interruptibility, we analyzed interruptibility across days of the week and parts of the day. Similar to a previous study [10], we considered the period between 9:00 and 24:00 for parts of the day.

Table 14 shows the interruptibility across days of the week and parts of the day for the three duration types. As shown in Table 15, only the main effects were significant. The post-hoc analysis suggested that in the morning (9:00–12:00), our learners were less interruptible than during other parts of the day ($p < 0.001$), which could be due to sleeping (36%). Regarding days of the week, there were minimal differences between Saturday and Wednesday. Namely, our learners were more interruptible for Wednesday than Saturday ($p < 0.001$). For the interaction duration types, shorter interactions were associated with higher interruptibility ($p < 0.001$).

Table 14: Heatmap for the interruptibility across parts of the day, days of the week, and duration types.

Parts of the days	Interruptibility for moment interactions (%)								Interruptibility for 3-min interactions (%)								Interruptibility for 10-min interactions (%)							
	Days of the week							Overall	Days of the week							Overall	Days of the week							Overall
	Mon	Tue	Wed	Thu	Fri	Sat	Sun		Mon	Tue	Wed	Thu	Fri	Sat	Sun		Mon	Tue	Wed	Thu	Fri	Sat	Sun	
Morning (9:00 ~12:00)	69.1	65.2	77.7	76.3	74.0	68.2	62.7	70.4	36.5	40.4	54.4	39.9	37.8	30.9	37.7	39.6	18.8	17.2	30.0	21.6	19.9	14.8	20.4	20.4
Afternoon (12:00 ~16:00)	78.4	83.2	80.9	87.5	89.1	78.7	78.4	82.3	56.7	55.9	50.2	51.2	65.4	48.7	52.8	54.4	30.5	23.9	26.2	24.5	33.9	22.0	35.1	28.0
Evening (16:00 ~20:00)	91.9	79.7	93.6	87.3	78.2	74.8	78.3	83.4	50.8	42.3	57.0	52.3	53.6	44.3	55.0	50.8	27.2	11.6	27.8	27.2	32.2	22.4	27.7	25.1
Night (20:00 ~24:00)	81.7	80.4	88.7	91.8	79.8	80.4	77.0	82.8	52.9	54.9	55.7	58.8	49.1	47.4	50.8	52.8	26.7	29.7	28.4	25.9	18.9	27.7	32.2	27.1
Overall	80.3	77.1	85.2	85.7	80.3	75.5	74.1	79.8	49.2	48.4	54.3	50.6	51.5	42.8	49.1	49.4	25.8	20.6	28.1	24.8	26.2	21.7	28.9	25.2

Table 15: Statistical result for the interruptibility across parts of the day (morning vs. afternoon vs. evening vs. night), days of the week, and duration types. PoD = Parts of the day, DoW = Days of the week, and DT = Duration types.

Effects	F-value	df1	df2	*p*-value
Parts of the day (PoD)	16.227	3	1755	< 0.001
Days of the week (DoW)	3.153	6	1755	< 0.01
Duration types (DT)	582.606	2	1755	< 0.001
PoD × DoW	1.274	18	1755	0.195
PoD × DT	0.875	6	1755	0.513
DoW × DT	0.754	12	1755	0.698
PoD × DoW × DT	0.34	36	1755	1

7 RQ3: CONTEXTUAL FACTORS ASSOCIATING INTERRUPTIBILITY

Similar to RQ1 (Section 5), we conducted a qualitative analysis and identified six contextual factors that influenced interruptibility: (1) *productivity*, (2) *concentration*, (3) *pausable duration*, (4) *auditory/verbal channel availability*, (5) *audible range*, and (6) *voice recognition range.* In this section, each factor is discussed in detail.

7.1 Productivity

The productivity level of learners' ongoing work was strongly linked to their interruptibility (i.e., whether they will engage in microlearning or not). Our learners reported that they were more likely to participate in microlearning tasks during less productive activities such as using media and resting. For instance, P23 remarked, *"I was just listening to music or sitting at my desk, fully awake but not really involved in any significant activity. [...] so I thought it was a good time to do something productive."* Similarly, our quantitative results indicated high interruptibility of these activities. When referring these activities, learners often described themselves as *"messing around"* or *"wasting time."* P22 commented, *"When I was wasting my time alone with Instagram or YouTube and this (SpeechMaster) rang, then I stopped for a while to participate in microlearning."* Therefore, they felt more inclined to participate in longer microlearning sessions to use their time more productively, as, for example, P21 shared *"I did it when I was really just playing around. Thinking, 'What's the point of just playing around?' then I went for a maximum of 10 minutes."*

7.2 Concentration

Concentration is also a crucial factor. The learners reported that it was easy and most comfortable to engage in microlearning tasks when they were involved in activities that required low concentration, such as house chores and resting. However, they experienced difficulties when their ongoing activities required high concentrations. For instance, P22 mentioned that *"Folding laundry was good for microlearning. If I'm too intensely focused on something else, I may struggle to concentrate on learning English."* In addition, learners reported that it was not easy to notice the start of the microlearning tasks while concentrating on their ongoing tasks. These activities include playing games and doing assignments. For example, as P27 noted, *"When I was playing games for a long time, I think I might have missed a few times because I was just caught to the game."* However, according to our quantitative analysis results, the study and work exhibited high interruptibility, despite typically requiring high concentrations. This paradox is explained by learners' frequent use of microlearning as an opportunity for refreshment or a short break, as discussed in Section 5.2.2.

7.3 Pausable Duration of Ongoing Work

Pausable duration – the duration for which learners can pause their ongoing tasks significantly influences their interruptibility (i.e., how long they will engage in microlearning). During the interviews, the learners reported that the longer they could pause their current activities, the longer they could engage in microlearning, indicating higher interruptibility. Our quantitative results revealed that interruptibility was consistently low for hygiene across all durations. Learners commonly identified 'taking showers' and 'bathroom functions' under hygiene as non-pausable activities. For instance, P15 illustrated the difficulty of pausing hygiene by comparing it to more pausable activities: *"When reading a book or listening to music, I could stop and engage in microlearning. But when I'm taking shower, I have to dry quickly. It's just quite difficult to leave immediately."*

Learners typically identified visiting outside/other room as examples of activities that they could pause for a moment (e.g., ≤ 1 minute), but not for extended periods. Our quantitative results showed a notable decrease in interruptibility from the moment interactions to 3-minute interactions. Learners commonly expressed concerns about the potential disruption caused by engaging in microlearning tasks during their preparation to leave or upcoming schedules. For instance, P15 said *"Engaging in microlearning might lead me to forget things or delay my departure, so I often choose to stop microlearning in the middle. It's like I have the opportunity to do it, but being unable to because of time constraints."*

Social interaction is a typical example of pausability for 1–3 min. They mentioned that, in the middle of social interaction, they considered it socially inappropriate to engage in another activity for a relatively long time. For example, P4 stated, *"When talking to friends or doing something with someone, I can't just detach myself for long."* In contrast, using media, resting and returning from outside/other room were identified as pausable for longer duration, as these activities often did not involve immediate follow-up tasks. For example, P4 reflected, *"Once I got back home, my schedule was more likely open. So, I think that allowed me to use it for a longer duration."* Similarly, in our quantitative results, these activities consistently had high interruptibility across duration types.

7.4 Auditory/verbal Channel Availability

The availability of the auditory and verbal channels plays a significant role in interruptibility. In the case of auditory channel availability, during house chores such as using a vacuum cleaner, learners were not able to hear the speaker's sound due to noise. For example, P21 mentioned, *"It's hard to recognize, because the vacuum cleaner is noisy [...]"* Regarding verbal channel availability, while learners were eating, their pronunciation accuracy lowered, making interactions more difficult than usual. For example, P25 stated, *"I tend to chew food for a long time. [...] I found it difficult to continue while doing it, so I just did a few [sentences] (and stopped.)"*

7.5 Audible Range

Interruptibility across locations in non-installed space was closely related to the audible range, which is the range within which learners can hear or notice the speaker's sound. Our quantitative results showed higher interruptibility in the living room, which is usually proximate to the installed space, thereby allowing learners to be within the audible range. Consequently, they can hear the speaker's sounds and move to the installed space to engage in the microlearning task. For example, P26 stated, *"When I was in the living room, I could hear. I immediately went there."* Conversely, in the rest room and other rooms, the learners were less likely to hear the speaker's sound. For example, P21, living in a studio, said *"The kitchen side was not that audible, and when the restroom door was closed, I couldn't hear the speaker. This made the restroom inconvenient either for engaging with the speaker."*

7.6 Voice Recognition Range

Interruptibility in locations within installed space was closely tied to the 'voice recognition range' – the distance within which smart speakers can effectively recognize and process spoken commands. Learners commonly mentioned that when they were within range, they could comfortably talk to the speaker, as it ensured accurate voice recognition. For instance, P21 mentioned, *"I could hear the speaker's sound from 1 to 2 meters away, but beyond that, like when on the bed, it couldn't recognize my voice. This made conversations difficult from farther distances."* Our qualitative analysis revealed that learners were more likely to be interrupted on their desks, the most common installation site for speakers, and less so on their beds, which often fell outside the voice recognition range. When in bed, they often choose not to engage in microlearning. P27 stated, *"It felt like the speaker wasn't good at picking up my voice from the bed, so I rarely used it there. I mainly use it near the desk area."*

8 DISCUSSION

8.1 Summary of Major Findings

We observed that proactive microlearning tasks increased the learning opportunities. While prior studies have shown that randomly assigned proactive tasks have been reported to cause disruptions [23], it is interesting to note that our learners generally perceived proactive microlearning at random intervals positively, namely, as non-distracting most times. This perception was attributed to the learners' strong motivation to achieve their learning goals, which led them to consider random provisions as additional learning opportunities. Fischer et al. also found similar findings in their study on the interruptibility (or receptivity) of mobile interruptions (i.e., receiving text messages at random times), that interruptibility can vary based on the usefulness and interest of the provided content [17].

In half of the instances (49%), the learners did not participate in the learning sessions (i.e., the speech shadowing step in the microlearning tasks). This highlights the importance of delivering microlearning at opportune moments to maximize learner participation in daily learning tasks (or sessions). While prior studies have considered interrupting tasks involving a short duration (typically ≤ 1 min) [10, 60], we explored opportune moments for interrupting tasks (i.e., microlearning) in various durations (i.e., 1, 3, and 10 minutes), and found that contextual factors (e.g., activity, location) relevant to opportune moments can be varied depending on task duration. Therefore, important contextual factors relevant to opportune moments in prior studies may not be applicable to interrupting tasks involving longer-than-one-minute interactions.

In our study, interruptibility varied significantly across the activity contexts. Using media, resting, studying/working, and returning from outside/other room were opportune moments (or highly interruptible) for microlearning. Prior studies have typically identified studying/working as highly uninterruptible owing to the high concentration required and potential disruption from interruptions [10, 60]. However, our study found studying/working to be highly interruptible in microlearning, possibly because of learners' high motivation and perception of microlearning as a productive short break. Indeed, while in our study, learners considered microlearning more entertaining than studying/working, interruptions could be viewed as opportunities that deviate from learning for a short time [27, 28].

Spatial context (or indoor location) also influences interruptibility. In the installed space, learners showed higher interruptibility when close to the smart speaker (e.g., at a desk) because of the speaker's better speech recognition. By contrast, in non-installed space, interruptibility was higher in areas close to the installed space (e.g., the living room), where learners could hear the speaker's sounds more clearly. This aligns with prior studies emphasizing the importance of speakers' communication range in determining interruptibility for proactively smart speakers [10, 60]. Our findings extend this by highlighting that the voice recognition range (or inbound communication range to the speaker) is crucial in installed space, whereas the audible range (or outbound communication range from the speaker) is crucial in non-installed space.

8.2 Privacy Concerns in Vision-based Activity Recognition

We collected visual contextual information (i.e., images) around the speakers to identify learners' activities and indoor locations. In the orientation, we informed the learners about our approach. Both our main and pilot studies generally expressed low privacy concerns, aligning with recent studies that reported a similar lack of concern regarding data collection using smart home devices (e.g., Nest cam indoors) [35, 55]. Indeed, our approach empowered learners with full control over the images, allowing them to review and delete any images that they preferred not to share with researchers. In our main study, 21 learners performed 727 deletions (34.6 deletion per learner). Literature indicates that maintaining ownership and control over data can alleviate privacy concerns during data collection [36]. However, in real-world applications, continuous management of contextual information can be challenging, leading to significant data accumulation for reviews. This suggests the importance of understanding privacy-control preferences and perspectives.

In our pilot studies, we analyzed the learners' reasons for image deletion (detailed in Section A.3.3) to understand their privacy concerns. The most common reason (31% of deletions) was discomfort with sharing images due to awkward faces or postures (e.g., *"The face turned out way too big in the picture, so I don't want to show it to anyone."*) This suggests privacy concerns regarding images being shown to others (e.g., researchers). Automatic user context recognition without human intervention may help reduce these concerns. In addition, 13% of the deletions were images showing personal items without people, underscoring privacy concerns even when individuals were not directly captured. As an alternative to vision-based context recognition, employing different technologies, such as internal microphone sensors or existing IoT home devices and sensors, can mitigate these privacy concerns. We discussed the technical mitigation strategies in the following section.

8.3 Utilizing Opportune Moments at Home

Our results showed that learners were highly interrupted when they engaged in using media, studying/working, resting or returning from outside/other room. In our study, cameras were used to collect activity and spatial context data, which may raise privacy concerns for real-world applications. Alternatively, smart speakers can utilize an internal microphone sensor and analyze surrounding sounds (e.g., door opening or media sound) to detect specific activities (e.g., returning from outside/other room, using media) [26, 34]. Speakers can also leverage external devices and sensors to infer activities. For instance, the activation of electronic devices associated with using media and studying/working could indicate relevant activities.

Interruptibility also varied across indoor locations, being higher near speakers in installed space, or closer to installed space in non-installed space. Speakers can use Bluetooth connections with their smartphones to determine their proximity to the speaker. Indoor localization techniques using smartphones or multiple WiFi devices can infer indoor locations and activities [49, 56]. Furthermore, entrance detection sensors (e.g., motion sensors) can be used to detect movements in specific rooms (e.g., entering or exiting the restroom or living room). Instead of conventional cameras, thermal cameras can be used to detect the learners' presence. Given that activities (e.g., using media and resting) requiring low concentration generally correspond to higher interruptibility, thermal images from a thermal camera can be used to detect low cognitive (or concentration) activities (e.g., cognitive heat [1]).

8.4 Design Implications to Facilitate Proactive Conversational Microlearning Services

In our study, the provision of microlearning was positively perceived, and it mostly did not distract learners. However, our participants engaged in such microlearning tasks in 51% of the instances. In addition to strategically delivering microlearning tasks at opportune moments, we propose that refining the design of these microlearning tasks can further enhance learning opportunities, and suggest three design implications for proactive conversational microlearning services.

8.4.1 Starter with a Content. Our results suggest that incorporating a starter into content can increase user interruptibility. In this study, the starter was a greeting with an activity inquiry (*"Hi, what are you doing now?"*). After presenting the inquiry, our service waited approximately 35 seconds for a response. The waiting period helps learners prepare for the learning task by completing their ongoing tasks. Other content can be provided alternatively as a starter to enhance experience. For example, in real-world classrooms, teachers often use formative assessments to monitor learners' learning progress [7]. Similarly, the starter contents could be formative assessment questions (e.g., *"How was the difficulty of the last learning material?"*) to gather information about the users' learning experiences. The responses to these questions could be further utilized by learning services to optimize learning outcomes (e.g., adjusting the difficulty level). The starter content can also serve as a motivational tool. For instance, we can provide recent learning summaries and encouragement to improve learning progress (e.g., *"The average score for last week was 90 points. You're doing great!"*). By leveraging these starter options, agents can collect and/or provide additional valuable information while users prepare for the learning task and take advantage of the provided valuable information.

8.4.2 Service-activation Reminder. Our results suggest that in addition to the starter with content, interruptibility could be further increased by a service-activation reminder – executing or prompting users to engage in a certain service at a user-specific time (e.g., 10 minutes later, a speaker proactively ask for engaging a microlearning task: *"Reminding from Speech Shadow. Would you like to practice your pronunciation?"*). However, existing models can deliver only user-defined information (e.g., 10 minutes later, a speaker prompts *"Remind you to practice your pronunciation"*) [22]. Although a starter provided a short preparation time, participants often required additional time to complete ongoing tasks such as finishing their meals or using the restroom. For example, P15 mentioned that *"There were some situations when the timing didn't match, like when I was dealing with an urgent or important task. I was certain that I would have done it if the timing was a few seconds or minutes later."* In this context, it could be beneficial to provide a command to terminate the interaction immediately and automatically create a service-activation reminder. With the aid of reminders, users are less likely to forget and more likely to engage in learning tasks after completing their current ongoing task.

8.4.3 Fine-grained Content Scheduling with Interruption Management. Our findings indicate that by incorporating support for adjustable operating times and learning intervals, learning opportunities can be expanded and the overall learning experience can be enhanced for learning services. Although users generally follow daily routines, they occasionally encounter variations in their routines. For example, as expressed by P21: *"When I stay up late the previous night, the next day, it's tough to stick to my usual sleep and wake-up routine."* In such cases, it can be helpful to empower users to customize their own operating times and learning intervals verbally via a human-in-the-loop approach. For example, we can support specific adjustments by providing commands like *"Give me more service for an hour from now on."* and *"Don't disturb me for the next 30 minutes."* This approach enables users to efficiently manage their learning opportunities. In existing models, virtual assistant settings for smart speakers are automatically synchronized with those of smartphones (e.g., Google Assistant) [21]. Given the widespread use of smartphone alarms, to eliminate the need for manual adjustment, it could also be possible to automate the adjustment of working time by importing alarm settings from a user's smartphone.

8.5 Design Implications for Voice-based Conversational Microlearning Services

Based on the findings of our pilot and main studies, we propose three design implications for voice-based conversational microlearning services.

8.5.1 Supporting Learning Time Awareness. Our results indicate that such services need to support learners in maintaining their ongoing awareness of learning time. To enable learners to end their learning according to their decision for in situ learning time allocation (See Section 3.1), we provided a continue-to-next inquiry step every two minutes to help the learners' time estimation. Alternatively, learners can be nudged with a simple earcon at regular intervals (e.g., one minute), similar to a metronome, a device that produces audible sound at regular intervals. Different earcon patterns could be provided to further enhance time awareness. For example, we can increase the number of earcons provided at each interval as the learning time increases. We can also enable learners to explicitly control the length of upcoming learning tasks by supporting duration-setting commands (e.g., *"I want to use 10 minutes from now on."*). Note that this approach was feasible for SpeechMaster; however, we did not implement it because of unintentional termination. We discussed details of unintentional termination in the following section.

8.5.2 Error Handling. Our results indicate that such services require support options to handle misunderstandings or transcription errors in responses. In our pilot studies, unintentional terminations occurred because of a misunderstanding of the learner's speech-shadowing sentences as generic termination commands (n = 73; see Appendix A.3.1). In our main study, unexpected terminations also occurred because of misrecognition caused by pronunciation, similar to termination commands (n = 226). This is the same as a transcription error, in that AI misinterprets and provides incorrect results. For example, in a single instance, *"skirt"* was recognized as *"stop"* and caused termination. This may be because speech recognition engines are more likely to be trained using datasets that are mostly based on native language speakers [50]. Thus, the pronunciation of non-native speakers may have a higher probability of being misrecognized [40]. However, it is necessary not only to improve the speech recognition engines of the speaker to enhance AI fairness, but also to handle transcription errors. Although existing smart speaker models recognize all user responses as commands, one potential approach is to empower developers to implement an option that skips the recognition of commands in response to a certain speaker question. Alternatively, smart speakers could provide a command (*"go-back"* or *"back to the previous conversation"*) that allows users to return to the previous conversation, which is similar to the previous page function in a web browser.

8.5.3 Supporting Multi-modality. In our interview, for enhanced speech shadowing experiences, several participants shared their insights about providing the following additional information: translations and spellings of sentences and history of past sentences and scores. In particular, they commonly expressed an interest in having access to the spelling of sentences when they could not repeat pronunciations. Although it is feasible to deliver this additional information via voice, it could be more effective to deliver it via visual modality (e.g., display). Recently, some smart speakers have built-in displays (e.g., Google Nest Hub and Amazon Echo Shows). This information can be visually displayed to such speakers. For traditional speakers (no display), multidevice interactions can be coordinated. Given that virtual assistants are systematically shared between smartphones and speakers (e.g., Google Assistant) [21], information can be visually displayed on the smartphone screen via shared assistants between smart speakers and smartphones.

8.6 Limitations and Future Research Directions

Although our study demonstrates that the interruptibility of proactive microlearning services can vary across learner contexts prior to engaging in the service, our results should be carefully interpreted and generalized for practical applications. First, we assumed that when learners engage in microlearning for a certain duration in a certain context, they would also be able to engage in learning for shorter durations in the same context. However, further studies are needed to confirm this hypothesis. Recently, some commercial smart speakers have a built-in display that also supports visual-manual modalities. Given that the visual-manual modality requires shorter distances to interact (e.g., touch input) than the auditory-verbal modality, our findings may not be generalizable to such multimodal smart speakers. Finally, although in our study we exclusively provided a smart speaker for a microlearning service (i.e., speech shadowing), a speaker is a multifunctional platform capable of providing a wide range of services beyond microlearning. However, smart speakers do not support concurrent execution of multiple services. For example, when using the same smart speaker, learners cannot access music and microlearning services simultaneously. In such scenarios, a microlearning service may compete with other smart speaker services. Consequently, further studies are required to explore the interruptibility of microlearning tasks under such competitive scenarios.

9 CONCLUSION

Advances in intelligent agents and the widespread adoption of smart speakers in domestic settings present new opportunities to expand opportune moments from mobile and computer environments to daily life by proactively providing conversational interactions. Our study provides initial insights into how the duration of conversational services and the user context prior to engaging in the services influence opportune moments when such services are proactively delivered in domestic settings. We hope that our findings will provide a primary step toward enabling various proactive conversational services, particularly those requiring interactions exceeding one minute, within domestic settings.

ACKNOWLEDGMENTS

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2023-00260267).

REFERENCES

[1] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 33 (sep 2017), 20 pages. https://doi.org/10.1145/3130898

[2] Lisa Aharon. 2020. What is the appropriate length of a microlesson? https://www.edapp.com/blog/what-is-the-appropriate-length-of-a-microlesson/

[3] Erik M. Altmann and J. Gregory Trafton. 2002. Memory for goals: an activation-based model. *Cognitive Science* 26, 1 (2002), 39–83. https://doi.org/10.1207/s15516709cog2601_2

[4] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (2006), 685–708. https://doi.org/10.1016/j.chb.2005.12.009 Attention aware systems.

[5] Brian P Bailey, Joseph A Konstan, and John V Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface.. In *Interact*, Vol. 1. 593–601.

[6] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (sep 2018), 24 pages. https://doi.org/10.1145/3264901

[7] Carol Boston. 2002. The concept of formative assessment. *Practical assessment, research, and evaluation* 8, 1 (2002), 9.

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[9] Ilona Buchem and Henrike Hamelmann. 2010. Microlearning: a strategy for ongoing professional development. *eLearning Papers* 21, 7 (2010), 1–15.

[10] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 74 (sep 2020), 28 pages. https://doi.org/10.1145/3411810

[11] Tamlin Conner Christensen, Lisa Feldman Barrett, Eliza Bliss-Moreau, Kirsten Lebo, and Cynthia Kaschub. 2003. A practical guide to experience-sampling procedures. *Journal of Happiness Studies* 4, 1 (2003), 53–78. https://doi.org/10.1023/A:1023609306024

[12] Tilman Dingler, Dominik Weber, Martin Pielot, Jennifer Cooper, Chung-Cheng Chang, and Niels Henze. 2017. Language Learning On-the-Go: Opportune Moments and Design of Mobile Microlearning Sessions. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 28, 12 pages. https://doi.org/10.1145/3098279.3098565

[13] Gilbert Dizon, Daniel Tang, and Yumi Yamamoto. 2022. A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers and Education: Artificial Intelligence* 3 (2022), 100088. https://doi.org/10.1016/j.caeai.2022.100088

[14] Fiona Draxler, Audrey Labrie, Albrecht Schmidt, and Lewis L. Chuang. 2020. Augmented Reality to Enable Users in Learning Case Grammar from Their Real-World Interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376537

[15] Darren Edge, Stephen Fitchett, Michael Whitney, and James Landay. 2012. MemReflex: Adaptive Flashcards for Mobile Microlearning. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (San Francisco, California, USA) *(MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 431–440. https://doi.org/10.1145/2371574.2371641

[16] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A. Landay. 2011. MicroMandarin: Mobile Language Learning in Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3169–3178. https://doi.org/10.1145/1978942.1979413

[17] Joel E. Fischer, Nick Yee, Victoria Bellotti, Nathan Good, Steve Benford, and Chris Greenhalgh. 2010. Effects of Content and Time of Delivery on Receptivity to Mobile Interruptions. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services* (Lisbon, Portugal) *(MobileHCI '10)*. Association for Computing Machinery, New York, NY, USA, 103–112. https://doi.org/10.1145/1851600.1851620

[18] Gerhard Gassler, Theo Hug, and Christian Glahn. 2004. Integrated Micro Learning–An outline of the basic method and first results. *Interactive computer aided learning* 4 (2004), 1–7.

[19] Luminiţa Giurgiu. 2017. Microlearning an Evolving Elearning Trend. *Scientific Bulletin - Nicolae Balcescu Land Forces Academy* 22, 1 (2017), 18–23. https://www.proquest.com/scholarly-journals/microlearning-evolving-elearning-trend/docview/2100359104/se-2 Copyright - Copyright Nicolae Balcescu 2017; Last updated - 2019-07-02.

[20] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C. Schultz, and Jue Wang. 2005. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1338–1343. https://doi.org/10.1109/IROS.2005.1545303

[21] Google. 2022. Set up your google nest or home speaker or display - android. https://support.google.com/googlenest/answer/7029485?hl=en&amp;co=GENIE.Platform%3DAndroid

[22] Google. 2023. Set & manage reminders - android. https://support.google.com/assistant/answer/9387035?hl=en&amp;co=GENIE.Platform%3DAndroid

[23] Joyce Ho and Stephen S. Intille. 2005. Using Context-Aware Computing to Reduce the Perceived Burden of Interruptions from Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 909–918. https://doi.org/10.1145/1054972.1055100

[24] Theo Hug. 2005. Micro Learning and Narration. Exploring possibilities of utilization of narrations and storytelling for the designing of "micro units" and didactical micro-learning arrangements. In *Fourth Media in Transition Conference*, Vol. 6.

[25] Shamsi T. Iqbal and Brian P. Bailey. 2005. Investigating the Effectiveness of Mental Workload as a Predictor of Opportune Moments for Interruption. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1489–1492. https://doi.org/10.1145/1056808.1056948

[26] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. PrivacyMic: Utilizing Inaudible Frequencies for Privacy Preserving Daily Activity Recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 198, 13 pages. https://doi.org/10.1145/3411764.3445169

[27] Hemin Jiang, Mikko Siponen, and Aggeliki Tsohou. 2021. Personal use of technology at work: a literature review and a theoretical model for understanding how it affects employee job performance. *European Journal of Information Systems* 32, 2 (2021), 331–345. https://doi.org/10.1080/0960085X.2021.1963193

[28] Soowon Kang, Cheul Young Park, Auk Kim, Narae Cha, and Uichin Lee. 2022. Understanding Emotion Changes in Mobile Experience Sampling. In *Proceedings*

of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 198, 14 pages. https://doi.org/10.1145/3491102.3501944

[29] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-Vehicle Proactive Auditory-Verbal Tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 175 (dec 2018), 28 pages. https://doi.org/10.1145/3287053

[30] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for In-Vehicle Multitasking: Influence of Voice Task Demands and Adaptive Behaviors. 4, 1, Article 14 (mar 2020), 22 pages. https://doi.org/10.1145/3381009

[31] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975).

[32] Judit Kormos and Kata Csizér. 2008. Age-Related Differences in the Motivation of Learning English as a Foreign Language: Attitudes, Selves, and Motivated Learning Behavior. *Language Learning* 58, 2 (2008), 327–355. https://doi.org/10.1111/j.1467-9922.2008.00443.x

[33] Innovative Language. 2023. Daily Dose by Innovative Language. https://www.englishclass101.com/alexa/

[34] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-Play Acoustic Activity Recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 213–224. https://doi.org/10.1145/3242587.3242609

[35] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 102 (nov 2018), 31 pages. https://doi.org/10.1145/3274371

[36] Hyunsoo Lee, Soowon Kang, and Uichin Lee. 2022. Understanding Privacy Risks and Perceived Benefits in Open Dataset Collection for Mobile Affective Computing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 61 (jul 2022), 26 pages. https://doi.org/10.1145/3534623

[37] Nick Leonard. 2023. English in 10 Minutes. https://podcasts.apple.com/us/podcast/english-in-10-minutes/id1159455275

[38] Abhinav Mehrotra, Mirco Musolesi, Robert Hendley, and Veljko Pejovic. 2015. Designing Content-Driven Intelligent Notification Mechanisms for Mobile Applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 813–824. https://doi.org/10.1145/2750858.2807544

[39] Hazel Morton and Mervyn A Jack. 2005. Scenario-Based Spoken Interaction with Virtual Agents. *Computer Assisted Language Learning* 18, 3 (2005), 171–191. https://doi.org/10.1080/09588220500173344

[40] Cristian Muñoz. 2023. Insightful resources for uncovering bias in English speech recognition. https://www.holisticai.com/blog/uncovering-bias-english-speech-recognition

[41] Kristine S. Nagel, James M. Hudson, and Gregory D. Abowd. 2004. Predictors of Availability in Home Life Context-Mediated Communication. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work* (Chicago, Illinois, USA) *(CSCW '04)*. Association for Computing Machinery, New York, NY, USA, 497–506. https://doi.org/10.1145/1031607.1031689

[42] Gonzalo Navarro. 2001. A Guided Tour to Approximate String Matching. *ACM Comput. Surv.* 33, 1 (mar 2001), 31–88. https://doi.org/10.1145/375360.375365

[43] Ambra Neri, Ornella Mich, Matteo Gerosa, and Diego Giuliani. 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning* 21, 5 (2008), 393–408. https://doi.org/10.1080/09588220802447651

[44] Jeungmin Oh, Darren Edge, and Uichin Lee. 2020. ScriptFree: Designing Speech Preparation Systems with Adaptive Visual Reliance Control on Script. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3334480.3382896

[45] Christina Pavlou. 2022. What is microlearning and how does it benefit your training? https://www.talentlms.com/blog/what-is-microlearning-and-its-benefits/

[46] Veljko Pejovic and Mirco Musolesi. 2014. InterruptMe: Designing Intelligent Prompting Mechanisms for Pervasive Applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 897–908. https://doi.org/10.1145/2632048.2632062

[47] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Serrà, Aleksandar Matic, and Nuria Oliver. 2017. Beyond Interruptibility: Predicting Opportune Moments to Engage Mobile Phone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 91 (sep 2017), 25 pages. https://doi.org/10.1145/3130956

[48] Martin Pielot, Rodrigo de Oliveira, Haewoon Kwak, and Nuria Oliver. 2014. Didn't You See My Message? Predicting Attentiveness to Mobile Instant Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3319–3328. https://doi.org/10.1145/2556288.2556973

[49] Jiuchao Qian, Jiabin Ma, Rendong Ying, Peilin Liu, and Ling Pei. 2013. An improved indoor localization method using smartphone inertial sensors. In *International Conference on Indoor Positioning and Indoor Navigation*. 1–7. https://doi.org/10.1109/IPIN.2013.6817854

[50] Kacper Radzikowski, Le Wang, Osamu Yoshie, and Robert Nowak. 2021. Accent modification for speech recognition of non-native speakers using neural style transfer. *EURASIP Journal on Audio, Speech, and Music Processing* 2021, 1 (2021). https://doi.org/10.1186/s13636-021-00199-3

[51] Leon Reicherts, Nima Zargham, Michael Bonfert, Yvonne Rogers, and Rainer Malaka. 2021. May I Interrupt? Diverging Opinions on Proactive Smart Speakers. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 34, 10 pages. https://doi.org/10.1145/3469595.3469629

[52] Gona Sirwan Mohammed, Karzan Wakil, and Sarkhell Sirwan Nawroly. 2018. The Effectiveness of Microlearning to Improve Students' Learning Ability. *International Journal of Educational Research Review* 3, 3 (2018), 32 – 38. https://doi.org/10.24331/ijere.415824

[53] Lucy Skidmore and Roger K. Moore. 2019. Using Alexa for Flashcard-Based Learning. In *Proc. Interspeech 2019*. 1846–1850. https://doi.org/10.21437/Interspeech.2019-2893

[54] U.S. DEPARTMENT OF STATE. 2023. Foreign language training. https://www.state.gov/foreign-language-training/

[55] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. 2019. "I don't own the data": End User Perceptions of Smart Home Device Data Practices and Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, 435–450. https://www.usenix.org/conference/soups2019/presentation/tabassum

[56] Sheng Tan, Linghan Zhang, Zi Wang, and Jie Yang. 2019. MultiTrack: Multi-User Tracking and Activity Recognition Using Commodity WiFi. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300766

[57] Association Tatoeba. 2006. What is tatoeba? https://tatoeba.org/en/about

[58] Andrew Trusty and Khai N. Truong. 2011. Augmenting the Web for Second Language Vocabulary Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 3179–3188. https://doi.org/10.1145/1978942.1979414

[59] Stuart Webb. 2007. The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics* 28, 1 (03 2007), 46–65. https://doi.org/10.1093/applin/aml048

[60] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2022. Understanding User Perceptions of Proactive Smart Speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 185 (dec 2022), 28 pages. https://doi.org/10.1145/3494965

[61] Jing Wei, Benjamin Tag, Johanne R Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 276, 15 pages. https://doi.org/10.1145/3491102.3517432

[62] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors* 50, 3 (2008), 449–455. https://doi.org/10.1518/001872008X288394 PMID: 18689052.

[63] Takashi Yamashita, Thomas J. Smith, Shalini Sahoo, and Phyllis A. Cummins. 2022. Motivation to learn by age, education, and literacy skills among working-age adults in the United States. *Large-scale Assessments in Education* 10, 1 (07 Mar 2022), 1. https://doi.org/10.1186/s40536-022-00119-7

[64] Fengpeng Yuan, Xianyi Gao, and Janne Lindqvist. 2017. How Busy Are You? Predicting the Interruptibility Intensity of Mobile Users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5346–5360. https://doi.org/10.1145/3025453.3025946

[65] Nima Zargham, Leon Reicherts, Michael Bonfert, Sarah Theres Voelkel, Johannes Schoening, Rainer Malaka, and Yvonne Rogers. 2022. Understanding Circumstances for Desirable Proactive Behaviour of Voice Assistants: The Proactivity Dilemma. In *Proceedings of the 4th Conference on Conversational User Interfaces* (Glasgow, United Kingdom) *(CUI '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3543829.3543834

[66] Manuela Züger and Thomas Fritz. 2015. Interruptibility of Software Developers and Its Prediction Using Psycho-Physiological Sensors. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2981–2990. https://doi.org/10.1145/2702123.2702593

A ITERATIVE DEVELOPMENT OF SPEECHMASTER

As shown in Table 16, we iteratively designed and developed SpeechMaster via six pilot studies with 29 learners. The procedures used in each study were consistent with those used in the main study. We first conducted an online orientation and then a field study in the learners' homes. After the field study, semi-structured interviews were conducted to gather in-depth feedback. In this section, we describe the key findings and relevant changes made during the pilot studies that were not covered in the main paper.

A.1 Conversational Microlearning Service

During the design process of the conversational microlearning service, the initial designs for the activity inquiry and *learning availability inquiry* remained the same. The reason-for-stop inquiry was provided in Study 3 to collect the learners' context for stopping microlearning. In this section, we explain the findings and related changes not covered in section 3.1.

A.1.1 Continue-to-next Inquiry. In the initial design, the service provided continue-to-next inquiry after every speech shadowing sentence, with no limit on the maximum number of sentences (or maximum duration of the interaction). This approach received feedback that repeated inquiries after every sentence made the learners feel fatigued. As a response, in pilot studies 3 and 4, we provided only 12 sentences without a continue-to-next inquiry. However, this approach led to feedback from learners desiring a longer learning duration, as the 12 sentences amounted to only approximately three minutes.

Consequently, for pilot study 5, we provided a microlearning task continuously up to 10 minutes without providing the continue-to-next inquiry step and allowed learners to stop their learning by saying *"Stop."* However, learners faced challenges when estimating their elapsed learning time. In the interviews, we found that prior to engaging in a microlearning task (e.g., at the moment of noticing the start of the task), learners pre-established an in-situ microlearning goal–whether and how long they would allocate their time for the current upcoming microlearning task–by considering their contextual factors. In addition, while engaging in a microlearning task, they preferred to be aware of their learning time to engage in microlearning for a specific duration. Therefore, for pilot study 6 and main study, we designed our service to provide a continue-to-next inquiry 3, 5, 7, and 9 min after the start.

A.1.2 Speech Shadowing Feedback. Once the learners repeated a given speech-shadowing sentence, feedback was provided in the following order:

- *Score:* Scores (0–100) were provided to allow learners to understand the similarity between their pronunciation and the original pronunciation. As in a previous study [44], the score was derived using an edit distance algorithm [42]. The algorithm counts the minimal number of transformations (e.g., inserting, removing, or substituting a letter) to eliminate discrepancies between the recognized and original pronunciations.
- *Compliment/encouragement:* After providing the score, SpeechMaster compliments (or encourages) the learners' effort. We designed our service to randomly select comments from a pool within a given score range. 43 comments were available for the final design. In pilot study 3, we provided compliments/encouragement in response to the opinion that learners want achievement. In pilot studies 3 to 6, a single comment was provided for each score range (e.g., 100s, 90s, 80s, etc.), but learners reported that they felt robotic rather than natural conversations with humans. For example, P4 from pilot study 6 stated, *"I remember the comment changed depending on the score, like 'Perfect!!'. But the same comments kept coming out (when I got similar scores) so I didn't feel like I'm having a conversation, and felt like talking to a machine. I wish there were various comments."* To enhance conversation experiences, we designed our service to randomly provide comments for a given score range. For instance, there are two pools of tokens for 90s, and the service selects a token from each pool and combines the two tokens to make a complementary comment.
- *Recognized pronunciation:* When the score is not perfect (or 100), our service provides recognized pronunciation as follows: *"Your speech sounds like <recognized pronunciation>."* This was because in our pilots, we found that learners are interested in how their pronunciation is heard.

A.2 Speaker Add-on Device

To enable existing smart speakers to operate proactively, we designed and developed a speaker add-on device (See Section 3.2 for its final design). The device was applied in pilot study 3, and its performance was confirmed in pilot studies 4–6. The performance results can be obtained later by reviewing the voice-command volume-adjustment features.

In pilot studies 1 and 2, similar to a previous study [60], we initially attempted to proactively operate a smart speaker using two earbuds. However, this attempt often fails when the ambient noise is loud (e.g., high volume music). In such noisy environments, the volume of voice commands must be higher than that of earbuds. Therefore, we have developed an add-on device to increase the volume of voice commands in noisy environments. In addition, we developed voice-command volume adjustments and retriggering features.

A.3 Applications for Proactive Microlearning Triggering

We also iteratively developed three apps (triggering, sensing, and image deleter apps) that operate with a speaker add-on device. In this section, we describe the findings and related changes in triggering and sensing apps that were not discussed in the main paper.

A.3.1 Triggering App.

- *Operating Hours:* In initial design of the app, learners could set only a single operating time. However, learners expressed a preference for setting distinct operating times on weekdays and weekends owing to different daily patterns. Therefore, in pilot study 3, the app incorporated such a setting. In the final design, learners could also select specific operating days of

Table 16: Iterative development of SpeechMaster and Proactive Smart Speaker.

				Study 1	Study 2	Study 3	Study 4	Study 5	Study 6	Final Design
Conversational Microlearning Service	Activity Inquiry	Ask learners what they were doing: "Hi, what are you doing now?"		O	O	O	O	O	O	O
	Learning availability inquiry	Ask if the learner wants to use service: "Would you like to practice your pronunciation? Please answer yes or no."		O	O	O	O	O	O	O
	Continue-to-next inquiry	Ask if the learner wants to continue: "Would you like to continue? Please answer yes or no."		per sentence		X			per period (3m, 2m, 2m, 2m, 1m)	per period (3m, 2m, 2m, 2m, 1m)
	Reason-for-stop inquiry	Ask learners why they are leaving: "Why do you want to stop?"		X	X	O	O	O	O	O
	Speech shadowing feedback	Speech shadowing ability scoring		Simple comparison (Good or bad)	Edit Distance Algorithm (0-100 point)					Edit Distance Algorithm (0-100 point)
		Providing reactive tokens for compliment / encouragement		X	X	Different comment by score range (e.g., "Excellent!", "Great!", "Good job!")				Various reactive tokens according to score
		Tell the learner how the pronunciation sounds like: "Your speech sounds like <recognized pronunciation>."		X	X	O	O	O	O	O
		Sentence levels		X	10 Levels (By sentence length)		20 Levels (By number of syllables)			20 Levels (By number of syllables)
	Maximum service duration			unlimited		12 sentences		10 minutes		10 minutes
Schedule	Test schedule			1 day	2 days	7 days	7 days	7 days	7 days	21 days
	Participant number			4	4	5	6	4	6	27
Speaker Add-on Device and Related Applications	Holder (Triggering device)			Earbuds + holded with silicone sealant	Earbuds + 3D-printed holder	Audio power amplifier with an external speaker + 3D-printed holer				Audio power amplifier with an external speaker + 3D-printed holder
	Triggering app	Operating hours		Set single working time		Set multiple working times (Weekdays / Weekends)				Set working days and multiple working times (Weekdays / Weekends)
		Random triggering interval		avg 45 minutes (30-60m)			avg 2 hours (1-3h)			avg 1 hour (30-90m)
		Voice-command volume adjustment and Retriggering		X	X	X	O	O	O	Can be retriggered up to 3 times
		Contextual data collection	Image capturing	X	X	X	O (Back camera)			O (Front camera)
			Wide angle lens	X	X	X	O	O	O	O
			Audio recording	O	O	O	O	O	O	O
		Recovery mode		X	X	X	X	X	X	O
	Image deleter			X	X	X	O	O	O (Zoom function)	O (Zoom function)
	Sensing app			X	X	SSID	BSSID			BSSID
Kit	Added to kit			-	-	Long selfie stand			Small stand	Guiding note

the week by considering individual preferences (e.g., taking rest on Saturdays).

- *Triggering Interval:* Up to pilot study 3, the average triggering interval was initially set to 45 min (range = 30–60 min). However, learners expressed that the microlearning tasks were provided more frequently than expected. Consequently, we adjusted the interval to an average of 120 min (range, 60–180 min) from the pilot study 4. In the final design, we set a triggering interval as a random interval with an average of 60 minutes (range = 30–90 minutes) by considering learners' feedback that they desire more frequent learning opportunities.
- *Voice-command Volume Adjustment and Retriggering:* To mitigate triggering failures, we added volume adjustment and retriggering features from the pilot study 4. Our results from pilot studies 4 to 6 show that these additions significantly reduced recognition failures; overall, microlearning tasks were successfully triggered in 98.5% of cases (n = 664 / 674). In the final design, to prevent further failures, the maximum number of retriggering attempts was increased to three.
- *Recovery Mode:* We incorporated a restart button on the smartphone screen to enable learners to resume conversations when conversations with the service were unintentionally terminated. In our pilot study, such unintentional termination occurred under the following three circumstances: (1) sentences containing termination words or expressions (e.g., *"Stop whispering."*, *"I want to forget it"*), (2) sentences not containing explicit termination words or expressions but were misinterpreted (e.g., due to learner pronunciation – 'stuck' in *"We got stuck."* or 'I-sTom' in *"Is Tom black?"* recognized as 'stop'), and (3) sentences including words or expressions that call other services. For example, when the learner repeated after the sentence *"Thank you both,"* the service responded with *"I'm honor to serve."* and terminated.

A.3.2 Sensing App. To provide tasks when learners are at home, we developed and provided a sensing app from pilot study 3 that detects WiFi signals. The app initially compared the SSID of the surrounding WiFi signals with those of the home WiFi signals. However, SSID are often duplicated (i.e., duplicated network names). Therefore, our final app compared the WiFi MAC address (or BSSID) from the pilot study 4.

A.3.3 Image Deleter App. In pilot study 4, to protect learners' privacy, we introduced an image deleter that allowed learners to quickly review and delete collected images that they did not want to share with the researchers. In pilot study 6, considering learners' need to review images in more detail, we introduced an image zoom function, allowing learners to enlarge images. In the app, learners could swipe left or right to see a previous or subsequent set of images captured within the same two-minute period. Similar to a gallery app, learners can scroll through images, zoom in and out, and select or delete multiple images. In our pilot study, learners had low privacy concerns because they had full control over the collected images. For example, in study 6, P3 stated, *"I wasn't bothered because I knew I could delete images, like those where I took off my clothes."*

In pilot study 4–6, upon deleting images, learners were asked to provide deletion reasons (e.g., *"In these photos, I was undressed."*). During these studies, nine learners executed 90 deletions, while seven chose not to delete the images. For these reasons, we identified four types of privacy concerns that led to deletions: awkward face or posture (31%, n = 28), changing clothes or undressing (29%, n = 26), showing other people (27%, n = 24), and private items in sight (e.g., underclothes, monitor screens) (13%, n = 12). Specifically, the learners deleted images in which their faces or postures appeared awkward (31%, n = 28). These images included close-up facial shots (n = 11), awkward postures (n = 11), images captured during meals or makeup application (n = 4), and yawning or nose-blowing moments (n = 2). They also deleted images in which they were changing clothes or undressing (29%, n = 26) or featuring other people, such as friends or family, due to a lack of consent (27%, n = 24). Even when no individuals were visible, they deleted images showing private items (e.g., underclothes and electronic devices) because of concerns about personal information (13%, n = 12).

A.4 Additional Figures

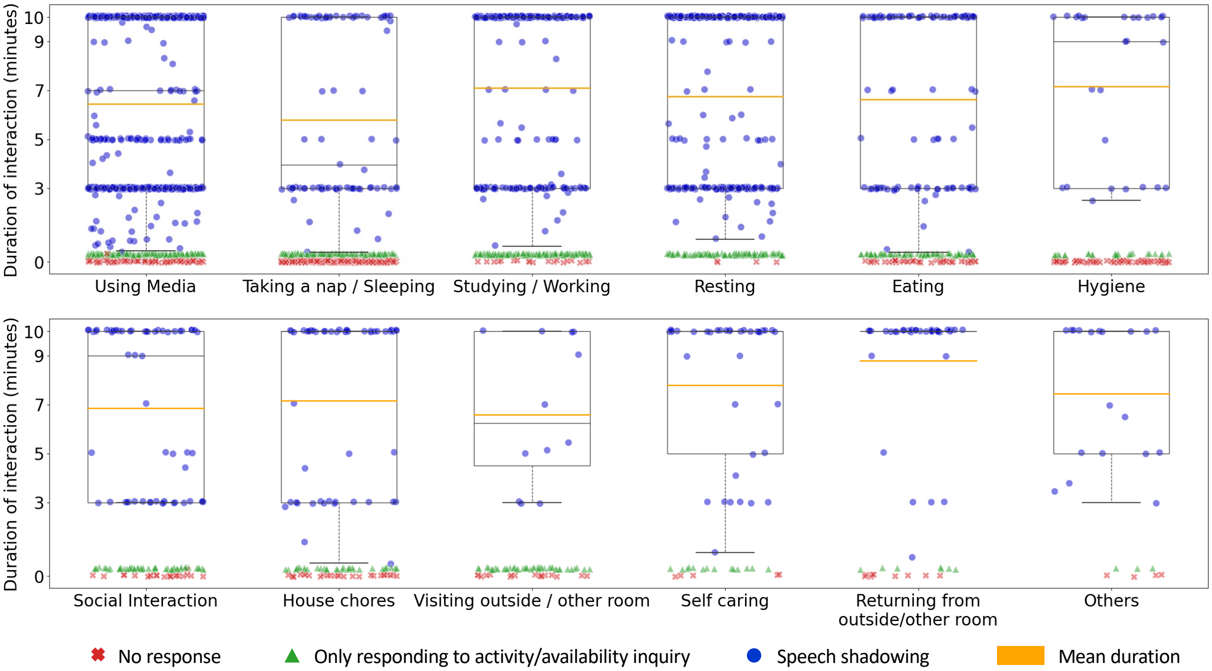


Figure 5: Distribution of microlearning tasks across activity categories.

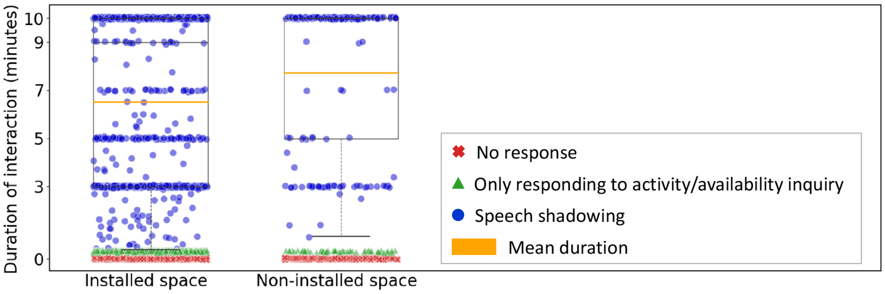


Figure 6: Distribution of microlearning tasks across location types (installed space vs. non-installed space).

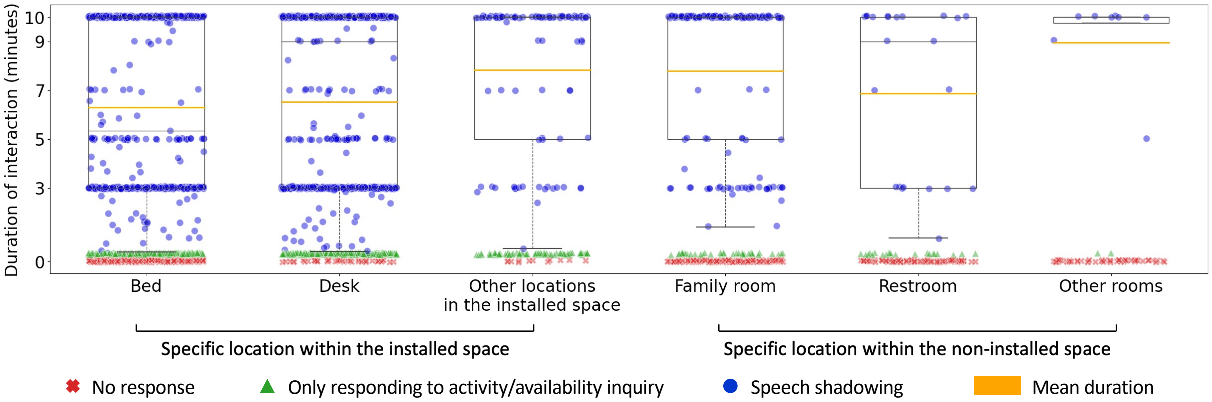


Figure 7: Distribution of microlearning tasks across specific locations within the installed and non-installed spaces.