Scientific Data

https://doi.org/10.1038/s41597-026-07618-0

Article in Press

Multimodal-context Interruptibility Dataset for Proactive Services on Smart Speakers at Home

Received: 15 September 2025

Accepted: 5 June 2026

Published online: 13 August 2026

Cite this article as: Sunatullaev, G., Hwang, J., Lee, J. *et al.* Multimodal-context Interruptibility Dataset for Proactive Services on Smart Speakers at Home. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07618-0

Golibjon Sunatullaev, Jiwoo Hwang, Jungmin Lee, Dongju Son, Kyung Ho Son & Auk Kim

A Multimodal-context Interruptibility Dataset for Proactive Services on Smart Speakers at Home

Golibjon Sunatullaev[1], Jiwoo Hwang[1], Jungmin Lee[1], Dongju Son[1], Kyung Ho Son[1], and Auk Kim[1,*]

[1]Kangwon National University, Chuncheon, 24341, South Korea
[*]corresponding author(s): Auk Kim (kimauk@kangwon.ac.kr)

ABSTRACT

As smart speakers become increasingly integrated into everyday home life, there is growing interest in leveraging these speakers to proactively deliver personalized proactive services. To enhance user experience and engagement of proactive services, a key challenge is identifying opportune moments, user contexts when users are mostly interruptible to engage in proactive services. Researchers have identified such moments by exploring interruptibility across various user contexts (in which contexts users are more interruptible), and several datasets have been released with interruptibility labels. However, no publicly available dataset has focused specifically on interruptibility within home environments. To fill this gap, we present INPROSH, a dataset collected from 26 participants over a three-week in-the-wild field study. Participants used proactive services via smart speakers in their home. INPROSH comprises 2,830 cases, each annotated with interruptibility labels and enriched with contextual information, including temporal, spatial, and behavioral contexts, as well as surrounding image and sound recordings near the smart speakers. We believe INPROSH will support deeper understanding and more accurate detection of interruptibility in home environments.

Background and Summary

Proactive services delivered by smart speakers in home environments have recently gained significant popularity[1,2]. In contrast to reactive services that are delivered upon explicit user requests, proactive services deliver personalized services without such requests. Thereby proactive services enhance the overall user experience and satisfaction. For example, when traffic congestion is expected, to ensure the user departs on time, a proactive service can be delivered proactively to notify the user to depart earlier than scheduled[3]. Smart speakers are particularly well suited to delivering these proactive services in home environments, because they are typically installed in accessible locations that facilitate user interaction[4–7]. Furthermore, since these devices are based on auditory-verbal interactions, their proactive interactions do not interfere with household chores, which are often visual-manual tasks (e.g., cooking, cleaning, or doing laundry)[8,9].

Proactive services can enhance user experience and satisfaction. However, such services can negatively impact both when the services were delivered at an inopportune moment, in which users are unavailable (or uninterruptible) to engage with the services[10,11]. Specifically, providing proactive services at inopportune moments can cause stress and discomfort[10], contribute to increased human error[12,13], and prolong task completion times[11]. For example, delivering a proactive service while a user is engaged in a conversation with others (e.g., talking on the phone or in a meeting) could disturb their conversation and reduce communication effectiveness.

To identify opportune moments, researchers have investigated and measured interruptibility across various user contexts (i.e., in which contexts whether users are more available or interruptible). For the investigations, several open datasets have been released. Such datasets include interruptibility labels (e.g., interruptible or uninterruptible) and user contexts at the time of measuring interruptibility. As shown in Table 1, 2, prior open datasets have primarily focused on smartphone and driving environments. For example, Pielot et al. release a dataset for investigating interruptibility of smartphone users toward proactive notifications of smartphones[14]. Their dataset contains interruptibility data and user contexts at the time of measuring interruptibility. The user contexts include temporal contexts (e.g., notification time), spatial contexts (e.g., work, home, other), and behavioral contexts (e.g., outdoor user activity). Another dataset, INAGT dataset focuses on interruptibility of driver

toward proactive agents on the infotainment system in driving environments[15] and provides interruptibility data (i.e., drivers' responses to questions asking whether it is good time to talk with), and contextual data (e.g., car GPS, driving behaviors, etc.).

While prior open datasets are useful to identify opportune moments or to examine interruptibility across different user contexts toward proactive services in smartphone or vehicular environments, they are not directly applicable to identifying

opportune moments in home environments. This is because their contextual data do not include details on domestic contexts (e.g., home activities, such as cooking, cleaning, taking a nap, or doing laundry, etc.; fine-grained user location at home, such as bedroom or living room, etc.).

In this paper, we introduce INPROSH[16], a multimodal dataset designed to support the identification of opportune moments for proactive services in smart home environments. Specifically, our dataset includes interruptibility labels (i.e., interruptible or uninterruptible to engage in proactive long- or short-interaction with smart speakers) and user contextual data at the time of measuring interruptibility. The contextual data includes (1) temporal contexts (e.g., service delivering and interaction time), (2) spatial contexts (e.g., participants' location within home), (3) behavioral contexts (e.g., user activity). For the data collection, we conducted a three-week in-the-wild field study with 26 participants. The participants interacted with an English-language learning proactive service via smart speakers in their homes. To the best of our knowledge, INPROSH[16] is the first publicly available interruptibility dataset that focuses on user interruptibility and contexts in domestic or home environments. We anticipate that this dataset will be a valuable resource for future research on identifying opportune moments for proactive services in smart-home environments.

Methods

Ethics Approval

The research procedure was reviewed and approved by the Institutional Review Board of Kangwon National University (KWNUIRB-2021-08-007). We explained to participants about the background, purpose, and procedures of the data collection, and the potential privacy risks associated with identifiable data, including image and audio recordings. After the explanation, we obtained written consent for the use and sharing of their identifiable data for research purposes. In addition, during the data collection, participants deleted any images they did not want to share with others using the image deleter app (for the details, see Section Add-on Device and Application).

Data Collection

To collect contextual data, we implemented a proactive service, SpeechMaster, in commercial smart speakers (i.e., Google Nest Mini) and developed a speaker add-on device and applications (i.e. triggering app, image deleter app).

SpeechMaster: proactive conversational microlearning service

SpeechMaster is a proactive, conversational microlearning service for smart speakers. It is designed to improve English pronunciation through speech shadowing techniques[17–19]. As illustrated in Figure 1, the process of SpeechMaster consists of four steps: (1) activity inquiry, (2) availability inquiry, (3) speech shadowing, and (4) continue-to-next inquiry. At any step, the user may say "Stop" to immediately terminate the current session. The following describes each step in detail:

- *Activity inquiry*: This step collects contextual information about the user's current activity before initiating the proactive service (i.e., SpeechMaster). SpeechMaster begins the interaction with the prompt: "Hi, what are you doing now?" This serves both to gather user activity context and to provide an auditory cue that initiates the conversation[6].

- *Availability inquiry*: This step determines whether the user is available (or interruptible) to engage in a speech shadowing practice session. SpeechMaster asks: "Would you like to practice your pronunciation? Please answer yes or no." A "No" response ends the session, while a "Yes" response advances to the *Speech shadowing* step.

- *Speech shadowing*: In this step, SpeechMaster delivers an English sentence aloud and prompts the user to repeat it (e.g., "Repeat after me. <Sentence>"). Upon repetition, the system analyzes pronunciation accuracy and provides feedback, such as a numerical score (0–100). The process then continues with the next sentence (e.g., "Next sentence. Repeat after me. <Sentence>").

- *Continue-to-next inquiry*: This step is designed to collect information on how long the user has participated in the service and whether they wish to continue. Beginning 3 minutes after the interaction starts, and every 2 minutes thereafter (i.e., at 3, 5, 7, and 9 minutes), SpeechMaster asks: "Would you like to continue? Please

answer yes or no." A "Yes" response resumes the *Speech shadowing* step. A "No" response ends the session. If the session reaches the maximum duration of 10 minutes, SpeechMaster announces: "You have reached the maximum time for now. See you next time," and concludes the session.

Add-on Device and Application

To enable commercial smart speakers to deliver proactive services, capture rich contextual data, and safeguard participant privacy, we developed an add-on device, a triggering app, and an image deleter app. Detailed descriptions of each are provided below.

- *Add-on device*: As shown in Figure 2, the device attaches directly to the smart speaker. It consists of a 3D-printed circular frame, an amplifier, and an external 3W speaker, all connected to a smartphone. The device delivers prerecorded voice commands to the speaker's microphone, from the triggering app installed on the smartphone, and the speaker is activated by listening to the voice commands.
- *Triggering app*: The triggering app triggers SpeechMaster by delivering prerecorded commands to the smart speaker via the add-on device. To deliver successfully, it dynamically adjusts the volume of voice commands based on the ambient noise levels measured two minutes before delivering proactive service. Furthermore, it captures surrounding visual contextual data by taking images and sound recordings (e.g., human voice and ambient noise) at one-second intervals for two minutes before delivering proactive service. To avoid triggering the service when the user is not at home, the app scans nearby MAC addresses of WiFi signals around the smartphone. This allows the app to detect the user's smartphone, and thus infer the user's presence at home.
- *Image deleter app*: Although participants consented in advance to the release of their image data for research purposes, they might still not want to share certain images. To address this, we developed the image deleter app, which enabled participants to review and selectively remove such images before data release. The image deleter app displayed captured images in a grid layout, allowing users to browse, select, and delete any images.

Contextual Data Collection

SpeechMaster was triggered at random intervals of 30 to 90 minutes (on average, every 60 minutes) within participant-defined operating hours (e.g., 09:00–24:00) while the participant was at home. The operating hours were configured separately for weekdays and weekends to accommodate each participant's daily patterns (e.g., waking and sleeping times). Each time SpeechMaster was triggered, the triggering app collected sound and image data at one-second intervals for two minutes prior to the proactive service delivery, as described above. After the three-week study, participants reviewed the collected images using the image deleter app and removed any images they did not wish to share.

Recruitment

For our data collection, we recruited 26 participants from local communities. As SpeechMaster is a conversational microlearning application designed to help improve English pronunciation, we selected participants aligned with this objective to ensure SpeechMaster would serve as a useful and personalized tool for them. For participants selection, we considered two key eligibility criteria: (1) interest in improving their English pronunciation skills, and (2) spending more than four hours per day at home (excluding sleep) on at least five days per week. The participants received approximately 80 US dollars for their participations. The average age of the participants was 23.5 years (SD = 9.28, range = 18-59 years), and 10 participants (38.5%) were female.

Procedure of data collection

Before data collection, participants visited the laboratory to sign informed consent forms and receive experiment kits. To ensure a smooth data collection process, participants underwent an orientation session detailing the procedure of installing experiment kits. Next, data were collected over three weeks through an in-the-wild field study. During this period, six researchers periodically reviewed participant response data to monitor completeness and consistency.

During data collection, participants responded with detailed contextual information to the provided contextual inquiries, including the *Activity inquiries* (for the details, see Section SpeechMaster). Responses to the *Activity inquiries* detailed their current activity (e.g., what they were doing), location (e.g., specific room and position or space within the house), method (e.g., whether tools were involved), presence of others (if anyone), and rationale (e.g., "I was watching a movie on my laptop at my desk").

If a participant provided incomplete responses (e.g., failing to mention watching a movie while eating) or did not respond to contextual inquiries, additional responses were collected via instant messaging once a day. For incomplete responses, participants were prompted to clarify their activities at the time. For non-responses, participants were asked to explain why they had not responded. Consistent with prior research, we provided recall cues to facilitate participants' memory, including response timestamps, sound recordings, and text-based responses[6].

Throughout the data collection period, the six researchers annotated activities on each delivered service by

cross-examining the participants' SpeechMaster responses, daily supplementary reports, images, and audio recordings. After the data collection, five researchers then analyzed and categorized the annotated activities into 12 distinct categories, as detailed in ActivityCat-egories.pdf included in our dataset. Specifically, each researcher initially developed a codebook independently, and they collaboratively refined the codebook through continuous discussion and iterative review until full agreement was reached on the final codebook. Then, based on this codebook, each researcher independently categorized the annotated activities, and they refined the categorizations through the same collaborative process. Instances where participants described multiple activities in

a single response were categorized into multiple relevant categories. For example, if a participant reported, "I was eating dinner while watching YouTube on [my] laptop at [my] desk," the activity was classified under both the 'eating' and 'using media' categories.

Data Records

The INPROSH dataset[16] is available upon application and signing of a Data Usage Agreement (DUA) for research purposes in the online Zenodo repository. In the following sections, we present detailed descriptions of our dataset, including *UserInfo.csv*, *ServiceSession.csv*, *Interruptibility.csv* and *UserContexts.csv*. All data were formatted as Comma-Separated Values (CSV) and Portable Document Format (PDF), images in Joint Photographic Experts Group (JPG) format, and sound in Third Generation Partnership Project (3GP) format. Figure 3 presents an overview of our dataset.

User Information

UserInfo.csv contains demographic information of participants, including participant-specific data collection settings (i.e., the type of their house and the installed position of smart speakers).

- uid: unique identifier of each participant.
- age: numerical variable indicating international age of each participant at the time of the experiment.
- gender: categorical variable indicating each participant's gender: (1) 'M' or (2) 'F', corresponding to male and female, respectively.
- homeType: categorical variable indicating the type of residence: 'more-than-two-bedroom house', 'two-bedroom house', and 'one-bedroom house'.
- speakerLocation: categorical variable indicating the room where the smart speaker was installed: 'Bed Room' or 'Study / Working Room'.
- speakerPosition: categorical variable indicating the smart speaker's placement within the room: 'Desk' or 'Bed'.

Service Session Data

ServiceSession.csv contains service session information, such as the service session delivery time.

- uid: unique identifier of each participant. This column matches with the uid in *UserInfo.csv*.
- sid: a unique identifier of each service session.
- weekOfExperiment: integer variable indicating the week within the data-collection period, ranging from 1 to 3.
- dayOfWeek: categorical variable indicating the day of the week: 'Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', or 'Sun'.
- startTime: time-of-day variable indicating when each service session started, recorded in 24-hour format as HH:mm:ss.sss.
- activityInquiry: continuous variable indicating the elapsed time from startTime to the onset of *Activity inquiry* step, recorded in seconds-milliseconds format as ss.sss. This value is empty if the participant did not respond to this step.
- availabilityInquiry: continuous variable indicating the elapsed time from startTime to the onset of *Availability inquiry* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if service session terminated before this step began.
- speechShadowing1: continuous variable indicating the elapsed time from startTime to the onset of *Speech shadowing* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if the service session terminated before this step began.
- continue-to-nextInquiry1: continuous variable indicating the elapsed time from startTime to the onset of *Continue-to-next inquiry 1* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if

the participant did not respond to this step or (2) if the service session terminated before this step began.

- speechShadowing2: continuous variable indicating the elapsed time from startTime to the onset of *Speech shadowing* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if the service session terminated before this step began.
- continue-to-nextInquiry2: continuous variable indicating the elapsed time from startTime to the onset of *Continue-to-next inquiry 2* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or (2) if the service session terminated before this step began.

- speechShadowing3: continuous variable indicating the elapsed time from startTime to the onset of *Speech shadowing* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if the service session terminated before this step began.
- continue-to-nextInquiry3: continuous variable indicating the elapsed time from startTime to the onset of *Continue-to-next inquiry 3* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or (2) if the service session terminated before this step began.
- speechShadowing4: continuous variable indicating the elapsed time from startTime to the onset of *Speech shadowing* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if the service session terminated before this step began.
- continue-to-nextInquiry4: continuous variable indicating the elapsed time from startTime to the onset of *Continue-to-next inquiry 4* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or (2) if the service session terminated before this step began.
- speechShadowing5: continuous variable indicating the elapsed time from startTime to the onset of *Speech shadowing* step, recorded in seconds-milliseconds format as ss.sss. This value is empty (1) if the participant did not respond to this step or
  (2) if the service session terminated before this step began.
- endTime: continuous variable indicating the elapsed time from startTime to the end of the proactive service, recorded as seconds-milliseconds format as ss.sss.
- endType: categorical variable indicating how and at which step the service session terminated. Possible values are 'max' and codes of the form '<step name>_<end type>' (e.g., 'continue-to-nextInquiry1_no'). 'max' indicates that the service reached the maximum duration of 10 minutes and terminated automatically. Codes of the form '<step name>_<end type>' indicate that the service stopped during <step name> step due to <end type>. Specifically, <step name> includes *Activity inquiry*, *Availability inquiry*, *Speech shadowing 1–5*, and *Continue-to-next inquiry 1–4*. In addition, <end type> includes no (the participant answered "No."), stop (the participant answered "Stop."), no-response (the participant did not respond to at that step and a session terminated automatically).

Interruptibility Data

Interruptibility.csv contains interruptibility information (i.e., whether participants were available or interruptible) for each proac-tive service session. The interruptibility value is binary (interruptible or uninterruptible). Two distinct interruptibility labels are provided to reflect different interaction lengths: *SHORT_INTERACTION* and *LONG_INTERACTION*. *SHORT_INTERACTION* represents interruptibility for short interactions (i.e., less than one minute service). This label is particularly useful for identifying opportune moments to deliver proactive services that involve no more than two conversational turns (i.e., a single-turn interaction in which the smart speaker issues one prompt and receives one response without further follow-up)[20]. *LONG_INTERACTION* represents interruptibility for extended interactions (i.e., longer than three minutes). This label is well suited for identifying opportune moments to deliver proactive services that require more than two conversational turns (i.e., multi-turn interactions comprising a sequence of related prompts and responses)[20]. Figure 4 illustrates examples of these two interaction types.

- uid: unique identifier of each participant. This column matches with the uid in *UserInfo.csv*.
- sid: a unique identifier of each proactive service session. This column matches with the sid in *ServiceSession.csv*.
- SHORT_INTERACTION_interruptibility: binary variable indicating interruptibility for the *SHORT_INTERACTION*: 'Interruptible' or 'Uninterruptible'. It takes the value 'Interruptible' if the participant interacts with proactive services (e.g., responding at the *Activity inquiry* step).
- LONG_INTERACTION_interruptibility: binary variable indicating interruptibility for the *LONG_INTERACTION*. It takes the value 'Interruptible' if the participant interacts with the proactive service for more than three minutes.

User Context Data

UserContexts.csv contains information about user context information before proactive service delivery. The user contexts include behavioral context (i.e., user activities) and spatial contexts (i.e., user location and position. see Figure 5).

- uid: unique identifier of each participant. This column matches with the uid in *UserInfo.csv*.
- sid: a unique identifier of each service session. This column matches with the sid in *ServiceSession.csv*.
- activity1: categorical variable indicating each participant's primary activity before proactive service delivery. Possible values correspond to the categories listed in *ActivityCategories.pdf*.

- activity2: categorical variable indicating each participant's secondary activity (i.e., simultaneously engaged in with activity1; e.g., eating while using media). Possible values correspond to the categories listed in *ActivityCategories.pdf* and an empty value. An empty value denotes that no secondary activity was performed.
- userLocation: categorical variable indicating the room in which each participant was located within the home: 'Bed Room', 'Living Room', 'Rest Room', 'Study / Working Room', 'Other Bed Room', 'Laundry Room', 'Kitchen', 'Dress Room', 'Mud Room', or 'Balcony'.
- roomType: categorical variable indicating the type of each room within the home: 'Bed Room', 'Living Room', 'Rest Room', 'Study / Working Room', or 'Other Room'.
- userPosition: categorical variable indicating each participant's position within speakerLocation (i.e., the room where the smart speaker was installed): 'Bed', 'Desk', 'Other Position', or empty. An empty value denotes that the participant was not in the speakerLocation.

ActivityCategories.pdf contains information about 12 activity categories and detailed descriptions for each activity.

The multimodal data, consisting of surrounding images and sound recordings, are provided as separate participant-level archive files (uid1.zip–uid27.zip). These data were collected during the two minutes until each proactive service began. As illustrated in Figure 3, the data are organized hierarchically: each top-level uid directory contains sid subdirectories. Each sid subdirectory includes (1) a sound recording file (sound_recording.3gp) and (2) an images subdirectory containing 120 images captured at one-second intervals. For privacy reasons, six image sets were removed (uid5: sid581; uid15: sid1629, sid1639, sid1644, sid1685, sid1723) because participants chose not to share content containing sensitive information.

Technical Validation

Machine Learning Analysis

To ensure that our dataset is technically convincing, we built and evaluated machine learning models to predict users' interruptibility (i.e., whether the users are available or interruptible to engage with) toward proactive services in either *SHORT_INTERACTION* or *LONG_INTERACTION* (see Section Interruptibility Data). To build and evaluate the machine learning models, we first preprocessed our data, then extracted features. These features were used to build and evaluate various machine learning algorithms. For the evaluation, we considered 5-fold cross-validation and leave-one-subject-out (LOSO) cross validation. The pipeline of our machine learning analysis is illustrated in Figure 6. Detailed procedures are described in the following sections.

Preprocessing and Feature Extraction

We first preprocessed our contextual data (i.e., *UserInfo.csv*, *UserContexts.csv* and *ServiceSession.csv*). From these data, we extracted the following types of features to capture relationships between interruptibility and various types of user contexts: temporal features (i.e., startTime and dayOfWeek), spatial features (i.e., homeType, roomType, userPosition, speakerLocation, and speakerPosition), and behavioral features (i.e., activity1 and activity2).

For the temporal features, we utilized a cyclical encoding method for both starTime and dayOfWeek to capture their cyclical pattern[21]. Each feature was transformed into two continuous variables (sine and cosine), ranging from -1 to 1. For the spatial features, in addition to the variables listed above, we calculated a proximity feature representing the distance between the user and the smart speaker based on roomType, userPosition, speakerLocation, and speakerPosition. This feature is an ordinal variable ranging from 0 to 2; higher values indicate a closer distance. For example, if the user and the speaker are in different rooms, the proximity is 0. If they are in the same room but in different positions, the proximity is 1. If they are in the same room and the same position, the proximity is 2.

For the behavioral features, since users could be engaged in multiple activities at the same time, we applied one-hot encoding to represent each activity independently in a binary format, enabling the machine learning models to recognize them as independent categorical features[22]. For example, if the user was engaged in one activity, 'Hygiene', when the proactive service was provided, the corresponding activity feature 'act_Hygiene' is marked as 1, and all other activity features are marked as 0. Similarly, if the user was engaged in two activities, such as 'Eating' and 'Using a Media' at the same time, the corresponding features 'act_Eating' and 'act_Using a Media'

are marked as 1, and all other activity features are marked as 0.

For class, we considered two types of interruptibility classes: *SHORT_INTERACTION* and *LONG_INTERACTION*. For each class type, we build separate models. Each class predicts interruptibility toward proactive services with different interaction lengths. Models associated with *SHORT_INTERACTION* can be used to identify opportune moments to deliver proactive services that involve no more than two conversational turns. Other models associated with *LONG_INTERACTION* can be used for identifying opportune moments to deliver proactive services that require more than two conversational turns.

Model Training and Evaluation

For the machine learning algorithms, we considered Random Forest[23], Gradient Boosting[24], XGBoost[25], LightGBM[26], CatBoost[27], and Support Vector Machine (SVM)[28]. Random Forest, Gradient Boosting, XGBoost, LightGBM, and CatBoost are tree-based ensemble learning methods capable of handling non-linear relationships between features. They have been widely considered in prior studies on prediction of users' willingness to engage[29–31]. SVM was included due to its robustness in handling both linear and non-linear relationships. For a baseline model, similar to prior studies[32,33], we trained a dummy classifier that always predicts the majority class, and compared performances with our machine learning models.

In real-world deployment scenarios, deployed models are often not trained on data from the actual users of the system. For example, when a model is deployed on Google Nest Mini speakers, it is unlikely that the training data includes usage data from those specific end users. To account for this gap and better assess generalizability, we also considered LOSO cross-validation (CV), in addition to 5-fold CV. LOSO CV evaluates performance on entirely unseen users by partitioning the dataset into folds such that each fold contains all samples from a single participant. In each iteration, one fold was used as the test set while the remaining folds served as the training set, and this process was repeated until every participant's data had been used once for testing. For 5-fold CV, the dataset was randomly divided into five equal-sized folds. Across five iterations, one fold was held out for testing and the remaining four were used for training. To mitigate class imbalance within the training data, we applied the Synthetic Minority Over-sampling Technique[34] in every iteration of both CVs.

Model Evaluation

Model evaluation results are shown in Table 3. For performance metrics, accuracy and macro F1-score were used. In both short interaction and long interaction prediction, all models outperformed the baseline. For *SHORT_INTERACTION*, LightGBM showed the highest performance under LOSO CV, with an accuracy of 0.809 and Random Forest with the highest F1-score of 0.727. Similarly, in 5-fold CV, Random Forest led with an accuracy of 0.814 and an F1-score of 0.777, highlighting the effectiveness of both boosting-based models and ensemble-based models for *SHORT_INTERACTION* prediction. For *LONG_INTERACTION*, CatBoost obtained the highest accuracy of 0.695 and F1-score of 0.630 in LOSO CV. Under 5-fold CV, SVM outperformed others, achieving the highest accuracy of 0.682 and achieved the highest F1-score with 0.680. These results underscore the robust and consistent performance of CatBoost as a boosting-based model and SVM as a strong standalone model for *LONG_INTERACTION* prediction.

Usage Notes

The INPROSH dataset[16] is available upon application and signing of a Data Usage Agreement (DUA) for research purposes in the online Zenodo repository.

Potential Applications

Our dataset is designed to support researchers in identifying and analyzing opportune moments to engage in proactive services in home environments. To achieve this, our dataset provides interruptibility labels with user context information at the time of measuring the interruptibility. The user context information includes behavioral context, temporal context, spatial context, and surrounding image and sound recordings. While prior studies have typically attempted to identify opportune moments for short interactions (e.g., less than 1 minute services)[5,6,35], our dataset provides up to ten minutes of interaction time, making it well-suited for investigating longer proactive services. Researchers using our dataset can redefine the *LONG_INTERACTION* threshold to fit their needs (e.g., 5min, 7min, 9min).

Beyond the primary intent of our dataset, we anticipate its applicability in broader research areas. First, the dataset can be leveraged to detect opportune moments for household appliances (e.g., TV, refrigerator, microwave, etc.) other than smart speakers, in order to facilitate proactive interactions between users and household appliances. For example, Sarkar et al. proposed a smart refrigerator with proactive inventory management that proactively interacts with users and informs users about food status, expiry dates, and spoilage risks[36]. Our dataset can be used to build machine learning models to detect opportune moments for interactions between users and such household appliances, enabling more context-aware and timely proactive services.

Next, our dataset includes image data at one-second intervals for two minutes before providing a proactive service, along with user activities at the time of image capture. This dataset can be utilized for research on human-object interaction (HOI) detection. HOI detection is a computer vision task that aims to identify people, objects,

and their specific interactions (e.g., holding, riding) in images or videos[37]. Each two-minute sample in our dataset comprises (1) a sequence of images that serve as visual features, (2) user activity labels that provide ground-truth annotations.

Furthermore, our dataset includes surrounding sound data (i.e., human voice and ambient noise) at the time of image capture. This sound data can enrich the input modality, in addition to image data. This makes our dataset more valuable for training multimodal models to understand complex HOIs, ultimately improving detection performance. For example, a model can be

trained to detect whether a user is 'studying' or 'using media' by recognizing visual cues (e.g., a person sitting at a desk with a laptop) and auditory cues (e.g., typing sounds and laptop speaker sounds). Therefore, this dataset facilitates the development of robust multimodal models for accurate HOI detection.

Limitations

Our dataset excludes data during nocturnal periods, as proactive services were not delivered while participants were asleep. However, given the low interruptibility observed during daytime naps, this omission may not compromise the dataset's overall utility. Furthermore, the dataset is confined to a single proactive service (microlearning), which precludes analysis of how service type influences opportune moments. Instead, our dataset is suitable for capturing the impact of context by excluding those of service type.

Data availability

The INPROSH dataset[16] is available upon application and signing of a Data Usage Agreement (DUA) in the online Zenodo repository at https://doi.org/10.5281/zenodo.20487071 for research purposes.

Code availability

The Jupyter Notebook, including our technical validation process, is available at https://github.com/HAI-lab-KNU/INPROSH-SupplementaryCodes. The SpeechMaster source code is available for download at https://github.com/HAI-lab-KNU/SpeechMaster.

References

1. TechRepublic. *Google Assistant: A cheat sheet*. TechRepublic, https://www.techrepublic.com/article/google-assistant-the -smart-persons-guide (2021).
2. Kellner, T. *Amazon's new head of Alexa shares his vision for the future, including new shopping and entertainment features*. Amazon, https://www.aboutamazon.com/news/devices/amazons-new-head-of-alexa-rohit-prasad (2022).
3. Arrow. *Home Technology —From Connected to Proactive*. Arrow, https://www.arrow.com/en/research-and-events/articles /home-technology (2020).
4. Bentley, F. *et al.* Understanding the long-term use of smart speaker assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, https://doi.org/10.1145/3264901 (2018).
5. Wei, J., Dingler, T. & Kostakos, V. Understanding user perceptions of proactive smart speakers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **5**, https://doi.org/10.1145/3494965 (2022).
6. Cha, N. *et al.* 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**, https://doi.org/10.1145/3411810 (2020).
7. Lim, J., Koh, Y., Kim, A. & Lee, U. Exploring context-aware mental health self-tracking using multimodal smart speakers in home environments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, https://doi.org/10.1145/3613904.3642846 (Association for Computing Machinery, New York, NY, USA, 2024).
8. Wickens, C. D. Multiple resources and performance prediction. *Theor. Issues Ergonomics Sci.* **3**, 159–177, https: //doi.org/10.1080/14639220210123806 (2002).
9. Kim, M. *et al.* Interrupting for microlearning: Understanding perceptions and interruptibility of proactive conversational microlearning services. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, https://doi.org/10.1145/3613904.3642778 (Association for Computing Machinery, New York, NY, USA, 2024).
10. Iqbal, S. T. & Bailey, B. P. 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*, CHI EA '05, 1489–1492, https://doi.org/10.1145/1056808.1056948 (Association for Computing Machinery, New York, NY, USA, 2005).

11. Bailey, B. P. & Konstan, J. A. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Comput. Hum. Behav.* **22**, 685–708, https://doi.org/10.1016/j.chb.2005.12.009 (2006). Attention aware systems.

12. Kim, A., Choi, W., Park, J., Kim, K. & Lee, U. Interrupting drivers for interactions: Predicting opportune moments for in-vehicle proactive auditory-verbal tasks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2**, https://doi.org/10.1 145/3287053 (2018).

13. Kim, A., Park, J.-M. & Lee, U. Interruptibility for in-vehicle multitasking: Influence of voice task demands and adaptive behaviors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **4**, https://doi.org/10.1145/3381009 (2020).

14. Pielot, M. *CRAWDAD telefonica/mobilephoneuse*. IEEE Dataport, https://doi.org/10.15783/jh3s-h006 (2022).

15. Wu, T., Martelaro, N., Stent, S., Ortiz, J. & Ju, W. Learning when agents can talk to drivers using the inagt dataset and multisensor fusion. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **5**, https://doi.org/10.1145/3478125 (2021).

16. Zenodo. https://doi.org/10.5281/zenodo.20487071 (2025).

17. Hamada, Y. Shadowing: Who benefits and how? uncovering a booming efl teaching technique for listening comprehension.
Lang. Teach. Res. **20**, 35–52, https://doi.org/10.1177/1362168815597504 (2016).

18. Hamada, Y. Shadowing: What is it? how to use it. where will it go? *RELC J.* **50**, 386–393, https://doi.org/10.1177/003368 8218771380 (2019).

19. Newton, J. M. & Nation, I. S. *Teaching ESL/EFL listening and speaking* (Routledge, 2020).

20. Chen, B., Wang, G., Guo, H., Wang, Y. & Yan, Q. Understanding multi-turn toxic behaviors in open-domain chatbots. In *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, RAID '23, 282–296, https://doi.org/10.1145/3607199.3607237 (Association for Computing Machinery, New York, NY, USA, 2023).

21. Mahajan, T. *et al.* An experimental assessment of treatments for cyclical data. In *Proceedings of the 2021 computer science conference for csu undergraduates, virtual*, vol. 6, 22 (2021).

22. Zhu, W., Qiu, R. & Fu, Y. Comparative study on the performance of categorical variable encoders in classification and regression tasks, https://arxiv.org/abs/2401.09682 (2024).

23. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).

24. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. & Data Analysis* **38**, 367–378, https://doi.org/10.1016/S016 7-9473(01)00065-2 (2002). Nonlinear Methods and Data Mining.

25. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, KDD '16, 785–794, https://doi.org/10.1145/2939672.2939785 (Association for Computing Machinery, New York, NY, USA, 2016).

26. Ke, G. *et al.* Lightgbm: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 3149–3157 (Curran Associates Inc., Red Hook, NY, USA, 2017).

27. Dorogush, A. V., Ershov, V. & Gulin, A. Catboost: gradient boosting with categorical features support, https://arxiv.org/ab s/1810.11363 (2018).

28. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297, https://doi.org/10.1023/A:1022627411411 (1995).

29. Pielot, M. *et al.* Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **1**, https://doi.org/10.1145/3130956 (2017).

30. Künzler, F. *et al.* Exploring the state-of-receptivity for mhealth interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **3**, https://doi.org/10.1145/3369805 (2020).

31. Sano, A., Johns, P. & Czerwinski, M. Designing opportune stress intervention delivery timing using multi-modal data. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 346–353, https://doi.org/10.1109/ACII.2017.8273623 (2017).

32. Kang, S. *et al.* K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Sci. Data* **10**, 351, https://doi.org/10.1038/s41597-023-02248-2 (2023).

33. Hwang, J. *et al.* A dataset on takeover during distracted l2 automated driving. *Sci. Data* **12**, 539, https://doi.org/10.1038/s4 1597-025-04781-8 (2025).

34. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**, 321–357 (2002).

35. Wei, J., Tag, B., Trippas, J. R., Dingler, T. & Kostakos, V. 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*, CHI '22, https://doi.org/10.1145/3491102.3517432 (Association for Computing Machinery, New York, NY, USA, 2022).

36. Sarkar, S., R., J., S., J. H. & Prasad, M. Smart refrigeration system with proactive inventory management. In *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1631–1635, https://doi.org/10.1109/ICESC60852.2024.10690137 (2024).

37. Antoun, M. & Asmar, D. Human object interaction detection: Design and survey. *Image Vis. Comput.* **130**, 104617, https://doi.org/10.1016/j.imavis.2022.104617 (2023).

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2026-RS-2023-00260267) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2023-00242528). We gratefully acknowledge Jiwook Lee and Minyeong Kim for their invaluable contributions and active involvement in the experimental procedures and data collection.

Author contributions statement

G.S. prepared the collected data for publication and wrote the manuscript. A.K. conducted the experiment and collected the data.
K.S. J.L. and D.S. provided general support and feedback on the manuscript. J.H. reviewed the manuscript. A.K. supervised the overall study and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Figures & Tables

Table 1. Comparison of data collection setting between INPROSH and prior datasets.

Dataset	Participant (M/F)	Target Device (Environment)	Interruption modality
INPROSH (Ours)[16]	26 (16/10)	Smart speaker (at home)	auditory-verbal
Pielot et al. (2017)[14]	337 (159/178)	Smartphone (not specified)	visual-manual
INAGT (2021)[15]	61 (37/23)	Infotainment system (in-car)	auditory-verbal
TD2D (2025)[33]	50 (25/25)	Infotainment system (in-car)	auditory-verbal
K-EmoPhone (2023)[32]	77 (53/24)	Smartphone, wearable device (not specified)	visual-manual

Table 2. Comparison of data between INPROSH and prior datasets. PPG = photoplethysmography, HR = heart rate, EDA = electrodermal activity.

Dataset	Data Types						
	Interruptibility	Temporal Context Physiological	Spatial Context		Behavioral Context		
			Indoor user location	Outdoor user location	Indoor(in-car) user activity	Outdoor user activity	
INPROSH (Ours)[16]	interruptibility (e.g., interruptible, uninterruptible)	service start time, service end time, etc.	place information (e.g., bedroom), position information (e.g., desk)		indoor activity (e.g., laundry, studying), surrounding images and sound recordings		
Pielot et al. (2017)[14]	interruptibility (e.g., distracted, non-distacted)	service start time, service end time		place information (e.g., work, home)		outdoor activity (e.g., on bycicle, in vehicle)	
INAGT (2021)[15]	interruptibility (e.g, interruptible, uninterruptible, no response)	service start time, etc.		place information (GPS coordinates)	in-car activity (e.g., steering wheel angle), driver videos		PPG, HR, EDA, etc.
TD2D (2025)[33]	interruptibility (e.g., takeover success or failure)	adversarial scenario duration, critical event occurrence time			in-car activity (e.g., texting, audiobook listening)		HR, pupil diameter, EDA, etc.
K-EmoPhone (2023)[32]	interruptibility (e.g., disturbance on a 7-point Likert scale)	service start time, service end time, etc.		place information (GPS coordinates)		outdoor activity (e.g., on bycicle, in vehicle), etc.	HR, skin temperature, EDA, etc.

Table 3. Machine learning model performance.

Model	LOSO CV				5-fold CV			
	ShortInteraction		LongInteraction		ShortInteraction		LongInteraction	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Random Forest	0.808	**0.727**	0.664	0.620	**0.814**	**0.777**	0.656	0.654
Gradient Boosting	0.800	0.706	0.682	0.622	0.796	0.744	0.673	0.669
XGBoost	0.808	0.716	0.685	0.622	0.803	0.761	0.668	0.663
LightGBM	**0.809**	0.722	0.688	0.623	0.805	0.764	0.664	0.659
CatBoost	0.806	0.720	**0.695**	**0.630**	0.809	0.769	0.676	0.671
Support Vector Machine	0.807	0.725	0.682	0.626	0.811	0.773	**0.682**	**0.680**
Baseline	0.493	0.448	0.512	0.479	0.515	0.476	0.504	0.503

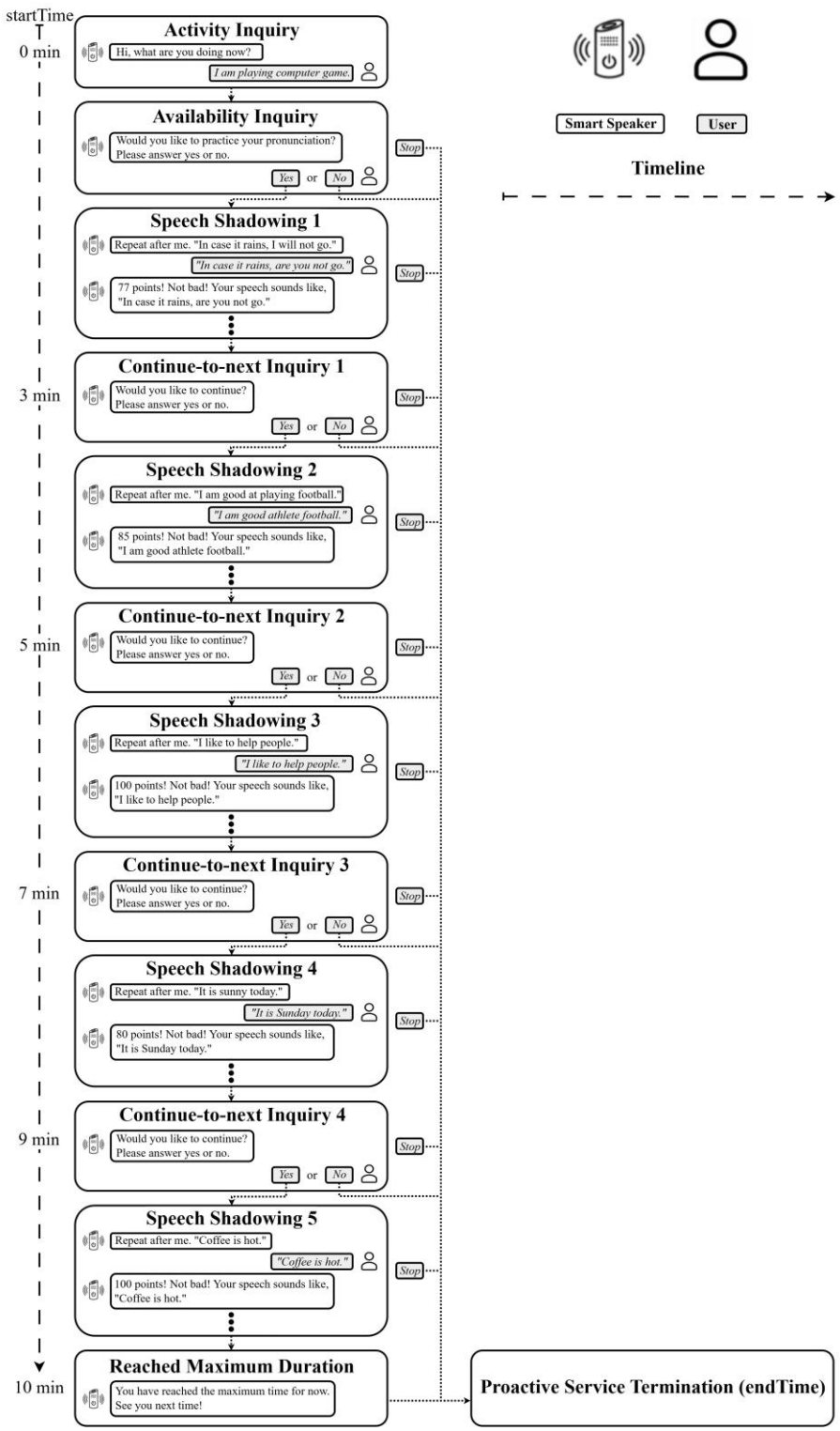

Figure 1. Procedure of the SpeechMaster (microlearning proactive service).

Figure 2. Components of SpeechMaster: a smart speaker (e.g., Google Nest Mini) and a speaker add-on device.

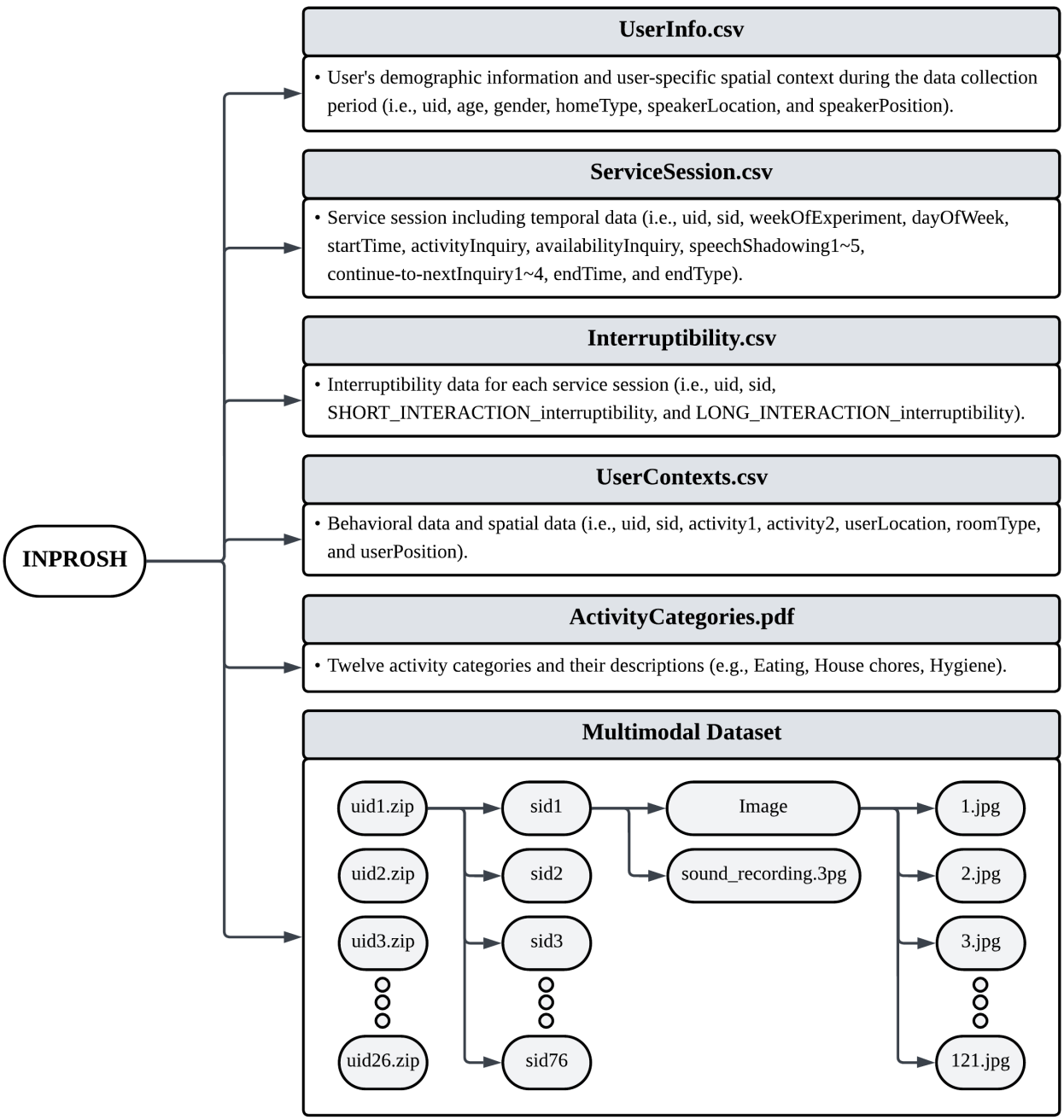

Figure 3. Overview of our dataset.

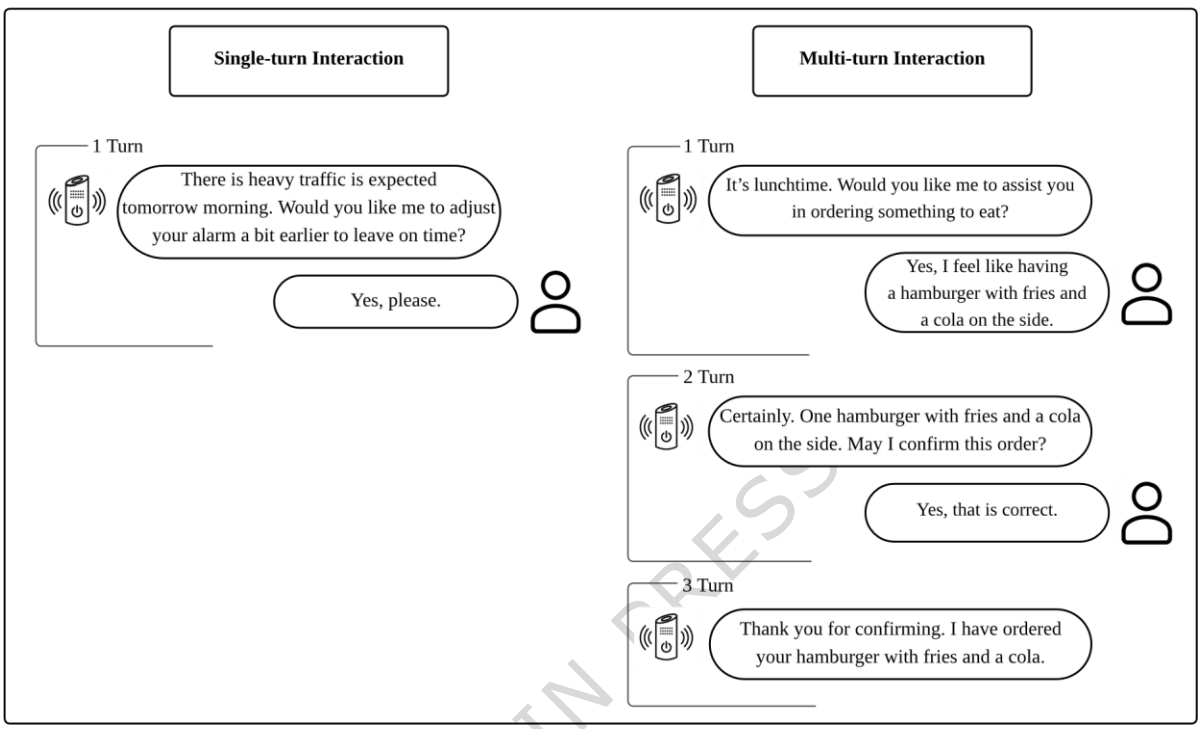

Figure 4. Examples of Single-turn and Multi-turn interactions.

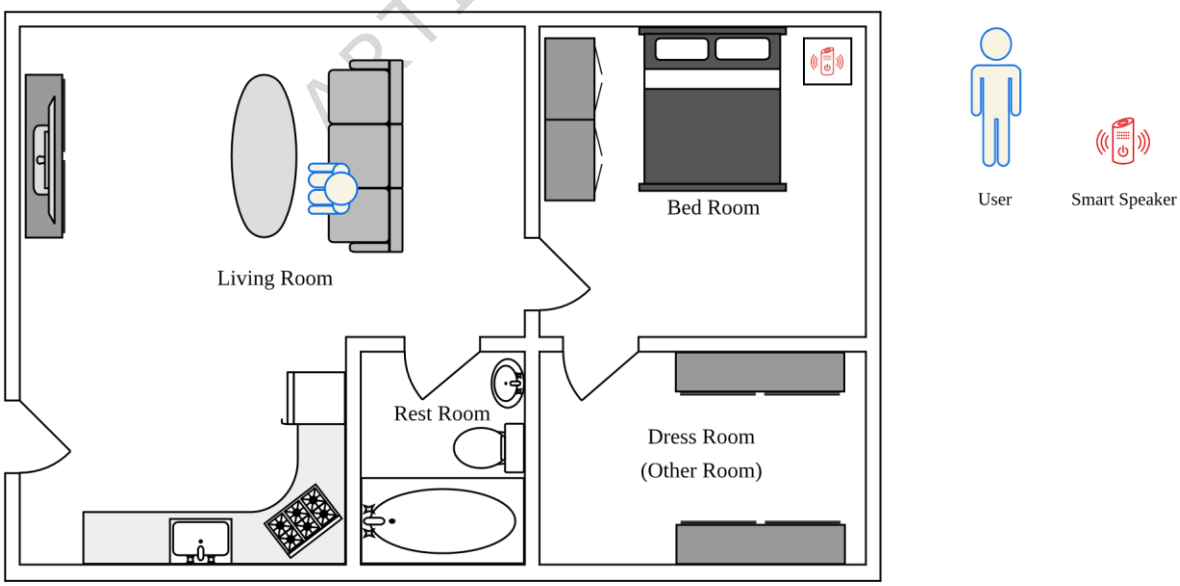

Figure 5. An example of the spatial configuration of the user and the smart speaker within a home. In this scenario, the user is located in the Living Room, whereas the smart speaker is located in the Bed Room, positioned on the bed.

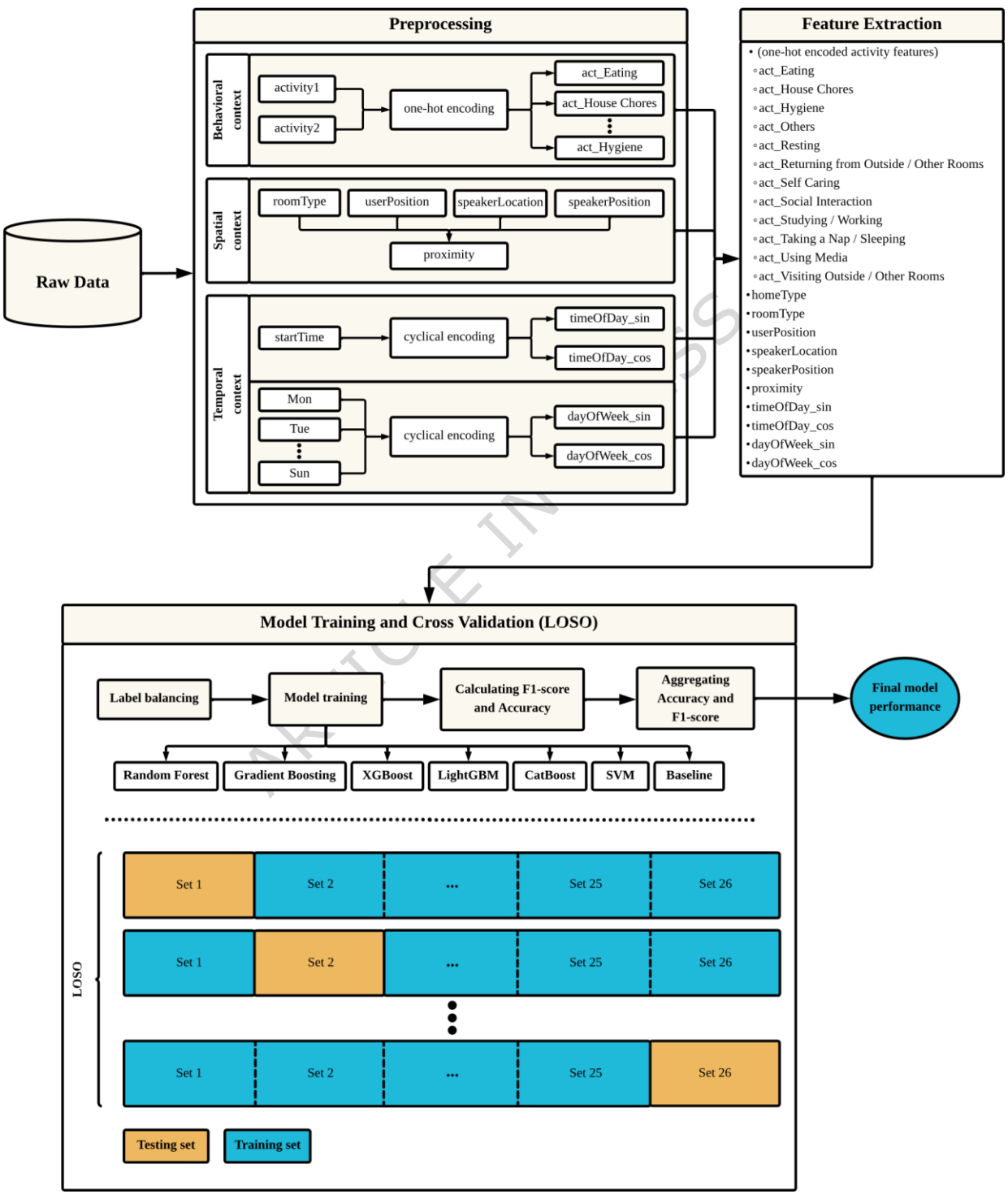

Figure 6. Machine learning pipeline. The pipeline illustrates preprocessing and feature extraction followed by model training and cross-validation. The section labeled “Model Training and Cross-Validation” in the figure

specifically shows the procedure of LOSO CV. For simplicity, the five-fold CV process is not separately illustrated.